ABSTRACT

Title of dissertation: STATISTICAL DEGRADATION MODELS FOR ELECTRONICS

Vasilis A. Sotiris, Doctor of Philosophy, 2011

Dissertation directed by: Professor Michael Pecht
Department of Mechanical Enginerring
Professor Eric Slud
Department of Mathematics

With increasing presence of electronics in modern systems and in every-day products, their reliability is inextricably dependent on that of their electronics. We develop reliability models for failure-time prediction under small failure-time samples and information on individual degradation history. The development of the model extends the work of Whitmore et al. 1998, to incorporate two new data-structures common to reliability testing. Reliability models traditionally use lifetime information to evaluate the reliability of a device or system. To analyze small failure-time samples within dynamic environments where failure mechanisms are unknown, there is a need for models that make use of auxiliary reliability information. In this thesis we present models suitable for reliability data, where degradation variables are latent and can be tracked by related observable variables we call markers.

We provide an engineering justification for our model and develop parametric and predictive inference equations for a data-structure that includes terminal observations of the degradation variable and longitudinal marker measurements. We

compare maximum likelihood estimation and prediction results obtained by Whitmore et. al. 1998 and show improvement in inference under small sample sizes. We introduce modeling of variable failure thresholds within the framework of bivariate degradation models and discuss ways of incorporating covariates.

In the second part of the thesis we investigate anomaly detection through a Bayesian support vector machine and discuss its place in degradation modeling. We compute posterior class probabilities for time-indexed covariate observations, which we use as measures of degradation. Lastly, we present a multistate model used to model a recurrent event process and failure-times. We compute the expected time to failure using counting process theory and investigate the effect of the event process on the expected failure-time estimates.

Statistical Degradation Models for Electronics

by

Vasilis A. Sotiris

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2011

Dissertation Committee:
Professor Michael G. Pecht, Chair/Advisor
Professor Eric Slud, Chair/Co-Advisor
Professor Konstantina Trivisa
Professor Abhijit Dasgupta
Professor Peter Sandborn, Dean's representative

© Copyright by
Vasilis A. Sotiris
2011

Dedication

I dedicate this work to my wife Katya. I owe a lot to my wife, who has wholeheartedly stood by me, both in good and in bad times, who has helped me, grounded me, inspired me and believed in me. I would like to thank professor Eric Slud for his support and guidance throughout my research and for his dedicated and persistent pursuit to teach and instill mathematical rigor. It's been an honor to know and work with professor Slud. I would also like to thank prof. Michael Pecht for motivating my research, believing in me and for holding me to high research standards. Prof. Pecht has taught me the value of knowing the bigger picture, of public speaking and more importantly the value of skillfull communication of technical and abstract ideas. I would like to thank Dr. Michael Azarian for always opening up his time for me and for all his guidance and suggestions. I would like to thank Dr. Diganta Das for his support, encouragement and always useful insight and advice. Last but not least, I want to thank all my friends and family for their support.

Acknowledgments

I would like to acknowledge professor Mei-Ling Ting Lee for her role in introducing me to first hitting time models and their role in reliability. I would like to acknowledge Ed Tinsley and Nikhil Vichare at Dell Inc., and Kai Goebel, Abhinav Saxena and Jose Celaya at NASA, for their support of my research at CALCE.

CONTENTS

1. INTRODUCTION 1

1.1 Problem Setting 3

1.2 Reliability Models using degradation and lifetime data 7

1.3 First Hitting Time Degradation Models 8

1.4 Literature Review 11

1.5 Contributions 22

2. DATA STRUCTURE AND NOTATION 24

2.1 Failure and degradation data in electronics 24

2.1.1 Examples of direct failures 25

2.1.2 Examples of indirect failures 28

2.2 Data-Structures 29

2.2.1 Terminal Structure 30

2.2.2 Longitudinal Structure 31

3. ESTIMATION THEORY 32

3.1 Maximum Likelihood Estimators 32

3.2 Methods of Evaluating Estimators 33

3.2.1 Finite Sample Measures 33

3.2.2 Asymptotic Evaluations 36

3.3 Asymptotic Properties of Maximum Likelihood Estimators 38

3.4 Computing 39

3.4.1 Observed Fisher Information Matrix 39
3.4.2 Multivariate Integration using Gaussian Quadratures 39
3.4.3 Nonlinear Optimization 40

4. WIENER PROCESS AS A DEGRADATION MODEL 41
4.1 Mixed-Type Densities 41
4.2 Degradation Model 43
4.3 Definition of a Wiener Process 44
4.4 The Inverse Gaussian Distribution - Lifetime Model 47
4.5 Marker Model 48
4.6 Bivariate Wiener Model 49
4.6.1 Conditional Independence in the Bivariate Wiener Model . . . 51
4.6.2 An Extended Bivariate Wiener Model 53
4.6.3 The Likelihood Function 53
4.7 Parametric Inference for the TMD Data-Structure 53
4.7.1 Contribution to likelihood from failed devices 54
4.7.2 Contribution to likelihood from surviving devices 56
4.7.3 Derivation of the Complimentary Wiener Term 57
4.8 Predictive Inference 61

5. LONGITUDINAL MARKER MODEL 64
5.1 Modeling Correlated Data 65
5.2 Scheduling Longitudinal Measurements 67
5.3 Parametric Inference 68
5.4 One Intermediate Marker Observation - TMDL1 69
5.4.1 Contribution to likelihood from failed devices 69
5.4.2 Contribution to likelihood from surviving devices 71
5.5 Two Intermediate Marker Observations - TMDL2 73

5.5.1 Contribution to likelihood from failed devices 74

5.5.2 Contribution to likelihood from surviving devices 77

5.6 General Longitudinal Case - GENL 79

5.6.1 Contribution to likelihood from failed devices 79

5.6.2 Contribution to likelihood from surviving devices 81

5.7 Summary and Conclusions 83

6. *CASE STUDIES - TERMINAL AND LONGITUDINAL* 84

6.1 Lifetime and Degradation Simulation Model 87

6.2 Simulation Design 88

6.3 Simulation Results 89

6.3.1 Simulation 1 - TM vs. TMD 89

6.3.2 General Patterns for ARE of $\hat{\mu}$ 92

6.3.3 How can general ARE patterns be used in a practical setting? 95

6.3.4 Predictions on the failure-time distribution under TMD 98

6.3.5 Simulation 2 - TMD vs. TMDL1 98

6.3.6 Simulation 3 - TMD vs. TMDL1 vs TMDL2 101

6.4 Degradation data on Aircraft Gas-Turbine Engines 102

7. *COVARIATES AND REGRESSION STRUCTURES* 108

7.1 Multiplicative Hazards Regression Models 110

7.2 Accelerated Failure Time Regression Models 110

7.3 The Marker Variable as a Special Covariate 111

7.4 Gaussian Process Regression 112

7.4.1 Linear Model 113

7.4.2 Function Space View 114

7.4.3 Connection to FHT models 117

7.5 Support Vector Machines 118

7.5.1 Connection to FHT models 119

8. VARIABLE THRESHOLD MODEL 121

8.1 Uncertain Failure Thresholds 122

8.2 Parametric Inference 122

8.2.1 Contribution to likelihood from failed devices 122

8.2.2 Contribution to likelihood from surviving devices 124

9. SUPPORT VECTOR DEGRADATION MODEL 128

9.1 Introduction 128

9.2 Data Notation and Algorithm Overview 130

9.3 Data Pre-Processing - Principal Component Projections 132

9.4 Two-Class Classifier 133

9.5 Statistical Properties of SVMs and Their Connection to the Evidence Framework 134

9.6 One-Class Classifier 136

9.7 Posterior Class Probabilities 141

9.8 Posterior Class Probabilities as a Marker Variable 144

10. CASE STUDIES 146

10.1 Lockheed Martin Data 146

10.2 Simulated Degradation Data 150

10.3 Degradation Data on Gas-Turbine Engines 154

11. MULTISTATE MODELS AS DEGRADATION MODELS 160

11.1 Introduction 160

11.2 Failure Criteria and Degradation Model 164

11.3 Data Structure and Notation 165

11.4 The Multistate Model 166

11.5 Multistate Markov Chain 168
11.6 Estimator for the Transition Probability Matrix 171
11.6.1 Evaluating the Probability Transition Matrix 174
11.6.2 Example of a two-dimensional multistate model 175
11.7 Case Study 179
11.8 Summary and Conclusions 186

12. SUMMARY AND CONCLUSIONS 190

LIST OF TABLES

1.1 Life Cycle Loads for Electronics 2

1.2 Failure Mechanisms in Electronics 3

1.3 Failure modes and mechanisms analysis for the circuit card assembly 4

2.1 Summary of variables, their names, description and realizations, under each data-structure 30

4.1 Conditional-density terms under TMD for i^{th} device, $s < \tau$ 54

5.1 Key conditional-density terms under TMDL1 68

6.1 MLEs under four data-structures 84

6.2 Asymptotic Standard Errors for ν_X under TM and TMD, from Observed Information vs. Empirical 90

6.3 Central Limit Theorem-based confidence intervals for asymptotic empirical Standard Deviations under TM and TMD 91

6.4 Asymptotic relative efficiency of $\hat{\mu}_{TMD}$ versus $\hat{\mu}_{TM}$ for different (ρ, τ) combinations, with $\theta = (0.1, 1.0, 0.2, 0.1)$ 92

6.5 Asymptotic relative efficiency of $\hat{\mu}_{TMD}$ versus $\hat{\mu}_{TM}$ for different (ρ, τ) combinations, with $\theta = (0.1, 1.0, 0.4, 0.1)$ 92

6.6 Combinations of parameter vector $\boldsymbol{\theta}^*$ 93

6.7 Asymptotic Standard Errors for ν_X under TMD and TMDL1, from Observed Information vs. Empirical 100

6.8 Central Limit Theorem-based confidence intervals for asymptotic empirical variances under TMD and TMDL1 100
6.9 Relative efficiency of $\hat{\mu}_{TMDL1}$ versus $\hat{\mu}_{TMD}$, for different (ρ, τ) combinations 101
6.10 Asymptotic Standard Errors for ν_X under TMD, TMDL1 and TMDL2 from Observed Information 102
6.11 Illustration of ST3-0 data-structure with augmented degradation data 104
6.12 Asymptotic Standard Error for ν_X under TM and TMD from Observed Information. Empirical drift ~ 0.005 106
6.13 Asymptotic Standard Error for ρ under TM and TMD from Observed Information. Empirical correlation $= 0.42$ 106

10.1 SVM Optimization Results 149
10.2 Comparison of CALCEsvm and LibSVM Detection Accuracy Against Lockheed Data 150
10.3 LibSVM Accuracy Results for Simulated Data 152
10.4 CALCEsvm Accuracy Results for Simulated Data 153

11.1 Lifetime data and survival probability estimates 176
11.2 Sample of failure times for collected from a computer 180
11.3 Life Table 180
11.4 Simulated Data for Case Study 189

LIST OF FIGURES

2.1 Crack formation in a solder joint 25

2.2 Whisker growth, seen at initial stage 27

2.3 Pre and post aging of an IGBT part. Increased reflectivity picked up by SAM indicates degradation 28

5.1 Illustration of an observation on the degradation process under a longitudinal data structure. Rectangular boxes represent density factors given by equation (4.21), and the oval shape the factor given by equation (4.13) 73

6.1 General ARE patterns of $\hat{\mu}_{TMD}$ versus $\hat{\mu}_{TM}$, under increasing σ_X . . . 94

6.2 General ARE patterns of $\hat{\mu}_{TMD}$ versus $\hat{\mu}_{TM}$, sunder increasing σ_Y . . 95

6.3 General ARE patterns of $\hat{\mu}_{TMD}$ versus $\hat{\mu}_{TM}$, under increasing ν_Y . . . 96

6.4 General ARE patterns of $\hat{\mu}_{TMD}$ versus $\hat{\mu}_{TM}$, under increasing ν_Y . . . 96

6.5 Predicted degradation and survival density as a function of time . . . 99

6.6 Correlation matrix between covariates (1 through 17) and degradation (18) 105

6.7 Relative efficiency of $\hat{\mu}$ from G_2 vs. G_1 for different (ρ, n) combinations evaluated with gas-turbine engine degradation data 107

7.1 Function space view. Left: nine functions drawn at random from a GP prior, and the dot plots a single observation x. Right: Nine random functions drawn from the posterior. In both plots the shaded area represents point wise mean plus and minus two times the standard deviation. . . . 115
7.2 GP mean function $m(\boldsymbol{z}_*)$. The blue dots are the training data $(\boldsymbol{x}, \boldsymbol{z})$. 117

8.1 Relationship between threshold and degradation variables at time τ . 125

9.1 Algorithm flow diagram showing the processing of the data. 132
9.2 Logistic distribution model for posterior class probabilities. 142

10.1 Joint posterior class probability vs. observation for Lockheed Martin test data set . . . 147
10.2 Joint posterior class probabilities for CALCEsvm, and the open source support vector classification software called LibSVM. . . . 147
10.3 Joint positive posterior class probability for simulated data set. . . . 152
10.4 Joint positive posterior class probability for simulated data set in P2. 153
10.5 Joint positive posterior class probability for simulated data set in P3. 154
10.6 Distribution of projected multivariate data on to first two principal components . . . 156
10.7 Estimate of unit health as a function of time . . . 157
10.8 Variation in estimated health probability across all training units . . 158
10.9 Cross unit variation in the estimated health probability . . . 158
10.10 GP model fit to the health estimate time series of a test unit 159

11.1 Illustration of sample paths from a bivariate stochastic process $\{X(t), Y(t)\}$ 166
11.2 Multistate model representation of a bivariate stochastic process . . . 167
11.3 Shape of probability transition matrix . . . 170

11.4 Survival probability estimates from the AJ estimator 178
11.5 Expected time to failure 179
11.6 Simulation of the degradation process conditioned on an error event process 182
11.7 Expected time to failure distribution for devices with 0,1 and 2 error events 183
11.8 Transition probability matrix used in case study 184
11.9 Aalen-Johansen survival probability estimate starting from states 0, 1 and 2 185
11.10 Expected time to failure starting from state 0,1 and 2 respectively . . 186

1. INTRODUCTION

The Center for Advanced Life Cycle Engineering (CALCE) at the University of Maryland has over the past 30 years pioneered methodologies for reliability analysis of electronic products. With increasing presence of electronics in modern systems and in every-day products, their reliability is inextricably dependent on that of their electronics [2]. Reliability methodology for electronics went from simple standard based assessments, to using physics of failure (PoF) models in the late 1980s, and more recently prognostics and health management (PHM) models.

The fundamental aim of PoF modeling is to postulate, based on the physics and mechanics of the failure mechanisms, a set of generic functional relationships between the mean fatigue life and the operational loads [3]. In their 1990 paper, Dasgupta et al. [3] are among the first to stress the importance of modeling failure-times in conjunction with PoF models. Pecht et al. 1990 [4], point out the importance of considering PoF, especially for material properties, for modeling failures of electronics obtained from laboratory tests. Hu et al. 1991 [5], point out limitations in conducting accelerated failure tests without knowledge of the failure mechanisms. They point out the need for ensuring that accelerated failure tests target and therefore induce the intended failure mechanism.

During use, electronics are exposed to a variety of loading conditions such as temperature or power excursions, shock and vibration. Interconnects, such as solder joints, printed circuit board traces, component leads, and connectors are vulnerable to these loading conditions and are susceptible to failures by mechanisms such as fatigue, creep, corrosion, and mechanical over-stress [58]. Life cycle loads, either

individually or in various combinations may lead to performance or physical degradation and reduce its service life [7]. Table 1.1 lists life cycle loads experienced by electronics. The extent and rate of degradation depends on the nature, magnitude, and duration of exposure to such loads [8].

Tab. 1.1: Life Cycle Loads for Electronics

Loads	Examples
Thermal	Steady-state temperature, temperature ranges, temperature cycles, spatial temperature gradients, temperature ramp rates, heat dissipation
Mechanical	Pressure magnitude, pressure gradient, vibration, shock load, acoustic level, strain, stress
Chemical	Aggressive versus inert environment, humidity level, contamination, pollution, fuel spills
Physical	Radiation, electromagnetic interference, altitude
Electrical	Current, voltage, power

Failure modes, mechanisms and effects (FMMEA) analysis is a process developed at CALCE, that characterizes the product on all levels, i.e., parts, systems and physical interfaces [9]. Failure mechanisms are the physical process by which stresses cause damage to the elements comprising the system, ultimately leading to failure [10]. Table 1.2 lists generic failure mechanisms which can serve as potential agents of failure [10]. A failure mode is the means by which a failure manifests, or by which degradation is measured. Table 1.3 lists failure modes and mechanisms analysis for the circuit card assembly, which represents a typical electronic part or device. Generally failures in electronics are thought to be a result of either, *over-stress*, or *wear-out* failure mechanisms. Over-stress failures occur when the stress exceeds the device strength, and failures occur suddenly. Wear-out failures occur due to the accumulation of damage with repeated stress or generally usage.

Recently, with the advent of powerful and accurate sensor technology, there is interest in *real-time* or *in-situ* reliability analysis. Device-specific degradation can be explained and predicted, based on sensor data on individual devices, rather than

Tab. 1.2: Failure Mechanisms in Electronics

Over Stress Failures	Wear-out Failure
Brittle Fracture	Wear
Ductile Fracture	Corrosion
Yield	Dentritic Growth
Buckling	Interdiffusion
Large Elastic deformation	Fatigue crack propagation
Interfacial De-adhesion	Diffusion
	Radiation
	Fatigue crack initiation
	creep

on sample averages. There is also economic and business strategic needs that can be solved using real-time reliability analysis results. PHM is a method that permits the assessment of the reliability of a component (or system) under its actual application conditions [2]. PHM aims to provide advanced warning of failures, enable optimal maintenance actions, reduce life-cycle costs, and aid in mission critical decisions.

The key element to PHM is its prognostic element, which as we show can play the role of fault detection, degradation estimation and failure-time prediction. The measures provided by these predictive outcomes are central to useful implementation of PHM technology. Pecht 2010 presents a PHM road-map and an assessment of the state of practice for information and electronic-rich systems [17].

1.1 Problem Setting

Implementing PHM requires us to collect data and build appropriate models for the various kinds of questions we are asking of a PHM program. In this work we are interested in addressing questions related to failure-time prediction and all the data and modeling assumptions necessary to get there.

Mathematically failure-time predictions can be evaluated using the conditional probability of a future failure time given survival up until current time. Traditionally the only data collected on tested devices were lifetime data. The survival status of

Tab. 1.3: Failure modes and mechanisms analysis for the circuit card assembly

Category	Site	Mode	Mechanism	Stress
Printed circuit board (PCB)	PTH	Electrical open	thermal fatigue	Temperature cycling
		Electrical short (between PTHs)	Conductive filament formation	Voltage, high RH small PTH spacing
	Metaliz. traces	Electrical short (between traces) and degradation in resistance open traces	Electro-migration Corrosion ionic contamination	High current density High RH, electrical bias
Compon.	Inductors	Short between windings and the core	Thermal fatigue	Overheating due to excessive current and prolonged use at high temperature
		Short between windings	Thermal fatigue	Overheating due to excessive current and prolonged use at high temperature
		Open circuit inside the inductor	Thermal fatigue	Prolonged use at high temperatures
Inter connect	Solder joints	Intermittent change in electrical resistance	Thermal fatigue, creep, high-cycle fatigue	Temperature cycling and vibration

the device therefore constitutes a binary process, which is equal to zero while the device is not failed and equal to one at the failure-time.

This data-structure is deficient in detail and inference models lack predictive power because of the following reasons:

- A model that only uses lifetime data cannot account for dynamic environments that products are exposed to in the field. In other words, predictive inference does not account for real-time environmental or usage conditions.

- Often in tests we collect few failures. This can be because we do not have many devices to test in the first place, typically because they are expensive. This can also be a result of short test times, constrained by money and time.

- In engineering applications, lifetime data are typically collected under "accelerated" conditions, which makes them sometimes inappropriate for inference procedures in a PHM setting, as we discuss later on.

Singpurwalla [22] and Sobcyzk [12] among others motivate the use of stochastic processes in models in order to account for the dynamic environments that fielded devices experience. In this context, there is therefore a need for data that can describe the evolution of the stochastic process. In this case PHM prediction models will become a little more complicated because now they have to accommodate data measured on the stochastic process. However, these models are presumably better suited at capturing the changing environments.

Failure tests often result in few failure-time samples. When the failure-time sample is small, the traditional inference and prediction models suffer because the available data are not ample enough to fit the model parameters with adequate precision. There is a need therefore for auxiliary reliability information. Auxiliary reliability information can come from observing the *degradation variable*(s) of the device over time, i.e, the degradation process of the device. The degradation process consists of a collection of degradation variables indexed by time, and can be thought of as a stochastic process of accumulating damage. Because failures in electronics can be strongly linked to known degradation variables (failure modes), and because the response to stress (even under constant stress) is random (due to variations in material properties), it is not unreasonable to model degradation as a stochastic process, and failure as its first hitting time of a threshold.

Due to "self-healing" of materials in electronics, the damage is not generally considered non-decreasing over time, but instead, it can "heal" or recover. Under

these physical conditions, a Gaussian process can be appropriate for modeling the fluctuations in the degradation process. If we can observe this process then we have access to valuable "auxiliary" information that can presumably explain the rate of degradation and therefore improve inference on failure-times. Auxiliary reliability information can also come from covariates collected on each device, and when degradation is latent, as we discuss, it can also come from marker variables.

- The degradation variable is a time-varying random variable that defines the failure-time
- A marker variable is a time-varying random variable which co-varies with the degradation variable, and assists in tracking its progress. When degradation is latent, the marker variable forms the basis for inference about degradation and its progress towards a threshold.
- Covariates are time-varying, possibly time-dependent deterministic variables specific to each device, and are not assumed to follow any distribution. Covariates form an important part of degradation models, and we discuss them in chapter 7.

In failure tests of engineered products or systems, failure-times are most often attained under higher than normal stress conditions, using what are called accelerated life tests (ALT). The problem with such tests is that they can change the failure mechanisms and cause the device to fail in a way that it would ordinarily never experience under normal usage conditions. This naturally brings up the question of what is considered normal operating/usage conditions. We assume here that the levels of stress in ALTs will never be experience in the field. In other words the failure-time information resulting from ALTs needs to be related to failure-time scales in the field.

So we see that the use of stochastic processes in modeling lifetimes is motivated from *i*) an intuitive representation of degradation, *ii*) from a data limitation; namely the lack of large failure-time samples, and *iii*) from the uncertainties generated in using ALT lifetime data. ALTs do help increase the failure-time sample size, however they also introduce uncertainty that is difficult to handle in models.

In conclusion, there is a need for PHM reliability models that perform "well" under small sample sizes, and we have above pointed out three reasons why we would have small sample sizes:

- Not many devices to start off with
- Time/money constraints
- Accelerated test conditions are reduced in order to preserve the failure generating mechanism. Less stress means fewer samples fail in a fixed period of time.

The first contribution in my thesis is to address the small failure-time sample problem with a model that draws strength from surviving devices in addition to failed ones to improve inference under smaller failure-time samples.

1.2 Reliability Models using degradation and lifetime data

It has been noted by Chown, Pullum and Whitmore 1994 [13], that reliance on lifetime data is becoming less and less practical in engineering, and there exists a pressing need for reliability models that capture the degradation response of a device over time. Nair 1988 [15] states: "... Degradation data are a much richer source of information than time-to-failure data. The lack of statistical methods for analyzing them prevents users from exploiting this valuable source of information [21].

Degradation models are based on lifetimes, degradation and covariate measurements that can be collected in the same failure-test. Failure-tests are usually performed over a fixed time period and some devices may survive. In fact as discussed in the previous section, it is more common that most devices do survive. The information on covariates, degradation and lifetimes, collected on both surviving and failed devices can arguably make PHM technology possible.

1.3 First Hitting Time Degradation Models

When failure is believed to result from wearout, or damage accumulation or as we refer to *degradation*, then failure-times can be defined as its *first hitting time* (FHT) to a degradation threshold level, which can be known or unknown. The class of degradation models, therefore, that we discuss we call FHT degradation models. In FHT degradation models the definition of failure is strictly defined by the degradation variable, which as we see next can create some ambiguity depending on the data.

The next modeling complication therefore, is related to the definition of failure. Under wear-out failure mechanisms, we define two categories of failure definitions:

Definition 1.1 (**Direct failures**). *Direct failures are defined as the time when an observable degradation variable first violates a fixed and known failure threshold, so that the terminal level of degradation is the same for all failed devices.*

Definition 1.2 (**Indirect failures**). *Indirect failures are defined as the time when a latent degradation variable first violates an unknown, possibly random failure threshold, so that the terminal level of degradation varies across devices.*

In failure-tests of electronic devices and products we observe both direct and indirect failures. Direct failures are more commonly used for electronic devices or components (with small number of parts), such as for power semiconductor devices

like IGBTs, or more basic devices such as capacitors, inductors, resistors, diodes and transistors. These devices have in common simple structures, for which there exists an understanding of the physics of failure. Because we understand their PoF, we know how and why they fail, and therefore can select useful degradation variables with which to define failure-time.

In failure-tests, indirect failures, are commonly used for electronic systems or products that are composed of many parts that can interact in complicated and typically unknown ways. Most modern engineered products depend on sophisticated electronics, which are housed on densely populated boards or chassis. For electronic products there are generally no suitable PoF models, and therefore no suitable degradation variables that can define failure. Failure instead is observed as an external process, typically by observing the performance of the system, and failure is defined as the lack of performance to some predefined degree. For example, a computer freezes, a car stalls, onset of heart attack, etc. In each of these examples there exists a complicated host system, the computer, the car the human body, where the true health/degradation is unknown.

We are interested in reliability models that are motivated by both direct and indirect failures. We are interested in direct failure data because at the component level we can use PoF models to enhance the predictive power of the data-driven models. We are interested in indirect failures because they are of greater commercial and application level importance.

Although direct failures are based on an observable degradation variable, that variable may not be predictive of failure, i.e., it attains the failure-threshold level suddenly without any preceding trend. In such cases, the observable degradation variable is called a *surrogate* degradation variable, and inference is based on unobservable (latent) degradation variables, which we also call *true degradation variable*. Some examples of latent degradation variables are: the length of a crack in a solder

joint, the surface roughness in a ball bearing fan, etc. These variables are latent because they cannot be measured during the failure-test.

We are however, interested in the situation (experimental setup) where the latent degradation variable can be measured/determined at failure, potentially through some intrusive (postmortem or terminal) examination. Access to terminal degradation measurements is important, especially when the test is designed to induce specific failure mechanisms, and when there are known PoF models to work with. It is difficult to observe the latent degradation variable during the test without intrusive and often destructive procedures, however we can measure it at the failure-time without interfering. In the example of the solder joint, the crack length can be measured by cross sectioning and x-ray microscopy analysis, and similarly to determine the surface roughness on the ball bearing.

Because the latent degradation is only observable at termination, a degradation model must account for its latency at any other time, motivating what we call *Latent degradation models*. We believe that access to true degradation data can help better estimate the rate and variability of degradation in fielded devices. Using the true degradation data we gain stronger insight into the effects of the environment and usage on the rate of degradation. The drawback of using true instead of surrogate degradation data is we only have one such measurement, whereas we have many observations on the surrogate, on each device. In this case the motivation for a latent degradation model is only as good as the value/importance of the true degradation data relative to the surrogate.

For indirect failure data, the need for a latent degradation model is more obvious. The failure is assumed to occur due to wear-out, but the failure-time is not determined based on anything we can observe. At the failure-time, however, we are again interested in the situation where we can measure the terminal degradation level, on one or more degradation variables.

Degradation models should also account for uncertain or unknown failure thresholds. It is not uncommon in failure-tests to use arbitrary thresholds or thresholds based on antiquated standards. Failure thresholds represent the strength of a device to sustain stress, and is therefore a heterogeneous quality across devices. Failure-thresholds may not only vary across devices, but may also vary with time. In other words, the strength of a device can itself decay/degrade with time.

In failure tests of electronics, reliability data can be observed frequently in time on each device, and in small failure-time sample situations, it behooves us to use it. Longitudinal time-indexed observations on device reliability, whether observed directly on the degradation variable or its surrogate or some higher level marker can give strong insight on the wear-out and the strength of the device as a function of time and therefore improve estimation. When the degradation variable is latent, as we consider here, longitudinal measurements are made on degradation markers that track the progress of the degradation towards a threshold, giving rise to what we call *bivariate latent degradation models.*

1.4 Literature Review

Reliability models based on stochastic processes are discussed by many, and here we mention the main references used in this thesis: Desmond 1985 [16], K. Sobczyk 1985 and 1992, Lu and Meeker 1993 [18], Kahle 1993 [19], G. Whitmore 1995 [118], J. Lu 1995 [21], Singpurwalla 1995 [22], Doksum and Normand 1995 [117], Whitmore et al. 1998 [24], Petit and Young 1999 [25], Lee et al 2000 [26], Lawless and Crowder 2004 [27], and more recently Lee and Whitmore 2006 [28], Lehmann 2008, Tang and Su 2008 [29], Kahle and Lehmann 2010 [30], Wang and Xu 2010 [31], Singpurwalla 2010 [32], and Lee et al. 2010 [33], among others. These sources have in common the use of Gaussian processes to model measurements on, or estimation of a degradation process leading to failure, in what are generally called

degradation models. Few have investigated a bivariate stochastic process setting to model lifetime and degradation data. Predominantly the bivariate structure has been used to model latent degradation processes, and is found mostly in medical studies, specifically in immunological and epidemiological studies. There are by far much fewer literary references to bivariate stochastic processes in the reliability field. We discuss some of these papers further below.

Sobczyk 1987 [34] and in his 1992 book, presents an exposition of methods of modeling and analyzing fatigue fracture of engineering materials. He thinks of fatigue to be random and therefore considers stochastic models for fatigue processes. He argues that early probabilistic treatment of fatigue was mainly concerned with the statistics of dispersed data, fitted by various probability distributions, such as the lognormal and Weibull. He introduces the limitation that such approaches do not provide any direct relationship to the basic fatigue mechanisms. His modeling approach of fatigue consists of three basic steps: (i) choosing an appropriate stochastic model-process for fatigue accumulation; (ii) determining the probabilistic properties of the model-process (iii) relating the model process to empirical data and parameter estimation.

Lu and Meeker 1993 acknowledge small failure samples in engineering failure tests, specifically in electronics systems. They motivate the need for degradation models that can be used to define time-to-failure distributions. Based on this idea they develop statistical methods for using degradation measures to estimate a time-to-failure distribution for a broad class of degradation models. They use fatigue-crack-growth data to motivate the model. For each device, they assume that degradation measurements y_j are available for pre-specified times $t_j = t_1, \ldots, t_s$, until y_j crosses the pre-specified critical level D or until a pre-specified censoring time t_s. Sample paths are modeled by a parametric general path model $y_j = \eta_j + \epsilon_j$, $\epsilon_j \sim \mathcal{N}(0, \sigma_\epsilon^2)$, and the failure distribution is written in terms of $(t_j, D, \eta_j, \epsilon_j)$. From

here they consider several examples where the model parameters are given specific parametric forms, such as Weibull, Bernstein, lognormal and multivariate normal.

Kahle 1993 considers the Wiener process as a degradation model of a damage process. He shows that for independent, not necessary identically distributed observations of process increments, for observations of lifetime-distributions and for a mixture of these observations the assumptions of asymptotic normality of Maximum-Likelihood-Estimations (MLE) are fulfilled. The asymptotic normality of MLEs is used to find simultaneous confidence regions for parameters of the damage process.

Whitmore 1995 et al. model the degradation process by a Wiener process with drift. They also model measurement errors as independent normal random outcomes that are also independent of the degradation process. The true degradation process is therefore separated from the observed by an error term, that they incorporate into the model.

In her thesis J. Lu 1995 does an excellent job at motivating the need for models that use both lifetime and degradation data for estimation and prediction. She introduces the Wiener process as a degradation model and discusses it for a mixed data-structure consisting of degradation observations at a set of fixed time points $0 < t_1 < t_2 < \ldots < t_n$, and failure-times s_i. For a failed device i, the mixed data-structure has the following form: $(x_{i1}, x_{i2}, \ldots, x_{in}, s_i)$, $i = 1, \ldots, p$ and for a surviving device j: $(x_{j1}, x_{j2}, \ldots, x_{jn})$, $j = p+1, p+2, \ldots, p+q$. She derives a likelihood function $L(\delta, \nu)$, and inference equations for the mixed data-structure. She also touches on inference when the failure threshold (a) is unknown. Inference in this case is accomplished by fixing the failure threshold level and estimating the process parameters for many different such levels. The most suited threshold level, given the data, can be determined by maximizing the likelihood function $L(\hat{\delta}(a), \hat{\nu}(a))$ over the chosen threshold levels.

In an influential expository paper, Singpurwalla 1995, provides an overview of

failure models, based on stochastic processes, that are suitable for describing the lifetime of items that operate in dynamic environments.They signal a new philosophy of life-testing experiments wherein one also monitors the environmental factors that govern tests, and sets a tone for work in the development of models for survival wherein the physics of failure and the characteristics of the operating environment play a central role.

Doksum and Normand 1995, present two stochastic models that describe the relationship between biomarker process values at random time points, event times, and a vector of covariates. In the first model the biomarker process is a Wiener process whose drift is a function of the covariate vector. In the second model the biomarker process is taken to be the difference between a stationary Gaussian process and a time drift whose drift parameter is a function of the covariates. They present the methods principally in the context of conducting inference in a population of HIV infected individuals.

In their 1998 paper, Whitmore et al., present a bivariate Wiener process model for degradation processes, applied to a terminal data-structure, where degradation is entirely unobservable. Their paper is one of the earliest sources that use a bivariate Wiener process to model degradation. This work is the main inspiration for the model development in part-I of my thesis. Inference is based on observations on a marker variable that can be used to track the progress of the latent degradation. They derive joint densities for the likelihood and apply the model to simulated lifetime and marker data.

Pettit and Young 1999 consider both lifetime and degradation data on failure and survived devices. They model degradation by a Wiener process, and lifetime as its first hitting time to a fixed threshold.They extend the analysis in J. Lu 1995 by using a fully Bayesian approach to estimation and prediction.

Lee et al. 2000, extend the bivariate Wiener process considered by Whitmore

and co-workers 198, and model the joint process of a marker and a latent health status. Covariates are related to the model parameters through generalized linear regression functions.They derive formulas for predicting residual survival time and discuss model validation on clinical trial data.

Lawless and Crowder 2004 argue that for certain types of degradation processes a model involving independent non-negative increments is appropriate. They use, therefore, the Gamma process as a model for degradation processes. They construct a tractable gamma-process model incorporating a random effect and fit the model to data on crack growth. Covariates are incorporated via an accelerated life model by replacing process parameters with a functional of the covariates and a new set of unknown parameters.

Lee et al. 2006 review first hitting time (FHT) models for survival data, and introduce threshold regression for survival analysis. They argue that FHT models can only be valuable in applications if they can include regression structures. Regression structures allow effects of covariates to explain the inherent dispersion of the data, thereby taking account of variability and sharpening inferences. Threshold-regression refers to FHT models with regression structures that accommodate covariate data. The parameters of the process, threshold and time scales may depend on the covariates.

Lehman 2008 et al. survey some approaches to model the relationship between failure time data and covariate data. In particular they consider a class of degradation-threshold-shock models in which failure is due to the competing causes of degradation and trauma. They express the failure time in terms of degradation and covariates, where degradation is modeled by a process with stationary independent increments and related to covariates through a random time scale.

Tang and Su 2008, propose to obtain the first hitting times of a degradation process, modeled by a Wiener process with drift, over certain non-failure thresholds.

Based on only these intermediate data, they obtain the uniformly minimum variance unbiased estimator for the mean lifetime.

Kahle and Lehmann 2010 describe a simple degradation model based on the Wiener process with drift. They consider the case that each realization of the degradation process, both process increments and failure time are observable, and they estimate the process parameters. They observe each sample path of the degradation process to either a failure time or to a censoring time. They develop a likelihood function based on the conditional distribution of the process under the condition that the threshold level is not exceeded and the joint distribution of the conditional process increment and lifetime variable.

Wang and Xu 2010 discuss a class of inverse Gaussian process models for degradation data and associated maximum likelihood inferences. They use an expectation maximization (EM) algorithm to obtain MLEs of the unknown parameters and the bootstrap method to assess the variability of the MLEs.

Singpurwalla 2010 provides an interesting perspective on damage accumulation and marker processes, a perspective and thoughts that are much related to the ideas developed in this thesis. He talks about damage being an abstract concept, which is not measurable, but its surrogates can be measured. With this in mind he highlights a probabilistic architecture based on a bivariate stochastic process with one component that is non-decreasing and the other that may fluctuate around some mean. The non-decreasing process leads to the fluctuating observable process. He argues that the failure threshold is random with an exponential(1) distribution, and he calls this threshold the *hazard potential* of an item.

Lee et al. 2010, consider sequential observations on degradation and/or on covariates prior to failure. They argue there is a need for simple regression methods to handle longitudinal data, and present the use of the Markov property to do this. They outline a model that can handle a longitudinal process with an unobservable

health status (degradation) as well as time-varying covariates.

A large portion of the literature on degradation models in reliability is found under accelerated degradation models. The reason as discussed earlier is due to the need for early failures. Accelerated degradation models also use lifetime and degradation data collected in ALTs, but differ from "non-accelerated" degradation models in that the data structure often includes extra complications that we address in this work. Work on accelerated degradation models in reliability can safely find its way back to the middle of the 20th century with early work by Epstein and Sobel 1963, Singpurwalla 1970, 1971 and 1973 [36] [37] [38], Mann et al. 1974 [39] and a few others, then followed by Bhattacharyya and Fries 1982 [40], Nelson 1990 [41], Carey and Koenig 1991 [42], Doksum and Hoyland 1992 [116], Meeker and Escobar 1993 [44], Whitmore and Schenkelberg 1997 [45], Lu, Park and Yang 1997 [46], Meeker, Escobar and Lu 1998 [47], Owen and Padgett 1999 [48] ,Onar and Padgett 2000 [49], Bagdonavicius and Nikulin 2001 and 2004 [50] [51], and more recently, Padgett and Tomlinson 2004 [52], Park and Padgett 2005 [53], Park and Padgett 2006 [54], Bae, Kuo and Kvam 2007 [55], and Meeker et al. 2009 [56].

Singpurwalla 1970 proposed to investigate the functional relationship between parameter vector $\boldsymbol{\theta}$ for the probability density function of the time-to-failure random variable and stress vector $\boldsymbol{S}$. With the above relationship he was interested in making inference about the failure behavior of the device at environmental conditions which cannot be simulated in a test. In this work, he assumes that i) the device fails due to single failure mode and ii) that the severity of the stress level does not change the type of life distribution, but that the stress level influences the values of its parameters. He assumes a linear stress-failure relationship and an exponential failure time pdf, with the hazard rate parametrized as: $\lambda_i = BS_i$, with B unknown parameter.

Singpurwalla 1973 discusses the problem of inference when both the location

and the scale parameter of the time-to-failure distribution are re-parameterized, the former as a linear function of stress and the latter according to the Arrhenius re-action rate model. The failure time distribution is again exponential with two parameters λ_i and γ_i, $f(t|\lambda_i, \gamma_i) = \lambda_i exp(t - \gamma_i)$ and $\lambda_i = exp(A - B/V_i)$, and $\gamma_i = \alpha - \beta V_i$, where A, B, α and β are unknown parameters. The objective is to predict the mean time-to-failure at use conditions $\mu_u = \gamma_u + \lambda_u^{-1}$.

In their book Mann, Schafer and Singpurwalla 1974 discuss accelerated life testing, models and some results. They are interested in making inference from accelerated life tests when certain relationships between parameters of a failure time distribution and the environmental conditions can be reasonably hypothesized. These relationships or models are derived from an understanding of the physics of failure (PoF) of the device under discussion. The time-to-failure random variable is given by $f(t|\boldsymbol{\theta})$ where $\boldsymbol{\theta} = g(\boldsymbol{S}|a, b, \ldots)$ is known except for $a, b, \ldots$ and is valid for certain ranges of stress $\boldsymbol{S}$. Their objective is to obtain estimates of $a, b, \ldots$ based on life tests conducted at elevated values/levels of stress, and then use these estimates to make inference about $\boldsymbol{\theta}$ in use environment stress $\boldsymbol{S}_u$.

Bhattacharyya and Fries 1982 focus on the inverse Gaussian $IG(\theta, \lambda)$ as the failure time distribution. They motivate that the genesis of the IG can be cast in the context of cumulative fatigue, or depletion of strength. They point out the relationship to a Wiener process crossing a fixed threshold ω, $\theta = \omega\mu^{-1}$ and $\lambda = \omega^2\delta^{-2}$, where μ and δ are the Wiener process with drift parameters. In their accelerated degradation model, the mean of the Wiener process is parameterized as a linear function of stress as $\mu = \alpha + \beta x$, ensuring a direct relation between the cumulative fatigue/wear/degradation and stress levels. For inference they observe stress and failure time pairs across a range of stress settings.

Carey and Koenig 1991 describe an analysis strategy to extract reliability information from measured degradation of devices submitted to elevated stress. Degra-

dation is the propagation delay in an integrated logic family device, which increases with age and temperature. The degradation model on a single device from a given temperature group is: $y_n - y_0 = \theta(1 - exp(-\sqrt{\lambda t_n})) + \epsilon_n$, where $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$, θ is related to the concentration of impurities in the device and therefore to the maximum change in propagation delay, $y_n - y_0$ is the change observed in the propagation delay between times t_n and t_0. The effect of temperature on the maximum degradation θ is given by $log(\theta) = A - (B/kT) + \eta$, where $\eta \sim \mathcal{N}(0, \sigma_\eta^2)$, T_i is the absolute temperature, k the Boltzmans constant and h a random effect representing unobserved variability.

Doksum and Hoyland 1992 consider step stress accelerated testing, where failure is modeled in terms of accumulated decay reaching a threshold ω. Accumulated decay is assumed governed by a Wiener Process $W(y)$. The distribution of $W(y)$ depends on stress level $s(y)$. The stress level $s(y)$, in turn, is assigned to the device at each time point y. Time-to-failure is given by $Y = IG(y|\mu, \lambda)$. Their accelerated degradation model is given by: $W(y) = W_0(t + \alpha[y - t])$ if $y \geq t$ and $W(y) = W_0$ if $y < t$, where W_0 is the Wiener process under the nominal stress level 0. This model has two stress levels and a decay rate changing from η to $\alpha\eta$ as y crosses the stress change point at time t. In the two stress level case, the distribution of the failure time is $IG(\tau(y)|\mu, \lambda)$, $\tau(y) = y$ if $y \leq t$ and equal to $t + \alpha(y - t)$ if $y > t$. The model corresponds to making a monotonic transformation of time Y. They think of Y as the true (calendar) time, and $Z = \tau(Y)$ as the effective (non-accelerated) time. Lastly in their model they further parameterize α as a function of stress.

Meeker and Escobar 1993 review research and issues in accelerated testing, and make the point that there are two types of accelerated tests: i) Accelerated Life Tests (ALT) and ii) Accelerated Degradation Tests (ADTS). In ALTs one observes time-to-failure information and typically assumes a time-to-failure distribution. In ADTS one observes at one or more points in time the amount of degradation for a device

and typically assumes a model for degradation as a function of time. Traditional accelerated test statistical models assume a relationship(s) between the constant stress model parameters.

Whitmore and Schenkelberg 1997 consider a degradation process to be a Wiener diffusion process with a time scale transformation. The model incorporates Arrhenius extrapolation for high stress testing. A time transformation accommodates for a time dependent Wiener process drift parameter. The time transformation depends on the particular degradation mechanism. They encode a relationship between the parameters and stress level through a functional an Arrhenius model.

Meeker, Escobar and Lu 1998 give a review article on degradation modeling with accelerated life and degradation data. They assume that the degradation follows a path defined by: $y_{ij} = D_{ij} + \epsilon_{ij}$,$i = 1, \ldots, n$ (item) and $j = 1, \ldots, m_i$ (number of observations). In this model y is the predicted degradation path, D is the actual degradation path and $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ the error term. They define $D_{ij} = h(t_{ij}, \boldsymbol{\beta}_i)$; $\boldsymbol{\beta}_i = (\beta_{1i}, \ldots, \beta_{ki})$, and an Arrhenius model to describe the effect of temperature on the rate $R(temp) = fun(temp)$ of a simple first order chemical reaction. They define the acceleration factor $AF = R(temp)/R(temp_U)$. The chemical reaction model is given as $D(ttemp) = D_\infty \times (1 - exp[-R_U \times AF(temp) \times t])$. Solving for the failure time at temperature $T(temp) = T(temp_U)/AF(temp)$. Therefore if $T(remp_U) \sim Weib(\alpha_U, \beta)$ then $T(temp) \sim Weib(\alpha_U/AF(temp), \beta)$.

Owen and Padgett 1999 model the strength of materials with cumulative damage models. They assume that at i) each increment in stress (can be thought of as time) causes a random amount of non-negative damage D, subject to some distribution function f_D and ii) the system has a fixed theoretical strength (threshold) ψ, but the initial strength W is a random quantity. The cumulative damage after $n+1$ increments in stress is represented by: $X_{n+1} = X_n + D_n g(X_n)$, with $g(u)$ being the damage model, for example $g(u) = 1$ gives the additive damage model. They

let N be the number of increments of stress applied to the system until failure and express the survival probability as: $P(N > n) = \int_0^\infty F_n(w) f_W(w) dw$. The acceleration variable in their work is the gauge length of the specimen L, knowing that longer specimens fail faster or equivalently show smaller tensile strength. Therefore, the distribution of the initial strength variable is parameterized with L.

Onar and Padgett 2001 consider models for the strength of systems based on cumulative damage arguments. The models are based on a three-parameter inverse Gaussian distribution, and incorporate system size as a known acceleration variable. The stress level is denoted as L, the lifetime at stress level L as X_L, with cdf $F_{XL}(x)$ and $X \sim IG(\mu, \lambda)$. Under a cumulative damage model, a system is placed under a steadily increasing stress or load until failure occurs. It is assumed that the load is increased in small discrete increments and that each of these increments causes a random amount of non-negative damage $D > 0$, with cdf F_D. The systems initial strength is given by y, and the initial damage by X_0 (due to existing flaws or other damages). The cumulative damage is given as: $X_{n+1} = X_n + D_{n+1} h(X_n)$, and consider $h(\cdot) = 1$. Again the survival probability is expressed as a function of the initial strength W, $P(N > n) = \int_0^\infty F_n(w) f_W(w) dw$. The initial damage represents the reduction in strength (or reduction in lifetime) due to inherent flaws in the system. They assume that the flaw process can be described by a stationary Gaussian process which given the theoretical strength ψ, yields truncated Gaussian distribution function for W. They let S represent the total load after N increments (or calendar time), then $S \sim IG(\Lambda(\theta; L)/\zeta, \Lambda(\theta; L)/\sigma^2)$.

Bagdonavicius and Nikulin (20012) consider degradation models influenced by covariates under accelerated test conditions. They model degradation by a gamma process $Z(t) = \sigma^2 \gamma(t)$, where $\gamma(t) \sim Gamma(1, \nu(t)) = Gamma(1, m(t)/\sigma^2)$. They consider functional forms for the mean degradation $m(t)$ similar to Koenig and Carey (1991), where $m(t)$ is parameterized in some way $m(t) = m(t, g)$, where

$\boldsymbol{g} = (g_0, \ldots, g_m)^T$ are unknown parameters. They assume that the process has independent increments and therefore the moments are known. A failure caused by degradation occurs when $Z(t)$ reaches the value z_0, $T = inf(t|Z(t) \geq z_0)$. The stochastic process under the influence of covariates x, is given by: $Z_x(t) = \sigma^2\gamma \int_0^t exp(\beta^T x(s))ds$.

In addition to degradation models for latent degradation processes, there is also a need as we motivate in the thesis, for degradation models on partially observable degradation processes. In this context, there is no visible literature, and the work in this thesis we hope can help shed some light and inspire future research.

Bivariate degradation models have predominantly been used to analyze terminal data observations, mostly because degradation is latent. For longitudinal data, the bivariate model has not been extensively studied. Longitudinal treatment of markers in the context of bivariate Wiener models can be found in Sy et al. 1997, Henderson et al. 2000, and Guo and Carlin 2004, among a few selected others, and predominantly in HIV AIDS studies. No visible literature exists on bivariate longitudinal degradation models in reliability.

Failure thresholds in most of the work in bivariate degradation models are considered known and fixed for a given sample of devices. It is however desirable, and indeed a relevant topic in reliability analysis today, to accommodate random or uncertain failure thresholds. Literature on random failure thresholds within Wiener process degradation models, can be found in J. Lu 1995, Singpurwalla, 2010, Wang and Coit 2007. For bivariate degradation models, however, this is still an open area of research.

1.5 Contributions

We develop a new degradation model that extends the bivariate Wiener process model introduced by Whitmore et al. 1998 and allows us to incorporate:

1. Terminal degradation observations
2. Longitudinal marker observations
3. Variable failure thresholds

With the above extensions, we contribute to the improvement of bivariate latent degradation models, and more broadly to the field of reliability. Our new model draws strength from data on surviving devices in addition to failed ones and shows improved inference under a terminal data-structure. Results obtained in this thesis show that our model is suitable for:

1. Small failure-time samples
2. Reduced accelerated stress conditions
3. Non-predictive observable degradation data

2. DATA STRUCTURE AND NOTATION

The common characteristic in failure-tests of electronics is the generally latent degradation variable. This is why as discussed earlier, failure-test designs for electronics consider direct failures based on observable variables. Test centers typically invest in sophisticated failure analysis equipment that can be used to destructively and non-destructively examine the device. Typically destructive tests are used to determine the mode of failure and validate hypothesis regarding the targeted failure mechanism. Often when the there exists PoF models and identifiable degradation variables, these destructive tests can be used to measure the degradation, which ordinarily is latent.

Not all failure analysis tests are destructive. Non-destructive tests typically rely on more sophisticated and expensive equipment, and are therefore not used often. Generally, pre-failure tests are not common because they interfere too much with the controlled test environment and add undesirable unexplainable levels of bias and variation. The main idea is that modern failure testing facilities are equipped to be able to measure degradation, and we argue here that this information should be used in modeling lifetimes in electronics and overall for developing PHM technology.

2.1 Failure and degradation data in electronics

Next we present three examples of different types of direct failures observed in electronics during failure tests, and two examples of indirect failures. Definitions for direct and indirect failures are given in chapter 1.

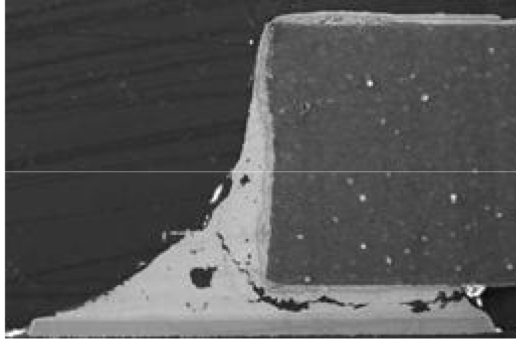


Fig. 2.1: Crack formation in a solder joint

2.1.1 Examples of direct failures

Direct failures are characteristic of electronics, in that although they are defined by an observable degradation variable, that variable is not predictive of failure. Therefore, through FMMEA and experimentation, more valuable degradation variables are determined. We consider in the following material, three types of failures in electronic components:

1. Interconnect failures due to solder joint crack formation,

2. Dielectric failures due to growth of tin whiskers

3. Insulated gate bipolar transistor (IGBT) failures due to die-attach or gate-oxide damage

1. Interconnect failures typically occur due to the formation of a cracks in solder joints, printed circuit board traces and connectors. Often the crack starts at the surface of the joint, and propagates inwards [57] (see figure 2.1). Failure is defined by a DC open circuit, which occurs when the solder joint first ruptures and current stops flowing. The DC resistance of the material is typically used as the degradation variable.

According to Kwon et al. [58], however, DC resistance, often responds too little (for example, changes in DC resistance are often obscured by the environmental noise in a real life situation) or too late (for example, after the crack is large enough to result in a DC open circuit). Therefore, they argue, DC resistance measurements would not be expected to provide early indications of interconnect degradation. Instead, they propose using RF impedance as the degradation variable. They argue that due to the *skin effect*, a phenomenon wherein signal propagation at high frequencies is concentrated near the surface of a conductor, RF impedance exhibits increased sensitivity to small cracks initiated at the surface of an interconnect. Time domain reflectrometry (TDR) is a method used to measure RF impedance.

2. A dielectric (or insulator) is a material that resists the flow of electric charge, and they fail when they collapse (typically due to high voltage) and start to conduct. The collapse of insulator is sometimes caused by the growth of protruding material called whiskers, that can grow to create a conductive path between two differently biased conductors. Tin whiskers are electrically conductive, crystalline structures of tin that sometimes grow from surfaces where tin (especially electroplated tin) is used as a final finish. Electronic system failures may occur due to short circuits caused by tin whiskers that bridge closely-spaced circuit elements maintained at different electrical potentials. Failure is defined by a DC open circuit, and the degradation variable used is again DC resistance.

DC resistance, is again not useful for failure prediction since it suffers from the same effects as in the previous example. Instead, degradation is explained in terms of the growth of whiskers, for example, the average length of all whiskers larger than a predefined nominal length (threshold), or the number of whiskers larger than a predefined threshold, etc. The density and length of whiskers can be measured in a laboratory setting [59].

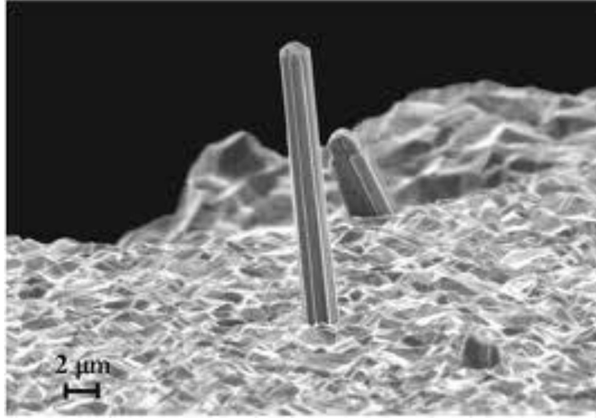


Fig. 2.2: Whisker growth, seen at initial stage

3. IGBTs are used in many modern appliances to regulate DC voltage. The failure modes for the IGBT include short circuits, increased leakage current, or loss of gate control. The identified failure mechanisms are the gate-oxide damage and the die-attach delamination. The potential failure causes are high electric fields and/or high temperatures. Failure is defined by *latch-up*, a term used for integrated circuits (ICs) to describe a particular type of short circuit which can occur in an improperly designed circuit. Failure is defined by a DC open circuit, or equivalently low DC resistance.

Patil et al. [60], found alternative degradation variables in the threshold voltage believed to be a response to gate-oxide damage and the collector-emitter-on voltage believed to be a response to die-attach damage. Both these variables show a trend as a function of time. In addition, a third degradation variable, can be determined using scanning acoustic microscopy (SAM) analysis. SAM is a non-destructive analysis technique that can be used to detect delaminations and voids in microelectronic packages. Figure 2.3 shows the SAM analysis on a pre and post aged IGBT part, with visible degradation. In the images, bright areas indicate increased reflection due to degradation. In this case, the amplitude of the reflected signal can be used as a measure of degradation.

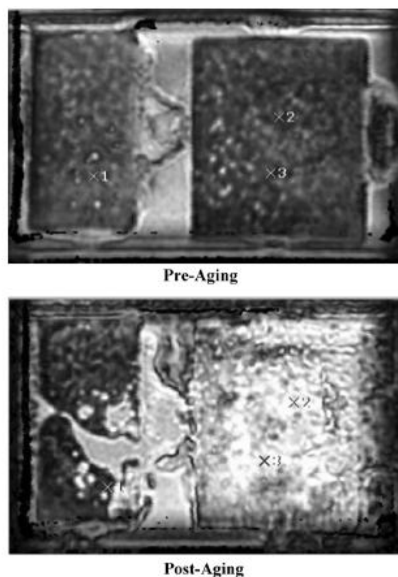


Fig. 2.3: Pre and post aging of an IGBT part. Increased reflectivity picked up by SAM indicates degradation

The above examples motivate the need for degradation variables better suited for failure-time prediction. In all three cases failure analysis methods are used to measure more useful degradation variables, which theoretically are more faithful to the underlying failure mechanism(s). Some failure analysis methods, like SAM, can only be performed after failure has been determined, while others, like TDR can be performed more frequently. In this thesis, we consider only terminal measurements on degradation and longitudinal observations on markers.

2.1.2 Examples of indirect failures

Computers, observed from the system level, which includes both hardware and software functionality, exhibit indirect failures. Examples of indirect failures in computers are: sudden shut-down or freeze (blue screen). Hardware variables, like the motherboard temperature, %CPU throttle, the fan speed, and many more are

readily measurable in most personal computers. Software variables are also readily available, specifically, event indicator variables that flag the occurrence of various types of programmed errors and warnings. Event processes can be used to model either the degradation or a marker to degradation over time.

Observed from the system level, Gas-Turbine engines, also exhibit indirect failures, like when the engine unexpectedly stops producing power. Before engines stop producing power, they exhibit other intermediate events, that are typically used as failures. One such event is called compressor surge or stall, which results when the compressor can no longer compress the incoming air. Typically most turbine engine failures result from blade degradation due to fatigue and creep of the materials.

2.2 Data-Structures

In chapter 1 we discussed the possible modeling complications that can arise due to the type of data and its structure. We talked about the lack of adequate failure time samples observed in failure-tests. We also discussed data situations where the degradation variable is unobserved and has to be tracked by a marker. In this chapter we present three data structures:

1. Terminal-marker only, abbreviated as (TM)
2. Terminal-marker and degradation, abbreviated as (TMD)
3. Longitudinal and terminal-marker with k intermediate marker-measurements, plus terminal degradation, abbreviated as (TMDL)

The TM data structure is based on Whitmore et al. 1998, and forms the basic data-structure to which we augment terminal degradation and longitudinal marker data. We consider one and two marker observations as separate cases. Table 2.1 summarizes the variables used under each data-structure.

Tab. 2.1: Summary of variables, their names, description and realizations, under each data-structure

Variable Name	Description	Terminal Structure		Longitudinal Structure
		TM	TMD	TMDL
T	Event time $= S \wedge C$	$s \wedge \tau$	$s \wedge \tau$	$s \wedge \tau$
S	Failure-time	s	s	s
C	Censored-time	τ	τ	τ
Δ	Failure indicator $=I_{[S \leq C]}$	(0,1)	(0,1)	(0,1)
Y	Marker	$y(T)$	$y(T)$	$\boldsymbol{y} = (y(t_0), \ldots, y(T))$
Z	Covariate	$\boldsymbol{z}(T)$	$\boldsymbol{z}(T)$	$\boldsymbol{z} = (\boldsymbol{z}(t_0), \ldots, \boldsymbol{z}(T))$
X	Degradation	$x(s)$	$x(T)$	$x(T)$

2.2.1 *Terminal Structure*

A cohort of n independent devices are monitored over a fixed time period $[0, \tau]$, where the end of testing at τ is considered nonrandom, and known, and typically chosen as a cost and time constraint. The number of devices that fail by time τ is denoted by $q = \sum_{i=1,\ldots,n} I(s_i < \tau)$, a binomial random variable that describes the natural proportion of failed to survived devices in a fixed test period $[0, \tau]$. The number of devices that survive by time τ is denoted by p, such that $p + q = n$. The failure threshold is again represented by a scalar a, and is assumed known and fixed.

Under the TM data structure, for each device we observe the lifetime $T = min(S, \tau)$, the marker $Y(T)$ and the degradation $X(T)$ at T. For failed devices we observe $T = s$, $Y(s)$ and $X(s) = a$. For surviving devices we observe $T = \tau$, and $Y(\tau)$. The TM data-structure is a subset of the data available under the TMD data-structure. In the TMD data structure we also observe terminal degradation on surviving devices, that is $X(\tau)$. In both TM and TMD we assume that the degradation process starts at time $t = 0$ is equal to zero, $X(0) = 0$.

2.2.2 Longitudinal Structure

Similarly, a cohort of n independent devices are monitored over a fixed time period $[0, \tau]$, where the end of testing at τ is considered nonrandom, and known, and typically chosen as a cost and time constraint. The number of devices that fail by time τ is again denoted by q, and the number that survive by p, $p + q = n$. The failure threshold is again represented by a scalar a, and is assumed known and fixed. For each device we collect information on a fixed covariate process $Z(t)$, and a marker process $Y(t)$ correlated with the latent degradation process $X(t)$, at a succession of scheduled increasing times $t = t_1, t_2, ...$ Equal spacing of the times t_j is not necessary; indeed, these times could be different for each device, but if random must be assumed independent of the processes $X(t), Y(t), Z(t)$. These observations can be made only up to the terminal observation time T, where $T = min(S, \tau)$ and the failure-time variable S is defined as the first (random) time at which the process $X(t)$ hits the threshold a.

For failed devices, vectors $\boldsymbol{Y} = (Y(t_1), \ldots, Y(t_K))$, $\boldsymbol{z} = (z(t_1), \ldots, z(t_K))$ and the degradation variable $X(S)$ at time S are observed, where $t_K = max\{t_j : t_j \leq T\}$, and K is the number of observations on the marker (random) before the failure. The degradation level for failed devices (those with $S < \tau$) is by definition equal to the failure threshold, that is, $X(S) = a$.

For surviving devices, vectors $\boldsymbol{Y} = (Y(t_1), \ldots, Y(t_m))$, $\boldsymbol{z} = (z(t_1), \ldots, z(t_m))$ and the degradation variable $X(\tau)$ at time τ are observed, where m is the total number of scheduled marker measurements in a test. The degradation level for surviving devices is expected to vary and be less than the threshold level, that is, $X(\tau) < a$.

3. ESTIMATION THEORY

3.1 Maximum Likelihood Estimators

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ be a set of estimator and $\mathbf{X} = (X_1, \ldots, X_n)$ be a sample of size n with an assumed *pdf* or pmf $f(\mathbf{X}|\boldsymbol{\theta})$. Then, knowledge of $\boldsymbol{\theta}$ yields knowledge of the entire population. Hence, it is natural to seek a method of estimating $\boldsymbol{\theta}$, i.e. a point estimator.

Definition 3.1 (Point Estimator). *A point estimator of a parameter θ, is a function $W(X_1, \ldots, X_n)$ of a sample, that is, any statistic is a point estimator of the same dimension as θ.*

In many cases, there is an obvious candidate for a point estimator, for example, a sample mean is a natural estimator of a population mean. However, when we leave a simple case like this, intuition will often lead us astray. Therefore, it is useful to have techniques of arriving at reasonable candidates for consideration.

The method of maximum likelihood is by far the most popular technique for deriving estimators. The following are some of the advantages of this estimation technique. Maximum likelihood provides a consistent approach to parameter estimation problems. This means that maximum likelihood estimates can be developed for a large variety of estimation situations. For example, they can be applied in reliability analysis to censored data under various censoring models. Also, it is well known that maximum likelihood methods have desirable mathematical and optimality properties. We will discuss these properties in the next section.

The disadvantages of maximum likelihood estimation include the following. The likelihood equations need to be specifically worked out for a given distribution and estimation problem. The mathematics is often non-trivial, particularly if confidence intervals for the parameters are desired. The numerical estimation is usually non-trivial. Maximum likelihood estimates can be heavily biased for small samples and can be sensitive to the choice of starting values. Maximum likelihood estimation requires the adoption of strong assumptions about the joint density of the data, if the assumptions fails, MLEs may be inconsistent.

Consider the likelihood function $L(\theta|\mathbf{X}) = L(\theta_1, \ldots, \theta_k|X_1, \ldots, X_n)$.

Definition 3.2 (Maximum Likelihood Estimator). *For each sample point* $\mathbf{X}$*, let* $\hat{\theta}(\mathbf{X})$ *be the value of the parameter vector at which* $L(\theta|\mathbf{X})$ *attains its maximum (over a neighborhood of the true parameter) as a function of* θ*, with* $\mathbf{X}$ *being held fixed. A maximum likelihood estimator (MLE) of the parameter vector* θ *based on a sample* $\mathbf{X}$ *is* $\hat{\theta}(\mathbf{X})$*.*

3.2 Methods of Evaluating Estimators

For ease of exposition, in this section we assume that the parameter vector consists of a single parameter θ. All of the results presented here extend to the case of a multi-parameter distribution.

3.2.1 Finite Sample Measures

We first discuss finite sample measures of the quality of an estimator, beginning with its mean squared error.

Definition 3.3 (Mean Squared Error). *The mean squared error (MSE) of an estimator* W *of a scalar parameter* θ *is the function of* θ *defined by:* $E_\theta(W - \theta)^2$*.*

Other distance measures between the parameter and its estimator can be used as a measure of performance of an estimator. In general, any increasing function of the absolute difference $|W - \theta|$ can be considered, but the MSE has at least two advantages over other distance measures. First, it is quite tractable analytically and, second, it has the interpretation:

$$E_\theta(W-\theta)^2 = Var_\theta W + (E_\theta W - \theta)^2 = Var_\theta W + (Bias_\theta W)^2$$

where we define bias of an estimator as follows:

Definition 3.4 (Bias of an Estimator). *The bias of a point estimator W of a parameter θ is the difference between the expected value of W and θ, that is $Bias_\theta W = E_\theta W - \theta$. An estimator whose bias is identically (in θ) equal to 0 is called unbiased and satisfies $E_\theta W = \theta \ \forall \ \theta$.*

Thus, the MSE incorporates two components, one measuring the variability of an estimator and the other its bias (accuracy). To find an estimator with good MSE properties, we need to find estimators that control both variance and bias.

Definition 3.5 (UMVUE Estimator). *An estimator W^* is the best unbiased estimator of θ if it is unbiased and for any other unbiased estimator W, we have $Var_\theta W^* \leq Var_\theta W \ \forall \ \theta$. W^* is also called a uniform minimum variance unbiased estimator (UMVUE) of θ.*

Theorem 3.1 (Cramer-Rao Inequality). *Let $X_1, \ldots, X_n$ be a sample with pdf $f(\mathbf{x}|\theta)$ and let $W(X_1, \ldots, X_n)$ be any estimator satisfying:*

$$\frac{d}{d\theta} E_\theta W(\mathbf{X}) = \int \frac{\partial}{\partial \theta}[W(x) f(x|\theta)]dx \tag{3.1}$$

and $Var_\theta W(\mathbf{X}) < \infty$, then

$$Var_\theta(W(\mathbf{X})) \geq \frac{\left(\frac{d}{d\theta}E_\theta W(\mathbf{X})\right)^2}{E_\theta\left(\left(\frac{\partial}{\partial\theta}logf(\mathbf{X}|\theta)\right)^2\right)} \tag{3.2}$$

This theorem specifies the lower bound on $Var_\theta W$, thus an estimator W which attains this variance is UMVUE. Such an estimator is also referred to as **finite-sample efficient**. The quantity $E_\theta((\partial/\partial\theta logf(\mathbf{X}|\theta))^2)$ is called the **Information number** or **Fisher information** of the sample. As the information number gets bigger and we have more information about θ, we have a smaller bound on the variance of the best unbiased estimator.

Remark 3.1. *If the estimator $W(X_1,\ldots,X_n)$ is unbiased, the Cramer-Rao lower bound is simply the reciprocal of the Fisher Information.*

Remark 3.2. *In the multi-parameter case, where $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_k)$, we have a matrix analogue of the Fisher Information, which we call the Information Matrix. The elements in the matrix are defined as follows:*

$$\left(\mathcal{I}(\theta)\right)_{i,j} = E\left[\left(\frac{\partial}{\partial\theta_i}logf(\mathbf{X}|\theta)\right)\left(\frac{\partial}{\partial\theta_j}logf(\mathbf{X}|\theta)\right)|\theta\right] \tag{3.3}$$

The Cramer-Rao lower bound for θ_i is then the ith diagonal element of the inverse of the Information Matrix.

Remark 3.3. *A sample-based estimate of the Fisher Information is the Observed Fisher Information, which is defined as $-\partial^2/\partial\theta^2 logL(\theta|\mathbf{X})|_{\theta=\hat{\theta}}$.*

Notice that the estimation of the Fisher information is a two-step process, first we approximate the expression for the Fisher information with the sample-based analogue and then we estimate the resulting approximation by replacing θ with $\hat{\theta}$. A computation result which aids in the application of the theorem is stated in the lemma below.

Lemma 3.1. *If $f(\mathbf{x}|\theta)$ satisfies*

$$\frac{d}{d\theta}E_\theta\left(\frac{\partial}{\partial\theta}logf(\mathbf{X}|\theta)\right) = \int\frac{\partial}{\partial\theta}\left[\left(\frac{\partial}{\partial\theta}logf(\mathbf{x}|\theta)\right)f(\mathbf{x}|\theta)\right]dx$$

then

$$E_\theta\left(\left(\partial/\partial\theta logf(\mathbf{X}|\theta)\right)^2\right) = -E_\theta\left(\frac{\partial^2}{\partial\theta^2}logf(\boldsymbol{X}|\theta)\right)$$

Suppose we have two estimators W_1 and W_2 of θ. W_1 is said to dominate W_2 if $E_\theta(W_1-\theta)^2 \leq E_\theta(W_2-\theta)^2$ holds for all θ, with strict inequality holding somewhere.

3.2.2 Asymptotic Evaluations

Asymptotic properties of estimators describe the behavior of estimators as the sample size becomes infinite. The power of asymptotic results is that when we let the sample size become infinite, calculations simplify. On a practical level, these results can be thought applicable to large, as opposed to infinite samples.

Definition 3.6 (Consistent Sequence of Estimators). *A sequence of estimators $W_n = W_n(X_1, \ldots, X_n)$ is a consistent sequence of estimators of the parameter θ if, for every $\epsilon > 0$ and every $\theta \in \Theta$,*

$$lim_{n\to\infty}P_\theta(|W_n-\theta| < \epsilon) = 1$$

Informally, this says that as the sample size becomes infinite, the estimator will be arbitrarily close to the parameter with high probability. Equivalently, we can say that the probability that a consistent sequence of estimators misses the true parameter is arbitrarily small (or converges to 0).

Theorem 3.2 (Asymptotic Normality). *An asymptotically normal sequence of estimators $W_n(X_1, \ldots, X_n)$ of θ is a consistent sequence whose distribution around the true parameter θ approaches a normal distribution with standard deviation shrink-*

ing in proportion to $1/\sqrt{n}$ *as the sample size* n *grows. Using* $\xrightarrow{D}$ *to denote convergence in distribution,* W_n *is an asymptotically normal sequence of estimators if*

$$\sqrt{n}(W_n - \theta) \xrightarrow{D} \mathcal{N}(0, V)$$

for some V*, which is called the asymptotic variance of the estimator.*

In the spirit of the Cramer-Rao lower bound, there is an optimal asymptotic variance.

Definition 3.7 (Asymptotic Efficiency). *A sequence of estimators* W_n *is asymptomatically efficient, if it is asymptotically normal and the asymptotic variance is identical to the Cramer-Rao lower bound.*

In practice, we do not deal with infinite samples and therefore it makes sense to talk instead of "large-sample efficiency". Informally, an estimator W_n of θ is large-sample efficient if, for a large enough sample size n, its empirical distribution is centered around the true value of θ and it's empirical variance is equal to V/n plus a remainder of smaller order than $1/n$, where V is the approximated Cramer-Rao lower bound. To approximate the Cramer Rao lower bound for a given sample size n, we approximate the expected information number with the observed information number, $-\partial^2/\partial\theta^2 logL(\theta|\mathbf{X})|_{\theta=\hat{\theta}}$.

Large-sample efficiency of a sequence of estimators $W_n(X_1, \ldots, X_n)$ can be checked via simulation by sampling k independent vectors $(\mathbf{X}_n^1, \ldots, \mathbf{X}_n^k)$ of size n, computing $\mathbf{W}_n = (W_n^1, \ldots, W_n^k)$ and studying the sampling distribution of $\mathbf{W}_n$ as n grows. For an asymptotically efficient sequence, we expect the sampling distribution to look more and more normal, centered around the true parameter, with the empirical estimate of the variance asymptotic to V/n where V is the approximated (via observed information number) Cramer Rao lower bound for large n.

Definition 3.8 (**Asymptotic Relative Efficiency**). *If two estimators W_n and V_n satisfy $\sqrt{n}[W_n - g(\theta)] \rightarrow \mathcal{N}[0, \sigma_W^2]$, and $\sqrt{n}[V_n - g(\theta)] \rightarrow \mathcal{N}[0, \sigma_V^2]$ in distribution, the asymptotic relative efficiency (ARE) of V_n with respect to W_n is:*

$$ARE(V_n, W_n) = \frac{\sigma_W^2}{\sigma_V^2} \tag{3.4}$$

Definition 3.9 (**Estimated Asymptotic Relative Efficiency**). *Estimated asymptotic relative efficiency is defined as the ratio of asymptotic variance estimates as follows:*

$$\widehat{ARE}(W_1, W_2) = \frac{\widehat{\sigma_W}^2}{\widehat{\sigma_V}^2} \tag{3.5}$$

By Definition 3.9, the asymptotic variances are estimated by the Cramer-Rao lower bound.

3.3 Asymptotic Properties of Maximum Likelihood Estimators

Theorem 3.3 (**Consistency of MLEs**). *Let $X_1, X_2, \ldots,$ be iid $f(x|\theta)$, and let $L(\theta|\mathbf{x}) = \prod_{i=1}^n f(x_i|\theta)$ be the likelihood function. Let $\hat{\theta}$ denote the MLE of θ. Under regularity conditions [61] on $f(x|\theta)$ and, hence, $L(\theta|\mathbf{x})$, for every $\epsilon > 0$ and every $\theta \in \Theta$,*

$$lim_{n\rightarrow\infty} P_\theta(|\hat{\theta} - \theta| \geq \epsilon) = 0$$

Theorem 3.4 (**Asymptotic Efficiency of MLEs**). *Let $X_1, X_2, \ldots,$ be iid $f(x|\theta)$, let $\hat{\theta}$ denote the MLE of θ. Under the regularity conditions [61] on $f(x|\theta)$, and, hence, $L(\theta|\mathbf{x})$,*

$$\sqrt{n}[\hat{\theta} - \theta] \rightarrow \mathcal{N}[0, V(\theta)]$$

where $V(\theta)$ is the Cramer-Rao Lower Bound. That is, $\hat{\theta}$ is a consistent and asymptotically efficient estimator of θ.

Note that asymptotic efficiency is defined only when the estimator is asymptotically normal and, asymptotic normality implies consistency.

3.4 Computing

3.4.1 Observed Fisher Information Matrix

The Cramer-Rao lower bound is approximated by the observed Fisher Information given by:

$$\hat{I}(\theta)_{i,j} = -\frac{1}{n}\sum_{i=1}^{n} \bigtriangledown^{\otimes 2} log f(\boldsymbol{X}|\hat{\boldsymbol{\theta}}(\boldsymbol{X})) \tag{3.6}$$

The matrix of second partial derivatives is approximated by the Hessian matrix $\boldsymbol{H}$, which is computed using finite difference schemes to numerically evaluate first and second order derivatives of the log-likelihood function.

3.4.2 Multivariate Integration using Gaussian Quadratures

Likelihood functions for longitudinal data as discussed in chapter 5, require the evaluation multivariate integrals. Specifically we are interested in numerically integrating analytically intractable joint densities. For smooth functions, however, like those formed by the products of smooth functions by jointly Gaussian densities, where smoothness is used to ensure that the integrands are well approximated by polynomials over most of the range of the density, Gaussian Quadrature rules are generally preferred. This is because the integrand is well approximated by a polynomial function.

We approximate the integral of a function f by the sum of its functional values at a set unequally spaced points x_i (nodes), multiplied by appropriately chosen weighting coefficients w_i, $i = 1, \ldots, N$, where is fixed and known. The basic idea of Gaussian quadratures is to allow flexibility in not only choosing the weighting coefficients, but also the location of the location of the nodes.

We calculate nodes $x_1, \ldots, x_n$ and coefficients $w_1, \ldots, w_n$ such that

$$\int_a^b f(x)dx \approx \sum_{i=1}^{N} w_i f(x_i) \tag{3.7}$$

The theory behind Gaussian quadratures is closely tied to that of orthogonal polynomials. We consider the most common and useful case: *Legendre* polynomials $P(x)$.

Theorem 3.5 (Fundamental theorem of Gaussian Quadratures). *The nodes of the N-point Gaussian quadrature formula given by equation* (3.7) *are precisely the roots of the orthogonal polynomial* $P(x)$ *for the same interval and weighting function.*

Given the nodes x_i, $i = 1, \ldots, N$ one can find the weights w_i by solving a set of linear equations for the weights such that the quadrature (3.7) gives the correct answer for the integral of the first $N - 1$ orthogonal polynomials. There are more efficient ways of computing weighting coefficients, such as through the eigenvalue decomposition of the symmetric, tridiagonal Jacobi matrix. For further details we point to the following references [62].

We implement Gaussian quadratures by first selecting N, and computing x_i and w_i, $i = 1, \ldots, N$, and storing the vectors into memory. When integration is required we apply (3.7) using the stored nodes and corresponding nodes.

3.4.3 Nonlinear Optimization

Maximum likelihood estimation involves finding the maxima of the multidimensional log-Likelihood function, constrained by the range of the parameter values $\boldsymbol{\theta}$. To find the maxima we need to solve a nonlinear constrained optimization problem. In this section we simply define the optimization problem we face, without any exposition of the relevant theory.

4. WIENER PROCESS AS A DEGRADATION MODEL

Degradation is typically thought to act on a device over time, either as a series of adverse events or as a continuous process. Degradation is also associated with failure or other intermediary health related events. In this work we model the evolution of degradation as a stochastic process, namely through Wiener process with drift $\{X(t)\}$, which is defined later in this chapter. The idea is that such a process has properties useful for modeling the accumulation of damage, represented by degradation, which starts at some nominal low level at the start of life and reaches a failure-threshold at end of life. In this chapter we work with a fixed and known failure-threshold, and later in chapter 8 investigate a variable failure-threshold.

A Wiener process model is not always appropriate as a degradation model, especially for strictly monotonically increasing degradation. However, due to the possibility of healing, the fluctuations of the Wiener process, are appropriate for modeling degradation in electronics. We extend Whitmore's bivariate Wiener model to accommodate a new data-structure that includes degradation, in addition to terminal marker observations. In chapter 5, we utilize this extension to model longitudinal marker observations.

4.1 Mixed-Type Densities

The joint densities in the Whitmore and our extended model are of mixed type. They require the joint relationship between continuous and discrete random variables. Degradation and marker variables are considered continuous type ran-

dom variables, while the failure-indicator variable is considered a discrete random variable. In this section we define some of the machinery for specifying mixed type joint densities.

Let U be a continuous random variable with density $f(u)$ and let V be a discrete random variable taking values v_i, $i = 1, 2, \ldots, n$, with probabilities $P(v_i)$. To characterize the relationship between U and V we specify the conditional density $f(u|v_i)$ or the conditional probability $P(v_i|u)$. For any $a \le b$ the conditional density $f(u|v_i)$ must satisfy:

$$\int_a^b f(u|v_i)dx = P(a \le U \le b|V = v_i)$$

Assuming that all continuous-variable densities and joint densities are continuous functions of their arguments u, the conditional probability $P(v_i|u)$ is defined as the limit of $P(v_i|u \le U \le u + \epsilon)$ as ϵ goes to zero.

$$\begin{aligned} P(v_i|u) &= \lim_{\epsilon\to 0} P(V = v_i|u \le U \le u + \epsilon) \\ &= \lim_{\epsilon\to 0} \frac{P(V = v_i, u \le U \le u + \epsilon)}{P(u \le U \le u + \epsilon)} \\ &= \lim_{\epsilon\to 0} \frac{P(u \le U \le u + \epsilon|V = v_i)P(V = v_i)}{P(u \le U \le u + \epsilon)} \\ &= \lim_{\epsilon\to 0} \frac{\int_u^{u+\epsilon} f(q|v_i)dqP(V = v_i)}{\int_u^{u+\epsilon} f(q)dq} \end{aligned}$$

If $f(u)$ and $f(u|v_i)$ are all continuous function of u, then by the mean value theorem:

$$P(v_i|u) = \frac{f(u|v_i)P(v_i)}{f(u)} \tag{4.1}$$

The joint distribution of U and V is specified through function $\psi(u, v_i)$ such that $P(a \le U \le b, V \in \mathcal{V}) = \int_a^b \sum_{i\in I} \psi(u, v_i)du$, where I is a subset of integers

$(1, 2, \ldots, n)$ and $\mathcal{V} = \{v_i | i \in I\}$.

$$\int_a^b \sum_{i \in I} f(u|v_i) P(v_i) dx = \sum_{i \in I} P(a \leq U \leq b | V = v_i) P(V = v_i)$$
$$= \sum_{i \in I} P(a \leq U \leq b, V = v_i) = P(a \leq U \leq b, V \in \mathcal{V})$$

The product $f(u|v_i)P(v_i)$ plays the role of the bivariate function $\psi(u, v_i)$. Hence by equation (4.1) so does $P(v_i|u)f(u)$. Material on mixed-type densities is taken from [71].

4.2 Degradation Model

The basic degradation model used in this work is given by the Wiener process with drift:

$$X(t) = x_0 + \nu_X t + \sigma_X W_X(t) \tag{4.2}$$

A Wiener process $X(t)$ with drift ν_X and variance σ_X^2 has stationary and independent Gaussian increments with probability density given by:

$$f_{X(t)}(x) = \frac{1}{\sigma_X \sqrt{t}} \phi \left(\frac{x - x_0 - \nu_X t}{\sigma_X \sqrt{t}} \right) \tag{4.3}$$

It is assumed that each device experiences its own degradation process which is independent of other devices. Devices from the same "family", having the same design, are assumed to have the same drift and variance parameters. Covariates can be used to model the heterogeneous drift in the degradation process as a function of some other parameters $\alpha, \beta, \ldots$. In electronics, accelerated lifetimes are most common, induced by higher than normal experimental stress factors z, such as temperature, humidity and pressure. It becomes important to model the influence of such stressors on the rate of degradation. We discuss this further in chapter 7.

4.3 Definition of a Wiener Process

The treatment of the Wiener process starts by defining it as a process on a probability space $(\Omega, \mathcal{F}, P)$. Assume also that for every $t \in [0, \infty)$ we are given a σ-algebra $\mathcal{F}_t \subset \mathcal{F}$ such that $\mathcal{F}_s \subset \mathcal{F}_t$ for $t \geq s$. We call such a collection of σ-algebras an (increasing) filtration of σ-algebras.

Definition 4.1. *A process W_t on a probability space $(\Omega, \mathcal{F}, P)$ is called a Wiener process if:*

- *Sample paths $W_t(\omega)$ are continuous function of t for almost all ω.*
- *For any $k \geq 1$ and $0 \leq t_1 \leq \ldots \leq t_k$, the random vector $(W_{t1}, \ldots, W_{tk})$ is Gaussian with covariance matrix $\Sigma(t_i, t_j) = E(W_{ti}, W_{tj}) = t_i \wedge t_j$, where $1 \leq i$, $j \leq k$, and where $s \wedge t = \min(s, t)$*

[63]

Lemma 4.1. *A process W_t on a probability space $(\Omega, \mathcal{F}, P)$ is a Wiener process if and only if:*

- *Sample paths $W_t(\omega)$ are continuous function of t for almost all ω*
- *$W_0(\omega) = 0$ for almost all ω*
- *For $0 \leq s \leq t$, the increment $W_t - W_s$ is a Gaussian random variable with mean zero and variance $t - s$*
- *Random variables $W_{t_0}, W_{t_1} - W_{t_0}, \ldots, W_{t_k} - W_{t_{k-1}}$ are independent for every $k \geq 1$ and $0 = t_0 \leq t_1 \leq \ldots \leq t_k$.*

[63]

Remark 4.1 (**Distribution**). *Since* $W_t - W_s \sim \mathcal{N}(0, |t-s|)$*, it follows that* $\xi := (W_t - W_s)|t-s|^{-1/2} \sim \mathcal{N}(0,1)$ *for* $t \neq s$ *and the distribution of* $W_t - W_s$ *has the density:*

$$f(x) = \frac{1}{\sqrt{2\pi|t-s|}} exp\left(-\frac{x^2}{2|t-s|}\right) \tag{4.4}$$

Definition 4.2 (**Independence of sigma-algebras**). *In the probability space* $(\Omega, \mathcal{F}, \boldsymbol{P})$*, let* $\mathcal{F}_1, \ldots, \mathcal{F}_k$ *be sub-sigma algebras of* $\mathcal{F}$*. We say that these sigma-algebras are independent if:*

$$Pr\left(\bigcap_{i=1}^{k} A_i\right) = \prod_{i=1}^{k} Pr(A_i)$$

for all $A_i \in \mathcal{F}_i$*.*

Definition 4.3 (**Independence of random variables**). *If* $\{W_1, \ldots, W_n\}$ *is some collection of random variables, we say that they are independent if* $\{\sigma\{W_i\}\}$ *are independent.*

Theorem 4.1. *Let* $\boldsymbol{W} = (W_1, \ldots, W_n)$ *be an absolutely continuous random vector. Then,* $W_1, \ldots, W_n$ *are independent if and only if the joint and marginal densities are related through the equation:* $f_{\boldsymbol{W}}(w_1, \ldots, w_n) = \prod_{i=1}^{n} f_{W_i}(w_i)$ [64].

Definition 4.4 (**Independent Random Vectors**). *The random vectors* W_1 *and* W_2 *are independent if for any Borel set* $\mathcal{B}_1$ *and* $\mathcal{B}_2$*, which are subsets of the respective value spaces of* W_1 *and* W_2*,* $P(W_1^{-1}(\mathcal{B}_1) \cap W_2^{-1}(\mathcal{B}_2)) = P(W_1^{-1}(\mathcal{B}_1))P(W_2^{-1}(\mathcal{B}_2))$ [65].

Definition 4.5 (**Uncorrelated Random Vectors**). *The random vectors* $\boldsymbol{W}_1$ *and* $\boldsymbol{W}_2$ *are uncorrelated if* $Cov(\boldsymbol{W}_1, \boldsymbol{W}_2) = 0$ [65].

According to lemma 4.1, there always exists a filtration with respect to which W_t is a Wiener process.

Definition 4.6. *Let W_t be a $\{\mathcal{F}_t\}$-adapted Wiener process; assume that for any t, $h \geq 0$ the random vector $W_{t+h} - W_t$ and σ-algebra $\mathcal{F}_t$ are independent. Then we will say that W_t is a Wiener process with respect to $\{\mathcal{F}_t\}$, or that $(W_t, \mathcal{F}_t)$ is a Wiener process.*

Theorem 4.2 (The Markov Property). *Let $(W_t, \mathcal{F}_t)$ be a Wiener process. Fix $t, h_1, \ldots, h_n \geq 0$. Then the vector $(W_{t+h_1} - W_t, \ldots, W_{t+h_n} - W_t)$ and the σ-algebra $\mathcal{F}_t$ are independent. Furthermore, $W_{t+s} - W_t$, $s \geq 0$, is a Wiener process* [67].

Theorem 4.2 says that, for every fixed time $t \geq 0$, the process $W_{t+s} - W_t$ starts fresh as a Wiener process "forgetting" everything that happened before time t. For a Wiener process, this property has a natural extension when t is replaced with a random time s, provided that s does not depend on the future in any way. To describe exactly what we mean by this, we continue with the following definition:

Definition 4.7 (Stopping time). *A stopping time is defined as a nonnegative random variable s such that for each (nonrandom) $t \geq 0$ the event $s \leq t$ is an element of the σ-algebra $\mathcal{F}_t$.*

Theorem 4.3 (Strong Markov Property). *Let $(W_t, \mathcal{F}_t)$ be a Wiener process and s an $\mathcal{F}_t$ stopping time. Assume that $P(s < \infty) = 1$. Let*

$$\mathcal{F}^W_{\leq s} = \sigma\{\{\omega : W_{u \wedge s} \in B\}, u \geq 0, B \in \mathcal{B}\}$$
$$\mathcal{F}^W_{\geq s} = \sigma\{\{\omega : W_{s+u} - W_s \in B\}, u \geq 0, B \in \mathcal{B}\}$$

Then the σ-algebras $\mathcal{F}^W_{\leq s}$ and $\mathcal{F}^W_{\geq s}$ are independent in the sense that for every $A \in \mathcal{F}^W_{\leq s}$ and $B \in \mathcal{F}^W_{\geq s}$ we have $P(AB) = P(A)P(B)$. Furthermore, $W_{s+t} - W_s$ is a Wiener process [67].

The strong Markov property gives justification to one of the most important properties of Wiener processes, a property that forms the basis for the likelihood

equations to follow; the *Reflection Principle*. The Reflection principle helps simplify otherwise complicated probability expressions related to the Wiener process.

Proposition 4.1 (**The Reflection Principle for a Wiener process with zero drift).** *Let $W(t)$ be a Wiener process with $\nu_X = 0$, $a > 0$, and $s_a = inf\{t : W(t) \geq a\}$, then:*

$$f_{W(t),I(S<t)}(w,1) = f_{W(t)}(2a - w)$$

The argument for proposition 4.1 is made by noticing that if $s_a < t$, then $W(t)$ is conditionally just as likely to be above or below level a by the same distance.

Proposition 4.2 (**The Reflection Principle for Wiener process with drift).** *Let $X(t)$ be a Wiener process with $X(0) = 0$, $\nu_X \neq 0$, $a > 0$, and $s_a = inf\{t : X(t) \geq a\}$, then:*

$$f_{X(t),I(S<t)}(x,1) = exp\left(\frac{2\nu_X(x-a)}{\sigma_X^2}\right) f_{X(t)}(2a - x) \tag{4.5}$$

The joint density in equation (4.5) is a building block for constructing likelihood equations under the longitudinal data-structures. We call this density the complimentary Wiener term, and we derive it in section 4.7.3.

4.4 The Inverse Gaussian Distribution - Lifetime Model

Like the Weibull and logNormal distributions the Inverse Gaussian distribution is used to model lifetime data. Chhikara and Folks 1977 [68] study the use of the inverse Gaussian distribution for a lifetime model and suggest its application for studying reliability aspects when there is a high occurrence of early failures. Unlike the Weibull and log Normal, the inverse Gaussian distribution is also physically justified as the first hitting time of a Wiener process to a threshold, which implies a natural applicability in studying degradation processes which lead to failure events.

This means that the lifetime distribution described by the inverse Gaussian density function depends on the drift and the variance of the degradation process. This relationship facilitates the development of degradation models that use mixed lifetime and degradation data and provide a natural framework of incorporating covariates.

Definition 4.8 (First Hitting Time). *Define S_a, the first time at which the process $X(u)$, starting from $x_0 = 0$, with drift $\nu_X \geq 0$ reaches level a, by: $S_a = inf\{u \geq 0 : X(u) \geq a\}$, $a > x_0$. If $a \neq 0$, then the distribution of S_a has the density $f(s) = 0$, for $s \leq 0$, and , $\forall\ s > 0$.*

$$f_{S_a}(s) = (2\pi\sigma_X^2)^{-1/2}|a|s^{-3/2}exp\left(-\frac{(a-\nu_X s)^2}{2\sigma_X^2 s}\right) \tag{4.6}$$

The density in equation (4.6) is called the Wald density or the inverse-Gaussian density. The inverse-Gaussian density can also be expressed in terms of its scale and shape parameters μ and λ.

$$f_{S_a}(s;\mu,\lambda) = \left(\frac{\lambda}{2\pi s^3}\right)^{1/2} exp\left(-\frac{\lambda(s-\mu)^2}{2\mu^2 s}\right) \tag{4.7}$$

As λ tends to infinity, the inverse-Gaussian starts to look more like a normal density. We know that in the case of positive drift $\nu_X > 0$ and $x_0 < a$, that: $\mu = (a - x_0)/\nu_X$ and $\lambda = (a - x_0)/\sigma_X^2$ [68]

4.5 Marker Model

With failure-time data samples drawn in dynamic environments where precise failure mechanisms are unknown, there is a need for degradation models that make use of auxiliary reliability information. Typically in electronics, failure-time data samples are small, and failure mechanisms are not well understood due to their complexity [77]. Auxiliary reliability information can be obtained in the laboratory

by measuring variables which characterize the *degradation* (aging) of each observed device over time.

In practice, degradation variables in electronics are typically latent (unobservable). For this reason, lifetime predictions are typically based on information from observed *marker* variables that track degradation. Markers, unlike covariates are random variables related to degradation through a parametric model with unknown parameters. For example, in printed circuit boards, the degradation variable for a degrading solder joint can be the length of a crack. It is impractical and often impossible to measure the length of a crack, and we depend therefore on markers to the crack-length, such as resistance and capacitance across the joint. From a cost perspective, degradation information is seen as being expensive, and Marker information as being cheap.

Our basic Marker model is given by a Wiener process with drift. The motivation for a Wiener marker process is mathematical convenience in expressing a bivariate marker-degradation relationship. The critical component of the marker-degradation model is the correlation coefficient, made available through the joint Gaussian relationship. Our basic marker model, like the degradation model, is given by a Wiener process with drift:

$$Y(t) = y_0 + \nu_Y t + \sigma_Y W_Y(t) \tag{4.8}$$

$$f_{Y(t)}(y) = \frac{1}{\sigma_Y \sqrt{t}} \phi \left(\frac{y - y_0 - \nu_X t}{\sigma_Y \sqrt{t}} \right) \tag{4.9}$$

4.6 Bivariate Wiener Model

Following [24] and [26], the basic analytical framework for a degradation-marker process is an independent-increments bivariate process, in which all paired increments $X(t) - X(s)$ and $Y(t) - Y(s)$ have correlation coefficient ρ, which does

not vary over time.

The vector $\{X(t_1), Y(t_2)\}$ has a bivariate normal distribution with mean vector $(\nu_X t_1, \nu_Y t_2)'$ and variance covariance matrix Σ_{XY}.

$$\Sigma_{XY} = \begin{pmatrix} t_1\sigma_X^2 & t_1 \wedge t_2 \rho\sigma_X\sigma_Y \\ t_1 \wedge t_2 \rho\sigma_Y\sigma_X & t_2\sigma_Y^2 \end{pmatrix} \tag{4.10}$$

We assume that Σ_{XY} is positive definite, $\nu_X \geq 0$ and $|\rho| > 0$ and close to 1. Weak correlation between the marker and the degradation variables will reduce the predictive efficiency of the model. With $c = \rho\sigma_Y/\sigma_X$, the conditional density of $Y(t)$ given $X(t)$ is:

$$f_{Y(t)|X(t)}(y|x) = \frac{1}{\sigma_Y\sqrt{(1-\rho^2)t}}\,\phi\left(\frac{y - y_0 - \nu_Y t - c(x - \nu_X t)}{\sigma_Y\sqrt{(1-\rho^2)t}}\right) \tag{4.11}$$

Proposition 4.3 (Multivariate Normal). *Define vectors* $\boldsymbol{X} = (X(t_1), X(t_2), \ldots, X(t_m))$ *and* $\boldsymbol{Y} = (X(r_1), X(r_2), \ldots, X(r_n))$*, evaluated at times* $\boldsymbol{t} = (t_1, t_2, \ldots, t_m)$ *and* $\boldsymbol{r} = (r_1, r_2, \ldots, r_n)$*,* $r_i \geq 0, t_i \geq 0$*. The joint density function of the vector* $(\boldsymbol{X}, \boldsymbol{Y})$ *is multivariate normal of the form:*

$$\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu_X} \\ \boldsymbol{\mu_Y} \end{bmatrix}, \begin{bmatrix} \Sigma_{\boldsymbol{XX}} & \Sigma_{\boldsymbol{XY}} \\ \Sigma_{\boldsymbol{YX}} & \Sigma_{\boldsymbol{YY}} \end{bmatrix}\right)$$

where

$\boldsymbol{\mu_X} \in \mathbb{R}^m$*,* $\boldsymbol{\mu_Y} \in \mathbb{R}^n$*, and the matrix blocks* Σ_{XX}*,* Σ_{XY}*,* Σ_{YX} *and* Σ_{YY}*. are respectively of size* $m \times m$*,* $m \times n$*,* $n \times m$*,* $n \times n$*.*

Proposition 4.4 (Conditional Multivariate Normal). *Using results from [70], the general formula for the conditional multivariate normal distribution of* $\boldsymbol{Y}$ *given* $\boldsymbol{X} = \boldsymbol{x}$ *is given by:*

$$\boldsymbol{Y}|\boldsymbol{X} \sim \mathcal{N}\left(\boldsymbol{\mu_{Y|X}}, \Sigma_{\boldsymbol{Y|X}}\right) \tag{4.12}$$

where, for $x_0 = 0, y_0 = 0$ *we have:*

$$\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{X}} = \nu_Y \boldsymbol{r} + \Sigma_{\boldsymbol{XY}} \Sigma_{\boldsymbol{XX}}^{-1} (\boldsymbol{x} - \nu_X \boldsymbol{t})$$

$$\Sigma_{\boldsymbol{Y}|\boldsymbol{X}} = \Sigma_{\boldsymbol{YY}} - \Sigma_{\boldsymbol{XY}} \Sigma_{\boldsymbol{XX}}^{-1} \Sigma_{\boldsymbol{YX}}$$

We are interested in a data-structure that we argue will provide sharper inference and more accurate lifetime predictions. The TMD data-structure introduced in section 2 uses degradation information on surviving devices. With the TMD data-structure, inference is based on the triplet of lifetime, marker, and degradation observations on each device. From now on the thesis we assume that $x_0 = 0, y_0 = 0$.

4.6.1 Conditional Independence in the Bivariate Wiener Model

In this section we define notation to prove conditional independence between sets of random vectors, and between random vectors and survival events, under a bivariate Wiener model. The results from this section will serve to simplify joint conditional densities in derivations to follow for parametric and predictive inference equations.

Definition 4.9. *Define vector* $\boldsymbol{t} = (t_j)_{j=1}^{\infty}$ *to be a discrete subset of* $\mathbb{R}^+$

Definition 4.10. *Let* $Q(t_j) = Y(t_j) - cX(t_j)$*, with* $c = \rho\sigma_y/\sigma_x \ \forall i$

Within the bivariate-Gaussian case $(X(t_j), Y(t_j))$, $Q(t_j) \sim \mathcal{N}(\mu_Q(t_j), \Sigma_Q(t_j))$ where $\mu_Q(t_j) = t_j(\nu_Y - c\nu_X)$, and $\Sigma_Q(t_j) = \sigma_Y^2 t_j (1 - \rho^2)$

Definition 4.11. *Let vector* $\boldsymbol{t}^l \subset \boldsymbol{t}$*. Define* $\boldsymbol{Q}^l = \{Q(u) : u \in \boldsymbol{t}^l\}$ *and* $\boldsymbol{X}^l = \{X(u) : u \in \boldsymbol{t}^l\}$

Lemma 4.2. *Let vectors* $\boldsymbol{t}^j, \boldsymbol{t}^i \subset \boldsymbol{t}$*. Then* $\boldsymbol{Q}^j \perp \boldsymbol{X}^j$*.*

Proof. Let $\boldsymbol{\Sigma^{j,i}} = Cov(\boldsymbol{Q}^j, \boldsymbol{X}^i)$. With $c = \rho\sigma_Y/\sigma_X$, $\boldsymbol{\Sigma^{j,i}} = \boldsymbol{0}$. Therefore, $\boldsymbol{Q}^j \perp \boldsymbol{X}^i$ due to the fact that a zero covariance implies independence for jointly Gaussian vectors. □

Corollary 4.2.1. *The entire process $Q(\cdot)$ is independent of the entire process $X(\cdot)$*

The corollary 4.2.1 of lemma 4.2 is stated in a more general way so that the finite number of evaluation points for $\boldsymbol{Q}$ are not necessarily the same as the finite number of evaluation points for $\boldsymbol{X}$.

Lemma 4.3. *Process $Q(\cdot) \perp (X(\cdot), I(S \geq t))$ where $I(S \geq t) = I(X(u) \leq a, u \leq t)$*

Proof. By definition, $I(S \geq t)$ is $\sigma(X(\cdot))$ measurable. Therefore, by corollary 4.2.1 $\sigma(X(\cdot)) \perp \sigma(Q(\cdot))$, and therefore $I(S \geq t) \perp Q(\cdot)$. This implies $Q(\cdot) \perp (X(\cdot), I(S \geq t))$.

□

Lemma 4.4. *Let $\boldsymbol{t}^j, \boldsymbol{t}^i \subset \boldsymbol{t}$. Let $S_a = inf\{u \in \boldsymbol{t}^i : X(u) \geq a\}$. Then $\boldsymbol{Q}^j \perp S_a$*

Proof. This follows from corollary 4.2.1 and measurability of S_a with respect to $\boldsymbol{X}^i$

□

Lemma 4.5. *Let $\tilde{S}_a$ be a shifted first-hitting time defined as $\tilde{S}_a(u) \equiv S_a(u - t_j)$, then, $f_{S_a|X(t_j),I(S\geq t_j)}(u|x,1) = f_{\tilde{S}_{a-x}}(u)$, $u \geq t_j$, where*

$$f_{\tilde{S}_{a-x}}(u) = \frac{a-x}{\sqrt{2\pi\sigma_X^2(u-t_j)^3}} exp\left(-\frac{(a-x-\nu_X(u-t_j))^2}{2\sigma_X^2(u-t_j)}\right) \tag{4.13}$$

Proof.

$$\begin{aligned}
&S_a|(X(t_j) = x, I(S \geq t_j) = 1) = \\
&= inf\{u \geq t_j : X(u) \geq a | X(t_j) = x, S \geq t_j\} \\
&= inf\{u \geq t_j : X(u - t_j) \geq a - x | X(t_j) = x, S \geq t_j\} \qquad (1) \\
&= inf\{v \geq 0 : X(t_j + v) - X(t_j) \geq a - x | X(t_j) = x, S \geq t_j\}
\end{aligned}$$

By theorem 4.3, because $X(t_j+v)-X(t_j)$ is $\mathcal{F}^X_{\geq t_j}$ measurable, it is independent of $(X(t_j), I(S \geq t_j))$ which is $\mathcal{F}^X_{\leq t_j}$ measurable. Therefore we can simplify (1) to $inf\{u \geq t_j : X(u-t_j) \geq a-x\} = inf\{u-t_j \geq 0 : X(u-t_j) \geq a-x\} \sim \tilde{S}_{a-x}$.

□

4.6.2 An Extended Bivariate Wiener Model

Generalizing Y away from being Wiener, is also of interest, especially when the relationship between $X(\cdot)$ and $Y(\cdot)$ is non-Gaussian. Such a generalization is reasonable as long as $Y(\cdot)-cX(\cdot)$ is selected to be any independent-increments process, independent of $X(\cdot)$, with $X(\cdot)$ Wiener with drift, but $Y(\cdot)-cX(\cdot)$ distributionally unspecified. We propose this extension as part of future research.

4.6.3 The Likelihood Function

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random vector and $\{f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ a statistical model parametrized by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, the parameter vector in the parameter space $\boldsymbol{\Theta}$.

Definition 4.12 (The Likelihood Function). *The likelihood function is a map $L(\boldsymbol{\theta}) = f_X(\boldsymbol{X}(\omega), \boldsymbol{\theta}) : \boldsymbol{\Theta} \to L_0(\boldsymbol{X}, P)$, where $L_0(\boldsymbol{X}, P)$ is the space of all $\sigma(\boldsymbol{X})$-measurable random variables*

4.7 Parametric Inference for the TMD Data-Structure

Parametric inference is based on observations on q failed and p surviving devices. To simplify exposition, covariates are not included in the following derivations. The unknown parameter is the vector $\boldsymbol{\theta} = (\nu_X, \nu_Y, \sigma_X, \sigma_Y, \rho)$, and the likelihood

function is given by:

$$L_{\boldsymbol{\theta}} = \prod_{i=1}^{q} f_{Y_{S_a},X_{S_a},T,I(S_a<\tau)}(Y_i(S_i), a, S_i, 1)\times \prod_{j=1}^{p} f_{Y_\tau,X_\tau,T,I(S<\tau)}(Y_j(\tau), X_j(\tau), \tau, 0) \quad (4.14)$$

For a given realization of the random variables above, the likelihood becomes a function of parameter vector $\boldsymbol{\theta}$. For failed devices we observe $(Y_i(S_i) = y_i, S_i = s_i; X_i(S_i) = a)$, $i = 1, \ldots, q$ and for surviving devices we observe $(Y_j(\tau) = y_j, X_j(\tau) = x_j)$, $j = 1, \ldots, p$.

Tab. 4.1: Conditional-density terms under TMD for i^{th} device, $s < \tau$

Failed devices:	$f_{Y_S,X_S,T,I(S<\tau)}(y, a, s, 1) =$
	$=f_{Y_S\|X_S,S}(y\|a, s)f_S(s)$
	$=f_{Q_s}(y - ca)f_S(s)$
Surviving devices:	$f_{Y_\tau,X_\tau,T,I(S<\tau)}(y, x, \tau, 0) =$
	$=f_{Q_\tau}(y - cx)f_{X_\tau,I(S<\tau)}(x, 0)$
where	$f_{X_\tau,I(S<\tau)}(x, 0) = f_{X_\tau}(x) - f_{X_\tau,I(S<\tau)}(x, 1)$
and	$f_{X_\tau,I(S<\tau)}(x, 1) = (4.5)$ by proposition 4.2

In the following sections we derive analytical equations for the joint density terms required for the likelihood function. Table 4.1 summarizes the required terms for failed and surviving devices. For short we use the notation S in place of S_a.

4.7.1 Contribution to likelihood from failed devices

The joint density for a failed device is given by:

$$f_{Y_S,X_S,S,I(S<\tau)}(y, a, s, 1) = f_{Q_s}(y - ca)f_S(s) \quad (4.15)$$

with the probability density function of S given by equation (4.6), and the probability density of Q_s given by:

$$f_{Q_s}(u) = \frac{1}{\sigma_Y \sqrt{(1-\rho^2)s}} \phi \left(\frac{u - y_0 - c(a - x_0) - (\nu_Y - c\nu_X)s}{\sigma_Y \sqrt{(1-\rho^2)s}} \right) \tag{4.16}$$

Proof. For failed devices, the terminal time random variable T and lifetime random variable S_a are equal, with values denoted $s_i < \tau$. Therefore, for $s < \tau$,

$$f_{Y_S,X_S,T,I(S<\tau)}(y, a, s, 1) = f_{Y_S|X_S,S,I(S<\tau)}(y|a, s, 1) f_{X_S,S,I(S<\tau)}(a, s, 1) \tag{4.17}$$

For $S = s < \tau$, $I(S < \tau) = 1$ is a degenerate random variable, the first term above can therefore be written as:

$$f_{Y_S|X_S,S,I(S<\tau)}(y|a, s, 1) = f_{Y_S|X_S,S}(y|a, s)$$

With definition 4.8, we have:

$$f_{Y_S|X_S,S}(y|a, s) = f_{Q_S|X_S,S}(y - ca|a, s) = f_{Q_S|S}(y - ca|s)$$

Due to lemma 4.4,

$$f_{Q_S|S}(y - ca|s) = f_{Q_S}(y - ca)$$

The second term in equation (4.17) can be simplified through the observation that on the event $(S < \tau)$, $(X_S, I(S < \tau))$ is a degenerate random-variable pair equal by definition to $(a, 1)$, and therefore the density $f_{X_S,S,I(S<\tau)}(a, s, 1)$ can be replaced by $f_S(s)$

$$f_{X_S,S,I(S<\tau)}(a, s, 1) = f_S(s)$$

Finally, the contribution to the likelihood from a failed device is given by:

$$f_{Y_S,X_S,T,I(S<\tau)}(y,a,s,1) = f_{Q_S}(y-ca)f_S(s)$$

□

4.7.2 *Contribution to likelihood from surviving devices*

The contribution to the likelihood from data on surviving devices is given by:

$$f_{Y_\tau,X_\tau,T,I(S<\tau)}(y,x,\tau,0) = ABC \tag{4.18}$$

$$A = f_{Q_\tau}(u) = \frac{1}{\sigma_Y\sqrt{(1-\rho^2)\tau}}\,\phi\left(\frac{u-y_0-c(x_\tau-x_0)-(\nu_Y-c\nu_X)\tau)}{\sigma_Y\,\sqrt{(1-\rho^2)\tau}}\right)$$

$$B = f_{X_\tau}(u) = \frac{1}{\sigma_X\,\sqrt{\tau}}\,\phi\left(\frac{u-x_0-\nu_X\tau}{\sigma_X\,\sqrt{\tau}}\right)$$

$$C = \left(1-exp\left(\frac{2\nu_X(x_\tau-a)}{\sigma_X^2}\right)\right)$$

Proof. For surviving devices, the non-degenerate observables are (y_j, x_j). Therefore, variable $T=\tau$ can be dropped from the density.

$$f_{Y_\tau,X_\tau,I(S<\tau)}(y,x,0) = f_{Y_\tau-cX_\tau+cX_\tau|X_\tau,I(S<\tau)}(y|x,0)f_{X_\tau,I(S<\tau)}(x,0) =$$

by lemma 4.3

$$= f_{Q_\tau|X_\tau,I(S<\tau)}(y-cx|x,0) = f_{Q_\tau}(y-cx)$$

therefore,

$$f_{Y_\tau,X_\tau,I(S<\tau)}(y,x,0) = f_{Q_\tau}(y-cx)f_{X_\tau,I(S<\tau)}(x,0) \tag{4.19}$$

The first density factor on the right-hand side of equation (4.19) is normal, with s replaced by τ. The second, is the probability density function at x (necessarily

$< a$) of the terminal value $X(\tau)$ for a device surviving at time τ. Any degradation sample path starting at $X(0) = x_0 < a$ and terminating at $X(\tau) = x < a$ either does or does not cross the failure threshold at $a > 0$ in the interval $(0, \tau)$. By the law of total probability:

$$f_{X_\tau, I(S<\tau)}(x, 0) = f_{X_\tau}(x) - f_{X_\tau, I(S<\tau)}(x, 1) \tag{4.20}$$

Equation (4.20) says that the probability of reaching a terminal value x for a device surviving at time τ is equal to the probability of a Wiener process with drift reaching a value of x at time τ minus the probability of reaching x and crossing the threshold at some time earlier.

Using equation (4.5) for the complimentary Wiener term, $f_{X_\tau, I(S<\tau)}(x, 1)$, we get:

$$f_{X_\tau, I(S<\tau)}(x, 0) = f_{X_\tau}(x) \left(1 - exp\left(\frac{2\nu_X(x-a)}{\sigma_X^2}\right)\right) \tag{4.21}$$

Plugging in equation (4.21) into equation (4.18) we get the final expression for the likelihood contribution from data on surviving devices:

$$f_{Y_\tau, X_\tau, I(S<\tau)}(y, x, 0) = f_{Q_\tau}(y - cx) f_{X_\tau}(x) \left(1 - exp\left(\frac{2\nu_X(x-a)}{\sigma_X^2}\right)\right) \tag{4.22}$$

□

4.7.3 Derivation of the Complimentary Wiener Term

We present two approaches to derive the complimentary Wiener term, one based on the definition of probability, and the other on the reflection principle for a Wiener process with drift.

Lemma 4.6 (Derivation of complimentary Wiener density based on definition of probability). *The probability density function of $f_{X_\tau, I(S<\tau)}(x, 1)$ is given*

by:

$$f_{X_\tau, I(S<\tau)}(x,1) = \int_0^\tau f_{X_\tau - X_s}(x-a) f_S(s) ds \tag{4.23}$$

$$f_{X_\tau - X_s}(x-a) = \frac{1}{\sqrt{2\pi\sigma_X^2(\tau - s)}} exp\left(-\frac{(x-a-\nu_X(\tau-s))^2}{2\sigma_X^2(\tau-s)}\right)$$

$$f_S(s) = \frac{a}{\sqrt{2\pi\sigma_X^2 s^3}} exp\left(-\frac{(a-\nu_X s)^2}{2\sigma_X^2 s}\right)$$

Proof. By definition, $f_{X_\tau, I(S<\tau)}(x,1)dx$ is the probability the degradation level belongs to a small interval $(x, x+dx)$ at time τ, and that it crossed the failure threshold some time earlier.

$$f_{X_\tau, I(S<\tau)}(x,1) = P(X_\tau \in (x, x+dx), S \leq \tau)/dx$$

$$= \int_0^\tau \frac{P(X_\tau \in (x, x+dx), S = s)}{dx} ds \tag{4.24}$$

The integrand in equation (4.24) can be expressed as:

$$f_{X_\tau, X_S, S}(x, a, s)$$

$$= f_{X_\tau | X_S, S}(x|a, s) f_{X_S, S}(a, s)$$

$$= f_{X_\tau - X_s | X_s, S}(x-a|a, s) f_S(s) \tag{4.25}$$

By theorem 4.3, because $X_\tau - X_s$ is $\mathcal{F}_{\geq s}^X$ measurable, it is independent of $(X(s), S)$ which is $\mathcal{F}_{\leq s}^X$ measurable. Therefore, we can simplify as follows:

$$= f_{X_\tau - X_S}(x-a) f_S(s) \tag{4.26}$$

Plugging in equation (4.26) into (4.24) we get the final analytical expression

for the complimentary Wiener density term:

$$f_{X_\tau, I(S<\tau)}(x,1) = \int_0^\tau f_{X_\tau - X_s}(x-a) f_S(s) ds \tag{4.27}$$

□

Lemma 4.7 (Derivation of proposition 4.2). *The probability density function of $f_{X_\tau, I(S<\tau)}(x,1)$ is given by:*

$$f_{X_\tau, I(S<\tau)}(x,1) = exp\left(\frac{2\nu_X(x-a)}{\sigma_X^2}\right) f_{X_\tau}(2a-x) \tag{4.28}$$

Proof.

$$f_{X_\tau, I(S\leq\tau)}(x,1) = \int_0^\tau f_{X_\tau|S}(x|s;\nu_X) f_S(s;a) ds \tag{4.29}$$

By theorem 4.3, the first factor in equation (4.29) is given by:

$$f_{X_\tau|S}(x|s;\nu_X) = f_{X_\tau - X_s}(x-a;\nu_X) =$$

$$= f_{X_\tau - X_s - \nu_X(\tau-s)}(x-a-\nu_X(\tau-s);0) = exp\left(-\frac{(x-a-\nu_X(\tau-s))^2}{2\sigma_X^2(\tau-s)}\right)$$

We take the ratio of the last expression at ν_X over the same expression with ν_X replaced by 0.

$$C_1(x,s) = \frac{exp\left(-\frac{(x-a-\nu_X(\tau-s))^2}{2\sigma_X^2(\tau-s)}\right)}{exp\left(-\frac{(x-a)^2}{2\sigma_X^2(\tau-s)}\right)} = exp\left(\frac{2\nu_X(x-a)-\nu_X^2(\tau-s)}{2\sigma_X^2}\right)$$

Therefore,

$$f_{X_\tau|S}(x|s;\nu_X) = exp\left(\frac{2\nu_X(x-a)-\nu_X^2(\tau-s)}{2\sigma_X^2}\right) f_{X_\tau|S}(x|s;0)$$

Above, by theorem 4.3, $f_{X_\tau|S}(x|s;0) = f_{X_\tau - X_s}(x-a;0)$, which by proposition

4.1 is equal to $f_{X_\tau|S}(2a-x|s;0)$. Then,

$$f_{X_\tau|S}(x|s;\nu_X) = exp\left(\frac{2\nu_X(x-a)-\nu_X^2(\tau-s)}{2\sigma_X^2}\right) f_{X_\tau|S}(2a-x|s;0) \tag{4.30}$$

Similarly we continue with

$$f_{X_\tau|S}(2a-x|s;0) = C_2(x,s) f_{X_\tau|S}(2a-x|s;\nu_X) \tag{4.31}$$

where

$$C_2(x,s) = \frac{f_{X_\tau|S}(2a-x|s;0)}{f_{X_\tau|S}(x|s;\nu_X)} = \frac{exp\left(-\frac{(a-x)^2}{2\sigma_X^2(\tau-s)}\right)}{exp\left(-\frac{(a-x-\nu_X(\tau-s))^2}{2\sigma_X^2(\tau-s)}\right)}$$

Plugging in C_2 into (4.30) and the resulting (4.30) into (4.29) by direct calculations we get:

$$f_{X_\tau|S}(x|s;\nu_X) = exp\left(\frac{2\nu_X(x-a)}{\sigma_X^2}\right) f_{X_\tau|S}(2a-x|s;\nu_X) \tag{4.32}$$

By plugging in (4.32) into (4.30) we get

$$f_{X_\tau,I(S<\tau)}(x,1) = exp\left(\frac{2\nu_X(x-a)}{\sigma_X^2}\right) \int_0^\tau f_{X_\tau|S}(2a-x|s;\nu_X) f_S(s) ds \tag{4.33}$$

The integral in equation (4.33) is equal to: $f_{X_\tau,I(S<\tau)}(2a-x,1)$. Since $2a-x>a$ we get finally the expression in equation (4.5):

$$f_{X_\tau,I(S<\tau)}(x,1) = exp\left(\frac{2\nu_X(x-a)}{\sigma_X^2}\right) f_{X(\tau)}(2a-x) \tag{4.34}$$

□

The exponential factor in equation (4.34) can be interpreted as the likelihood ratio of the Wiener process with drift on $[0,\tau]$ for a path with $X_\tau = 2a-x$ versus

a Wiener process path which is reflected symmetrically around the level a after the instant S of hitting a.

The likelihood function in equation (4.14) is computed with factors given by equation (4.15) for devices failing before τ, and by equation (4.18) for devices surviving past τ.

4.8 Predictive Inference

We aim to predict the degradation level and failure-time for a device surviving at time t, and whose marker and covariate vector are known at time t. The conditional density of the degradation variable at time t and of the failure-time density at future time $s \geq t$ are given by:

$$h(x|y;\boldsymbol{\theta}) = f_{X_t|Y_t,I(S<t)}(x|y,0) \tag{4.35}$$

and

$$g(s|y;\boldsymbol{\theta}) = f_{S|Y_t,I(S<t)}(s|y,0) \tag{4.36}$$

Related, are more complex expressions for density functions of X given a vector observation on Y. Some of this material is discussed in chapter 5. For now we derive analytical expressions for the above density functions given one marker observation. Predictive inferences are based on MLEs of the process parameters, and should therefore consider sampling error and predictive uncertainty. Nevertheless, useful insights are obtained by considering predictive inference when process parameters are assumed known. By varying the response variable across its domain we can compute the density at each discrete sample point. For example, we evaluate function $g(s|y;\boldsymbol{\theta})$, for a given $\boldsymbol{\theta}$ on the range $0 \leq s \leq \infty$ to get an understanding of the distribution of $g(\cdot)$.

Lemma 4.8 (Predicted degradation density). *The probability density function of the degradation random variable conditioned on the marker and survival at time* t *is given by:*

$$h(x|y;\boldsymbol{\theta}) = \frac{f_{Q_t}(y-cx)f_{X_t}(x)\left(1-exp\left(\frac{2\nu_X(x-a)}{\sigma_X^2 t}\right)\right)}{\int_{-\infty}^{a} f_{Q_t}(y-cu)f_{X_t}(u)\left(1-exp\left(\frac{2\nu_X(u-a)}{\sigma_X^2 t}\right)\right)du} \quad (4.37)$$

Proof. From Bayes rule equation (4.35) is expressed as:

$$f_{X_t|Y_t,I(S<t)}(x|y,0) = \frac{f_{X_t,Y_t,I(S\geq t)}(x,y,1)}{f_{Y_t,I(S\geq t)}(y,1)} \quad (4.38)$$

The numerator in equation (4.38) can be further expanded by conditioning on $X_t = x$ and the device surviving at time t

$$f_{X_t,Y_t,I(S\geq t)}(x,y,1) = f_{Y_t|X_t,I(S\geq t)}(y|x,1)f_{X_t,I(S\geq t)}(x,1) \quad (4.39)$$

The first factor is given by equation (4.11) due to lemma 4.4, and the second factor by equation (4.21). The denominator in equation (4.38) is the joint density of the marker and survival at time t, and is computed by integrating out $X(t)$.

$$f_{Y_t,I(S\geq t)}(y,1) = \int_{-\infty}^{a} f_{Y(t)|X(t),I(S\geq t)}(y|x,1)f_{X_t,I(S\geq t)}(x,1)dx \quad (4.40)$$

then we have

$$h(x|y;\boldsymbol{\theta}) = \frac{f_{Q_t}(y-cx)f_{X_t}(x)\left(1-exp\left(\frac{2a(x-a)}{\sigma_X^2 t}\right)\right)}{\int_{-\infty}^{a} f_{Q_t}(y-cu)f_{X_t}(u)\left(1-exp\left(\frac{2a(u-a)}{\sigma_X^2 t}\right)\right)du} \quad (4.41)$$

□

Lemma 4.9 (Predicted failure-time density). *The probability density function*

of the future failure-time random variable, conditioned on the marker and survival at time t is given by:

$$g(s|y;\boldsymbol{\theta}) = \frac{\int_{-\infty}^{a} f_S(s-t;a-u) f_{Q_t}(y-cu) f_{X_t}(u)\left(1-exp\left(\frac{2a(u-a)}{\sigma_X^2 t}\right)\right)du}{\int_{-\infty}^{a} f_{Q_t}(y-cu) f_{X_t}(u)\left(1-exp\left(\frac{2a(u-a)}{\sigma_X^2 t}\right)\right)du}$$

$$f_S(s-t;a-x) = \frac{a-x}{\sqrt{2\pi\sigma_X^2(s-t)^3}} exp\left(-\frac{(a-x-\nu_X(s-t))^2}{2\sigma_X^2(s-t)}\right) \tag{4.42}$$

Proof. From Bayes rule we have:

$$f_{S|Y_t,I(S<t)}(s|y,0) = \frac{f_{S,Y_t,I(S\geq t)}(s,y,1)}{f_{Y_t,I(S\geq t)}(y,1)} \tag{4.43}$$

The numerator can be conditioned on $X(t)$ to get

$$\int_{-\infty}^{a} f_{S,Y_t,X_t,I(S\geq t)}(s,y,x,1)dx =$$

$$= \int_{-\infty}^{a} f_{S|X_t,Y_t,I(S\geq t)}(s|x,y,1) f_{X_t,Y_t,I(S\geq t)}(x,y,1)dx \tag{4.44}$$

In the first factor in equation (4.44), the marker Y is dropped because S is only a function of X.

$$f_{S|X_t,Y_t,I(S\geq t)}(s|x,y,1) = f_{S|X_t,I(S\geq t)}(s|x,1)$$

Due to the conditioning on $X(t)$ we can invoke lemma 4.5 to get:

$$f_{S|X_t,I(S\geq t)}(s|x,y,1) = f_{\tilde{S}_{a-x}}(s-t;a-x) \tag{4.45}$$

where $\tilde{S}_{a-x}$ is given by equation (4.13)

The second term in equation (4.44) is given by equation (4.39) □

5. LONGITUDINAL MARKER MODEL

Under a bivariate latent degradation process, an extension to the terminal data-structure is the inclusion of longitudinal observations on the marker variable. The longitudinal data-structure is more representative of information collected in reliability tests. In electronics, for example, one can monitor the resistance, capacitance and current level on individual components on a printed circuit board over time. In addition, temperature, humidity, and other environmental variables can also be monitored. Longitudinal marker information can help explain the marker variability across devices and across time, and in turn help improve inference and event-time predictions.

The primary interest of longitudinal data analysis lies in the mechanism of change over time, including growth, aging, time profiles or effects of covariates [72]. Longitudinal studies allow researchers to investigate how the variability of the response varies in time with covariates. For instance, in a clinical trial that presumably aims to investigate the effectiveness of a new drug treating a disease, it is often of interest to examine pharmacokinetic behavior of the drug when it is applied to experimental animals or patients. Most drugs do not have constant efficacy over time, possibly due to drug resistance. Such time-varying effectiveness can be examined through a longitudinal study in which responses to the drug are monitored over time. Some features of longitudinal data:

1. The presence of repeated measurements for each device implies that the observations from the same device are autocorrelated or serially correlated. This

requires us to develop statistical methodology that takes the serial correlation into account.

2. In practice it is often true that at a given time, a multi-dimensional measurement is recorded, giving rise to data of repeated response vectors. The complication associated with such data arises from the fact that there exists two levels of correlation to be accounted for, namely the serial correlation and the correlation across the components of the response vector

3. Most longitudinal data from practical studies contain missing data. Dealing with missing data, when the missing data mechanism is informative, is generally nontrivial. To make proper statistical inference, one has to rely on the information that is supposed to be, but actually not, observed.

5.1 Modeling Correlated Data

A parametric modeling framework assumes that response observations are drawn from a certain population with a certain form, with a parameter set $\boldsymbol{\theta}$. The primary objective is to estimate and infer the model parameters $\boldsymbol{\theta}$. Explicitly specifying such a parametric distribution for nonnormal data is not trivial. When we have more than one response variable at each time point, we need to consider their joint distribution. Note that the multivariate normal is the distribution that can be fully determined when the first and second moments are given. In nonnormal distributions, it is generally difficult to determine a joint distribution based only on few low-order moments.

One way to overcome the difficulty of directly specifying the joint distribution is to consider conditional distributions, which are essentially one dimensional [73]. When one of the response variables is latent, like the degradation variable $X(t)$ in this work, the distribution of the observed response variable is obtained by inte-

grating out the latent variable. With the availability of the joint distribution, the full maximum likelihood estimation and inference can be developed. Albert (1999) points out that this modeling approach is particularly useful for analyzing longitudinal data in which there is a sizable number of missing observations either due to missed visits, loss to follow-up, or death. Conditional modeling via a latent variable can pose some challenges:

1. If the dimension of the latent variable is high, numerical evaluation of the joint distribution in the likelihood function can be intricate, which typically leads to computational problems.

2. The conditional approach relies on the conditional independence assumption, which needs to be validated.

In this data setting, time plays a more important role in the likelihood function and contributes therefore more strongly to parameter estimation. With highly reliable manufactured products, like modern electronic components, for example, very few failures are observed in tests. On the other hand, rich longitudinal covariate histories are typically collected and stored during the test. Longitudinal marker models, are therefore becoming increasingly more relevant and in demand in the area of reliability.

The intuition and hypothesis is that additional information on degradation markers will improve the accuracy of parametric and predictive inference as compared to inference in the terminal data-structure case, discussed in chapter 4. We derive the joint densities for failed and surviving devices under TMDL, just as in the TMD data-structure. In this case, however, the joint density must account for the dependency between marker observations in time. After all, the value of the marker at a certain time point is certainly dependent on the history of the marker process up until that time.

5.2 Scheduling Longitudinal Measurements

Typically marker variables are monitored using appropriate sensors, that can collect data at very high frequencies. When marker observations are cost or time prohibitive, failure-tests must be designed accordingly. One of the natural questions that arises is: "How many marker observations are needed for this test?". This question assumes there exists a minimum number of necessary marker observations, after which any additional marker observations do not improve parametric inference.

This is an important question in many fields. From a medical stand point, in designing clinical trials, for example, practitioners may ask patients to make scheduled return visits to the hospital in order to collect marker data over a certain period of time. Complications arise when patients don't return on schedule, or miss their scheduled appointments. In addition, data on markers can require complicated, intrusive and expensive procedures. It is therefore of interest to intelligently plan for the minimum number of scheduled visits. In reliability tests the idea is very similar. Data on degradation markers can also require complicated, intrusive and expensive measurement procedures, and examples abound in various applications. In electronics, however, although marker measurements are typically easy to collect, there is still interest in reducing the size of the collected data in order to expedite computations.

Another important question is when to schedule marker measurements. For a given device, is it better to collect data on the marker(s) with a fixed frequency, or given a fixed number of possible observations, is it better to observe the marker more frequently later or earlier in the device's life?

5.3 Parametric Inference

In this section we develop parametric inference for three longitudinal data-structures:

1. With 1 intermediate marker observation at time t_1. This data-structure is called TMDL1.

2. With 2 intermediate marker observations at times t_1, t_2. This data-structure is called TMDL2.

3. With m scheduled or K random marker observations on surviving or failed devices respectively. This data-structure is called GENL, which stands for general longitudinal. Table 5.1 illustrates this structure.

We show that expressions for the joint density between marker observations have the same form for both failed and survived items. What differs is only the last term, which accounts for the lifetime observation $T = min(S, \tau)$. Table 5.1 summarizes the key density terms used in the likelihood function for the TMDL1 data-structure.

Tab. 5.1: Key conditional-density terms under TMDL1

For both failed and surviving:	$t_j < t_i < \tau$
$f_{Y_i,Y_j\|X_j,X_i,I(S\geq t_j)}(y_i, y_j\|x_j, x_i, 1) =$	$f_{Q_i,Q_j}(y_i - cx_i, y_j - cx_j) \sim \mathcal{MVN}$
For failed devices:	$t_1 < s < \tau$
$f_{Y_1,Y_S,X_S,S}(y_1, y_s, a, s)$	$f_{X_S,I(S\geq t_1)\|X_1}(a, 1\|x_1) =$
	$f_{\tilde{S}_{a-X_1}}(s - t_1; a - x_1)$ =(4.13)
	$f_{X_1,I(S\geq t_1)}(x_1, 1)$=(4.21)
For surviving devices:	$t_1 < \tau$
$f_{Y_1,Y_\tau,X_\tau,I(S\geq\tau)}(y_1, y_\tau, x_\tau, 1)$	$f_{X_\tau,I(S\geq\tau)\|X_1,I(S\geq t_1)}(x_\tau, 1\|x_1, 1)$ =(4.21)
	$f_{X_1,I(S\geq t_1)}(x_1, 1)$=(4.21)

5.4 *One Intermediate Marker Observation - TMDL1*

Parametric inference under TMDL1, is again based on observations on q failed and p surviving devices. The unknown parameter vector is the vector $\boldsymbol{\theta} = (\nu_X, \nu_Y, \sigma_X, \sigma_Y, \rho)$, and the likelihood is given by:

$$L_{\boldsymbol{\theta}} = \prod_{i=1}^{q} f_{Y_S,Y_1,X_S,T,I(S_a<\tau)}(Y_i(S_i), Y_i(t_1), a, S_i, 1)\times \prod_{j=1}^{p} f_{Y_\tau,Y_1,X_\tau,T,I(S<\tau)}(Y_j(\tau), Y_j(t_1), X_j(\tau), \tau, 0)$$

5.4.1 *Contribution to likelihood from failed devices*

In the TMDL1 data-structure, for a failed device we observe (y_1, y_s, x_s, s), where y's are the observations on the marker at times t_1 and s respectively, $(t_1 < s)$, $x_s = a$, and s is the failure-time observation. Time t_1 is treated as a fixed and known scheduled time-point. Devices that fail before the scheduled intermediate marker observation, that is for cases where $s \leq t_1$, the joint density reduces to the TMD case. Under the TMDL1 data-structure, the joint density for a failed device is given by:

$$f_{Y_1,Y_S,X_S,T,I(S\leq\tau)}(y_1, y_s, a, s, 1) = \int_{x_1} E_1 E_2 \left\{E_3 - C_1 E_4\right\} dx_1 \tag{5.1}$$

where

$$E_1 = f_{Q_1,Q_s}(y_1 - cx_1, y_s - ca) = \frac{1}{(2\pi)^{-(k+1)/2}|V_f|^{-1/2}} exp\left\{-\frac{1}{2}\boldsymbol{q}_c V_f^{-1}\boldsymbol{q}_c'\right\}$$

k=1, the number of marker observations

$$\boldsymbol{q}_c = \begin{pmatrix} y_1 - cx_1 - t_1(\nu_Y - c\nu_X) \\ y_s - ca - s(\nu_Y - c\nu_X) \end{pmatrix} \qquad V_f = \begin{pmatrix} \sigma_Y^2 t_1(1-\rho^2) & \sigma_Y^2 t_1(1-\rho^2) \\ \sigma_Y^2 t_1(1-\rho^2) & \sigma_Y^2 s(1-\rho^2) \end{pmatrix}$$

$$E_2 = f_{\tilde{S}_{a-X_1}}(s-t_1; a-x_1) = \frac{(a-x_1)}{(2\pi\sigma_X^2(s-t_1)^2)^{1/2}} exp\left(-\frac{(a-x_1-\nu_X(s-t_1)^2)}{2\sigma_X^2(s-t_1)}\right)$$

$$E_3 = f_{X_1}(x_1) = (2\pi\sigma_X^2 t_1)^{-1/2} exp\left(-\frac{(x_1-\nu_X t_1)}{2\sigma_X^2 t_1}\right)$$

$$E_4 = f_{X_1}(2a-x_1) = (2\pi\sigma_X^2 t_1)^{-1/2} exp\left(-\frac{(2a-x_1-\nu_X t_1)}{2\sigma_X^2 t_1}\right)$$

$$C_1 = exp\left(\frac{2\nu_X(x_1-a)}{\sigma_X^2}\right)$$

Proof. Using arguments from chapter 4, for failed devices we can drop the indicator variable. The joint density can then be expressed as:

$$f_{Y_1,Y_S,X_S,T,I(S\leq\tau)}(y_1, y_s, a, s, 1) = f_{Y_1,Y_S,X_S,S}(y_1, y_s, a, s) =$$

Integrate out X_1, the degradation level at time t_1

$$\begin{aligned} &= \int_{x_1} f_{Y_1,Y_S,X_1,X_S,S}(y_1, y_s, x_1, a, s) = \\ &= \int_{x_1} f_{Y_1,Y_S|X_1,X_S,S}(y_1, y_s|x_1, a, s) * \\ &\quad f_{X_1,X_S,S}(x_1, a, s) dx_1 \end{aligned} \tag{5.2}$$

The first factor in (5.2) is the term E_1, and is equal to $f_{Q_1,Q_s}(y_1 - cx_1, y_s - ca)$ by lemma 4.3 and lemma 4.4. The second factor is given by:

$$\begin{aligned} &f_{X_1,X_S,S}(x_1, a, s) = f_{X_1,I(S\geq t_1),X_S,S}(x_1, 1, a, s) \\ &= f_{S|X_1,I(S\geq t_1)}(s|x_1, 1) f_{X_1,I(S\geq t_1)}(x_1, 1) \end{aligned} \tag{5.3}$$

The first factor in equation (5.3) represents the term E_2 in equation (5.1), and is given by equation (4.13). The second factor in equation (5.3) is given by equation (4.21), and represents the term $\{E_3 - C_1 E_4\}$ in equation (5.1). □

5.4.2 *Contribution to likelihood from surviving devices*

For surviving devices we observe (y_1, y_τ, x_τ), where y's are the observations on the marker at times t_1 and τ respectively, $(t_1 < \tau)$, and $x_\tau < a$. Time t_1 is again fixed and known. Under the TMDL1 data-structure, the joint density for a survived device is given by:

$$f_{Y_1,Y_\tau,X_\tau,I(S>\tau)}(y_1, y_\tau, x_\tau, 1) = \tag{5.4}$$

$$\int_{x_1} f_{Q_1,Q_\tau}(q_1, q_\tau) \left\{ f_{X(\tau-t_1)}(x_\tau - x_1) - f_{X(\tau-t_1)}(2a - x_\tau - x_1) exp\left(\frac{2\nu_X(x_\tau - a)}{\sigma_X^2}\right)\right\} \times \left\{ f_{X_1}(x_1) - f_{X_1}(2a - x_1) exp\left(\frac{2\nu_X(x_1 - a)}{\sigma_X^2}\right)\right\} dx_1 \tag{5.5}$$

where

$$f_{Q_1,Q_\tau}(q_1, q_\tau) = \frac{1}{(2\pi)^{-(k+1)/2}|V_s|^{-1/2}} exp\left\{-\frac{1}{2}\boldsymbol{q}_c V_s^{-1} \boldsymbol{q}_c'\right\}$$

$$q_1 = y_1 - cx_1$$

$$q_\tau = y_\tau - cx_\tau$$

and, k=1, is the number of marker observations, V_s is the variance covariance matrix. For example, the upper right element of V_s is given by:

$$cov(Q_1, Q_\tau) = cov(Y_1 - cX_1, Y_\tau - cX_\tau) =$$

$$cov(Y_1, Y_\tau) - ccov(Y_1, X_\tau) - cov(X_1, Y_\tau) + c^2 cov(X_1, X_\tau) =$$

$$= \rho\sigma_X\sigma_Y t_1 - \rho^2\sigma_Y^2 t_1 - \rho^2\sigma_Y^2 t_1 + \rho^2\sigma_Y^2 t_1$$

$$\therefore cov(Q_1, Q_\tau) = cov(Q_\tau, Q_1) = \sigma_Y^2 t_1(1 - \rho^2)$$

$$V_s = \begin{pmatrix} \sigma_Y^2 t_1(1-\rho^2) & \sigma_Y^2 t_1(1-\rho^2) \\ \sigma_Y^2 t_1(1-\rho^2) & \sigma_Y^2 \tau(1-\rho^2) \end{pmatrix}$$

$$\boldsymbol{q}_c = \begin{pmatrix} y_1 - cx_1 - t_1(\nu_Y - c\nu_X) \\ y_\tau - cx_\tau - \tau(\nu_Y - c\nu_X) \end{pmatrix}$$

$$f_{X_{\tau-t_1}}(x_\tau - x_1) = (2\pi\sigma_X^2(\tau - t_1))^{-1/2} exp\left(-\frac{(x_\tau - x_1 - \nu_X(\tau - t_1)^2}{2\sigma_X^2(\tau - t_1)}\right)$$
$$f_{X_{\tau-t_1}}(2a - x_\tau - x_1) = (2\pi\sigma_X^2(\tau - t_1))^{-1/2} exp\left(-\frac{(2a - x_\tau - x_1 - \nu_X(\tau - t_1)^2}{2\sigma_X^2(\tau - t_1)}\right)$$
$$f_{X_1}(x_1) = (2\pi\sigma_X^2 t_1)^{-1/2} exp\left(-\frac{(x_1 - \nu_X t_1)^2}{2\sigma_X^2 t_1)}\right)$$

Proof.

$$\begin{aligned} & f_{Y_1,Y_\tau,X_\tau,I(S\geq\tau)}(y_1, y_\tau, x_\tau, 1) = \\ & = \int_{x_1} f_{Y_1,Y_\tau|X_1,X_\tau,I(S\geq\tau)}(y_1, y_\tau|x_1, x_\tau, 1) f_{X_1,X_\tau,I(S\geq\tau)}(x_1, x_\tau, 1) dx_1 \end{aligned} \quad (5.6)$$

It follows directly from lemma 4.4 that the first factor in equation (5.6) is equal to $f_{Q_1,Q_\tau}(q_1, q_\tau) \sim \mathcal{MVN}$. The second factor is given by:

$$\begin{aligned} & f_{X_1,X_\tau,I(S>\tau)}(x_1, x_\tau, 1) = \\ & = f_{X_\tau,I(S\geq\tau)|X_1,I(S\geq t_1)}(x_\tau, 1|x_1, 1) f_{X_1,I(S\geq t_1)}(x_1, 1) \end{aligned}$$

By theorem 4.2, both factor above have the form of equation (4.21), each given by:

$$\begin{aligned} & f_{X_{\tau-t_1},I(S>\tau-t_1)}(x_\tau - x_1, 1) = \\ & = f_{X_\tau - X_1}(x_\tau - x_1) - f_{X_\tau - X_1}(2a - x_\tau - x_1) exp\left(\frac{2\nu_X(x_\tau - a)}{\sigma_X^2}\right) \\ & f_{X_1,I(S>t_1)}(x_1, 1) = f_{X_1}(x_1) - f_{X_1}(2a - x_1) exp\left(\frac{2\nu_X(x_1 - a)}{\sigma_X^2}\right) \end{aligned}$$

□

For both failed and survived devices, the joint density between marker observation time-points should have the same form, as we see in the proof above. The only contributing factor that should differ (in the likelihood) is the one between the last marker observation time and the failure-time. This term, as we showed is a re-started Inverse-Gaussian factor. This repetitive structure will help set the framework for later deriving the more general longitudinal joint densities. Figure 5.1 illustrates this result.

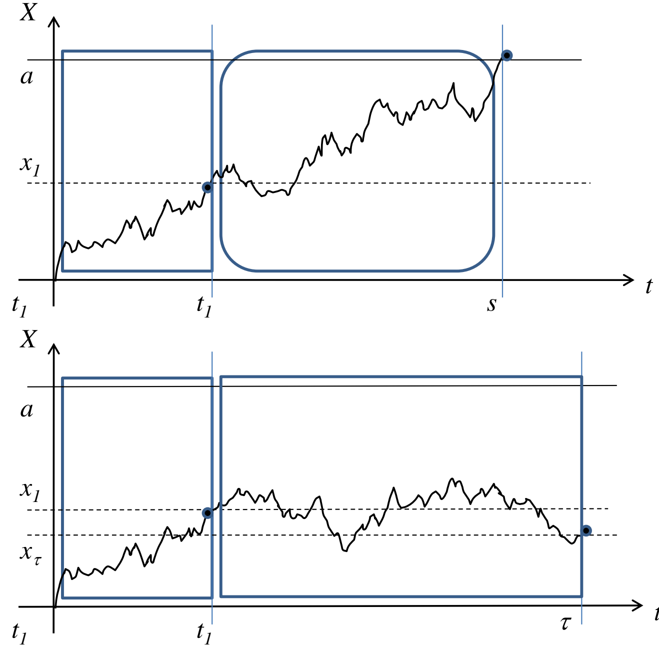


Fig. 5.1: Illustration of an observation on the degradation process under a longitudinal data structure. Rectangular boxes represent density factors given by equation (4.21), and the oval shape the factor given by equation (4.13)

5.5 Two Intermediate Marker Observations - TMDL2

Parametric inference under TMDL2, is again based on observations on q failed and p surviving devices. The unknown parameter vector is the vector $\boldsymbol{\theta} = (\nu_X, \nu_Y, \sigma_X, \sigma_Y, \rho)$, and the likelihood is given by:

$$
\begin{aligned}
L_{\boldsymbol{\theta}} = \prod_{i=1}^{q} & f_{Y_S,Y_2,Y_1,X_S,T,I(S_a<\tau)}(Y_i(S_i), Y_i(t_2), Y_i(t_1), a, S_i, 1)\times \\
& \prod_{j=1}^{p} f_{Y_\tau,Y_2,Y_1,X_\tau,T,I(S<\tau)}(Y_j(\tau), Y_j(t_2), Y_j(t_1), X_j(\tau), \tau, 0)
\end{aligned}
$$

We now consider the situation of two scheduled marker observations. In this case we observe (y_1, y_2, y_s, a, s) on each failed device and $(y_1, y_2, y_\tau, x_\tau)$ for each surviving device. For a failed device, the marker is observed in total three times: $t_1 < t_2 < s$. Similarly for a surviving device, the marker is observe at times: $t_1 < t_2 < \tau$. If $s < t_1$ or $t_1 < s < t_2$, we rely respectively on previously established

expressions under the TMD and TMDL1 data-structures.

The derivation of the joint densities in the TMDL1 data-structure call for integration over latent variable X_1. When more than one intermediate marker is observed, higher dimensional nested integrals will be needed. High dimensional integration is not only undesirable from an approximation perspective, it also poses a serious computational problem. In this section, therefore, we derive a computationally more efficient alternative.

5.5.1 Contribution to likelihood from failed devices

Under the TMDL2 data-structure, the joint density for a failed device is given by:

$$\begin{aligned} & f_{Y_1,Y_2,Y_S,X_S,S}(y_1, y_2, y_s, a, s) = \\ & = \int_{x_2} \int_{x_1} E(s) G_1(x_2) G_2(x_1) H_1(x_2) H_2(x_1, x_2) H_3(x_1) dx_1 dx_2 = \\ & = E(s) \int_{x_2} G_1(x_2) H_1(x_2) \left(\int_{x_1} G_2(x_1) H_3(x_1) H_2(x_1, x_2) dx_1 \right) dx_2 \end{aligned} \tag{5.7}$$

In equation (5.7), the terms E, G and H represent density terms introduced next. To solve for the joint density in this case we need to integrate over x_1 and x_2. To simplify computations, we factor out terms from the nested integral, as shown in equation (5.7) above. In this case, factor $E(s)$ is not a function of either x_1 or x_2 so its taken outside of both integrals. Factors $G_1(x_2)$ and $H_1(x_2)$ are not functions of x_1 so they are taken out of the second integral. Factors $G_2(x_1), H_3(x_1)$ and $H_2(x_1, x_2)$ are functions of both x_1 and x_2 and therefore need to be integrated over twice.

Definition 5.1. *Let the correlation coefficient between Q_i and Y_j as ρ_{QY}, given by:*

$$\rho_{Q_iY_j} = ((t_i \wedge t_j)(1-\rho^2))^{1/2}(t_i \vee t_j)^{-1/2} \tag{5.8}$$

$$E(s) \sim \mathcal{N}\left(\mu_{Q_s} + \rho_{Q_sY_2}\frac{\sigma_{Q_s}}{\sigma_{Y_2}}(y_2 - \mu_{Y_2}), (1-\rho^2_{Q_sY_2})\sigma^2_{Q_s}\right)$$

$$\mu_{Q_s} = E(Y_s - cX_s) = s(\nu_Y - c\nu_X)$$

$$\mu_{Y_2} = \nu_Y t_2$$

$$\sigma_{Q_s} = \sqrt{Var(Y_s - cX_s)} =$$

$$= (Var(Y_s) + c^2Var(X_s) - 2cCov(Y_s, X_s))^{1/2} = (\sigma^2_Y s(1-\rho^2))^{1/2}$$

$$\sigma_{Y_2} = \sigma_Y\sqrt{t_2}$$

$$\rho_{Q_sY_2} = \frac{Cov(Q_s, Y_2)}{\sigma_{Q_s}\sigma_{Y_2}} = \frac{Cov(Y_s, Y_2) - cCov(X_s, Y_2)}{\sigma_{Q_s}\sigma_{Y_2}} =$$

$$= \frac{t_2\sigma^2_Y(1-\rho^2)}{\sigma_Y\sqrt{s(1-\rho^2)}\sigma_Y\sqrt{t_2}} \Rightarrow \rho_{Q_sY_2} = (t_2(1-\rho^2))^{1/2}s^{-1/2}$$

Then,

$$F(s) \sim \mathcal{N}\left(s(\nu_Y - c\nu_X) + (1-\rho^2)(y_2 - \nu_Y t_2), \sigma^2_Y(1-\rho^2)(s - t_2(1-\rho^2))\right)$$

$$G_2(x_1) \sim \mathcal{N}\left(\mu_{Q_1}, \sigma^2_{Q_1}\right)$$

$$\mu_{Q_1} = t_1(\nu_Y - c\nu_X)$$

$$\sigma_{Q_1} = t_1\sigma^2_Y(1-\rho^2)$$

$$G_1(x_2) \sim \mathcal{N}\left(\mu_{Q_2} + \rho_{Q_2Y_1}\frac{\sigma_{Q_2}}{\sigma_{Y_1}}(y_1 - \mu_{Y_1}), (1-\rho^2_{Q_2Y_1})\sigma^2_{Q_2}\right) \Rightarrow$$

$$G_1(x_2) \sim \mathcal{N}\left(t_2(\nu_Y - c\nu_X) + (1-\rho^2)(y_1 - \nu_Y t_1), \sigma^2_Y(1-\rho^2)(t_2 - t_1(1-\rho^2))\right)$$

$$H_1(x_2) = f_{\tilde{S}_{a-x_2}}(s - t_2; a - x_2) \text{ given by equation (4.13)}$$

$$H_2(x_1, x_2) = f_{X_{t_2-t_1}, I(S>t_2-t_1)}(x_2 - x_1, 1) =$$

$$= f_{X_{t_2-t_1}}(x_2 - x_1) - exp\left\{2\nu_X(x_2 - a)/\sigma^2_X\right\} f_{X_{t_2-t_1}}(2a - x_2 - x_1)$$

where

$$f_{X_{t_2-t_1}}(x_2 - x_1) \sim \mathcal{N}\left(\mu_{X_{t_2-t_1}}, \sigma^2_{X_{t_2-t_1}}\right)$$

$$\mu_{X_{t_2-t_1}} = \nu_X(t_2 - t_1)$$

$$\sigma^2_{X_{t_2-t_1}} = \sigma^2_X(t_2 - t_1)$$

and similarly,

$$H_3(x_1) = f_{X_1}(x_1) - exp\left\{2\nu_X(x_1 - a)/\sigma_X^2\right\} f_{X_1}(2a - x_1)$$
$$f_{X_1}(x_1) \sim \mathcal{N}(\nu_X t_1, \sigma_X^2 t_1)$$

and similarly for $f_{X_1}(2a - x_1)$

Proof.

$$\begin{aligned}
f_{Y_1,Y_2,Y_S,X_S,S}(y_1, y_2, y_s, a, s) = \\
\int_{x_2}\int_{x_1} f_{Y_S|Y_2,Y_1,X_S,X_2,X_1,S}(y_s|y_2, y_1, a, x_2, x_1, s) * \quad (1) \\
* f_{Y_2|Y_1,X_S,X_2,X_1,S}(y_2|y_1, a, x_2, x_1, s) * \quad (2) \\
* f_{Y_1|X_S,X_2,X_1,S}(y_1|a, x_2, x_1, s) * \quad (3) \\
* f_{X_S,X_2,X_1,S}(a, x_2, x_1, s)dx_1dx_2 \quad (4)
\end{aligned}$$

Due to theorem 4.2, in equation (1), Y_1 can be dropped, so

$$(1) = f_{Y_S|Y_2,X_S,X_2,X_1,S}(y_s|y_2, a, x_2, x_1, s) =$$

Due to lemma 4.3 we get:

$$(1) = f_{Q_s|Y_2}(q_s|y_2) = E(s)$$

where $Q_s = Y_s - cX_s = Y_s - ca$, and $c = \rho\sigma_Y/\sigma_X$. Note that $Q(\cdot)$ is not independent

of $Y(\cdot)$. By the same logic, we have:

$$
\begin{aligned}
(2) &= f_{Q_2|Y_1}(q_2|y_1){=}G_1(x_2), \text{ where } Q_2 = Y_2 - cX_2 \\
(3) &= f_{Q_1}(q_1){=}G_2(x_1), \text{ where } Q_1 = Y_1 - cX_1 \\
(4) &= f_{X_S,X_2,X_1,I(S\geq t_2),I(S\geq t_1),S}(a, x_2, x_1, 1, 1, s) = \\
&\quad = f_{X_S,S|X_2,X_1,I(S\geq t_2),I(S\geq t_1)}(a, s|x_2, x_1, 1, 1)* \\
&\qquad * f_{X_2,I(S\geq t_2)|X_1,I(S\geq t_1)}(x_2, 1|x_1, 1) f_{X_1,I(S\geq t_1)}(x_1, 1) = \\
&\quad = f_{\tilde{S}_{a-x_2}}(s - t_2; a - x_2) f_{X_2-X_1,I(S\geq t_2-t_1)}(x_2 - x_1, 1) f_{X_1,I(S>t_1)}(x_1, 1)
\end{aligned}
$$

First factor above is given by $H_1(x_2)$, the second by $H_2(x_1, x_2)$ and third by $H_3(x_1)$. □

5.5.2 *Contribution to likelihood from surviving devices*

For a surviving device we observe $(y_1, y_2, y_\tau, x_\tau)$ at times t_1, t_2, and at the end-of-test time τ for any known and fixed $t_1 < t_2 < \tau$. Under the TMDL2 data-structure, the joint density for a surviving device is given by:

$$
E(\tau)\int_{x_2} G_1(x_2)H_1(x_2)\left(\int_{x_1} G_2(x_1)H_3(x_1)H_2(x_1, x_2)dx_1\right)dx_2 \tag{5.9}
$$

The factors in (5.9) are expressed as follows:

$$F(\tau) \sim \mathcal{N}\left(\mu_{Q_\tau} - \rho_{Q_\tau Y_2}\frac{\sigma_{Q_\tau}}{\sigma_{Y_2}}(y_2 - \mu_{Y_2}), (1-\rho^2_{Q_\tau Y_2})\sigma^2_{Q_\tau}\right)$$

$$\mu_{Q_\tau} = E(Y_\tau - cX_\tau) = \tau(\nu_Y - c\nu_X)$$

$$\mu_{Y_2} = \nu_Y t_2$$

$$\sigma_{Q_\tau} = (\sigma_Y^2 \tau(1-\rho^2))^{1/2}$$

$$\sigma_{Y_2} = \sigma_Y \sqrt{t_2}$$

$$\rho_{Q_\tau Y_2} = \frac{Cov(Q_\tau, Y_2)}{\sigma_{Q_\tau}\sigma_{Y_2}} = \frac{cov(Y_\tau, Y_2) - cCov(X_\tau, Y_2)}{\sigma_{Q_\tau}\sigma_{Y_2}} =$$

$$= \frac{t_2\sigma_Y^2(1-\rho^2)}{\sigma_Y\sqrt{\tau(1-\rho^2)}\sigma_Y\sqrt{t_2}}$$

$$\Rightarrow \rho_{Q_\tau Y_2} = (t_2(1-\rho^2))^{1/2}\tau^{-1/2}$$

and,

$$E(\tau) \sim \mathcal{N}\left(\tau(\nu_Y - c\nu_X) + (1-\rho^2)(y_2 - \nu_Y t_2), \sigma_Y^2(1-\rho^2)(\tau - t_2(1-\rho^2))\right)$$

$$G_1(x_2) \sim \mathcal{N}\left(t_2(\nu_Y - c\nu_X) + (1-\rho^2)(y_1 - \nu_Y t_1), \sigma_Y^2(1-\rho^2)(t_2 - t_1(1-\rho^2))\right)$$

$$G_2(x_1) \sim \mathcal{N}\left(t_1(\nu_Y - c\nu_X), \sigma_Y^2 t_1(1-\rho^2)\right)$$

$$H_1(x_2) = f_{X_{\tau - t_2}}(x_\tau - x_2) - exp\left\{2\nu_X(x_\tau - a)/\sigma_X^2\right\} f_{X_{\tau-t_2}}(2a - x_\tau - x_2)$$

$$H_2(x_1, x_2) = f_{X_{t_2 - t_1}}(x_2 - x_1) - exp\left\{2\nu_X(x_2 - a)/\sigma_X^2\right\} f_{X_{t_2-t_1}}(2a - x_2 - x_1)$$

$$H_3(x_1) = f_{X_1}(x_1) - exp\left\{2\nu_X(x_1 - a)\sigma_X^2\right\} f_{X_1}(2a - x_1)$$

Proof.

$$f_{Y_1,Y_2,Y_\tau,X_\tau,I(S\geq\tau)}(y_1, y_2, y_\tau, x_\tau, 1) =$$

$$= \int_{x_2}\int_{x_1} f_{Y_1,Y_2,Y_\tau,X_\tau,X_1,X_2,I(S\geq\tau)}(y_1, y_2, y_\tau, x_\tau, x_2, x_1, 1) =$$

$$= \int_{x_2}\int_{x_1} f_{Y_\tau|Y_2,X_\tau,X_2,X_1,I(S\geq\tau)}(y_\tau|y_2, x_\tau, x_2, x_1, 1) * \qquad (1)$$

$$* f_{Y_2|Y_1,X_\tau,X_2,X_1,I(S\geq\tau)}(y_2|y_1, x_\tau, x_2, x_1, 1) * \qquad (2)$$

$$* f_{Y_1|X_\tau,X_2,X_1,I(S\geq\tau)}(y_1|x_\tau, x_2, x_1, 1) * \qquad (3)$$

$$* f_{X_\tau,X_2,X_1,I(S\geq\tau)}(x_\tau, x_2, x_1, 1)dx_1dx_2 \quad (4)$$

Due to theorem 4.2 and lemma 4.3 we get:

$$(1) = f_{Q_\tau|Y_2}(q_\tau|y_2) = F(\tau)$$
$$(2) = f_{Q_2|Y_1}(q_2|y_1) = G_1(x_2)$$
$$(3) = f_{Q_1}(q_1) = G_2(x_1)$$

In factor (4) we condition respectively on survival past times t_1 and t_2 to get:

$$\begin{aligned}
& f_{X_\tau,X_2,X_1,I(S\geq\tau)}(x_\tau,x_2,x_1,1) = \\
& = f_{X_\tau,I(S\geq\tau),X_2,I(S\geq t_2),X_1,I(S\geq t_1)}(x_\tau,1x_2,1,x_1,1) = \\
& = f_{X_\tau,I(S\geq\tau)|X_2,X_1,I(S\geq t_2)}(x_\tau,1|x_2,x_1,1) * \qquad (a) \\
& \quad f_{X_2,I(S\geq t_2)|X_1,I(S\geq t_1)}(x_2,1|x_1,1) * \qquad (b) \\
& \quad f_{X_1,I(S\geq t_1)}(x_1,1) \qquad (c)
\end{aligned}$$

In factor (a) above, we drop $I(S \geq t_1)$, because $I(S \geq t_1) = 1 \subset I(S \geq t_2) = 1$. Due to theorem 4.2 we drop X_1 in factor (a), and we can then show that it reduces to the term given by $H_1(x_2)$. Factors (b) and (c) above are of the same form given respectively by $H_2(x_1, x_2)$ and $H_3(x_1)$. □

5.6 General Longitudinal Case - GENL

In the general case we are interested again in scheduled marker observations made at $t_1, t_2, \ldots$. We will present similar derivations for the joint densities of failed and survived devices.

5.6.1 Contribution to likelihood from failed devices

For a failed device we observe: $(y_1, y_2, \ldots, y_k, y_s, x_s, s, k)$, where k is random, and represents the number of total marker observations before the failure-time. The marker is observed at times $t_1, t_2, \ldots, t_k, s$, where $t_1 < t_2 < \ldots < t_k < s$.

In a general case, under the GENL data-structure, for a failed device, the joint density is given by:

$f_{Y_1,\ldots,Y_K,Y_S,X_S,T,I(S<\tau),K}(y_1,\ldots,y_k,y_s,a,s,1,k) =$

$$\int_{x_k}\cdots\int_{x_1} f_{Q_1,\ldots,Q_k,Q_s}(q_1,\ldots,q_k,q_s) f_S(s-t_k;a-x_k) *$$
$$* \prod_{j=1}^{k}\left[f_{X_j-X_{j-1}}(x_j-x_{j-1}) - exp\left(\frac{2\nu_X(x_j-a)}{\sigma_X^2}\right) f_{X_j-X_{j-1}}(2a-x_j-x_{j-1})\right] \tag{5.10}$$

Proof.

$$\begin{aligned} & f_{Y_1,\ldots,Y_K,Y_S,X_S,T,I(S<\tau),K}(y_1,\ldots,y_k,y_s,a,s,1,k) = \\ & = \int\cdots\int f_{Y_1,\ldots,Y_K,Y_S|X_1,\ldots,X_K,X_S,S,K}(y_1,\ldots,y_k|x_1,\ldots,x_k,a,s,k) * -(a) \\ & \qquad * f_{X_1,\ldots,X_K,X_S,S,K}(x_1,\ldots,x_k,a,s,k)dx_1\ldots dx_k - (b) \end{aligned} \tag{5.11}$$

Factor (a) can be expressed as:

$$(a) = f_{Q_1,\ldots,Q_K|X_1,\ldots,X_K,X_S,S,K}(q_1,\ldots,q_k|x_1,\ldots,x_k,a,s,k) \tag{5.12}$$

Due to lemma 4.3 and lemma 4.4, we have $\{Q_1,\ldots,Q_K\} \perp K$ because K is a function of S. For a given $K=k$, we factor (a) simplifies to:

$$(a) = f_{Q_1,\ldots,Q_k,Q_s}(q_1,\ldots,q_k,q_s) \sim \mathcal{MVN} \tag{5.13}$$

Let $\boldsymbol{X}^{k-1} = \{X_1,\ldots,X_{k-1}\}$, then factor (b) in (5.11) is written as:

$$(b) = f_{\boldsymbol{X}^{k-1},X_k,X_s,S,K}(\boldsymbol{x},x_k,a,s,k)$$

The randomness of $K = k$ is entirely captured by $S = s$. Therefore

$$
\begin{aligned}
(b) = & f_{\boldsymbol{X}^{k-1},X_k,X_S,S}(\boldsymbol{x}, x_k, a, s) = \\
& = f_{\boldsymbol{X}^{k-1},X_k,X_S,I(S\geq t_1),I(S\geq t_2),\ldots,I(S\geq t_k),S}(\boldsymbol{x}, x_k, a, 1, 1, \ldots, 1, s) = \\
& = f_{S|X_k,I(S\geq t_k)}(s|x_k, 1) * \qquad\qquad (b1) \\
& \quad * \textstyle\prod_{j=1}^{k} f_{X_j,I(S\geq t_j)|\boldsymbol{X}^{j-1},I(S\geq t_{j-1})}(x_j, 1|\boldsymbol{x}, 1) \quad (b2)
\end{aligned}
$$

where $\boldsymbol{X}^{j-1} = \{X_1, \ldots, X_{j-1}\}$. Due to lemma 4.5

$$(b1) = f_{\tilde{S}_{a-x_k}}(s - t_k; a - x_k)$$

Take one term in factor $(b2)$. Due to theorem 4.2 we only condition on the last observation on X at time t_{j-1}. Then,

$$
\begin{aligned}
(b2) = & f_{X_j,I(S\geq t_j)|X_{j-1},I(S\geq t_{j-1})}(x_j, 1|x_{j-1}, 1) = \\
& = f_{X_j-X_{j-1},I(S>t_j-t_{j-1})}(x_j - x_{j-1}, 1) = \\
& = f_{X_j-X_{j-1}}(x_j - x_{j-1}) - exp\,(2\nu_X(x_j - a)/\sigma_X^2)\, f_{X_j-X_{j-1}}(2a - x_j - x_{j-1})
\end{aligned}
$$

Therefore, putting things together, and plugging in factor $(b1)$ and $(b2)$ into (b) we get:

$$
\begin{aligned}
(b) = & f_{\tilde{S}_{a-x_k}}(s - t_k; a - x_k) * \\
* & \textstyle\prod_{j=1}^{k}\{f_{X_j-X_{j-1}}(x_j - x_{j-1}) - exp\left(\dfrac{2\nu_X(x_j - a)}{\sigma_X^2}\right) f_{X_j-X_{j-1}}(2a - x_j - x_{j-1})\}
\end{aligned}
$$

where $x_0 = 0$, $t_0 = 0$. □

5.6.2 Contribution to likelihood from surviving devices

For surviving devices we observe $(y_1, \ldots, y_n, x_n)$ at scheduled n time points $t_1 < \ldots < t_n$. The last scheduled observation is made at the end-of-test time τ, that is $t_n = \tau$, and n is the total number of scheduled observations, and is known,

fixed and non-random. In the general case, under the GENL data-structure, the joint density for surviving device is given by:

$$
\begin{aligned}
&f_{Y_1,\ldots,Y_n,X_n,I(S>\tau)}(y_1,\ldots,y_n,x_n,1) = \\
&\qquad \int_{x_{n-1}} \cdots \int_{x_1} f_{Q_1,\ldots,Q_n}(q_1,\ldots,q_n) \prod_{j=1}^{n} [f_{X_j-X_{j-1}}(x_j - x_{j-1}) - \\
&\qquad f_{X_j-X_{j-1}}(2a - x_j - x_{j-1}) exp\left(\frac{2\nu_X(x_j - a)}{\sigma_X^2}\right)] dx_1 \ldots dx_{n-1}
\end{aligned}
\tag{5.14}
$$

Proof.

$$
\begin{aligned}
&f_{Y_1,\ldots,Y_n,X_n,I(S\geq\tau)}(y_1,\ldots,y_n,x_n,1) = \\
&\quad = \textstyle\int_{x_{n-1}} \cdots \int_{x_1} f_{\boldsymbol{Y}^n|\boldsymbol{X}^n,I(S\geq\tau)}(\boldsymbol{y}|\boldsymbol{x},1) * \qquad (a) \\
&\qquad\qquad * f_{\boldsymbol{X}^n,I(S\geq\tau)}(\boldsymbol{x},1) dx_1 \ldots dx_{n-1} \qquad (b)
\end{aligned}
$$

where $\boldsymbol{Y}^n = \{Y_1,\ldots,Y_n\}$ and $\boldsymbol{X}^n = \{X_1,\ldots,X_n\}$. By lemma 4.4, we have:

$$
\begin{aligned}
&(a) = f_{\boldsymbol{Q}_n}(\boldsymbol{q}_n) \sim \mathcal{MVN} \\
&\boldsymbol{Q}_n = \{Q_1,\ldots,Q_n\}
\end{aligned}
$$

Factor (b) is expanded like before, using theorem 4.2 we can proceed as follows:

$$
\begin{aligned}
(b) &= f_{X_1,\ldots,X_\tau,I(S\geq t_1),\ldots,I(S\geq\tau)}(x_1,\ldots,x_\tau,1,\ldots,1) = \\
&= \prod_{j=1}^{n} f_{X_j,I(S\geq t_j)|X_{j-1},I(S\geq t_{j-1})}(x_j,1|x_{j-1},1) = \\
&= \prod_{j=1}^{n} f_{X_j-X_{j-1},I(S\geq t_j-t_{j-1})}(x_j - x_{j-1},1)
\end{aligned}
$$

Therefore, finally factor (b) can be written as the following product:

$$
(b) = \prod_{j=1}^{n} \left[f_{X_j-X_{j-1}}(x_j - x_{j-1}) - f_{X_j-X_{j-1}}(2a - x_j - x_{j-1}) exp\left(\frac{2\nu_X(x_j - a)}{\sigma_X^2}\right)\right]
$$

□

5.7 Summary and Conclusions

We were able to identify the two density factors needed to analytically express the joint densities for both failed and surviving devices in the likelihood function. These are namely, (1) the re-started inverse-Gaussian $f_{\tilde{S}_{a-X(t_j)}}(s - t_j; a - x_j)$, given by equation (4.13) and (2) $f_{X_j, I(S \leq t_j)}(x_j, 0)$, given by equation (4.21) .

Although the analytical structure is simple, larger multidimensional marker observations require an equal number of integrations, a fact that will burden computations. Some simplification can be achieved, as we show, by factoring out terms that are not functions of the space over which we integrate. However, high dimensional nested integrals still remains a computational limitation.

6. CASE STUDIES - TERMINAL AND LONGITUDINAL

In chapters 4, 5 and later in chapter 8 we present extensions to Whitmore's basic first hitting time model. The first extension augments the Whitmore data-structure with terminal degradation observations for surviving devices. The second extension incorporates longitudinal measurements on the marker variable. In chapter 8 we present a third extension that uses a variable instead of a fixed failure-threshold. All three extensions build on the Whitmore data structure, and each one adds a level of complexity to the likelihood function.

In this chapter we apply the degradation models with the first two extensions to: *i*) simulated lifetime and degradation data, and *ii*) Aircraft Gas-Turbine Engine degradation data. The objectives of the case studies are to:

- Confirm the likelihood-based calculations for MLEs
- Compare inference results with and without extended data-structures
- Validate the model on simulated and real data

Table 6.1 presents the MLE notations under the four data-structures we examine in this chapter.

Tab. 6.1: MLEs under four data-structures

TM	$\hat{\boldsymbol{\theta}}_{TM} = G_1(\{Y_i(T)\}_{i=1}^n)$	
TMD	$\hat{\boldsymbol{\theta}}_{TMD} = G_2(\{Y_i(T)\}_{i=1}^n, \{X_j(\tau)\}_{j=1}^p)$	
TMDL1	$\hat{\boldsymbol{\theta}}_{TMDL1} = G_3(\{\boldsymbol{Y}_i\}_{i=1}^n, \{X_j(\tau)\}_{j=1}^p)$	$\boldsymbol{Y}_i = (Y_i(t_1), Y_i(T))$
TMDL2	$\hat{\boldsymbol{\theta}}_{TMDL2} = G_4(\{\boldsymbol{Y}_i\}_{i=1}^n, \{X_j(\tau)\}_{j=1}^p)$	$\boldsymbol{Y}_i = (Y_i(t_1), Y_i(t_2), Y_i(T))$

The likelihood-based calculations of MLEs are confirmed on simulated and a real data set as we discuss next. With simulated data, the ML estimates are compared against the true parameter values using ML theory for them. In the real data, MLEs are compared to simpler method-of-moments parameter estimates. For example, in a real data set, we may compute an empirical estimate of the drift by dividing the threshold level over the average lifetime of a sample of tested devices: $na/\sum_i T_i$

Specifically under each of the TM and TMD data-structures, we expect ML parameters to be consistent and asymptotically efficient [Section 3.3]. We first compare inference under the TM and TMD data-structures. Typically the comparison is made using distance measures such as the MSE [Definition 3.3]. Given that both TM and TMD MLEs are asymptotically unbiased up to $o(1/\sqrt{n})$ remainders, for large sample sizes [Theorem 3.5], the improvement in estimation can be evaluated based on comparing their large-sample variances. Under both data-structures, MLEs are expected to be asymptotically efficient and thus their large sample empirical variances should approach the corresponding Cramer Rao lower bounds [Def 3.9]. However, for a fixed sample size, the Cramer Rao lower bounds for TM and TMD may differ, making one model more efficient than the other. We would like to examine the following points:

- At what sample sizes are estimators under the two-data-structures finite sample efficient? In other words, at what sample size does the empirical variance approach the Cramer Rao lower bound? Is this sample size different for the two data-structures?

- What is the Asymptotic Relative Efficiency (ARE) between the two data-structures? In other words, for a large enough sample size, what is the ratio of the variances for the corresponding data-structures.

- Even though we do not expect MLEs to be efficient under small samples, the relative performance of the two estimators for small sample sizes is of interest. We calculate therefore, and analyze the ARE for small samples

Because the TM data-structure is a subset of the TMD data-structure, we expect that MLEs under TMD will outperform those under TM on all three criteria above because ML estimation improves with enhanced data, when sample sizes are large. MLEs from the TMD data-structure are represented by vector $\hat{\boldsymbol{\theta}}_{TMD}$, and from the TM data-structure by $\hat{\boldsymbol{\theta}}_{TM}$. The TM data-structure is a subset of the TMD data-structure, where $X(\tau)$ is observed at τ under TMD but not under TM. Therefore the TMD data is considered enhanced in comparison to TM.

We compute the ARE of MLE's from the two data-structures in estimating the expected failure-time parameter μ. Parameter μ is a function of $\hat{\boldsymbol{\theta}}$. According to Chhikara and Folks 1989, for a Wiener process with positive drift, the scale parameter $\mu = g(\nu_X) = (a - x_0)/\nu_X$. The scale parameter μ of an inverse-Gaussian density function represents the expected first hitting time of the degradation process to a. Therefore, since a, x_0 are fixed, the ARE for estimating μ is the same as that for estimating ν_X.

As per definition 3.8 the *ARE* is defined as the ratio of asymptotic variances of the estimators for $\hat{\mu}$: $V(\hat{\mu}_2)/V(\hat{\mu}_1)$, and the the estimated ARE, $\widehat{ARE}$, as per definition 3.9 is defined as the ratio of estimated asymptotic variances: $\hat{V}(\hat{\mu}_2)/\hat{V}(\hat{\mu}_1)$. The asymptotic variance of $\hat{\mu}$, $V(\hat{\mu})$, is derived through the *Delta* method, which says that if $\sqrt{n}(\nu_X - \hat{\nu}_X) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$, where ν_X and σ^2 are finite valued constants and $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution, then

$$\sqrt{n}(g(\nu_X) - g(\hat{\nu}_X)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, g'(\nu_X)^2\sigma^2)$$
$$V(g(\hat{\nu}_X)) = V(\hat{\mu}) = g'(\nu_X)^2\sigma^2$$

In the ratio defined by the ARE, the term $g'(\nu_X)^2$ cancels out, and we are left with

the ratio of asymptotic variances for ν_X of TMD vs TM.

For the MLE's obtained under each data-structure, the estimated asymptotic variance $n\hat{V}(\hat{\nu_X})$ is the upper-left element $(I_{obs}^{-1})_{1,1}$ of the inverse of the observed information matrix [Remarks 3.2 & 3.3], averaged over R simulations to obtain the estimated asymptotic relative average efficiency $\widehat{ARE_R}$, with $\hat{\nu}_X^r$ the MLE for ν_X, obtained under the rth simulated data-set.

$$\widehat{ARE}(\hat{\nu}_X^{TMD}, \hat{\nu}_X^{TM}) = \frac{\frac{1}{R}\sum_{r=1}^{R} \hat{V}(\hat{\nu}_X^{TM})_r}{\frac{1}{R}\sum_{r=1}^{R} \hat{V}(\hat{\nu}_X^{TMD})_r} \tag{6.1}$$

The ARE estimate in (6.1) is the ratio of averages of ML variances estimated from each of R replicated simulations, under each of the TM and TMD data-structure, computed from R different estimators ν_X^r, $r = 1, \ldots, R$. The asymptotic variance of the MLE $\hat{\boldsymbol{\theta}}$ is equal to the inverse of the observation matrix [Remark 3.1], and this information matrix is estimated consistently by the *observed Fisher information* I_{obs} matrix. In turn, the I_{obs} matrix is computed under the numerically approximated Hessian matrix, in which finite-difference methods estimate the first and second order derivatives of the likelihood function.

6.1 *Lifetime and Degradation Simulation Model*

Following Whitmore et al. 1998, we simulate sample observations for n devices with the parameter set $\boldsymbol{\theta} = (\nu_X, \nu_Y, \sigma_X, \sigma_Y, \rho) = (0.1, 1.0, 0.4, 0.1, 0.75)$. Lifetimes were generated by adding correlated normally distributed increments $(\Delta x, \Delta y)$ over small time increments Δt=0.01 or Δt=0.005, with a=1, and τ=10. To capture the causal relationship between degradation and marker at each time point, marker samples are are drawn through the process $Q = Y - cX$ defined in section 4.5.1, which is independent of the degradation process X. We are simulating the follow-

ing phenomena: degradation of the device causes changes in the marker variable distribution. Our inference model, however, works in the other direction: it takes marker observations and infers the distribution of the degradation variable at each time point. Next we describe the simulation model.

Physically the rationale behind the selected parameter values can be motivated from the degradation of solder joints. Solder joints hold electronic components (like a capacitor, or a resistor) to a printed circuit board, and can, with use (abuse), develop micro-cracks, which in time can grow and cause reliability problems. These cracks can grow or shrink depending on usage or environmental conditions. Then, $X(t)$ in the model can be thought of as the length or size of the crack. It is not unreasonable to have high variance in this process, especially since the crack can entirely close "heal" (given the right conditions), and then snap back to a fully "opened" state. So the signal to noise ratio (ν_X/σ_X) is reasonably less than 1. The drift and variability of the marker process are not as important (at least in its physical interpretation), but the correlation coefficient ρ of course is. We see the influence of varying ρ in studying the ARE of TMDL1 vs TMD. We also consider other parameter combinations to study more general patterns in estimation.

6.2 Simulation Design

Before getting to the simulation results, we first define the simulation design. We discuss the method of generating the degradation, marker and lifetime data for the TM data-structure. We start by partitioning the total time-on-test $[0, \tau]$ into $N = 200$ equally sized time intervals Δt, and construct a time vector $\boldsymbol{t}^s = (t_1^s, \ldots, t_N^s)$, such that $\sum_{\forall i} \Delta t^s = t_N^s = \tau$, $i = 1, \ldots, N$. We then generate an N-dimensional increments vector $\boldsymbol{\Delta} X = (\Delta X(t_1^s), \ldots, \Delta X(t_N^s))$, each distributed,

with common parameters, as:

$$\Delta X(t_i) \sim \mathcal{N}\left(\nu_X \Delta t^s, \sigma_X^2 \Delta t^s\right) \tag{6.2}$$

A degradation vector $\boldsymbol{X} = (X(1), \ldots, X(N))$ is constructed by adding up the increments $\Delta X(t_i^s)$. For example, $X(t_1^s) = X(t_0^s) + \Delta X(t_1^s) = \Delta X(t_1^s)$, and $X(t_2^s) = X(t_1^s) + \Delta X(t_2^s) = \Delta X(t_1^s) + \Delta X(t_2^s)$, etc. The degradation vector represents a degradation path with drift ν_X and variance σ_X^2, starting at time 0 and ending at time τ. The marker samples are drawn conditionally given degradation based on equation (4.11) to generate a marker vector $\boldsymbol{Y} = (Y(t_1^s), \ldots, Y(t_N^s))$, evaluated at the same time points as the degradation process.

Devices are determined to fail if their discrete-time degradation paths hit a before censoring at τ. Specifically the crossing-time is determined as the last time-point t_i where $X(t_i) < a$. The choice of discrete spacing Δt^s is chosen small enough that estimation is not improved when re-done with smaller spacings Δt^s. In this simulation we compared results under $\Delta t^s = 0.01$ and $\Delta t^s = 0.005$. The proportion of failed to surviving devices is random, and vary in each simulation. A larger proportion of failed to survived devices can be achieved if the censoring time or drift parameter are increased. This is because longer test periods give more time for the degradation variable to reach a, and higher drift parameters drive the degradation process to a faster.

6.3 Simulation Results

6.3.1 Simulation 1 - TM vs. TMD

We simulated R=2000 independent data sets $\boldsymbol{D}$, for each sample size $n = (20, 40, 80, 160, 320, 640, 1280, 2560)$, and computed the MLEs $\hat{\boldsymbol{\theta}}^r$, $r = 1, \ldots, R$ for each. Table 6.2 reports the average asymptotic standard errors for ν_X computed

from I_{obs} compared to the empirical sample variance, under both the TM and TMD data-structures. The close agreement between the I_{obs}-derived and empirical columns, especially for large samples, is predicted by asymptotic MLE theory. As expected the uncertainty in the drift parameter estimates are consistently smaller under the TMD data-structure. In particular the large-sample Cramer-Rao lower bound is smaller for the TMD than it is for TM. Both TM and TMD seem to attain efficiency, up to the accuracy of the simulation, at around a sample size of 100 (This can be seen by comparing Cramer Rao lower bound to the empirical variances).

Tab. 6.2: Asymptotic Standard Errors for ν_X under TM and TMD, from Observed Information vs. Empirical

	Obtained from I_{obs}		Empirical	
Sample Size (n)	TM	TMD	TM	TMD
20	0.01495	0.01207	0.01604	0.0127
40	0.01052	0.00866	0.01053	0.00889
80	0.00750	0.00616	0.0077	0.00616
160	0.00530	0.00437	0.00554	0.00446
320	0.00377	0.00309	0.00383	0.00317
640	0.00266	0.00219	0.00262	0.00219
1280	0.00188	0.00155	0.0019	0.00157
2560	0.00133	0.00110	0.00138	0.00108

Table 6.3 tabulates the central limit theorem based confidence intervals (CI) for the empirical asymptotic variances in table 6.2, and provides a measure of how precise the agreement is between I_{obs}-derived and empirically-derived columns in table 6.2. MLE's $\hat{\nu}_X^r$, $r = 1, \ldots, R$, are independent and asymptotically normal, or we can say approximately normal for large enough n. For large enough n, therefore, we have that $\hat{\nu}_X^r \sim \mathcal{N}(\nu_X, V(\nu_X))$, where $V(\nu_X)$ is the asymptotic variance of ν_X. The estimated asymptotic variance is given by: $\hat{V}(\nu_X) = \sum_r(\hat{\nu}_X^r - \bar{\hat{\nu}}_X)^2/(n-1)$, where $\bar{\hat{\nu}}_X$ is the empirical average of $\hat{\nu}_X$ over R samples, and $(n-1)\hat{V}(\nu_X)/V(\nu_X) \sim \chi^2_{n-1}$. Therefore, a $100(1-\alpha)\%$ CI for $V(\nu_X)$ is given by:

$$\frac{n-1}{\chi^2_{\alpha/2,n-1}}\hat{V}(\nu_X) < V(\nu_X) < \frac{n-1}{\chi^2_{1-\alpha/2,n-1}}\hat{V}(\nu_X)$$

We are interested in the CI for the standard deviations (SD), so we take the square root of the upper and lower limit of the CIs for the variance. Therefore we have:

$$\sqrt{\frac{n-1}{\chi^2_{\alpha/2,n-1}}}\widehat{SD}(\nu_X) < SD(\nu_X) < \sqrt{\frac{n-1}{\chi^2_{1-\alpha/2,n-1}}}\widehat{SD}(\nu_X) \quad (6.3)$$

where $\widehat{SD}(\nu_X) = \sqrt{\hat{V}(\nu_X)}$

Tab. 6.3: Central Limit Theorem-based confidence intervals for asymptotic empirical Standard Deviations under TM and TMD

	TM		TMD	
Sample Size (n)	Upper	Lower	Upper	Lower
20	0.01219	0.02342	0.00965	0.0185
40	0.00862	0.01352	0.00728	0.01141
80	0.00666	0.00912	0.00533	0.00729
160	0.00499	0.00622	0.00401	0.00501
320	0.00355	0.00415	0.00294	0.00343
640	0.00248	0.00277	0.00207	0.00231
1280	0.00182	0.00197	0.00151	0.00163
2560	0.00134	0.00141	0.00105	0.00111

In failure tests, engineers are often interested in reducing test durations and accelerating factors/conditions. Longer test durations cost more money and resources, and overly accelerating conditions can change targeted failure mechanisms. Failure test-designs that use low or no accelerating conditions are preferred. In addition, correlation between degradation X and the marker Y is often weaker than expected. There is interest, therefore, to investigate the efficacy of the TMD data-structure when τ is small and correlation ρ is weak. A simulation experiment computes $\widehat{ARE}(\hat{\nu}_X^{TMD}, \hat{\nu}_X^{TM})$ for pairs of (ρ, τ) from $\rho = (0, 0.3, 0.6, 0.9)$ and $\tau = (5, 10, 15, 20)$. The correlation ρ is varied from weak to strong and τ from short to long.

For each pair (ρ, τ), and fixed $n = 500$, we simulated $R = 100$ data-sets and computed MLEs $\hat{\boldsymbol{\theta}}^*_r = (\nu_X, \nu_Y, \sigma_X, \sigma_Y)_r$, and their asymptotic variance estimates

Tab. 6.4: Asymptotic relative efficiency of $\hat{\mu}_{TMD}$ versus $\hat{\mu}_{TM}$ for different (ρ, τ) combinations, with $\theta = (0.1, 1.0, 0.2, 0.1)$

ρ/τ	4	7	10	13
0	5.9614	2.1593	1.4759	1.2488
0.3	5.6893	2.1465	1.4505	1.2336
0.6	5.4270	2.0464	1.4106	1.2088
0.9	4.0123	1.6320	1.2355	1.1262

$\hat{V}(\hat{\boldsymbol{\theta}}^*_r)$, $r = 1, \ldots, R$, under both TM and TMD data separately. We calculate $\widehat{ARE}(\cdot)$, according to equation (6.1) by averaging $\hat{V}(\cdot)_r$. Table 6.4 shows the ARE results from an experiment with $\boldsymbol{\theta}^* = (0.1, 1.0, 0.2, 0.1)$. In table 6.4 we point out the improvement in inference for low (ρ, τ) combinations. This result indicates that under the bivariate Wiener model the TMD data-structure improves inference under smaller sample sizes and weaker correlation coefficients. This result is very promising because it gives preliminary justification for a TMD data-structure.

6.3.2 General Patterns for ARE of $\hat{\mu}$

To ascertain general patterns of ARE's as a function of ρ and τ, we ran the simulation experiment for different combinations of $\boldsymbol{\theta}^* = (\nu_X, \nu_Y, \sigma_X, \sigma_Y)$. We expect patterns of inference-improvement in TMD over TM to be insensitive to changes in $\boldsymbol{\theta}^*$. Table 6.5 shows AREs from an experiment using the same (ρ, τ) combinations,

Tab. 6.5: Asymptotic relative efficiency of $\hat{\mu}_{TMD}$ versus $\hat{\mu}_{TM}$ for different (ρ, τ) combinations, with $\theta = (0.1, 1.0, 0.4, 0.1)$

ρ/τ	4	7	10	13
0	2.44236	1.68550	1.43906	1.31727
0.3	2.40830	1.67795	1.41769	1.31341
0.6	2.29307	1.59947	1.394463	1.29086
0.9	1.80918	1.40500	1.263771	1.19975

and with a different $\boldsymbol{\theta}^* = (0.1, 1.0, 0.4, 0.1)$. We can see that the pattern of AREs across the (ρ, τ) grid is qualitatively similar. However, because the variance of the

simulated Wiener process is larger, now $\sigma_X = 0.4$, it causes a larger proportion of devices to fail, which increases the ARE for each (ρ, τ) combination.

The above results in tables 6.4 and 6.5 can be explained as follows. With small τ's there are more censored lifetime observations in each simulated dataset, which means there are more observations on the terminal degradation in the TMD data-structure. With more observations on degradation, in turn, we expect improved inference under the TMD data-structure, and therefore higher AREs. AREs greater than 1 indicate the efficacy of the TMD data-structure, and AREs close to 1 indicate that the two data-structures provide the same inference power. When $\boldsymbol{\theta}^* = (0.1, 1, 0.2, 0.1)$, the percentage of failed devices is: 11%, 34%, 74%, and 97%, respectively for the four τ levels. When $\boldsymbol{\theta}^* = (0.1, 1.0, 0.4, 0.1)$ the percentage of failed devices is: 33%, 68%, 89%, and 99%, respectively for the four levels of τ. We see that with higher variance in the process, we have more failures under each τ, and therefore, access to less degradation information on surviving device, and therefore, as expected, lower AREs.

Tab. 6.6: Combinations of parameter vector $\boldsymbol{\theta}^*$

No.	ν_X	ν_Y	σ_X	σ_Y	No.	ν_X	ν_Y	σ_X	σ_Y
1	0.1	0.9	0.2	0.1	12	0.1	1.0	0.2	0.4
2	0.1	0.5	0.2	0.1	13	0.1	1.0	0.2	0.8
3	0.1	0.1	0.1	0.1	14	0.1	1.0	0.2	1.0
4	0.1	2.0	0.2	0.1	15	0.1	1.0	0.2	1.5
5	0.1	4.0	0.2	0.1	16	0.1	1.0	0.2	2.0
8	0.1	1.0	0.2	0.1	17	0.4	1.0	0.2	0.1
6	0.1	1.0	0.4	0.1	18	0.8	1.0	0.2	0.1
7	0.1	1.0	0.8	0.1	19	1.0	1.0	0.2	0.1
9	0.1	1.0	1.0	0.1	20	1.5	1.0	0.2	0.1
10	0.1	1.0	1.5	0.1	21	2.0	1.0	0.2	0.1
11	0.1	1.0	2.0	0.1					

More generally, we compute the AREs for the same set of (ρ, τ) pairs, failure threshold level $a = 1$, and time increment $\Delta t = 0.01$, using various combinations of

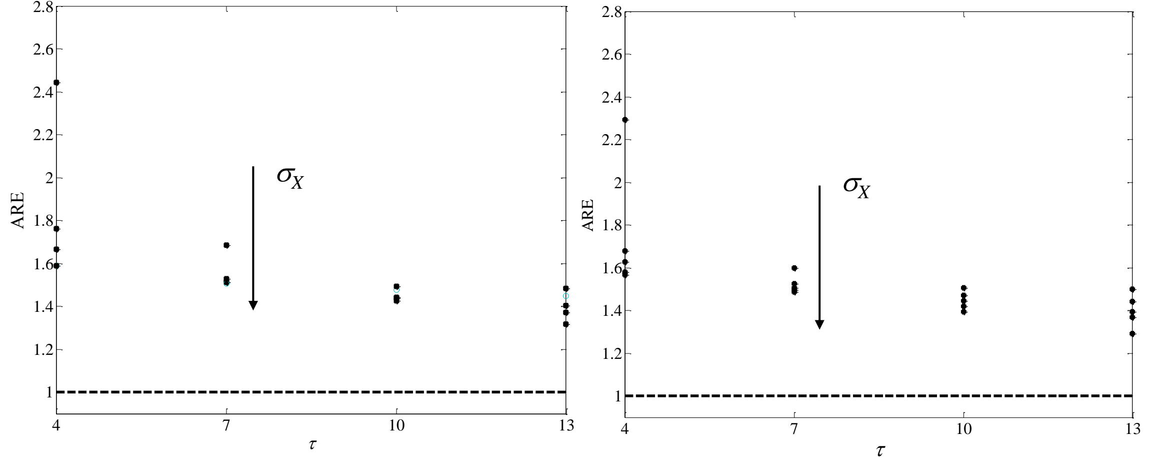


Fig. 6.1: General ARE patterns of $\hat{\mu}_{TMD}$ versus $\hat{\mu}_{TM}$, under increasing σ_X

$\boldsymbol{\theta}^*$ as listed in table 6.6. Each parameter θ_i in vector $\boldsymbol{\theta}^*$ is varied separately, keeping the others fixed. Parameter values are chosen to be reasonable in relationship to a and Δt. For example for values of $\nu_X > 2$ the mean path of the simulated Wiener process reaches the threshold at $t = 0.5$. Larger values of drift would simply cause all simulations to result in failure. As we already showed in table 6.5, an increase in the variance of the degradation process, decreases the AREs across all (ρ, τ) pairs.

Figure 6.1 (left) plots the general ARE patterns for $\rho = 0$, under increasing σ_X, as a function of τ. As expected, for increasing τ the AREs attenuate to 1, showing the diminishing efficiency of the TMD data-structure against the TM data-structure for longer test durations. We also see that for each fixed τ, larger σ_X generate lower ARE. On the right, we see similar general patterns under higher correlation $\rho = 0.6$. Figure 6.2 (left) plots the general ARE patterns for $\rho = 0$, under increasing σ_Y, as a function of τ. As τ increases, the ARE attenuates to 1. For each fixed τ, larger σ_Y generate higher ARE. With stronger correlation $\rho = 0.6$, the variance in the ARE is larger under each τ, however, for increasing τ, the AREs again attenuate to 1.

Increasing the drift of the marker process ν_Y does not seem to effect the ARE under the same conditions. Figure 6.3 plots the AREs for $\nu_Y = (0.1, 0.4, 0.8, 1.6, 3.2)$,

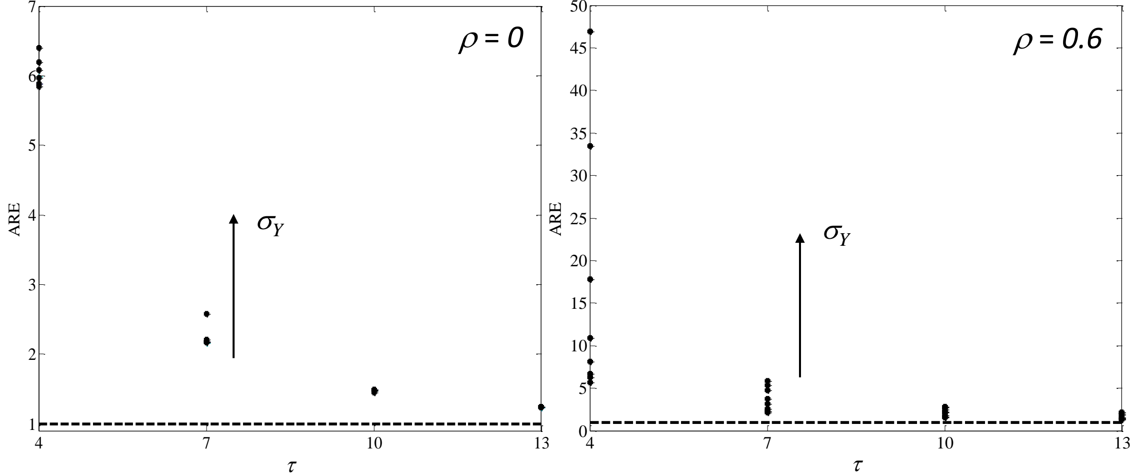


Fig. 6.2: General ARE patterns of $\hat{\mu}_{TMD}$ versus $\hat{\mu}_{TM}$, sunder increasing σ_Y

for $\rho = 0$ (left) and $\rho = 0.6$ (right). Here we see that under fixed τ, ν_Y does not influence estimation. The results from figures 6.2 and 6.3, indicate that inference is not affected by the marker magnitude at termination, but rather its variance. Although both data-structures include terminal marker observations, inference suffers more under TM than TMD because, under the TM data-structure, inference is only based on the marker observations, whereas under the TMD data-structure, inference is based on both marker and degradation data.

Increasing the drift of the degradation process on the other hand has more prominent consequences on AREs. Figure 6.4 plots the general ARE patterns for $\rho = 0$ (left) and $\rho = 0.6$ (right), under increasing ν_X, as a function of τ. As expected, under each fixed τ, larger ν_X, increases the ARE. This is highlighted especially for small values of τ. These results make sense, because under high drifts, we expect a larger proportion of failed devices, and therefore as per our hypothesis, improved inference under the TMD as opposed to the TM data-structure.

6.3.3 How can general ARE patterns be used in a practical setting?

Generally, these results validate our hypothesis that inference improves with enhanced data, and that the TMD data-structure is more efficient at ML-estimation under small failure-time samples. These results can be used to justify reducing

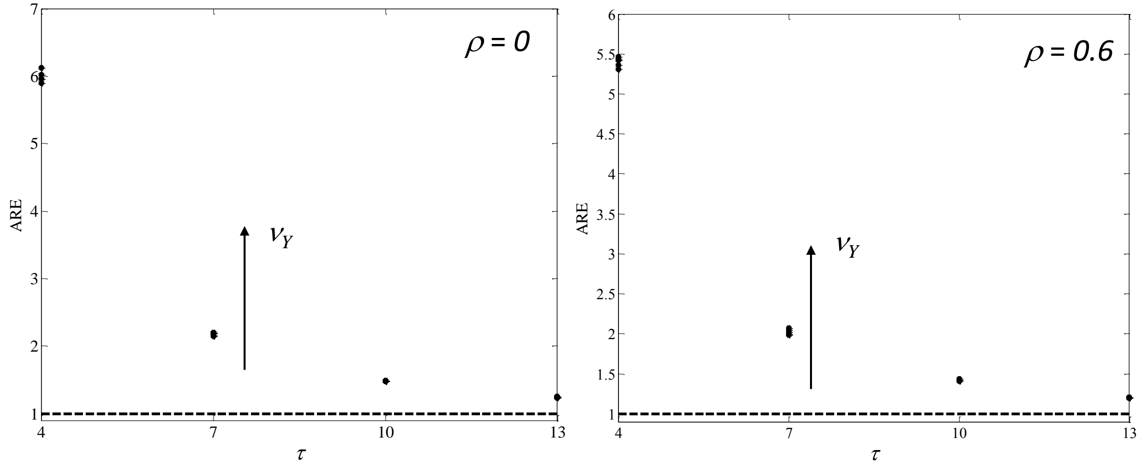


Fig. 6.3: General ARE patterns of $\hat{\mu}_{TMD}$ versus $\hat{\mu}_{TM}$, under increasing ν_Y

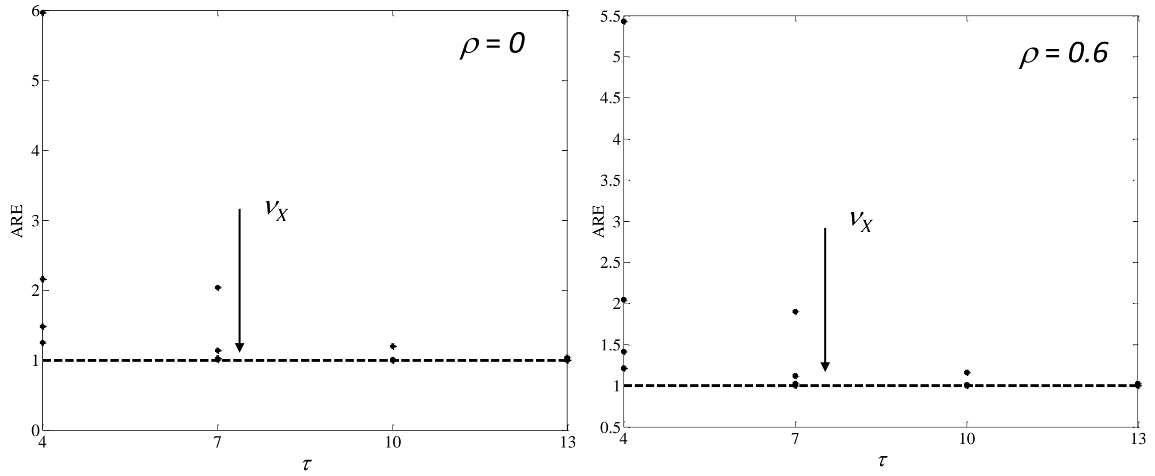


Fig. 6.4: General ARE patterns of $\hat{\mu}_{TMD}$ versus $\hat{\mu}_{TM}$, under increasing ν_Y

accelerating conditions in accelerated failure tests. By reducing the accelerating conditions, and keeping the test-duration τ fixed, we are likely to observe fewer failed devices by time τ. However, based on the above results, we need a smaller failure-time sample in the TMD data-structure to get the same predictive power as the TM data-structure, and therefore we can ”afford” to reduce the designed accelerating conditions in the planned failure-test. From an engineering perspective, reducing accelerating conditions in failure tests is a welcomed option, because, we we discuss in the introduction, highly accelerated test conditions may alter targeted failure-mechanisms.

These results can also be used to justify shorter test-durations for planned failure-tests. Typically, failure-tests are conducted over a time-period long enough to see enough devices fail. Shorter test-durations are not only less costly, they are sometimes required due to short product life-cycles.

High values of ρ make the marker data more closely associated to the failure mechanism, and therefore we expect lower AREs when ρ approaches 1. When the marker is perfectly correlated to the degradation variable, any observations on the degradation should not enhance inference. This argument is validated in the simulation results.In general we notice that for stronger correlations we get smaller ARE. For long test-times, most devices fail, and the TMD data-structure, therefore, looses valuable degradation information from surviving devices.

The improvement in estimation under the TMD data-structure is due to observations on terminal degradation. In tests where degradation is latent, as motivated in this thesis, we argued and proved with the results above that access to terminal degradation data reduces estimation variance in a bivariate Wiener model. In failure-test laboratories, terminal degradation is often measured using expensive equipment or proprietary techniques, and is therefore valuable information. Under latent degradation conditions, the above results show the efficacy of terminal

degradation data in reducing estimation variance. These results therefore, suggest investing in failure-analysis equipment capable of measuring terminal degradation.

6.3.4 Predictions on the failure-time distribution under TMD

Predictions are computed using the predictive inference equations (4.35) and (4.36) and the MLEs as plug-in estimators. Given survival and a marker observation at a sequence of time points t_i, Figure 6.5 plots the predicted conditional degradation (left) and future failure-time densities (right). The probability density function of the degradation variable at time t_i is plotted by connecting probabilities computed for a range of degradation values. Similarly the probability density function of the failure-time variable at time t_i is plotted by connecting the probabilities computed for a range of future failure-times. Predictions qualitatively capture the behavior/trend of the latent degradation process, and the decreasing uncertainty in failure-time predictions.

The shape of the density is plotted by connecting point probabilities evaluated for a vector-valued dependent variable. For example, we vary the level of degradation from -1 to 1 in small increments and evaluate the probability specified by the predictive inference equations for each x_i

6.3.5 Simulation 2 - TMD vs. TMDL1

We simulate $R = 250$ independent data-sets $\boldsymbol{D}$, for a range of sample sizes n, and compute the MLEs $\hat{\boldsymbol{\theta}}$ for each. MLEs from the TMD data-structure are represented by vector $\hat{\boldsymbol{\theta}}_{TMD}$, and from the TMDL1 data-structure by $\hat{\boldsymbol{\theta}}_{TMDL1}$. Under TMDL1, for failed devices we observe $\boldsymbol{Y} = (Y(t_1), Y(s))$, and for surviving devices $\boldsymbol{Y} = (Y(t_1), Y(\tau))$, and as always $0 \leq s \leq \tau$. The improvement in estimation from observing $Y(t_1)$ can again be seen by comparing standard errors of estimated parameters.

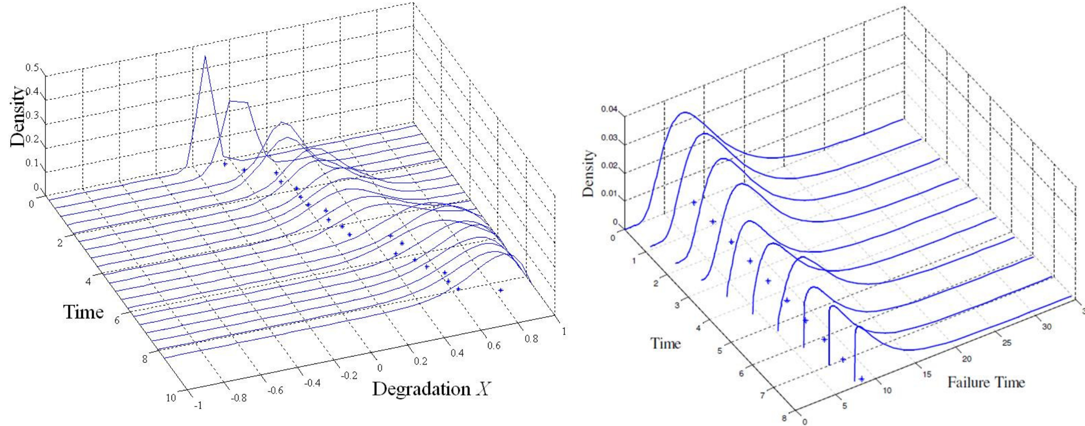


Fig. 6.5: Predicted degradation and survival density as a function of time

For each sample size n, we simulate sample observations with the parameter set $\boldsymbol{\theta} = (\nu_X, \nu_Y, \sigma_X, \sigma_Y, \rho) = (0.1, 1.0, 0.4, 0.1, 0.75)$. Lifetimes were generated in the same way as before (see section 6.1), with Δt=0.01, a=1, and τ=10. In each simulation run j, the intermediate marker observation is made at time-point $t_1 = \tau/2$. This way the marker observation time is independent of the failure-time and always less than the end-of-test time. When $t_1 > s$ the contribution of that failed device to the likelihood is made through the TMD data-structure.

Table 6.7 reports the asymptotic standard errors for ν_X computed from I_{obs} compared to the empirical sample variance-covariance matrix, under both the TMD and TMDL1 data-structures. We observe again a close agreement between the I_{obs}-derived and empirical columns. We observe, contrary to our expectation no improvement in estimation results under the TMDL1 data-structure in comparison to the TMD data-structure. This result is likely a result of simulation errors. However, under TMDL2 we actually see improvement in estimation as we discuss in the next section.

Asymptotic relative efficiency is used to test the efficacy of TMD vs TMDL1.

Tab. 6.7: Asymptotic Standard Errors for ν_X under TMD and TMDL1, from Observed Information vs. Empirical

	Obtained from I_{obs}		Empirical	
Sample Size (n)	TMD	TMDL1	TMD	TMDL1
20	0.00711	0.00621	0.00729	0.00828
40	0.00511	0.00533	0.00536	0.0058
80	0.00366	0.00369	0.00354	0.00354
160	0.00261	0.00263	0.00272	0.00272
320	0.00185	0.00187	0.00201	0.00201
640	0.00131	0.00131	0.00127	0.00127

Tab. 6.8: Central Limit Theorem-based confidence intervals for asymptotic empirical variances under TMD and TMDL1

	TMD		TMDL1	
Sample size (n)	Lower	Upper	Lower	Upper
20	0.00554	0.01064	0.00629	0.01209
40	0.00439	0.00688	0.00475	0.00744
80	0.00306	0.00419	0.00306	0.00419
160	0.00245	0.00305	0.00245	0.00305
320	0.00186	0.00217	0.00186	0.00217
640	0.00120	0.00134	0.00120	0.00134

We compute the ARE for estimating μ in an inverse-Gaussian lifetime density function for pairs of (ρ, τ), $\rho = (0, 0.3, 0.6, 0.9)$, and $\tau = (4, 7, 10, 13)$. For each pair of (ρ, τ) we simulated R=500 data-sets and computed $\hat{\boldsymbol{\theta}}^*_r$, $r = 1, \ldots, R$, for TMD and TMDL1 data, treating TMD as a subset of the data available under TMDL1. Table 6.9 presents the ARE's for all (ρ, τ) pairs.

From table 6.9 we see that the TMDL1 data-structure is more efficient for combinations of (ρ, τ) where ρ is strong, and τ is small. The rationale behind this results, can be arguably related to the model specifications. Both TMD and TMDL1 "benefit" from long test-times because more information on the degradation is revealed in surviving devices. Presumably information on the degradation variable, is more valuable than information on the marker variable, no matter how strongly correlated it is to degradation. When test-times are short, however, fewer failures are observed, and therefore little information is gained on the actual degra-

dation. In this case it matters how strongly correlated the marker variable is to the degradation, and table 6.9 verifies this hypothesis.

Once again, we have grounds to advertise the TMDL data-structures in application areas with few failure observations. Although direct information on the degradation variable is shown here to "outweigh" information on the marker variable (under certain conditions), it remains to be seen if there comes a point when this is no longer the case. In other words, for estimation purposes, does access to terminal degradation information become irrelevant when we have a lot of longitudinal marker information? One must also consider computational complexity and computational efficiency to make the comparison fair, after all, multivariate longitudinal marker observations require more complex multivariate (nested) integration.

Tab. 6.9: Relative efficiency of $\hat{\mu}_{TMDL1}$ versus $\hat{\mu}_{TMD}$, for different (ρ, τ) combinations

ρ/τ	4	7	10	13
0	1.00053	0.99984	0.99974	0.99983
0.3	1.01364	1.00681	1.00269	0.99854
0.6	1.07101	1.04112	1.01713	0.99849
0.9	1.24527	1.16814	1.06883	1.01328

6.3.6 *Simulation 3 - TMD vs. TMDL1 vs TMDL2*

In effort to gain further insight into the contribution that additional marker observations have on estimation, we investigate the TMDL2 data-structure. We again simulate R=250 independent data-sets $\boldsymbol{D}$, for $n = (20, 40, 80, 160)$, and computed the MLEs $\hat{\boldsymbol{\theta}}$ for each. MLEs from the TMD data-structure are represented by vector $\hat{\boldsymbol{\theta}}_{TMD}$, from the TMDL1 data-structure by $\hat{\boldsymbol{\theta}}_{TMDL1}$, and from the TMDL2 data-structure by $\hat{\boldsymbol{\theta}}_{TMDL2}$. Under TMDL2, for failed devices, we observe $\boldsymbol{Y} = (Y(t_1), Y(t_2), Y(s))$, and for surviving devices $\boldsymbol{Y} = (Y(t_1), Y(t_2), Y(\tau))$, and as always $0 \leq s \leq \tau$, and $t_1 < t_2 < \tau$. Devices that fail before t_2 contribute to the

likelihood function using the TMDL1 data-structure, and devices that fail before t_1 contribute through the TMD data-structure.

The same parameter set $\boldsymbol{\theta} = (0.1, 1.0, 0.4, 0.1, 0.75)$ is used, and lifetimes are generated in the same way as before, with Δt=0.01 and a=1. In each simulation run j, the intermediate marker observations are made at $t_1 = \tau/3$ and $t_2 = 2\tau/3$. Table 6.10 reports the asymptotic standard errors for ν_X computed from I_{obs}. The TMDL2 data-structure, on average, results in smaller asymptotic standard errors, than the TMDL and TMD structures, indicating therefore, without the aid of AREs, improved estimation. Further higher dimensional longitudinal data-structures are not investigated in this work, due to lack of computing power for evaluating higher dimensional integrals. Using the results in table 6.10 however, we can assume that inference will improve with more marker observations. It is not clear from these results if the influence of additional marker observations will attenuate.

Tab. 6.10: Asymptotic Standard Errors for ν_X under TMD, TMDL1 and TMDL2 from Observed Information

	Obtained from I_{obs}		
Sample Size	TMD	TMDL1	TMDL2
20	0.00612	0.00612	0.00604
40	0.00483	0.00483	0.00466
80	0.00365	0.00365	0.00360
160	0.00270	0.00270	0.00193

6.4 *Degradation data on Aircraft Gas-Turbine Engines*

In this section we present analysis of degradation data collected on aircraft gas-turbine engines. Multivariate time series observations on $m = 21$ covariates are made on a cohort of $n = 218$ independent engines. Each engine is randomly stressed under 8 usage conditions (settings): $ST - 0, 20, 40, 60, 80, 100$. We observed that multivariate projections of the 21 covariates from each setting form distinct clusters. Therefore, we consider data collected only under setting $ST3 - 0$ for our analysis.

We chose $ST3 - 0$ arbitrarily amongst the 6 settings.

One of the key features in the data is a latent degradation variable, which reflects the fact that the degradation variable is not observed. Because each device is tested to failure, however, we can assume that for each failed device, the degradation variable reached a fixed known threshold at the failure time. This assumption allows us to apply the FHT models presented in this thesis. In the simplest case, we simulate degradation to be linear in cycles, with heterogeneous drift and variance across devices. In the analysis we make the following assumptions:

- The degradation variable starts at level 0, and reaches a=1 at failure-time
- Degradation paths are assumed linear and proportional to the device's lifetime
- Censoring of failures is made at τ=200 cycles

Degradation is simulated proportional to each device's lifetime under all usage settings. The data is then filtered for setting $ST3 - 0$. Notice that the occurrence of setting $ST3 - 0$ is non-deterministic, that is the device is exposed to the usage condition defined by $ST3 - 0$ in a random way. In the filtered data we observe a nonlinear progression of cycles for each observation. This means that not every surviving device will have have its last observation made at 200 cycles. For consistency, therefore, we adopt the following rule for data collection on surviving devices, under setting $ST3 - 0$:

- If the device is not observed on the 200th cycle, then collect data on the cycle closest to cycle 200

For example, if under setting $ST3 - 0$ device i is observed on cycles 191 198 205 and 223, then we collect data associated with cycle 198. We augment simulated degradation values to the covariate data under setting $ST3 - 0$, as illustrated in table 6.11

Tab. 6.11: Illustration of ST3-0 data-structure with augmented degradation data

Usage Setting ST3-0						
Device	Cycles	z_1	z_2	...	z_{21}	Deg
1	4	491	601	...	0.22	0.002
1	6	491	607	...	0.19	0.03
1	23	490	607	...	0.24	0.009
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	31	491	607	...	0.24	0.012
1	45	490	608	...	0.24	0.018
1	75	496	607	...	0.29	0.023

To fit this data into the TM and TMD data-structures, we are interested in suitable marker variables that can be used. As a first data-cleaning step, variables z_1, z_5, z_{18} and z_{19} with zero standard deviation are removed from the data-set, leaving 17 covariates. Figure 6.6 plots a colormap surface of the empirical correlation matrix between each of the 17 covariates and the degradation variable. From examining the correlation matrix we see that variables 9,13, and 14 are strongly correlated with each other and therefore redundant. From the remaining variables, only four show considerable correlation to the degradation variable. We continue our analysis using covariates z_4, z_8, z_{11} and z_{15}.

Using principal component analysis (PCA), we reduce the dimensionality further, and ultimately derive one variable that can be used as the marker variable. PCA forms a new set of uncorrelated (not necessarily independent) variables that we denote by z'. The PCA scores on the most dominant eigenvector show the strongest correlation to the degradation variable (~ 0.42), and therefore this variable is used as the marker variable.

Table 6.12 compares the asymptotic standard errors for TM vs TMD under increasing sample sizes $n = (20, 40, \dots, 218)$. In these results we observe and validate that under the TMD data-structure, parameters are consistently estimated with lower uncertainty. Also its interesting to observe that under both models the estimates for ρ are in close agreement to the empirical correlation between marker

	s2	s3	s4	s6	s7	s8	s9	s10	s11	s12	s13	s14	s15	s16	s17	s20	s21	deg
s2		0.547	0.643	0.45	0.554	0.515	0.273	0.335	0.674	0.578	0.517	0.18	0.642	0.547	0.557	0.484	0.498	0.612
s3	0.547		0.619	0.433	0.537	0.5	0.264	0.323	0.649	0.558	0.495	0.176	0.62	0.509	0.533	0.483	0.481	0.583
s4	0.643	0.619		0.496	0.647	0.528	0.235	0.374	0.775	0.676	0.528	0.127	0.729	0.617	0.636	0.565	0.569	0.68
s6	0.45	0.433	0.496		0.43	0.373	0.18	0.231	0.525	0.456	0.374	0.109	0.492	0.344	0.428	0.38	0.381	0.502
s7	0.554	0.537	0.647	0.43		0.434	0.172	0.322	0.673	0.575	0.435	0.076	0.633	0.545	0.545	0.497	0.495	0.586
s8	0.515	0.5	0.528	0.373	0.434		0.87	0.369	0.534	0.436	0.945	0.818	0.569	0.49	0.515	0.411	0.412	0.653
s9	0.273	0.264	0.235	0.18	0.172	0.87		0.244	0.217	0.161	0.87	0.948	0.29	0.261	0.275	0.189	0.188	0.44
s10	0.335	0.323	0.374	0.231	0.322	0.369	0.244		0.389	0.335	0.371	0.191	0.387	0.338	0.346	0.298	0.284	0.381
s11	0.674	0.649	0.775	0.525	0.673	0.534	0.217	0.389		0.716	0.533	0.101	0.764	0.654	0.671	0.593	0.6	0.706
s12	0.578	0.558	0.676	0.456	0.575	0.436	0.161	0.335	0.716		0.438	0.057	0.666	0.562	0.576	0.511	0.519	0.613
s13	0.517	0.495	0.528	0.374	0.435	0.945	0.87	0.371	0.533	0.438		0.819	0.571	0.494	0.514	0.414	0.41	0.656
s14	0.18	0.176	0.127	0.109	0.076	0.818	0.948	0.191	0.101	0.057	0.819		0.183	0.17	0.185	0.102	0.102	0.348
s15	0.642	0.62	0.729	0.492	0.633	0.569	0.29	0.387	0.764	0.666	0.571	0.183		0.622	0.635	0.554	0.563	0.688
s16	0.547	0.509	0.617	0.344	0.545	0.49	0.261	0.338	0.654	0.562	0.494	0.17	0.622		0.544	0.479	0.481	0.604
s17	0.557	0.533	0.636	0.428	0.545	0.515	0.275	0.346	0.671	0.576	0.514	0.185	0.635	0.544		0.497	0.498	0.608
s20	0.484	0.483	0.565	0.38	0.497	0.411	0.189	0.298	0.593	0.511	0.414	0.102	0.554	0.479	0.497		0.433	0.533
s21	0.498	0.481	0.569	0.381	0.495	0.412	0.188	0.284	0.6	0.519	0.41	0.102	0.563	0.481	0.498	0.433		0.53
deg	0.612	0.583	0.68	0.502	0.586	0.653	0.44	0.381	0.706	0.613	0.656	0.348	0.688	0.604	0.608	0.533	0.53	

Fig. 6.6: Correlation matrix between covariates (1 through 17) and degradation (18)

and degradation (~ 0.42). Similarly we observe that the MLE for ν_X is in close agreement with the empirical rate, which is calculated by 1/(average number of cycles to failure) $\sim 200 = 0.005$. Table 6.13 compares the asymptotic standard errors and MLEs for ρ across the same range of sample sizes. Here again we observe an improved inference under the TMD data-structure. For larger n, although the variance is reduced (under both data-structures) the MLE for ρ seems to be relatively biased to the empirical correlation.

The proportion of failed to survived devices in each sample size is random and approximately equal to 50% as noted in Table 6.11. This is a result of constructing a sample by adding records from subsequently observed censored and failed devices, starting from the 1st device. For example, for a sample of size $n = 20$, we construct the data set by adding an additional row to the data-set for each device number, starting from device 1. In this way, the proportion of devices out of $n = 20$ that fail

Tab. 6.12: Asymptotic Standard Error for ν_X under TM and TMD from Observed Information. Empirical drift ~ 0.005

	MLE		SE		
Sample Size	TM	TMD	TM	TMD	Prop. Failed
20	0.00511	0.00526	0.0003445	0.000244	0.55
40	0.005	0.00505	0.0002116	0.0001598	0.525
60	0.00483	0.00497	0.0002000	0.0001305	0.466
79	0.00486	0.00501	0.0001757	0.0001150	0.482
99	0.00483	0.00499	0.000153	9.91E-05	0.484
119	0.00481	0.00498	0.0001385	8.79E-05	0.478
139	0.00485	0.00502	0.0001279	8.18E-05	0.489
159	0.00483	0.005	0.0001228	7.81E-05	0.484
179	0.00482	0.00498	0.0001128	7.22E-05	0.48
199	0.00481	0.00499	0.0001099	6.94E-05	0.482
218	0.0048	0.00499	0.000108	6.75E-05	0.481

or survive is random.

Tab. 6.13: Asymptotic Standard Error for ρ under TM and TMD from Observed Information. Empirical correlation $= 0.42$

	MLE		SE	
Sample Size	TM	TMD	TM	TMD
20	0.42394	0.44558	0.18111	0.16280
40	0.39581	0.44455	0.13869	0.11626
60	0.40546	0.41031	0.11424	0.09983
79	0.43247	0.42076	0.09574	0.08643
99	0.43147	0.41128	0.08417	0.07767
119	0.43723	0.40441	0.07613	0.07194
139	0.41361	0.38366	0.07177	0.06783
159	0.40293	0.38431	0.06872	0.06307
179	0.40918	0.39213	0.06449	0.05918
199	0.36231	0.35231	0.06408	0.058395
218	0.35679	0.33982	0.06128	0.056277

We are also interested in studying the ARE of the TM vs TMD data-structures when ρ is kept fixed and known and sample size is decreased. We are again expecting, based on the asymptotic standard errors presented in Tables 6.11 and 6.12 that the TMD with outperform the TM model. For each pair (ρ, n), $\rho = 0, 0.052, 0.104, \ldots 0.99$, $n = (20, 40, \ldots, 218)$, we computed $\hat{\theta}$, for TM and TMD data, treating again TM as a subset of the data available under TMD.

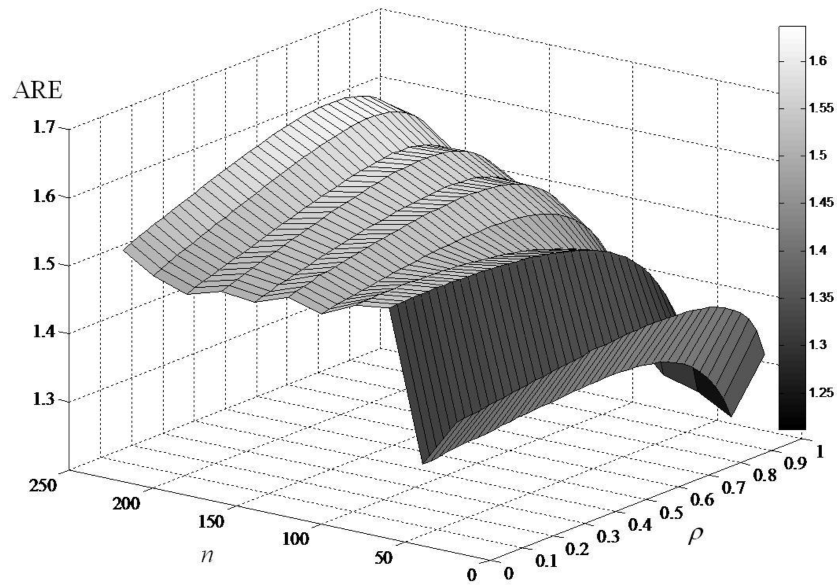


Fig. 6.7: Relative efficiency of $\hat{\mu}$ from G_2 vs. G_1 for different (ρ, n) combinations evaluated with gas-turbine engine degradation data

Figure 6.4 plots the results of the ARE experiment, and tells a different story from the ARE results in the simulations. In this case, the proportion of failed to surviving devices remains the same for all sample sizes. We observe that the efficacy of the TMD data-structure decreases with increasing ρ and decreasing sample size n. The rationale behind these results can be explained in terms of information content. When ρ is high it favors the TM data-structure because it can draw more information on surviving devices from the marker observations. From the TMD perspective, when ρ is high, therefore the marker information becomes more valuable, it takes away the advantage of observing terminal degradation. On the other hand, with large sample sizes the TMD data-structure slowly regains its advantage.

7. COVARIATES AND REGRESSION STRUCTURES

Emphasis in this thesis has been given to the development of parametric inference procedures for the bivariate Wiener model under various extensions to the TM data-structure. The effect of covariates still remains to be incorporated into these models. Covariates are measurable variables that contain information about device specific performance and environment. Covariate data in an engineering setting are usually collected from sensors embedded onto or near the device, and designed to measure targeted information related to the health/degradation of the device. Reliability models that incorporate covariate data together with lifetime data are anticipated to improve reliability predictions.

Much literature has been devoted to reliability models for lifetime data with covariates. This is especially true in what are known as accelerated failure time models, with early work by Epstein and Sobel 1963, Singpurwalla 1970, and many others, some of which we discuss later in this chapter. The main purpose of this chapter is to discuss the relevance of using covariates in degradation models, and specifically in FHT models.

Lee and Whitmore 2006 introduce *Threshold Regression* models, which are FHT models that include regression structures. Regression structures allow effects of covariates to explain some or all of the dispersion of the data, thereby taking account of variability ad sharpening inferences [Whit, Lee 2006]. The idea of using covariates to aid estimation and inference can be found in a broad range of literature in survival and reliability analysis. The main idea is that unknown distribution or process parameters are reparameterized as functions of covariates.

In FHT models, one can think about reparameterizing the drift parameter ν_X as $\nu_X = \alpha - \beta \boldsymbol{Z}$, where $\boldsymbol{Z}$ is a vector valued covariate, and α, β are new unknown parameters that need to be estimated. In electronics reliability experiments, accelerated lifetimes are most common, induced by higher than normal experimental stress conditions, such as high temperature, humidity and pressure. It becomes important to model the influence of covariates that measure the stresses and other environmental conditions, on the rate of degradation. In other words, they include the effect that covariate information has on the drift of the degradation process. This relationship is especially important during predictive inference calculations on test devices. In the example used above, instead of using the MLE $\hat{\nu_X}$ as a plug-in estimator in predictive inference equations, we can use $\hat{\alpha} - \hat{\beta} Z_i$ for device i, therefore explaining the drift accounting for heterogeneous covariate observations across the training sample.

One question that arises is what regression structure to use that best captures the functional relationship between covariates and the dependent variable of interest, like, drift or more generally, degradation. In cases where failure-mechanisms are well understood, then PoF models can motivate the appropriate regression structure. PoF models typically relate covariates at a point in time to either the degradation level or to the lifetime scale parameter. In case where lifetimes are attained under what are called accelerated test conditions, PoF models also try to account for the acceleration. When failure-mechanisms are not well understood, then data-driven approaches can help identify the statistical relationship. In the remainder of the chapter we discuss various regression structures that can be considered in the context of FHT models.

7.1 Multiplicative Hazards Regression Models

In survival analysis the hazard rate $\lambda(t)$ captures the instantaneous probability of failure at time t, and forms the basis for multiplicative hazard regression models:

$$\lambda_i(t) = \lambda_0(t) r(\boldsymbol{z}_i(t))$$

where $\lambda_0(t)$ is a *baseline hazard* function and $r(\boldsymbol{z}_i(t))$ is a positive valued device-specific *relative-risk multiplier*. Typically, the relative-risk is specified by $r(\boldsymbol{z}_i(t)) = exp(\boldsymbol{\theta}^T \boldsymbol{z}_i(t))$ with parameter vector $\boldsymbol{\theta}$. The hazard ratio between any two devices, $\lambda_i(t)/\lambda_j(t) = exp(\boldsymbol{\theta}^T(\boldsymbol{z}_i(t) - \boldsymbol{z}_j(t)))$, is constant if the difference in covariates is constant in time, specifying therefore a proportional hazard model.

In Cox's proportional hazard regression model $\lambda_i(t) = \lambda_0(t) exp(\boldsymbol{\theta}^T \boldsymbol{z}_i(t))$, the baseline hazard is left unspecified. Its semi-parametric form can help reduce estimation bias of covariate effects. The parameter vector $\boldsymbol{\theta}$ can be estimated using the partial likelihood [74].

7.2 Accelerated Failure Time Regression Models

Accelerated tests are typically used to collect information on the life distribution or performance over time of products. Meeker and Escobar 1993 provide an excellent review of research and issues in accelerated testing. In their 2002 book, Bagdonavicius and Nikulin present a comprehensive review of univariate accelerated life models, Nelson 1990 describes accelerated life models and life-stress relationships such as the Arrhenius, the inverse-power, and the fatigue relationships.

Accelerated failure time regression models, provide an alternative to the commonly used proportional hazards models. AFT regression models assume that the effect of a covariate is on the failure-time itself. AFT regression models are typically based on log transformations of the failure-time. For example, in a log-linear

failure-time model $log(s_i) = \mu_i + \boldsymbol{\beta}^T \boldsymbol{z}_i + \epsilon$, with $-\infty < \mu_i < \infty$, $\boldsymbol{\beta} > 0$ and ϵ is a normally i.i.d. error variable. The survival function and hazard rate are then given by:

$$S(t|\boldsymbol{z}_i) = S_0(texp(-\boldsymbol{\beta}^T \boldsymbol{z}_i))$$

$$\lambda(t|\boldsymbol{z}_i) = \lambda_0(texp(-\boldsymbol{\beta}^T \boldsymbol{z}_i))exp(-\boldsymbol{\beta}^T \boldsymbol{z}_i)$$

where S_0, the baseline survival function is given by:

$$S_0(t) = Pr(exp(\mu_i + \epsilon) > t)$$

The conditional survival function above, provides a mapping of survival function between baseline functions and functions at any given covariate level. Interestingly, when the baseline is Weibull, then $\lambda_i(t) = \lambda_0(t)(t\exp(-\alpha\boldsymbol{\beta}^T \boldsymbol{z}_i))$, where α is the Weibull probability density function shape parameter. This is equivalent to the multiplicative hazard form $\lambda_i(t) = \lambda_0(t)r(\boldsymbol{z}_i(t))$. When certain covariates are not observable, heterogeneity also leads to unexplained variation. Frailty models can be used to represent unobserved heterogeneity as a random variable.

7.3 The Marker Variable as a Special Covariate

In FHT models, marker variables, as discussed earlier, form the basis for inference in bivariate latent degradation models. Maker variables are generally chosen from the available covariates available on a device, making the bivariate model a type of regression model. The marker is considered to be a "special" covariate in that we attribute greater importance to it, because we believe it is more closely correlated to the degradation variable. At a fixed time point, markers, unlike covariates are treated as random variables related to degradation through a parametric model with unknown parameters. Indexed by time, therefore, a collection of marker

variables forms a stochastic process.

Markers, often are taken as functions of several covariates, so called *composite* markers. Composite markers can be derived, for example, from principal component analysis (PCA), factor analysis (FA), or more generally from generalized linear models. Covariate data scores onto the principal eigenvectors, or latent factors, in PCA and FA respectively, can be used to represent the composite marker. As another example, a composite marker can be constructed using the time-dependent cox-proportional hazards model. In this case, the composite marker variable represents the instantaneous risk of failure or the hazard rate.

Next we present two less traditional regression structures, that we argue, can be made available through machine learning methodology.

7.4 Gaussian Process Regression

Gaussian processes provide a computationally practical and tractable framework to deal with high dimensional covariate observations. We show that because a *Gaussian process* can be completely specified by its mean and covariance function, it can explain the variability of a dependent variable, such as degradation. Gaussian processes can be used when we are interested in making inference about the relationship between covariates and a dependent variable, i.e. the conditional distribution of the dependent variable given the covariates [75]. From a machine learning perspective *Gaussian process regression* is approached as an unsupervised learning problem, where the task is to find suitable properties for the covariance function. Many covariance functions, such as the squared exponential (SE), radial basis function (RBF), rational quadratic (RQ), mattern, can be used, among others, and each have parameters that need to be inferred or learned from the data.

Linear regression, where the output is the linear combination of inputs is simple and easy to interpret. It is, however, inflexible when the output cannot

reasonably be approximated by a linear function, like for example degradation or its drift. It is reasonable, for example, to assume that under higher stress conditions the degradation drift will increase, but not necessarily linearly.

Consider a cohort of n independent devices tested over a period $[0, \tau]$ as discussed in chapter 2. The training set D consists of n observations $D = [(\boldsymbol{z}_i, \boldsymbol{x}_i)|i = 1, \ldots, n]$, where $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots)$ denotes the input vector (covariates) and $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots)$ denotes the output vector for the ith device. In matrix form we can write $D = (\boldsymbol{Z}, \boldsymbol{x})$. Note as discussed in chapter 2 the number of input/output observations made on a device depends on whether it survives or fails in the period $[0, \tau]$.

7.4.1 Linear Model

The standard linear model with Gaussian noise is given by:

$$x = \boldsymbol{z}^T w + \epsilon \tag{7.1}$$

where $\boldsymbol{w}$ is the parameter vector of the linear model, $f(\boldsymbol{z}) = \boldsymbol{z}^T w$, and $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is the assumed error distribution between x and $f(z)$ and is assumed $i.i.d$ across devices. The likelihood of the data given the parameters is given by:

$$L(\boldsymbol{x}|\boldsymbol{Z}, \boldsymbol{w}) = \prod_{i=1}^{n} f_X(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{Z}^T\boldsymbol{w}, \sigma_\epsilon^2) \tag{7.2}$$

In a Bayesian fashion we put a prior on the parameters $\boldsymbol{w} \sim \mathcal{N}(0, \Sigma_z)$, where Σ_z is the covariance matrix on the parameters, which we discuss further in section 7.8. Inference in the Bayesian linear model is based on the posterior distribution over the parameters. The posterior distribution of the parameter vector is proportional

to the likelihood times the prior, given by:

$$f_{\boldsymbol{w}}(\boldsymbol{Z}, \boldsymbol{x}) \sim \mathcal{N}\left(\frac{\boldsymbol{Zx}}{\sigma_\epsilon^{-2} Z Z^T + \Sigma_z^{-1}}, \frac{1}{\sigma_\epsilon^{-2} Z Z^T + \Sigma_z^{-1}}\right) \tag{7.3}$$

To make predictions at an unobserved output vector (test vector) x_* conditioned on inputs z_*, we average over all possible parameter values weighted by their posterior probability:

$$f_{\boldsymbol{X}^*}(x^* | \boldsymbol{z}^*, \boldsymbol{Z}, \boldsymbol{x}) = \int f_{\boldsymbol{X}^*}(\boldsymbol{x}^* | \boldsymbol{z}^*, \boldsymbol{w}) f_{\boldsymbol{w}}(\boldsymbol{w} | \boldsymbol{Z}, \boldsymbol{x}) d\boldsymbol{w} \tag{7.4}$$

To overcome the limited expressiveness of linear models, we project covariates into a selected feature space. In summary, the covariates are transformed through a mapping function ϕ into some high dimensional space where the relationship between the covariates and the dependent variable is more linear. As long as the mapping is a fixed function, such as $\phi(z) = (1, z, z^2, z^3, \dots)^T$, i.e. independent of the parameters $\boldsymbol{w}$ the model is still linear in $\boldsymbol{w}$. The function $\phi(z)$ maps an $m-$dimensional covariate vector $\boldsymbol{z}$ into an $M-$dimensional feature space. The transformed model is now given by: $f(\boldsymbol{z}) = \phi(\boldsymbol{z})^T \boldsymbol{w}$, where and the predictive distribution becomes:

$$f_{\boldsymbol{X}^*}(x^* | \boldsymbol{z}^*, \boldsymbol{Z}, \boldsymbol{x}) \sim \mathcal{N}\left(\phi_*^T \Sigma_z \Phi (K + \sigma_\epsilon^2 I)^{-1} \boldsymbol{x}, \phi_*^T \Sigma_z \phi_* - \phi_*^T \Sigma_z \Phi (K + \sigma_\epsilon^2 I)^{-1} \Phi^T \Sigma_z \phi_*\right) \tag{7.5}$$

where $\phi_* = \phi(x_*)$, $\Phi(\boldsymbol{Z})$ is a matrix of columns $\phi(\boldsymbol{z})$. and $K = \Phi^T \Sigma_z \Phi$.

7.4.2 Function Space View

Before we looked at the weight space point of view, which keeps closer with the linear model perspective. An equivalent way to think about things is the function space perspective. Here, the weight vector becomes latent and the important concept is the function itself. So instead of trying to estimate the best posterior $\boldsymbol{w}$, we

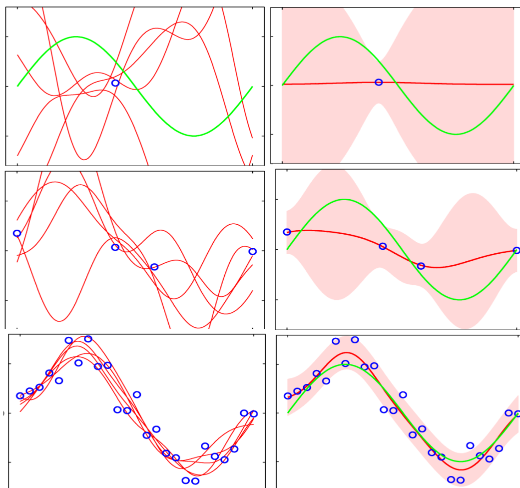


Fig. 7.1: Function space view. Left: nine functions drawn at random from a GP prior, and the dot plots a single observation x. Right: Nine random functions drawn from the posterior. In both plots the shaded area represents point wise mean plus and minus two times the standard deviation.

directly estimate the posterior distribution over the functions themselves. This is possible in GPs where the function is determined by its mean and variance, where the variance is also known as kernel function and is specified by the user. In the context of the GP, all random variables, including the r.v. at test data points are jointly Gaussian, and this means that there are an infinite number of functions (derived by a specific GP) that can be used to fit the data.

Here we use an example to illustrate the function-space inference process. In Figure 7.1 observations on x are generated by sampling from a sine function. The first tier in 7.1 shows the prior and the posterior generated after only one observation on x, the second tier after 4, and the third after many observations on x. With more covariate information in the third tier, we observe that the uncertainty in estimation is reduced.

Definition 7.1 (Gaussian Process). *A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution*

A Gaussian process is completely specified by its mean function $m(\boldsymbol{z})$ and covariance function, (or kernel function) $k(\boldsymbol{z}, \boldsymbol{z}')$ of a real process $f(\boldsymbol{z})$. From the

Bayesian linear regression model we can write $f(z) = \phi(\boldsymbol{z})^T \boldsymbol{w}$ with prior $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_z)$. Then

$$E[f(\boldsymbol{z})] = \phi(z)^T E[\boldsymbol{w}]$$

$$E[f(\boldsymbol{z})f(\boldsymbol{z}')] = \phi(\boldsymbol{z})^T E[\boldsymbol{w}\boldsymbol{w}^T]\phi(\boldsymbol{z}') = \phi(\boldsymbol{z})^T \Sigma_z \phi(\boldsymbol{z}') \tag{7.6}$$

The covariance function specifies the covariance between pairs of random variables, for example, $cov(f(\boldsymbol{z}_p), f(\boldsymbol{z}_q)) = k(\boldsymbol{z}_p, \boldsymbol{z}_q) = exp(-1/2|\boldsymbol{z}_p - \boldsymbol{z}_q|^2)$. The specification of the covariance function implies a distribution over functions. This can be seen in Figure 7.1, the prior as mentioned earlier is sampled from:

$$f(z_*) \sim \mathcal{N}\left(\mathbf{0}, K(Z_*, Z_*)\right) \tag{7.7}$$

where K is the matrix of kernel functions $k(\cdot)$. The posterior predictive distribution at test covariates $\boldsymbol{Z}_*$, i.e. the prior conditioned on training covariate observations $\boldsymbol{Z}$ and new test covariate observations $\boldsymbol{Z}_*$, is given by:

$$\begin{aligned} f(\boldsymbol{z}_*)|\boldsymbol{Z}_*, \boldsymbol{Z}, \boldsymbol{x} \sim \mathcal{N}\big(&K(\boldsymbol{Z}_*, \boldsymbol{Z})K(\boldsymbol{Z}, \boldsymbol{Z})^{-1}\boldsymbol{x}, \\ &K(\boldsymbol{Z}_*, \boldsymbol{Z}_*) - K(\boldsymbol{Z}_*, \boldsymbol{Z})K(\boldsymbol{Z}, \boldsymbol{Z})^{-1}K(\boldsymbol{Z}, \boldsymbol{Z}_*)\big) \end{aligned} \tag{7.8}$$

The expected value of the dependent variable, conditioned on training and test covariate observations, is given by: $K(\boldsymbol{Z}_*, \boldsymbol{Z})K(\boldsymbol{Z}, \boldsymbol{Z})^{-1}f(\boldsymbol{z})$, where $K(\boldsymbol{Z}_*, \boldsymbol{Z})$ is the kernel matrix whose elements define the the covariance between each test observation z_* and each training covariate observation z, and $\boldsymbol{x}$ is the vector of observations on the dependent variable. The conditional mean is a linear combination of observations $\boldsymbol{x}$:

$$m(\boldsymbol{z}_*) = \sum_{i=1}^{n} \alpha_i k(\boldsymbol{z}_i, \boldsymbol{z}_*) \tag{7.9}$$

where $\boldsymbol{\alpha} = K^{-1}\boldsymbol{x}$.

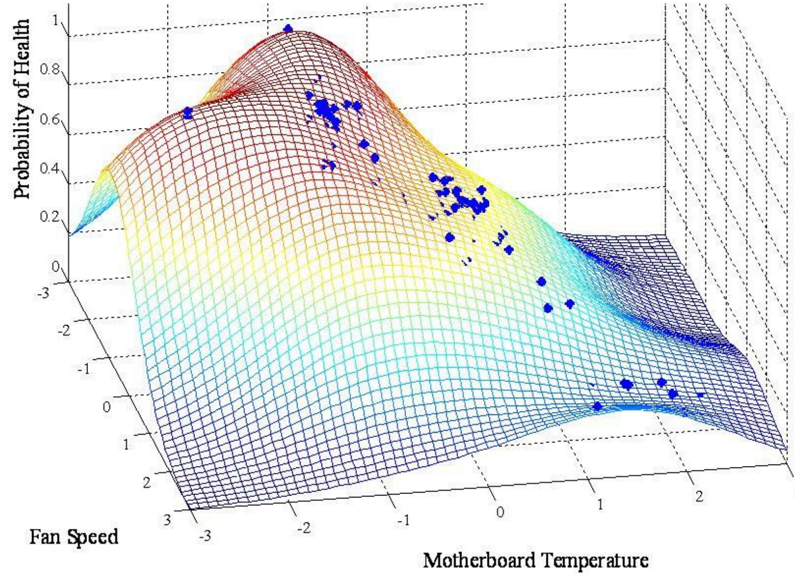


Fig. 7.2: GP mean function $m(\boldsymbol{z}_*)$. The blue dots are the training data $(\boldsymbol{x}, \boldsymbol{z})$.

7.4.3 Connection to FHT models

In the context of FHT models, we are interested in explaining heterogeneous drift conditioned on device specific covariate observations. In the case where degradation is observed we can examine the regression structure between it and the observed covariates. In this case the influence of the covariates is incorporated by re parameterizing the drift ν_X with the conditional mean $m(\boldsymbol{z}_*)$. Maximum likelihood optimization will be therefore performed over the parameter set $\boldsymbol{\alpha}$ instead of ν_X. Although covariates are empirically correlated to the degradation variable in this case, we stipulate, because degradation is a linear function of drift, then we can safely extend the relationship to exist between covariate and drift parameter.

Figure 7.2 plots the mean function of a GP regression model applied to a time series of a dependent variable and its associated covariates. We can see that covariates *motherboard temperature* and *fan speed* together with dependent variable observations are used to predict the functional dependency at other data combinations within the bivariate covariate range. Specifically, in this example, the de-

pendent variable represents the health probability of the device under observation. The health probability, is an output from the *Support Vector Degradation* model, discussed in chapter 10.

Because GPs are computationally very tractable and easy to compute, they can be used to handle large multivariate covariate observations. Typically covariates and dependent variables are indexed by time, and therefore, GP regression models can also be used later for event-time prediction, and we discuss this in more detail in chapter 10.

7.5 Support Vector Machines

Support vector machines (SVM) are based on the idea of large-margin linear discriminants that seek to find a function $f(data)$ to separate two or more classes of data by maximizing what is called the *hyperplane margin*. In this section we consider linear SVMs, and their possible connection to FHT models.

Consider a cohort of n independent devices tested over a period $[0, \tau]$, and define training data as: $(\boldsymbol{Z}, \boldsymbol{y})$ where $\boldsymbol{Z}$ is a collection of r covariate vector observations on n healthy devices and l covariate vector observations on q failed or severely degraded devices, such that $Z = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_r, \boldsymbol{z}_{r+1}, \ldots, \boldsymbol{z}_{r+l})$. Typically the healthy training data are collected by observing all n devices early in their life, up until some predefined time. The degraded training data are collected on all q failed devices from some predefined time before their observed failure time.

The class label of each $\boldsymbol{z}_i$ is given by $y_i \in (+1, -1)$, where $+1$ indicates the membership of $\boldsymbol{z}_i$ into the degraded/failed class. The training class label vector $\boldsymbol{y}$ consists of r healthy covariate observations and l degraded. We assume that $\boldsymbol{Z}$ can be separated by a decision function $f(\boldsymbol{z}; \boldsymbol{w}, b)$ with appropriate parameters $\boldsymbol{w}$ and

b, given by:

$$f(\boldsymbol{z}) = (\boldsymbol{w}^T\boldsymbol{z}) + b = \sum_{i=1}^{m} w_i z_i + b \tag{7.10}$$

subject to

$$y_i(\boldsymbol{w}^T\boldsymbol{z} + b) - 1 \geq 0$$

where $\boldsymbol{w} = [w_1, \ldots, w_m]^T$ is the weight vector and $\boldsymbol{z} = [z_1, \ldots, z_m]^T$. Training covariate observations $\boldsymbol{z}$ with $y_i = +1$ will fall into $f(\boldsymbol{z}; \boldsymbol{w}, b) > 0$ while the others with $y_i = -1$ will fall into $f(\boldsymbol{z}; \boldsymbol{w}, b) < 0$. A new (non-training) covariate vector measurement $\boldsymbol{z}_*$ is evaluated using:

$$f(\boldsymbol{z}_*) = \sum_{i=1}^{n} y_i \alpha_i \boldsymbol{z}_i^T \boldsymbol{z}_* + b \tag{7.11}$$

where

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

and

$$b = \frac{1}{n} \sum_{i=1}^{n} y_i \left(1 - \sum_{j=1}^{n} \boldsymbol{H}_{ji} \alpha_j \right) \tag{7.12}$$

where $\boldsymbol{H}$ is the Hessian matrix: $\boldsymbol{H} = y_i y_j \boldsymbol{z}_i^T \boldsymbol{z}_j$. The decision function $f(\boldsymbol{z}_*)$ uses training covariate vector data and their class label to evaluate the test observation $\boldsymbol{z}_*$.

7.5.1 *Connection to FHT models*

The drift parameter of the degradation process $X(t)$ can be reparametrized as a function of the distance of z_* to the decision boundary f specified by $\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{z}_i^T$. The perpendicular distance of $\boldsymbol{z}_*$ to f is given by:

$$d(\boldsymbol{z}_*) = \frac{f(\boldsymbol{z}_*)}{||\boldsymbol{w}||} \tag{7.13}$$

The drift parameter ν_X, can therefore, be expressed as a function $\boldsymbol{z}_*$ and its perpendicular distance to the decision boundary f that best linearly separates the two classes of covariate observations in the training set. One possible configuration is given by:

$$\nu_X = \beta + d(\boldsymbol{z}_*) \tag{7.14}$$

In equation (7.14) we see that when the distance $d(\boldsymbol{z}_*)$ is close to zero, ν_X is not influenced strongly by the covariate observation. For large positive and negative covariate observations however, ν_X increases or decreases respectively, and proportionally to $d(\cdot)$. This formulation therefore, incorporates the effects of covariates on drift, via a non-parametric classification function that discriminates between healthy and unhealthy/degraded covariate observations. In our opinion, this approach to including covariates in degradation models is useful in the case of large multivariate covariate data-sets. We anticipate further work on this area as part of future research.

8. VARIABLE THRESHOLD MODEL

Beyond the fixed-threshold model of Whitmore, there is also a need for models that account for uncertain failure thresholds, which arise in reliability data when failure is not defined deterministically from the degradation variable. Random thresholds can occur when failure is observed as the result of:

1. an externally observed process reaching a nominal state
2. a surrogate degradation variable X first hitting a threshold a

In either case, at the failure-time, the "true" degradation variable X^* will vary across devices. Some examples of (1) are: failure of electronics when system performance decreases, death due to illness, bankruptcy, etc. Some examples for (2) are: fan bearings fail when acoustic noise, X, first reaches a fixed threshold, however at this time, X^*, the bearing surface roughness, will vary from one failed device to another. A capacitor fails when resistance, X, reaches a fixed threshold level, but at this time (failure-time), X^*, the length of a micro-crack in its solder joint is observed to vary across devices.

Using data in case (2) setting, inference is based on the conditional true degradation estimate, $f_{X(t)|Y(t),X^*(t),A}(x|y,x^*,a)$. The threshold variable is denoted by A and takes a distribution of values that vary across devices. In the simplest case, $A \sim \mathcal{N}(\mu_A, \sigma_A^2) \perp X(\cdot), Y(\cdot)$, and is random but constant over time. The failure-time random variable is then given by $S = inf(t : X(t) = A)$. More general cases the threshold level is modeled by stochastic process with drift $A(t) \sim \mathcal{N}(a_0 + \nu_A t, \sigma_A^2 t)$,

where a_0, ν_A, σ_A are unknown parameters, and the failure-time is given by: $S = inf(t : X(t) = A(t))$.

8.1 Uncertain Failure Thresholds

In this chapter, we derive the parametric inference equations for the TMD data-structure using a variable failure threshold. The model is therefore misspecified, because it assumes degradation at failure can vary, when in fact its fixed and equal to a. It is however of practical interest to investigate the efficiency of a variable threshold model applied to TMD type data-structures. The reason is that failure thresholds are often arbitrarily chosen, or based on antiquated standards, or more generally not known. Sometimes engineers can know the range of failure thresholds, and can prescribe beliefs for the most likely thresholds. In such cases parametric models for the failure threshold are needed, eg., Gaussian as mentioned earlier.

8.2 Parametric Inference

Parametric inference is based on observations on q failed, and p surviving devices. To simplify exposition, covariates are not included in the following derivations. The parameter space is given by vector $\underline{\theta} = (\nu_X, \nu_Y, \sigma_X, \sigma_Y, \rho, \nu_A, \sigma_A)$, and the likelihood function is given by:

$$L_{\underline{\theta}} = \prod_{i=1}^{q} f_{Y(S),X(S),T,I(S<\tau)}(y_i, x_i, s_i, 1) \prod_{j=1}^{p} f_{Y(\tau),X(\tau),T,I(S<\tau)}(y_j, x_j, \tau, 0) \tag{8.1}$$

8.2.1 Contribution to likelihood from failed devices

The joint density for a failed device is given by:

$$f_{Y_S,X_S,T,I(S<\tau)}(y, x, s, 1) = C_1 C_2 C_3 \tag{8.2}$$

$$C_1(y,x,s) = \frac{1}{\sqrt{2\pi\sigma_Y^2(1-\rho^2)s}} exp\left(-\frac{(y-cx-(\nu_Y - c\nu_X)s)^2}{2\sigma_Y^2(1-\rho^2)s}\right)$$

$$C_2(x,s) = \frac{x}{\sqrt{2\pi\sigma_X^2 s^3}} exp\left(-\frac{(x-\nu_X s)^2}{2\sigma_X^2 s}\right)$$

$$C_3(x) = \frac{1}{\sqrt{2\pi\sigma_A^2}} exp\left(-\frac{(x-\nu_A)^2}{2\sigma_A^2}\right)$$

Proof.

$$f_{Y_S,X_S,T,I(S<\tau)}(y,x,s,1) = f_{Y_S,X_S,S,I(S<\tau)}(y,x,s,1)$$

Because $s \leq \tau$, $f_{S,I(S<\tau)}(s,1) = f_S(s)$, therefore

$$f_{Y_S,X_S,S,I(S<\tau)}(y,x,s,1) = f_{Y_S,X_S,S}(y,x,s) = f_{Y_S|X_S,S}(y|x,s) f_{X_S,S}(x,s)$$

Due to lemmas 4.3 and 4.4, $f_{Y_S|X_S,S}(y|x,s) = f_{Q_S}(y-cx)$. Therefore,

$$f_{Y_S,X_S,S}(y,x,s) = f_{Q_S}(y-cx) f_{X_S,S}(x,s) \tag{8.3}$$

For failed items, the level of degradation x at the failure-time s is equal to the threshold a. Because both $X(.)$ and A are Gaussian, we can replace $X(S)$ with A. From equation (8.3) we get:

$$f_{Y_S,X_S,S}(y,x,s) = f_{Q_S}(y-cx) f_{A,S}(x,s) = f_{Q_S}(y-cx) f_{S|A}(s|x) f_A(a)$$

In the last expression above, $f_{Q_S}(y-cx)$ is given by term C_1, $f_{S|A}$ by term C_2 and $f_A(a)$ by C_3. □

8.2.2 Contribution to likelihood from surviving devices

For each surviving device we observe the pair $(y(\tau), x(\tau))$, and the joint density for a surviving device is given by:

$$f_{Y(\tau),X(\tau),T,I(S<\tau)}(y, x, \tau, 0) = C_4 \left\{ C_5 - \int_{-\infty}^{\infty}\int_{0}^{\tau} C_6 C_7 C_8 ds da \right\} \tag{8.4}$$

$$C_4 = \frac{1}{\sqrt{2\pi\sigma_Y^2(1-\rho^2)\tau}} exp\left(-\frac{(y - cx - \tau(\nu_Y - c\nu_X))^2}{2\sigma_Y^2(1-\rho^2)\tau}\right)$$
$$C_5 = \frac{1}{\sqrt{2\pi\sigma_X^2\tau}} exp\left(-\frac{(x-\nu_X\tau)^2}{2\sigma_X^2\tau}\right)$$
$$C_6 = \frac{1}{\sqrt{2\pi\sigma_X^2(\tau - s)}} exp\left(-\frac{(x - a - \nu_X(\tau - s))^2}{2\sigma_X^2(\tau - s)}\right)$$
$$C_7 = a(2\pi\sigma_X^2 s^3)^{-1/2} exp\left(-\frac{(a-\nu_X s)^2}{2\sigma_X^2 s}\right)$$
$$C_8 = \frac{1}{\sqrt{2\pi\sigma_A^2}} exp\left(-\frac{(a-\nu_A)^2}{2\sigma_A^2}\right)$$

Proof.

$$f_{Y_\tau,X_\tau,T,I(S<\tau)}(y, x, \tau, 0) = f_{Y_\tau|X_\tau,I(S<\tau)}(y|x, 0) f_{X_\tau,I(S<\tau)}(x, 0)$$

Due to lemma 4.3 the first factor above is given by $f_{Q_\tau}(y - cx)$. Then,

$$f_{Y_\tau,X_\tau,T,I(S<\tau)}(y, x, \tau, 0) = f_{Q_\tau}(y - cx) f_{X_\tau,I(S<\tau)}(x, 0) \tag{8.5}$$

The probability density function of the degradation process $\{X(t), t \geq 0\}$ terminating at level x at time t is given by $f_{X_t}(x)$ and illustrated in figure 8.1. Figure 8.1 illustrates the relationship between the degradation and threshold variable at the end of test-time τ. This relationship holds for all times $t \geq 0$.

$$f_{X_\tau,I(S<\tau)}(x, 0) = f_{X_\tau}(x) - f_{X_\tau,I(S<t)}(x, 1) \tag{8.6}$$

Equation (8.6) represents the probability density function of a complimentary Wiener term as discussed in chapter 4. By definition, $f_{X(\tau),I[S\geq\tau]}(x, 0)dx$ is the

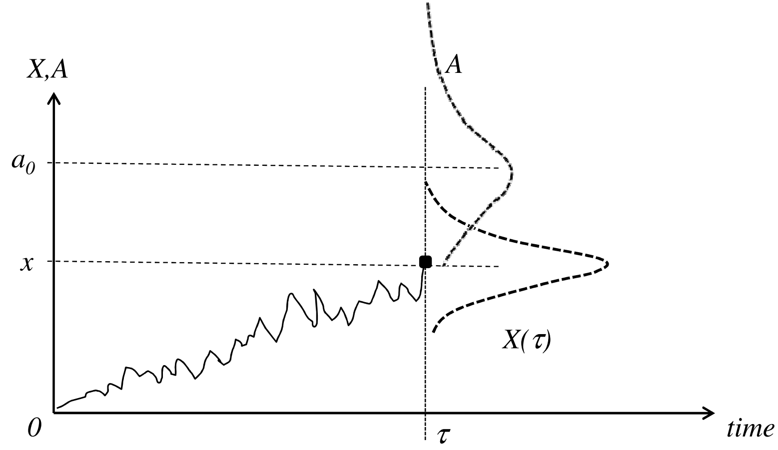


Fig. 8.1: Relationship between threshold and degradation variables at time τ

probability the degradation level belongs to a small interval $(x, x + dx)$ at time τ, and that it crossed the failure threshold some time earlier.

$$f_{X_\tau, I(S<\tau)}(x, 1) = P(X_\tau \in (x, x + dx), S \leq \tau)/dx$$

We condition the event $(X_\tau = x, I(S < \tau) = 1)$ on the threshold random variable, A. We integrate the resulting joint density over the measure of A. Figure 8.1 illustrates the relationship between the degradation and threshold variables for a surviving device, and the degradation path taken to reach an observed degradation level of x by time τ. The mixed joint density of the degradation and survival indicator variables is defined by integrating over a latent threshold variable A as follows:

$$f_{X_\tau, I(S<\tau)}(x, 1) = \int_{-\infty}^{\infty} f_{X_\tau, I(S<\tau), A_\tau}(x, 1, a) da$$

We integrate over the measure of the failure-time variable S conditioned on

the threshold variable A.

$$f_{X_\tau,I(S<\tau)}(x,1) = \int_{-\infty}^{\infty}\int_0^\tau f_{X_\tau|S,A}(x|s,a)f_{S,A}(s,a)dsda$$

$$f_{X_\tau|S,A}(x|s,a) =$$

$$= f_{X_\tau|S,X_s,A}(x|s,a,a)$$

$$= f_{X_\tau-X_s|S,X_s,A}(x-a|s,a,a)$$

Due to theorem 4.3 because $X_\tau - X_s$ is $\mathcal{F}^X_{\geq s}$ measurable it is independent of (S, X_s, A) which is $\mathcal{F}^X_{\leq s}$ measurable. Note, as mentioned earlier, the entire process $A(\cdot)$ is assumed independent of the entire process $X(\cdot)$. Therefore, we can continue as follows:

$$f_{X_\tau,I(S<\tau)}(x,1) = \int_{-\infty}^{\infty}\int_0^\tau f_{X_\tau-X_s}(x-a)f_{S|A}(sa)f_A(a)dsda \tag{8.7}$$

and then from equation (8.6) we get:

$$f_{X_\tau,I(S<\tau)}(x,0) = f_{X_\tau}(x) - \int_{-\infty}^{\infty}\int_0^\tau f_{X_\tau-X_s}(x-a)f_{S|A}(s|a)f_A(a)dsda \tag{8.8}$$

By plugging in equation (8.8) into equation (8.5) we get the joint density for a surviving device under a variable failure-threshold. The density $f_{Q_\tau}(y-cx)$ is given by term C_4, $f_{X_\tau}(x)$ by C_5, $f_{X_\tau-X_s}(x-a)$ by C_6, $f_{S|A}(s|a)$ by C_7 and $f_A(a)$ by C_8 in equation (8.4). □

The likelihood function in equation (8.1) is given by:

$$L_{\boldsymbol{\theta}} = \prod_{i=1}^{q} C_1(y_i,x_i,s_i)C_2(x_i,s_i)C_3(x_i)\prod_{j=1}^{p} C_4(y_j,x_j,s_j)\times$$

$$\left\{C_5(x_j) - \int_{-\infty}^{\infty}\int_0^\tau C_6(x_j)C_7(x_j)C_8dsda\right\}$$

Maximization of the log-likelihood function provides the maximum likelihood

estimates of the process parameters $\hat{\boldsymbol{\theta}}$.

9. SUPPORT VECTOR DEGRADATION MODEL

9.1 Introduction

Reliability is defined as the ability of a product to perform as intended (without failure, and within specified performance limits) for a specified time in its life cycle application environment. The accuracy of any reliability prediction depends upon both the prediction methodology used, and accurate knowledge of the product, generally including the structural architecture, material properties, fabrication process, and product life cycle conditions [76]. With the increasing functional complexity of on-board electronic systems and products, there is a growing demand for early system-level health assessment, failure diagnostics, and prognostics for electronics [77].

In this chapter, we analyze the reliability of a product from a health monitoring perspective, which allows a methodology that permits the reliability of a product to be evaluated in its actual application conditions [78]. We develop an algorithm in effort to evaluate the reliability of a system in the context of a prognostics and health management (PHM) framework. The value obtained from PHM can take the form of advance warning of failures; increased availability through extensions of maintenance cycles, or timely repair actions; lower life cycle costs of equipment from reductions in inspection costs, downtime, inventory, and no-fault-founds; or the improvement of system qualification, design, and logistical support of fielded, and future systems [79].

A product's health is defined as the extent of deviation or degradation from

its expected typical operating performance. Typical operation refers to the physical or performance-related conditions expected from the product [80]. We use this definition of "health" later in the chapter, and we see it applied in a case study of simulated degradation data. In the absence of suitable physics of failure (PoF) models, there is a need for data-driven approaches that can detect when electronic systems are degrading, or have sustained a failure that could be critical. In this chapter, we consider a data-driven approach for anomaly detection for electronic systems based on nonlinear classification. We argue that this approach can also be used to determine suitable marker variables for FHT models.

The resulting classifier gives the best estimate of the functional dependency of the system input data, X, such as resistance, capacitance, temperature, etc., on their class label, Y, a categorical variable that indicates the presence of an anomaly, through a mapping function, $D(X)$. The mapping function separates two classes of data, and is constructed from a sample of training data. If the training data only consists of examples from one class, and the test data contains examples from two or more classes, then the classification task is called novelty detection [81].

A critical part of novelty detection, and of health monitoring in general, is the evaluation of uncertainty in every decision. Due to incomplete training data, there is no mapping function that can be applied universally to all possible test data, and therefore decisions are not always completely correct. Incomplete training data refers to data that do not contain all possible healthy system performance states. Mapping functions, as we discuss in this chapter, constructed from larger, more densely distributed training sets convey greater confidence in their classification decisions as opposed to low population, and sparse training data.

We approach the problem of novelty detection and health evaluation based on support vector machine (SVM) classifiers. We use their connection to Bayesian linear models (BLM) to model the posterior class probability for future test data.

The Bayesian SVM algorithm is trained in the absence of failure data (negative class data), as is the case in many mission-critical systems. The contribution of this work to the field of reliability are the following:

1. It interprets reliability from a data-driven, machine-learning perspective.
2. It introduces a methodology that connects machine-learning analysis to FHT models by helping determine suitable marker variables that can track degradation
3. It solves a novelty detection problem with a one-class classification algorithm, and a Bayesian framework for uncertainty analysis
4. It connects SVM, BLM, and minimum volume sets (mvs).

9.2 Data Notation and Algorithm Overview

Consider the positive-class training data matrix, $\boldsymbol{Z} = [\boldsymbol{X}, \boldsymbol{Y}]$, where $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m]$ is the input data matrix. Each column $j = 1, \ldots, m$ in $\boldsymbol{X}$ contains data collected on separate covariates, and each row $i = 1, \ldots, n$ contains covariate measurements made at a specific time point t_i. Vector $\boldsymbol{Y}$ is the response vector which represents the class membership of row observations in $\boldsymbol{X}$, $Y = [Y_1, \ldots, Y_n]$, where $Y_i = (+1, -1)$.

Fig. 9.1 illustrates the detection algorithm, from left to right. The multivariate training data, $\boldsymbol{Z}$, is first pre-processed through a principal component analysis (PCA). The decomposition (projection) of the training data, $\boldsymbol{X}$, into more than two subspaces, as illustrated in Fig. 9.1, constructs m orthonormal subspaces, which can be used to estimate the joint posterior class probability, J_p, discussed later in this chapter. The benefit of the multiple models is that they separately capture a unique identifiable subset of information related to the covariance of the random variables

in $\boldsymbol{X}$. The dimension for each model can be chosen to be as low as 1, and as high as $m-1$, with m potential models for each, respectively, considering all the possible combinations. Here, for expositional simplicity, only two models are considered, $[M]$ and $[R]$, each of which is two-dimensional.

PCA was chosen to preprocess the original data to extract features related to changes in the variance of the data. In the context of statistical control theory, the variance and its changes are strong features indicating the onset of anomalies in multivariate systems [82]. Other options for this step include blind source separation (BSS), and independent component analysis (ICA), or, more generally, generalized linear models (GLMs). PCA is a special case of GLM, and although it suffers from the assumption of linearity and normality of the data (situations that are arguably not often encountered in real data sets), transformations can apply to the original data to approximate normality [83].

When failure data (negative class) are not available, a kernel density estimate (KDE) is computed for the projected (positive class) training data in the two subspaces $[M]$, and $[R]$ to estimate the likelihood of the projected data, and from it to construct the negative class. The SV classifier constructs two predictor models, D_1 and D_2, for each subspace. A soft decision boundary is constructed by fitting the training data with a model for posterior class probabilities using a logistic distribution that maps classified data to posterior classification probabilities: $P_M, P_{R1}, \ldots, P_{RM}$, respectively. The joint class probability J_p from the subspaces is used for the decision classification.

Support vectors produce an uncalibrated value that is not a probability. Therefore, the algorithm uses the support vector decision function, D, to produce a posterior probability, $P(class|input)$, according to a Bayesian formulation. Finally, the joint posterior class probability can be weighted with a weight vector $W = [w_1, \ldots, w_m]$ to emphasize some models as opposed to others. This weighting

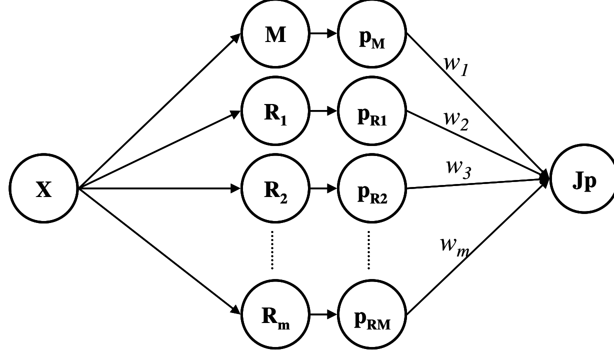


Fig. 9.1: Algorithm flow diagram showing the processing of the data.

could be beneficial for emphasizing the results from models, usually the principal model [M], which captures more of the data covariance information. In this paper, all models are weighted equally ($W = I$).

9.3 Data Pre-Processing - Principal Component Projections

In this chapter, we decompose the training data into two lower dimensional subspace models: the principal model [M], and the residual model [R]. We use singular value decomposition (SVD) of the input data, X [84], [85], [86], [87]. The SVD of data matrix X is expressed as $X = USV^T$, where $S = diag(s_1, \ldots, s_d) \in R^{n \times d}$, and $s_1 > s_2 > \ldots > s_d$ are the ordered singular values. The two orthogonal matrices U, and V are called the left, and right eigen matrices of X. Based on the SVD, the subspace decomposition of X is expressed as:

$$X = U_M \times S_M \times V_M^T + U_R \times S_R \times V_R^T \qquad (9.1)$$

The diagonal matrix, S_M, are the singular values $(s_1, \ldots, s_k)$, and $(s_{k+1}, \ldots, s_d)$ belonging to the diagonals of S_R. Any vector X can then be represented by a summation of two projection vectors as shown in equation (9.2), where $P_M = UU^T$, and $P_R = I - UU^T$ are the projection matrices for the principal, and residual model

subspaces, respectively. Both subspaces comprise the total data dimension. In this framework, we can apply a SV classifier, having oriented the data such that we can better capture system failures that are reflected in changes in variance, and so that we can "break down" the effects of multivariate data into separate models, each examining a different effect of the data on changes in variance. Then we can envision combining the results in the end to achieve a "global" detection result, as we demonstrate later in this paper.

$$X = P_M \times X + (I - P_R) \times X \tag{9.2}$$

9.4 Two-Class Classifier

SVMs alleviate the need for algorithms with statistically grounded frameworks, algorithms that require knowledge of the distribution of the random variables. SVMs are based on the idea of large-margin linear discriminants that seek optimum margin hyperplanes where the separating plane is chosen to minimize a risk bound motivated by structural risk minimization. Nonlinear extensions were introduced by the authors in [88], and [89] with a generalization often referred to as the "kernel trick", which builds on a direct consequence of Hilbert's space theory. Here, we review the linear SVMs to highlight certain concepts that we will use in this chapter. Given the data structure defined earlier, linear SVMs apply a linear model f that maps a m-dimensional real valued vector to a binary scalar as shown in equation (9.3).

$$f(x) = sign(w_{SVM}^T x + \epsilon(x)) \tag{9.3}$$

$$w_{SVM} = \sum_{i=0}^{n} \alpha_i y_i x_i \tag{9.4}$$

In (9.3), $w_{SVM}^T x = D(x)$, and its weights w_{SVM} are given by equation (9.4), are normal to D; $b/||w||$ is the perpendicular distance from D to the origin, and

$||w_{SVM}||$ is the Euclidean norm of w_{SVM}. The margin in classification is the distance between the nearest positive and negative labeled data points. For linearly separable cases, training the SVM is performed by solving the following optimization problem: $\arg_{\min} ||w_{SVM}||$, and the constraints are combined into a set of inequalities $\forall\ i$: $y_i(x_i w + b) - 1 \geq 0$.

Training the SVM becomes an optimization problem given by equation (9.5), and constrained by equation (9.6).

$$(\alpha_1, \ldots, \alpha_n, b) = argmin(1/2||w||^2 + C \sum_{i=1}^{n} \xi_i) \tag{9.5}$$

$$y_i(w^T x_i + b) + \xi_i \geq 1 \tag{9.6}$$

Lastly, in the case where a linear decision function is not suitable for the data, the above methods can be generalized using a transformation to another Euclidean space using a map function called Φ, where the training data are linearly separable. More reviews of SVMs can be found in [90], [91], [92], and [93].

9.5 Statistical Properties of SVMs and Their Connection to the Evidence Framework

From a Bayesian representation, $D(x)$ can be shown to be a relaxed maximum a posteriori solution (MAP) of the weights w_{MAP} in a Bayesian linear model, $y = f(x) + \epsilon$ discussed in detail by the authors in [94], [95], [96], and [97]. This connection is important because it motivates the use of a function centered on $D(x)$ to model the posterior class probabilities of test data.

This result is motivated under relaxed conditions that are based on the following assumptions. a) The functional dependency of Y on $\boldsymbol{X}$ is mapped through an unknown kernel function. b) The errors, and weights in the linear model are

normally distributed around zero with a certain variance, therefore modeling the conditional density of $Y|\boldsymbol{X}$ also as normal. c) Because of assumption b), we can express the posterior class density as a function of an SVM related term, namely ξ_i. To see this result, we consider the training data $\boldsymbol{Z}$, assume that the joint $P(\boldsymbol{Z}, w)$ exists, and assume that the conditional $P(w|\boldsymbol{Z})$ can be expressed as

$$\begin{aligned} P(w|\boldsymbol{Z}) &= \frac{P(\boldsymbol{Z}|w)P(w)}{P(\boldsymbol{Z})} \propto P(\boldsymbol{Z}|w)P(w) = \\ &= P(y|\boldsymbol{X}, w)P(\boldsymbol{X}, w)P(w) \end{aligned} \tag{9.7}$$

The posterior on the weights can be expressed as the product of three distributions, as shown above. The probability density over observations given the parameters is modeled through a binomial distribution to account for the possible states of the response random variable Y, and is given by (9.8) with $0 \leq q(x, w) \leq q1$, and $q(x, w) = Prob(y = +1|X, w)$.

$$P(y|\boldsymbol{X}, w) = \prod_{i=1}^{n} q(x_i, w)^{\dfrac{1+y_i}{2}} (1 - q(x_i), w))^{\dfrac{1-y_i}{2}} \tag{9.8}$$

If the errors are modeled to be statistically independent of x and w, and drawn from a Gaussian distribution, $\epsilon \sim \mathcal{N}(0, {\sigma_\epsilon}^2)$, then $Prob(y = +1|\boldsymbol{X}, w) = Prob_\epsilon(\epsilon \geq q - w^T x)$=

$$P(y|\boldsymbol{X}, w) = \int_{-w^T x}^{\infty} \frac{1}{\sqrt{2\pi\sigma_\epsilon}} e^{\left(-\dfrac{u^2}{2{\sigma_\epsilon}^2}\right)} du \tag{9.9}$$

If we further assume that the density of X is not parameterized by the model weights $P(\boldsymbol{X}|w) = P(\boldsymbol{X})$, and that the prior on the weights is drawn from a Gaussian distribution, $P(w) = C \exp(\phi/2||w||^2)$, then the posterior conditional density for the weights can be expressed proportional to the product of the error function given by (9.10). Taking the logarithm gives an expression that resembles the objec-

tive function for an SVM, and is given by (9.11).

$$P(w|\boldsymbol{Z}) \propto \prod_{i=1}^{n} (\rho^a (1-\rho)^b) e^{(-\frac{\phi}{2}||w||^2)} \tag{9.10}$$

$$-log(P(w|\boldsymbol{Z})) = C - \sum_{i=1}^{n} a log \rho - \sum_{i=1}^{n} b log(1-\rho) \tag{9.11}$$

$$\begin{aligned}
\rho &= erfc(-w^T x_i) \\
C &= \frac{\phi}{2}||w||^2 \\
a &= \frac{1+y_i}{2} \\
b &= \frac{1-y_i}{2}
\end{aligned}$$

By considering the asymptotic expansions for the error functions above, and if it can be shown that the expansions of the two sums reduce to a function of ξ_i, the log posterior on the weights of the linear model has an equivalent form to that of the SVM optimization in equation (9.5). This connection is useful because it effectively lays down a strong informative prior for modeling posterior class probabilities of future test data. This prior is implemented by treating $D(x)$ as the optimum classifier for the given training data. This fact will be used later in the chapter to provide rationale for the design of a posterior classification probability given $D(x)$.

9.6 One-Class Classifier

In many real world systems, especially mission critical systems, and components for which failures are not known, training data consists only of the positive class. To obtain estimates for the failure space (negative class), novelty detection

as discussed in [98], [99], [100], [101], and [102] among others (see [103], and [104] for more general review) is approached primarily as a data-versus-density problem, where the negative data are assumed to be generated from an unknown distribution, say $Q(\boldsymbol{X})$. The density of Q is intentionally left uninformative, and usually uniformly distributed to reflect the lack of any prior knowledge about anomalies.

Authors in references such as [98], and [102] discuss sampling schemes for Q that optimize supervised function estimation techniques (e.g., SVM) to best infer a general classification boundary for the given positive class training data. The sampling approaches depend on the choice of a prior for Q, which in the absence of any evidence is measured on the entire metric space spanned by the positive class training data, and suffers from high dimensionality. In [102], the authors discuss a negative class selection algorithm for data collected from various Internet sites. In this work, unlabeled data was made available by sampling the Internet, which is different from the situation we describe here. In this paper, there are no unlabeled data, and we cannot sample from a universal set (the Internet, for example).

Other approaches, as mentioned earlier, use the origin as the negative class in the applied feature space induced by some kernel function [105]. Others [106] extend this idea, and assume that all data points close enough to the origin are also considered as candidates for the negative class. Some of the critiques, however, of the one-class classification approach motivated by [105] focus on its sensitivity to specific choices of representation and kernel in ways that are not very transparent [106]. Further, its assumed homogeneous input feature space relies on comparable distances between data, which can lead to inaccurate classifications with non-Gaussian distributed data [83]. The authors of [83] propose a rescaling of the data in the kernel feature space to make it robust against large-scale differences in scaling of the input data. The data are rescaled such that the variances of the data are equal in all directions using kernel PCA.

The primary approach to one-class classification has been largely based on the work discussed above. In essence, the problem reduces to making the most of the information at hand, the positive class training data. As such, it becomes important to extract features from these data that can improve inference about potential anomalies. An important feature of the training data is its density, which can be estimated computationally, although this is an expensive task in high dimensions.

Therefore, on a practical level, the estimate of the negative class is seen as a conservative representation of a potential system failure space, an assumption that could lead to poor generalization of the algorithm in situations where the predictor model is not updated to reflect changes in the system performance characteristics. Such changes are plausible, for example, in a reliability setting in which the system has aged so that its performance signature has changed, but it is still functioning in a "healthy" state. Another example is a case where the original training data were not complete enough to represent the global system performance regimes (universal set), and in such situations the predictor model will naturally fall victim to large numbers of false alarms. Therefore, a one-class-classifier approach to novelty detection must be subject to complete, updated training data.

To utilize SVMs for classification, the negative class must be estimated first by considering the density of the positive class (training data) following similar reasoning as the authors of [107]. This work can be accomplished in several ways, one of which is to use a kernel density estimate (KDE) of the training data through the use of Gaussian kernel functions. For this work, the negative class was estimated based on assumptions on the failure space, summarized in 9.1

Definition 9.1. *The failure space is a) not linearly separable from the healthy training data, b) prevalent in the space not occupied by the healthy training data, and therefore c) assumed to conform to the distribution of the healthy training data.*

Through this definition, we aim to achieve minimum volume sets (mvs), similar

to work discussed in [108], [109], [110], and [105], that find sets of density functions that correspond to regions with the minimum volume or Lebesgue measure for a given error $1 - \alpha$ [109]. An *mvs* in a class of measurable sets C for an error α is defined in reference [110] as

$$M_c(\alpha) = argmin(\lambda(B) : B \in F_\alpha), \forall \alpha \in [0, 1] \quad (9.12)$$

where $F_\alpha = \{B : B \ \in C, \ P(B) \geq \alpha\}$, P is a probability measure, B is the positive class training set, and λ is the Lebesgue measure. For example, if P is a multivariate Gaussian distribution, and λ is the Lebesgue measure, then the *mvs* are ellipsoids. The parameter α is chosen by the user, and reflects a desired false alarm rate of $1 - \alpha$.

The *mvs* in our approach is implemented with an SVM given the positive and estimated negative training data. Therefore, the negative class training data are sampled from the subspaces in B^c (failure space), and designed to adhere to *Definition I*. Knowledge of the density of the positive training data should tell us something about where the most conservative boundary should exist. The inference of test data in areas of high density of positive training samples should have higher confidence, as opposed to areas with low density, and sparse information.

To estimate the negative class data, $X_n \in R^{m \times 2}$, in each subspace, we used the marginal kernel density estimate of the positive class, $X \in R^{n \times 2}$. This approach first partitions the data space $R^{m \times 2}$ into a grid of small square 2-dimensional blocks R^2, of length size h. A general parzen windowing approach with Gaussian kernels (among other alternatives, see reference [111]) was used to compute the density of each data point by centering a Gaussian kernel function, ϕ, on each point, x_i, with a bandwidth equal to the size of the grid length, h. All neighboring data x_i were evaluated against the Gaussian kernel centered at x, and their corresponding

influence weighted according to their Euclidean distance from x. One choice for a smooth ϕ is the standard normal distribution $\mathcal{N}(0,1)$.

To overcome over-parameterized density estimates that do not generalize well, the bandwidth, h, is determined through a nearest neighbor approach in which h is selected as the value that produces a volume around x_i containing $\sqrt{n}$ neighbors. This approach personalizes the value of h to each data point x_i, and effectively smoothes out the density in areas with sparse training data information.

The negative class data are constructed by selecting grid coordinates where the likelihood ratio ρ of the training data is below a threshold, τ, which is a grid center, and is labeled as a member of the negative class if $\rho \geq \tau$. The likelihood ratio is the ratio of negative to positive posterior class probabilities, as shown in (9.13). The denominator is computed by the KDE, and the numerator is modeled as a function of the gradient of the likelihood function.

In (9.13), we use $P(Y = -1|\boldsymbol{X} = \boldsymbol{x}_i)$ as p_i^-, and $P(Y = +1|\boldsymbol{X} = \boldsymbol{x}_i)$ as p_i^+. We note that the model favors the numerator proportional to the square of the likelihood function gradient, ∇L. Practically, this means that, in areas where the likelihood function changes faster, the negative class is more similar to the positive class.

$$\rho = \frac{P(Y = -1|\boldsymbol{X} = \boldsymbol{x})}{P(Y = +1|\boldsymbol{X} = \boldsymbol{x})} \tag{9.13}$$

$$p_i^- = (1 - p_i^+) + (1 - p_i^+)\nabla^2 L \tag{9.14}$$

Once the $D(x)$ is constructed through an SVM using the positive and estimated negative training data, the argument is, as motivated earlier by its statistical

properties, that $D(x)$ is the optimal classifier.

$$D(x) = \arg_{a=\{-1,+1\}} \min P(Y = a|\boldsymbol{X} = \boldsymbol{x}) = \tag{9.15}$$

$$= \left\{ \begin{array}{ll} -1 & \text{if } P(Y = +1|\boldsymbol{X} = \boldsymbol{x}) < 0.5 \\ +1 & \text{if } P(Y = +1|\boldsymbol{X} = \boldsymbol{x}) \geq 0.5 \end{array} \right\}$$

9.7 Posterior Class Probabilities

The objective is to classify data $\boldsymbol{X} = \boldsymbol{x}_i$ by comparing the probability that the class membership of $\boldsymbol{x}_i$ is +1, versus the probability that the class membership of $\boldsymbol{x}_i$ is -1. The larger probability classifies x_i into the corresponding class. Due to its connection to a Bayesian linear model (as discussed earlier), $D(x)$ can be thought of as a boundary, where classifications close to it will be associated with probabilities close to 0.5, and classifications far from it will be associated with probabilities closer to 1 or 0. Data that fall exactly on the boundary are randomly and fairly classified as either +1 or -1 with a classification probability of 0.5.

The classification problem defined by $P(y = +1|\boldsymbol{X} = \boldsymbol{x})$ can now be expressed as $P(y = +1|D(x))$, where $D(x)$ is the sufficient statistic to classify data $\boldsymbol{X} = \boldsymbol{x}$ into class +1 or -1. Intuitively, because $D(x)$ is the optimal classifier on which the probability of interest is exactly 0.5, distances to it can be calibrated to probabilities. The distribution of these posterior class probabilities is modeled by a logistic distribution [112], [113], [114] centered at $D(x) = 0$; see Fig. 9.2. The shape parameter for the distribution, as we discuss later, reflects the confidence in $D(x)$, and is a statistic dependent on the data. The positive posterior class probability for $\boldsymbol{X} = \boldsymbol{x}_i$ is given by (9.16), and the intuition that the distances of data $\boldsymbol{X} = \boldsymbol{x}_i$ to $D(x)$ can be calibrated to probabilities leads to the justification for using a logistic-type distribution to model these probabilities. From Bayes' rule, and the law of

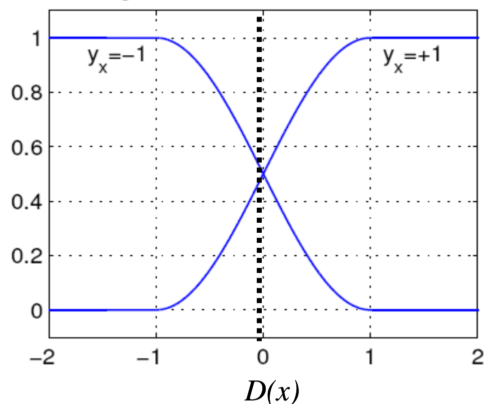


Fig. 9.2: Logistic distribution model for posterior class probabilities.

total probability, re-expressing the sum in the denominator, we get a function with parameter β that is a logistic-type distribution.

$$\begin{aligned} P(Y=+1|\boldsymbol{X}=\boldsymbol{x}_i) &= \frac{P(\boldsymbol{X}=\boldsymbol{x}_i|Y=+1)P(Y=+1)}{\sum\limits_{a=-1,+1} P(\boldsymbol{X}=\boldsymbol{x}_i|Y=a)P(Y=a)} \\ &= \frac{1}{(1+exp(\beta_i))} \end{aligned} \tag{9.16}$$

$$\beta_i = log\frac{P(\boldsymbol{X}=\boldsymbol{x}_i|Y=+1)P(Y=+1)}{P(\boldsymbol{X}=\boldsymbol{x}_i|Y=-1)P(Y=-1)} \tag{9.17}$$

The distribution scale parameter β affects the shape of the distribution by compressing it around $D(x)=0$ with large values of β, and stretching it for small values. The shape of the distribution reflects the level of uncertainty in the classifier, and should be estimated from the training data. From the resulting expression for β, all terms except for one are known, namely the probability $P(\boldsymbol{X}=\boldsymbol{x}_i|Y=+1)$, which was estimated previously. The unknown quantities are $P(\boldsymbol{X}=\boldsymbol{x}_i|Y=-1)$, and the priors $P(Y=+1)$ and $P(Y=-1)$. Replacing β by its intuitive interpretation, namely the data's relationship to $D(x)=0$ through their Lebesgue measure, we can evaluate the objective probability as:

$$P(Y=+1|X=x_i) = P(Y=+1|\lambda(x_i|D(x))) \equiv p_i \tag{9.18}$$

$$p_i = P(Y = +1|X = x_i) = \frac{1}{1 + e^{(-a_1 g(x_i) + a_2)}}$$

In equation (9.18), $\lambda(x_i|D(X)) \equiv g(x_i)$ is the Lebesgue measure of x_i in reference to D (or simply the perpendicular Euclidean distance to D). The parameters a_1, and a_2 are used to optimize the posterior class distribution [113] of the logistic form, and are estimated by maximizing the likelihood of class given the data over the parameter space a_1, a_2. The classification probability of a sequence of n data X into a binary classification $c = \{1, 0\}$ is given by a product Bernoulli distribution

$$P(c_1, \ldots, c_k) = \prod_{i=1}^{n} p_i^{c_i}(1 - p_i)^{1-c_i} \tag{9.19}$$

Here, p_i is the probability of classification when $c = 1 (y = +1)$, and $1 - \pi$ is the probability of classification when $c = 0$ $(y = -1)$. The evaluation of $sign(D(x_i))$ gives the class label for x_i, and $g(x_i)$ the distance to $D(x) = 0$.

The last step of the algorithm is to compute a joint posterior class probability based on the separate, statistically independent (assumed) results from each lower dimensional model (subspace), here the principal model $[M]$, and the residual model $[R]$. The joint result will provide a final classification with associated final positive, and negative posterior class probabilities. This result is anticipated to give a more accurate estimate of the classification of the data $X = x_i$ as compared to a treatment of the data in its original data space. The conditional joint posterior class probability is expressed in (9.20), with the assumption that the random variables X_M, and X_R are statistically independent, and uncorrelated. Due to PCA, the random variables can be shown to be uncorrelated, but not necessarily statistically independent.

$$P(X_M X_R|y) = P(X_M|y)P(X_R|y) \tag{9.20}$$

According to Bayes' rule, the conditional class probability is given by

$$P(Y|X_M, X_R) = \frac{P(X_M, X_R|Y)P(Y)}{P(X_M, X_R)} =$$

$$= \frac{P(X_M|Y)P(X_R|Y)P(Y)}{\sum\limits_{y=\{-1,+1\}} P(X_M, X_R|Y)P(Y)}$$

$$= \frac{P(Y|X_M)P(Y|X_R)P(Y)}{\sum\limits_{y} P(Y|X_M)P(Y|X_R)P(Y)} \tag{9.21}$$

where $X_R = [X_{R1}, X_{R2}, \ldots, X_{Rn}] \in R^{n\times 2}$, and $X_M = [X_{M1}, X_{M2}, \ldots, X_{Mn}]$. In the joint probability model, $P(y = a|X_M)$ is the probability that data point x_M is classified as class a in $[M]$, $P(Y = a|X_R)$ is the probability that data point x_R is classified as class a in $[R]$, and $P(Y = a|X_M, X_R)$ is the final conditional joint probability that $X = x$ is classified as class a, where $a \in A = \{-1, +1\}$. The main assumption is that the random variables in each subspace are statistically independent, which allows formulating the final joint probabilities of positive and negative classification, given by (9.22), and (9.23). Note that the same calculations apply for a sequence of test data.

$$\Pi_{(+)} = \frac{P(Y = +1|X_M)P(Y = +1|X_R)P(Y = +1)}{\sum\limits_{y} P(Y = y|X_M)P(Y = y|X_R)P(Y = y)} \tag{9.22}$$

$$\Pi_{(-)} = \frac{P(Y = -1|X_M)P(Y = -1|X_R)P(Y = -1)}{\sum\limits_{y} P(Y = y|X_M)P(Y = y|X_R)P(Y = y)} \tag{9.23}$$

9.8 Posterior Class Probabilities as a Marker Variable

The joint posterior class probability assigned to each test observation is a measure of health of the system, and can therefore be considered a marker to degra-

dation. In section 7.3 and 7.3.1 we consider SVMs as a potential regression structure in FHT models, here we argue that the posterior class probability (a direct result of SVM classification) can be used as a marker.

10. CASE STUDIES

The anomaly detection algorithm developed in chapter 10 is coded into a prototype called CALCEsvm. In this chapter we apply CALCEsvm in three case studies:

1. Anomalies in Lockheed Martin data
2. Simulated degradation data
3. Gas-turbine engine degradation data

Through the case studies we were interested in evaluating CALCEsvm on detecting anomalies and also in capturing the degradation process. Anomalies are flagged when the posterior class probability falls under a certain threshold, and the trend in degradation is captured by considering the resulting time series of posterior class probabilities outputed by CALCEsvm.

10.1 Lockheed Martin Data

To test the proposed algorithm, we used a data-set extracted from Lockheed Martin servers, $X \in R^{n \times d}$, where n=2471 observations, and d=22 covariates (p_1 through p_{22}). The first 800 observations were used as the positive training class, during which no failures occurred. The remaining data were used as the test data, which included three periods of failures. The failure periods were identified (by Lockheed) to occur during observations 912-1040, 1092-1106, and 1593-1651.

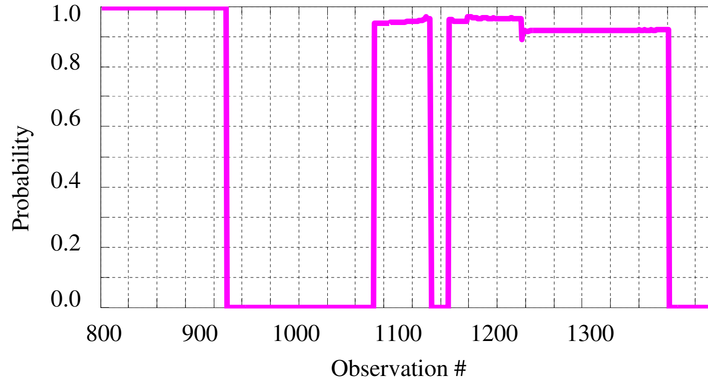


Fig. 10.1: Joint posterior class probability vs. observation for Lockheed Martin test data set

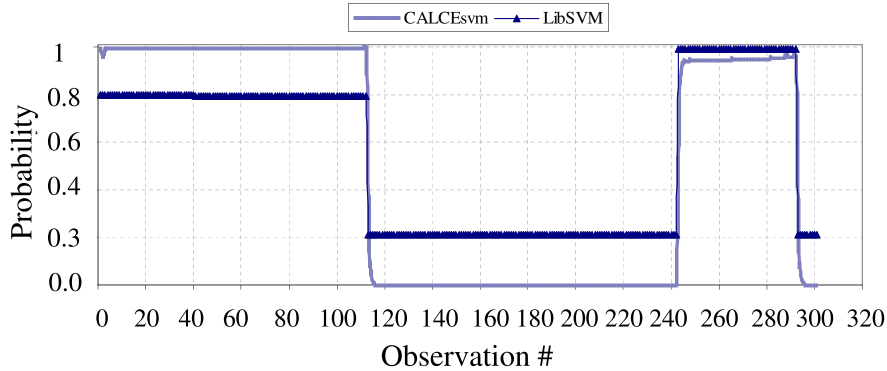


Fig. 10.2: Joint posterior class probabilities for CALCEsvm, and the open source support vector classification software called LibSVM.

CALCEsvm was used on these data. Fig. 10.1 shows the detection results. The algorithm detected the first two periods of anomalies, namely those between 912 and 1040, and between 1092 and 1106.

CALCEsvm was compared to the open source support vector classification software called LibSVM [93]. The setup for LibSVM used its two-class C-SVC

setting with input, the training data used in CALCEsvm. Because the one-class SVM in LibSVM does not provide posterior class probabilities for test data, we compared the two-class classification between CALCEsvm and LibSVM. For this comparison, the negative class training data were taken from the output estimate of CALCEsvm, and used as the negative class in LibSVM. Therefore, the actual comparison was made between the two-class SVM algorithms of CALCEsvm, and LibSVM. The option settings used for LibSVM are listed below with the margin penalty parameter, and tolerance setting of the termination criterion parameter chosen arbitrarily, and kept the same for both CALCEsvm and LibSVM.

1. s svm type : 0 – C-SVC
2. t kernel type : 2 – radial basis function
3. d degree : 1, degree in kernel
4. c cost : 150, margin penalty parameter
5. ϵ : setting for tolerance of termination criterion
6. b probability estimates: 1, outputs the class probabilities

The accuracy comparison was performed through three tests: 1) a direct comparison of the quadratic optimization results: the objective function, the sum of the Lagrange multipliers, and the number of support vectors; 2) detection accuracy based on class index only; and 3) detection accuracy based on the range of probabilities.

In Table 10.1, b_0 is the bias term, w_2 is the objective function equal to $\alpha^T H \alpha$ where $\alpha \in R^{1\times n}$ is the Lagrange multiplier vector, and $H \in R^{n\times n}$ is the Hessian matrix, where n is the length of the SVM training data. The parameter ϵ is the tolerance of the termination criterion, and nSV is the total number of support vectors. The results in table 10.1 show that the performance of the software is

Tab. 10.1: SVM Optimization Results

	Principal Model		Residual Model	
	CALCEsvm	LibSVM	CALCEsvm	LibSVM
b_0	0.181	181	0.099	0.099
w_2	15.40	7.70	14.70	5.60
ϵ	0.0001	0.0006	0.0001	0.0005
nSV	24	23	14	14

comparable to the difference found in the objective function. The number of support vectors, and the bias term were found to be the same.

The second, and third tests compared their detection accuracy against the known periods of anomaly. Each test file was coded with a column variable $z \in \{-1, +1\}$, indicating the known class of each observation, an index of +1 for the healthy data, and -1 for the anomalous data. LibSVM counted the number of misclassified observations based on the coded variable z. Table 10.2 shows the results comparing LibSVM to the CALCEsvm output. The first column in the table shows the detection accuracy based only on the class index, whereas the second column shows the detection accuracy based on a probability index. In the first comparison, both performed almost identically (see first column in table 10.2), but the second comparison (second column) clearly favors CALCEsvm. This result can be seen by comparing the accuracy of 98.1% for CALCEsvm vs. 30.5% for LibSVM given the criteria that the posterior class probability for a test observation should lie within the range specified in the algorithm, here 0.8 to 1.

The second comparison was performed based on a probability index reflecting an "expert" knowledge of system "health". This index therefore pertains to a belief, and is subjective to the user. Nonetheless, this index is based on an intuitive argument: because the posterior class probabilities reflect the certainty/uncertainty of the classification/detection, a known "healthy", and or known "unhealthy" observation should be associated with high, and low probabilities (or ranges of probabilities),

Tab. 10.2: Comparison of CALCEsvm and LibSVM Detection Accuracy Against Lockheed Data

Detection Accuracy Results Based On		
Class Index	Probability	Probability Index Range
CALCEsvm		
100.0%	98.1%	0.8-1.0
	100.0%	0.0-0.4
LibSVM		
99.6%	30.5%	0.8-1.0
	100.0%	0.0-0.4

respectively.

In the Lockheed data set, there are two system levels: "healthy" and "failed". Both system levels are known for the whole data set. The "healthy" level is set to be represented by posterior class probabilities between 0.8 and 1, and the anomalous level by probabilities between 0 and 0.4. Stronger restrictions can be modeled by expanding the range for the anomalous level, and shrinking the range for the "healthy" level. In light of these explanations, CALCEsvm had a 1.9% error rate in its detection accuracy as opposed to 69.5% for LibSVM. The reason LibSVM performed at 30.5% accuracy is because two out of three periods with "healthy" level operation were captured (by LibSVM) with a posterior class probability at around 0.75 to 0.78, therefore falling short of the user-defined "healthy" range of 0.8 to 1.0, and failing to correctly classify the healthy periods. LibSVM, as did CALCEsvm, captured the failed periods with 100% accuracy.

10.2 Simulated Degradation Data

A second case study was performed using simulated correlated data consisting of three random variables from three different but s-dependent distributions to construct the training data set. The objective in this case study was to test the algorithms on a system that was degrading, and in which the degradation took

place in the presence of considerable noise. Copulas were used to build a simulation model consisting of three random variables: $Gamma(2,1)$, $Beta(2,2)$, and $t(5)$. The family of bivariate Gaussian copulas is parameterized by $\rho = [1\ \rho;\ \rho 1]$, the linear correlation matrix. The random variables U_1, and U_2 approach linear s-dependence as ρ approaches +/-1, and approach complete statistically independence as ρ approaches zero. The Gaussian, and t copulas are known as elliptical copulas, and can generalize higher numbers of dimensions. Here we simulate data from a trivariate distribution with $Gamma(2,1)$, $Beta(2,2)$, and $t(5)$ marginals using a Gaussian copula.

Test data were generated from the trivariate distribution of $Gamma$, $Beta$, and t random variables; and were set up such that three degradation periods were generated. The first period was designed to be "healthy", the second introduced a shift in the mean for each variable separately while maintaining the correlation structure, and the third period introduced a larger shift in the mean.

The CALCEsvm results are shown in Fig. 10.3, with the four periods identified by breaking perforated lines and an index $P1$ through $P4$, where $P1$ is the identifier for the "healthy" period, with mean equal to nominal, and $P2$ through $P4$ having successively increasing changes in the mean.

The results of the algorithm show the ability to capture the trend of simulated degradation in the presence of noise. The beginning period that shows a dip in the probability estimate is a direct result of an initial over-smoothing (implementation of the exponential smoothing), and can be ignored for practical purposes. The larger result is the algorithm's ability to correctly classify the data for each period of operation, and to capture the expected trend. CALCEsvm results were compared to the results obtained from LibSVM, and are tabulated in table 10.3. The probabilities, as in the Lockheed Martin case study, again reflect a belief about the interpretation of the posterior class probabilities. In this case, posterior class probabilities between

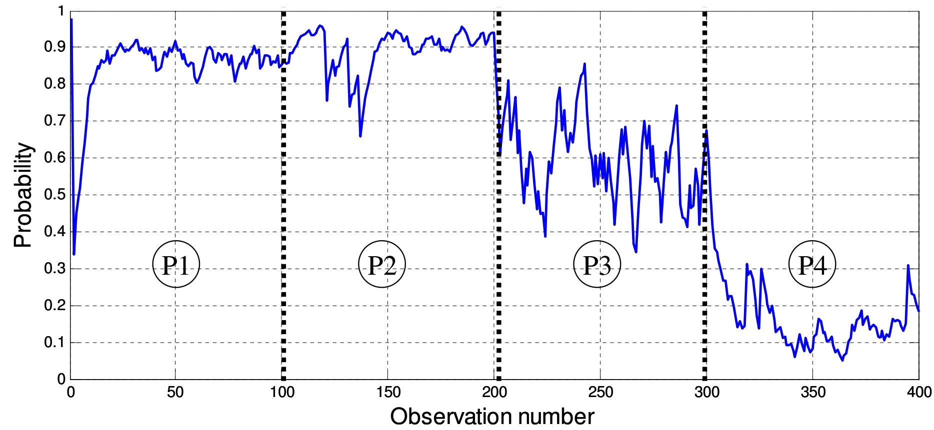


Fig. 10.3: Joint positive posterior class probability for simulated data set.

Tab. 10.3: LibSVM Accuracy Results for Simulated Data

LibSVM Output Based on		
Class Index	Probability	Probability Index Range
100.0%	100.0%	0.8 - 1.0
3.8%	14.0%	0.7 - 0.85
29.0%	46.0%	0.3 - 0.7
88.0%	79.0%	0.0 - 0.4

0.8 and 1 are acceptable for a "healthy" system, probabilities between 0.7 and 0.85 are acceptable for the next level of "health" allowing for some overlap, and so on until the range between 0 and, say 0.5 for example, are used to classify the system as failed.

The comparison of accuracy results based only on the class index shows that both algorithms performed virtually identically for the given probability ranges. Both CALCEsvm, and LibSVM had a detection accuracy rate of 100% in $P1$; in $P2$, both algorithms performed noticeably poorly; and both improved in $P3$ and $P4$ to 88% when the degradation became more distinct. A comparison of accuracy results based on the posterior class probabilities shows a slight improvement in the

Tab. 10.4: CALCEsvm Accuracy Results for Simulated Data

CALCEsvm Output Based on		
Class Index	Probability	Probability Index Range
100.0%	81.2%	0.8 - 1.0
3.8%	31.7%	0.7 - 0.85
29.0%	31.0%	0.3 - 0.7
88.0%	84.0%	0.0 - 0.4

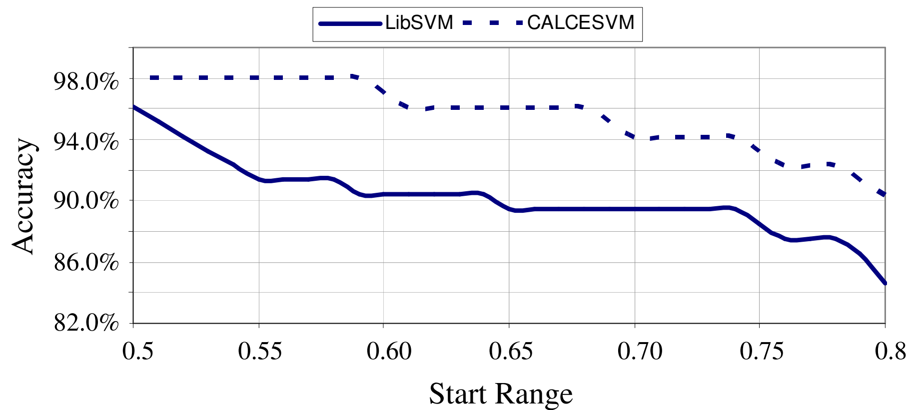


Fig. 10.4: Joint positive posterior class probability for simulated data set in P2.

performance of each algorithm for $P1$, and about the same performance for the other periods. Fig. 10.4 plots the detection accuracy of CALCEsvm and LibSVM vs. the start value for the probability index for levels 1, 2, and 3. For example, from the plot, it can be seen that when the lower bound on the probability index is 55%, and the upper limit is fixed at 100%, CALCEsvm has a detection accuracy of 96% vs. approximately 89% for LibSVM. This is a very liberal bound, as it says that any posterior class probability above 55% can be used to classify a test point as "healthy" instead of anomalous to some degree.

In Fig. 10.4, the x-axis shows the varying lower bound for period 2 ($P2$), and in Fig. 10.5 the varying lower bound for period 3 ($P3$). The y-axis shows the accuracy of the algorithms in classifying test data. As the lower bound on our belief becomes

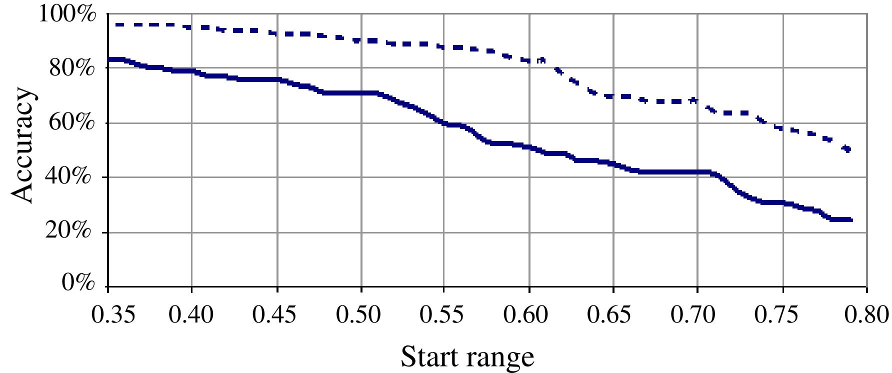


Fig. 10.5: Joint positive posterior class probability for simulated data set in P3.

more stringent (that is, we require higher certainty in the prediction), the accuracy of the algorithms falls. Because the anomalies are more distinct (due to stronger outliers, and reflected by lower posterior class probability values) in $P2$ than in $P3$, as the lower bound is "tightened" similarly for both periods, both CALCEsvm, and LibSVM perform better in $P3$. Here, for example, when the lower bound on the probability index was 70%, and the upper held at 100%, CALCEsvm had lower than 94% detection accuracy, whereas LibSVM had an accuracy of 89%.

10.3 Degradation Data on Gas-Turbine Engines

The last set of data that we used to test and validate the BSVM algorithm came from the NASA data repository of degradation data. The data was and is still available as part of a competition. The data is composed of training data and test data. The training data consists of multivariate time series of covariate observations from different enginesa total of 218 engines, which we call units. Each engine degraded due to wear based on the usage pattern of the engines and not necessarily due

to any particular fault mode. Each unit started with different unknown degrees of initial wear and failed at an unknown level of wear. Noise was injected into the data to represent manufacturing variation, process noise, and measurement noise. The engines were exposed to six operational settings that provided information about the operational mode and environmental conditions of the engine. The objective was to predict remaining life given a sequence of covariate measurements for a test engine.

Modeling degradation using a classification approach (as discussed here) we need to address some potential limitations (lessons learned). The classifier becomes inaccurate: 1) if the variance of the initial wear is large, and 2) if the difference in performance measurements between healthy and unhealthy engines is not significant. This is especially true with this algorithm, which is based on a binary classifier, i.e., one that discriminates on the basis of two classes. The data was first grouped according to the operational setting number 3 into six sets, each representing data collected at each setting respectively. Setting number 3 takes six possible levels: 0, 20, 40, 60, 80, 100 and represents some type of stress condition. To account for the six settings (stress levels), six classifiers were used based on the training data obtained from each group respectively, and each units level of health was estimated for each setting separately. For each setting the healthy training set was taken as a percentage of the initial observations, and the unhealthy/degraded data were taken as a percentage from the final observations across all units. The test data were similarly partitioned.

To get predictions, we first trained a classifier based on CALCEsvm. We used the output of CALCEsvm, the joint posterior class probability, to make predictions. As discussed above, CALCEsvm reduces the dimensionality and outputs a univariate time series of health estimates. In this case, CALCEsvm was used as a two-class classifier, using both healthy and unhealthy training data. Figure 10.6 plots the

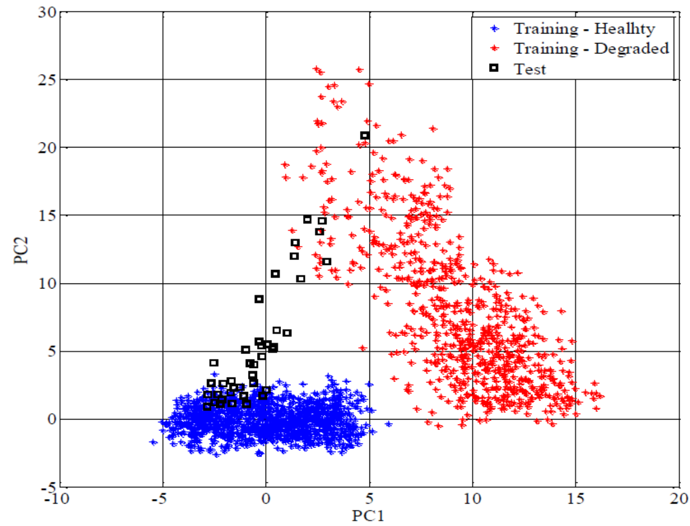


Fig. 10.6: Distribution of projected multivariate data on to first two principal components

distribution of the healthy, unhealthy, and test data for setting 0 projected onto the first two principal components. In Figure 10.6 we see that the two sets are suitable for classification-based detection approaches. We also see the trajectory of the test data, starting in the healthy set and moving towards the unhealthy. This pattern presents a time series of health estimates that exhibits a trend which is suitable for prediction.

CALCEsvm was applied to the training data for each setting and gave a sequence of health estimates indexed by a unique cycle number associated with each observation. By combining the health estimates from each setting into one vector and sorting by the cycle number, we got the final time series of health estimates for each unit. For example, figure 10.7 plots the probability of health for each test observation across all settings for training unit 200. In figure 10.7 we get the desired and expected drift in the health estimate as the unit ages. Health estimates close to 0 indicate a failing or failed unit.

A similar time series was estimated for each test unit from 1 to 218. There

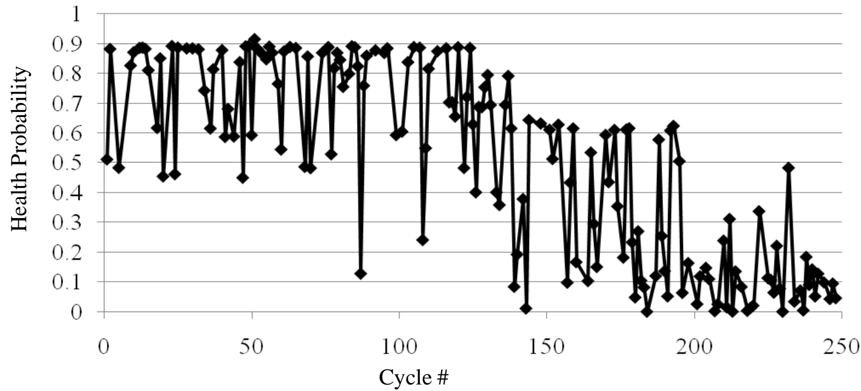


Fig. 10.7: Estimate of unit health as a function of time

were some test units with very short histories that had not started to degrade in health by last observation. For these units we did not get clear downward trends, and prediction in these cases was difficult. To predict the expected time to failure, we fit a GP model to the resulting time series. To account for health estimate variability, we fit the GP multiple times, each time perturbing the estimates with Gaussian noise with a mean and variance estimated from the training data across all units (cross unit variation). The cross unit variation for all test units is shown in figure 10.8. We see that the variance increases towards the end of life, validating the simulation design [38]. The drop in the expected variance, as seen in the lower plot of figure 10.8, is a result of a decreasing sample size during those time points (some test units survived longer than others).

Similar results were obtained for the training units. Using this variation profile, we injected noise into the original health estimates and fit the GP model multiple times. Each fit will give us the mean and the 2 standard deviation paths. We modeled the failure-time as the first time the mean path hit some fixed threshold; for example, a health level of 0. Figure 10.10 plots an example of such a procedure; it shows the GP fit to the health estimates for unit 200. It also overlays the empirical distribution of the 50 first hitting times to a threshold level set at 0. The expected

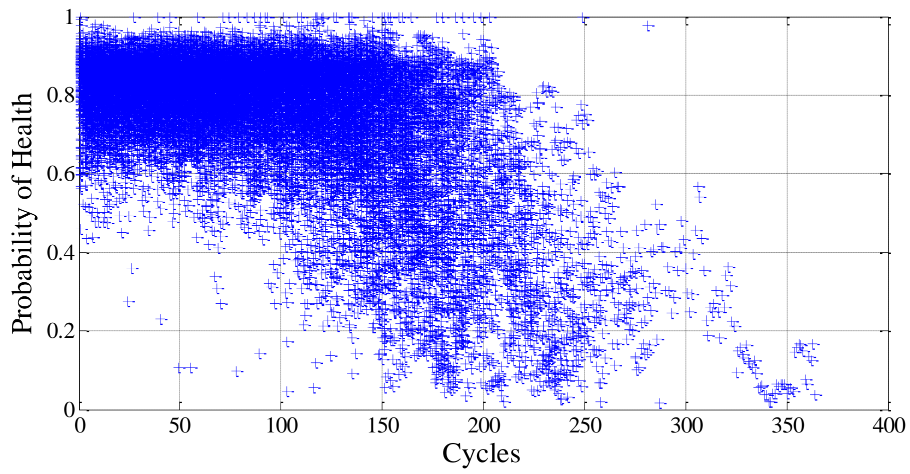


Fig. 10.8: Variation in estimated health probability across all training units

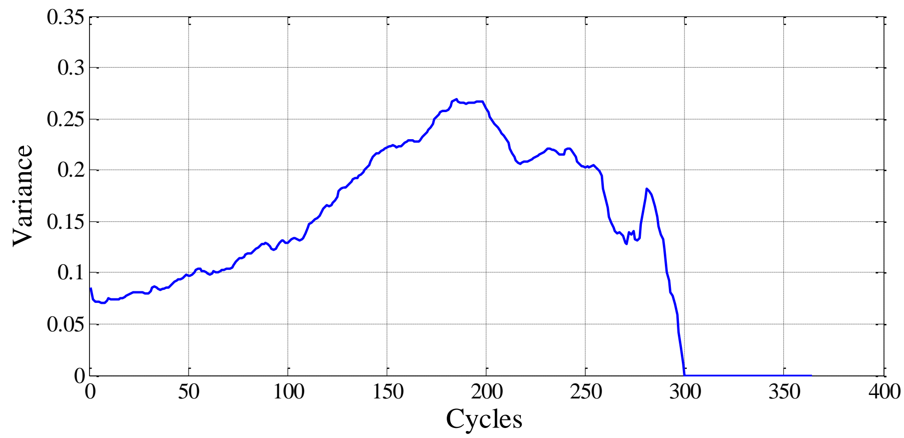


Fig. 10.9: Cross unit variation in the estimated health probability

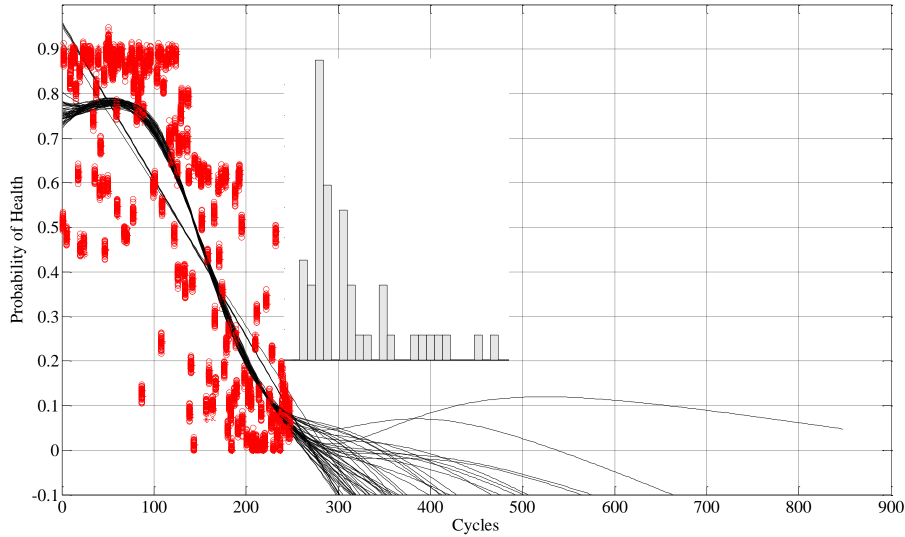


Fig. 10.10: GP model fit to the health estimate time series of a test unit

time to failure is computed by fitting a Weibull probability density function to the resulting n=50 first hitting times and computing its maximum likelihood parameters. Maximum likelihood estimates were obtained in the usual way, and we do not provide a discussion of this topic here. The expected time to failure estimates for all 430 test units are tabulated in Appendix A, together with the MLE confidence interval. The MLE confidence interval is estimated based on the asymptotically normal properties of MLEs. In this case the estimated expected time to failure is denoted by m, and the estimated variance by V. The confidence interval is calculated using: $CI = m + 1.96/\sqrt{n}V$.

11. MULTISTATE MODELS AS DEGRADATION MODELS

11.1 Introduction

In contrast to a continuous degradation process, where the degradation variable can take any real value, there is also interest in degradation variables that take on only discrete finite values. For example, computers often exhibit intermittent faults before failure. Intermittent faults may occur when the level of a performance variable exceeds a certain threshold and then recovers. Performance variables in electronics include variables such as resistance, voltage, current, etc. In computers, intermittent faults can take the form of discrete events, such as, error messages. Intermittent faults can be useful in tracking a latent (unobserved) degradation process [24], [136], [32], [137].

A *degradation process* in an electronic system is the process of degradation-wear-out that the system undergoes and that eventually causes it to fail. In general, a degradation process is a collection of degradation variables over time. A degradation process is considered latent when the *degradation variable* cannot be measured, or when it defines some unknown effect that we are interested in modeling. From chapter 1, recall, degradation variables in an electronic PCB can be, for example, the length of a crack in a solder joint, the level of corrosion on an inter-metallic lead, or the level of electro-migration of metal in a transistor. Due to the scale of electronic devices and their inaccessibility within larger host systems, such degradation variables are typically unobserved. Computers are a typical example of such host systems.

A degradation process, latent (unobservable) or not, typically also defines the time of failure, usually as a result of the degradation variable crossing a threshold. For example, in reliability studies of solder joints, material fatigue might be considered to be one of the failure mechanisms, and solder cracks as a failure mode. In this case, the degradation variable will be the crack length, which is usually unobservable, and time to failure will be the time at which the length of the crack reaches a certain threshold level. Thresholds for these variables are typically set based either on experience, industry standards, or estimations from past failures. In computers, for example, more general definitions of failure mechanisms and modes may apply. For example, failure might be defined from the user perspective, in which case failure is seen as the loss of functionality of the computer (Indirect failure, see Chapter 1). In this case, the degradation variable is unknown and therefore again considered latent.

Stress factors or covariates can be used to track the progress of the latent degradation variable, giving insight into the forces driving degradation. Covariates are typically collected from sensors that measure performance and/or environmental variables such as temperature, pressure, humidity, resistance, current, etc. We can also consider discrete event types as covariates; for example, the occurrence of an intermittent failure, or event. An intermittent failure can be thought of as a discrete event that can also help track the degradation process. When we assume that a covariate can be a stochastic time-varying variable, like the occurrence of an intermittent failure, then we can call that covariate a marker or precursor. When the marker is collected over time it forms a stochastic marker process. Together, the marker and degradation processes form a joint stochastic process that can be used in a model to predict the time-to-failure in new devices under observation.

In this chapter we present a degradation model that exploits information in a marker process to estimate 1) the expected time-to-failure, and 2) the survival

probability of an device given its marker process up until time t. Doksum and Hoyland (1992) [116], Doksum and Normand (1995) [117], Whitmore (1995) [118], and Whitmore et al (1998) [24], among others, model degradation by a Wiener diffusion process. Lee, Degruttola and Schoenfel (2000) [26], Henderson (2000) [120] consider extensions to the bivariate Wiener diffusion models with time-varying longitudinal marker processes. Longitudinal data are collected for each marker process not only at failure but also throughout the lifetime of the device. Satten and Longini (1996) and Hendriks (1996) use Markov models to combine a longitudinal trajectory with a survival event. In this work the authors develop parametric and predictive inference models that are generally analytically solvable. They make predictions of time to failure, expected time to failure, and other functionals of time by using the maximum likelihood estimates (MLEs) of the model parameters and plugging them into the predictive equations.

Meeker and Escobar (1998) [121], Commenges (1999) [122], Commenges (2002) [123], Bagdonavicius et al (2002) [124], Putter et al. (2006) [125], Pena (2006) [126], Machado et al. (2008) [128], Andersen and Perme (2008) [129], Cook, Lawless, Lakhal-Chaieb and Lee (2009) [130], Aalen, Borgan and Gjessing (2008) [131], Cook and Lawless (2007) [132] consider multistate models for survival and event history analysis based on counting processes. An excellent exposition, review and application of counting processes are given by Andersen, Borgan, Gill and Keiding (1993) [133], and Aalen and Johansen (1978) [134]. The counting process framework provides a non-parametric approach to inference and prediction, most famously through the *Kaplan Meier* (KM), *Nelson Aalen* and *Aalen-Johansen* (AJ) estimators. The AJ estimator or product-integral as it is otherwise known, estimates the transition probability matrix of a nonhomogeneous Markov process. This theory is useful to us in developing a simple markov multistate model to compute 1) the expected time-to-failure, and 2) the survival probability, for a surviving device given its marker

process up until time t.

The multistate model is a natural and simplifying representation of the bivariate marker-degradation process. With the multistate model, the multivariate state space of the marker and degradation processes is re-parameterized to a univariate Markov chain by expressing the state space as a list of all possible points of the marker-degradation vector pairs. Longitudinal marker observations are naturally incorporated into the model by directly contributing to the estimate the transition probability matrix, making predictions dependent on the process history. The time to failure is modeled as the FHT of the degradation variable to the failure threshold.

As discussed earlier, because degradation variables that define failure are typically not observable in computer systems, we examine the situation in which the degradation variable is unknown and unobservable. This might be the case in most complex electronic systems, such as computers, that can exhibit intermittent events indicating failure, but the underlying mechanism and therefore degradation variable is unknown. In this case predictive inference is based only on marker observations. In this work inference and prediction are based on data from one computer, and each failure time is modeled as independent and identically distributed. In our proposed model we assume that the system is as good as new after each failure. Critical error messages generated by internal performance monitoring software are considered as intermittent failure events correlated to failure.

Traditionally failure time models do not accommodate dynamic model parameters. In other words, most failure time models assume unknown but fixed parameters that are not influenced by changing environments. The focus in current literature is to incorporate information about the environment and or about the performance level of the system or device under observation, in order to model time dependent model parameters. Typically separate models are embedded into the time to failure probability density function to express the dependence on covariates

(sensor information) or markers like is the case in this study. Validating models and assumptions in such approaches becomes important and can be a limitation.

11.2 Failure Criteria and Degradation Model

Computer failures are defined from a user perspective, in which failure is seen as the loss of functionality of the computer. Failures are defined as automatic or forced restarts of the system. Automatic restart is triggered by the computer while forced restart is performed by the user. The failure time is defined as the time at which one of these events occurs. Once the computer is restarted, the process starts from time zero again, until the next failure. The time to the next failure can also be considered the time between failures.

The hypothesis is that each computer will fail (as defined above) as a result of usage and environment stresses. The hypothesis, therefore, is that there exists a measurable variable whose values correlate to the failure time, in other words, that there exists a marker/precursor to failure. Error messages generated by internal monitoring software are believed to be precursors to failure in computers. Fault events are simply called errors in the remainder of the chapter.

Lifetime and covariate data are only collected from one computer system. As mentioned earlier, the computer is assumed to be as good as new after each failure, and therefore we also assume that each failure-time is independent and identically distributed. For expositional simplicity, in this chapter, we use the term "device" to represent the information associated with the computer between each failure. For example, device 1 represents the computer between time zero and the time of the first failure. device 2 represents the computer from right after the first failure up to the second failure, and so on. Note also that this formulation is valid in the case when we have failure-time and covariate data from a sample of computer systems, in which case the assumption of independence is stronger, and the modeling holds

the same.

11.3 Data Structure and Notation

Internal computer software collects hardware data from a set of hardware variables indexed by time, and we denote the hardware variables by $X_{ij}, i = 1, \ldots, p$, $j = 1, \ldots, n$, where p is the total number of hardware variables, and n the total number of observations made on each variable. In other words, each hardware variable is observed n times. The hardware variables being monitored are: the Fan speed, CPU temperature, video temperature, memory temperature, motherboard temperature, %CPU usage, %CPU throttle, among other.

The computer also records information about background processes internal to each computer, also indexed by time. This information is observed in messages that are generated by the pre-installed Event Viewer program, and refer to errors, warnings or simply information about tasks that the computer succeeds or fails to accomplish. Events are represented by the binary variable Y_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, k$ with each Y_i representing a different event type. Each event variable Y_{ij} can take a value of 1 to indicate the occurrence of event j or 0 otherwise. Failure time data are collected as lifetimes T_q, defined as the time between failures, $q = 1, \ldots, r$, where r is the total number of failures observed in a test for a single computer.

As mentioned above, the Event Log Service records application, security, and system events in the Event Viewer. The Event Viewer records all events that occur into log files called event logs. Each event in a log can be classified into one of the following types: a) information, an event that describes the successful or failed operation of a task, such as an application, driver, or service. For example, an Information event is logged when a network driver loads successfully. b) Warning, an event that is not necessarily significant, however, may indicate the possible occur-

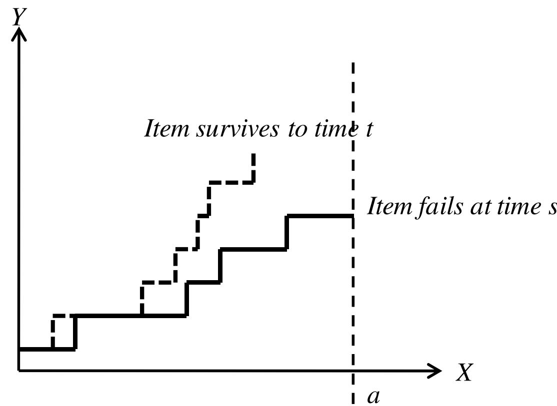


Fig. 11.1: Illustration of sample paths from a bivariate stochastic process $\{X(t), Y(t)\}$

rence of a future problem. For example, a Warning event message is logged when disk space starts to run low. c) Error, an event that describes a significant problem, such as the failure of a critical task. Error events may involve data loss or loss of functionality. For example, an Error event is logged if a service fails to load during startup. Examples of error events are: Application Hang, Memory Access Denied, etc.

11.4 The Multistate Model

A FHT model can describe the relationship between marker, degradation and lifetime variables. Typically, lifetimes are modeled by the FHT of the degradation process to a given threshold level a, namely $T = min(t|X(t) = a)$, in other words the minimum time at which the degradation process reaches the threshold level. Typically, the fht model for a joint stochastic process $\{X(t), Y(t)\}$ (degradation/marker) needs to condition on observations on $Y(t)$. Each stochastic process is defined on a state space $\mathcal{S}$ and time space $\mathcal{T}$. The state space of the marker variable is the set of all natural numbers, which record the cumulative number of errors.

Figure 11.1 illustrates the bivariate stochastic process $X(t), Y(t)$ with a

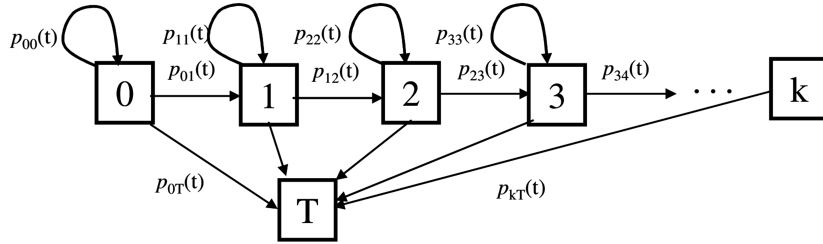


Fig. 11.2: Multistate model representation of a bivariate stochastic process

threshold level for $X(t)$ at a, and two possible paths, one resulting in a failed device and the other a surviving device. Because the state space of $Y(t)$ takes discrete states, the bivariate process looks like a step function. Later, we present a case study where we generate observations on $Y(t)$ by modeling it as a Poisson process. The discrete state space of $Y(t)$ can be expressed as a multistate representation as illustrated in figure 11.2.

Each state in the multistate model represents the cumulative number of errors experienced by the device, and the devices latent degradation level, with the exception of the last state which represents the cumulative number of errors and a degradation level of a. At any moment in time, therefore, the multistate model accounts for two possible events; the occurrence of an error or a failure. There are three possible transitions at any moment in time: a) to the failure state F, b) to the right adjacent state from 0 to 1 or from 1 to 2, etc., and c) remain in the same state, i.e., via $p_{11}(t)$ from state "1" to state "1". The characteristic probability transition matrix for this model is a sparse banded symmetric matrix P with elements along the primary and upper adjacent diagonal and along the last column of the matrix as illustrated in figure 11.3. Further description of the resulting matrix is given in section 12.6.

Although in principle there may not be a maximum number of states for the

marker variable, in this model a maximum number of states is used based on the sample data. For example, k in figure 11.2 is equal to the maximum number of errors across all devices for the computer under test. We assume that all devices begin in state 0 at $t = 0$, that state $0, \ldots, k$ are transient, and that state F is absorbing. With the multistate model, the bivariate state space $\{X(t), Y(t)\}$ is reparametrized to a univariate Markov chain by expressing the state space as a list of all possible points in the bivariate space. In other words, each state represents the cumulative number of observed errors and the associated state of health. A univariate representation of the state space allows for a model to treat the latent state of health (degradation) abstractly, without making any assumptions about the joint distribution of degradation and marker, and make predictions therefore about degradation purely through the marker.

11.5 Multistate Markov Chain

Consider a Markov chain Y_n for the multistate model, with state space $\Omega = \{\omega \in \mathcal{N}\}$ and transition probability matrix $P = \{p_{ij}\}$ $i, j \in \Omega$, with $p_{ij} \leq 0, \Sigma_{k \in \Omega}$ $p_{ik} = 1$ for all states i, j, $p_{ij} = P(Y_{n+1} = j | Y_n = i)$. The transition probability matrix is a sparse banded matrix with elements along the primary and upper adjacent diagonals and along the last column of the matrix as illustrated in Fig. 11.3.

Each element in the transition probability matrix represents the probability of transitioning into the corresponding state in the multistate model. Elements in row i represent states $j \neq i$ that the system transitions to from state i. For example, if we consider row $i = 1$, then element (1,2) of the matrix represents the probability of transitioning from state 1 to state 2, and element (1,1) the probability of transitioning from state 1 to state 1, in other words remaining in the same state. The matrix is sparse because most of it is populated with zero elements. Zero entries in the transition matrix are used to model a zero probability of transition into the

corresponding state. For example, in our model, element (1,3) is a zero element because the system cannot transition from state 1 to state 3; physically it cannot experience the third error before it experiences the second error. The diagonal entries then are the probabilities of transitioning along the marker states and the last column the probabilities of failure from each respective state. The multistate model is also an absorbing Markov chain.

Definition 11.1. *A state s_i of a Markov chain is called absorbing if it is impossible to leave it (i.e., $p_{ii} = 1$). A Markov chain is absorbing if it has at least one absorbing state, and if from every state it is possible to go to an absorbing state (not necessarily in one step)*

Definition 11.2. *In an absorbing Markov chain, a state which is not absorbing is called transient. The one-step transition probability matrix P is decomposed into the following block partitioned form, Bremaud (1998) [135], called the canonical form of P.*

$$P = \begin{pmatrix} Q & U \\ 0 & I \end{pmatrix} \tag{11.1}$$

Component Q is an k-by-k matrix of transition probabilities among the set of transient states $\mathcal{T}$ called the transient state matrix, U is a nonzero k-by-1 vector of transition probabilities from transient states to the failure state (absorbing state) F, and I a diagonal matrix of ones, in our case I=1, because we have only 1 absorbing state. The probability of being in state j after n steps, when the chain starts in state i is given by:

$$P^n = \begin{pmatrix} Q^n & U \\ 0 & I \end{pmatrix} \tag{11.2}$$

where E is a matrix written in terms of Q and U, but its expression is not needed here. The form P^n shows that the entries of Q^n give the probabilities of being in each transient state after n steps.

$$
\begin{bmatrix}
p_{11} & p_{12} & 0 & \cdot & \cdot & \cdot & p_{1n} \\
0 & p_{22} & p_{23} & & & & \cdot \\
0 & 0 & p_{33} & \cdot & & & \cdot \\
\cdot & & & \cdot & \cdot & & \cdot \\
\cdot & & & & \cdot & \cdot & \\
\cdot & & & & & \cdot & p_{n-1n} \\
0 & \cdot & \cdot & \cdot & & 0 & p_{nn}
\end{bmatrix}
$$

Fig. 11.3: Shape of probability transition matrix

Theorem 11.1. *In an absorbing Markov chain, the probability that the process will be absorbed is 1 (i.e., $Q^n \to 0$ as $n \to \infty$).*

The results from equation (11.2) concern the probability of remaining forever in the transient set, or alternatively, the probability of never being absorbed by the absorbing set. It is of interest to compute the probability of being absorbed by a given absorbing set when starting from an initial state $i \in T$. For this we use the idea of the fundamental matrix, which is related to the number of visits to a particular state j when starting from a state i.

Theorem 11.2. *For a homogenous Markov Chain with transition probability matrix P, the probability of absorption by the absorbing set starting from transient state i is*

$$P_F = HU \tag{11.3}$$

where H is the fundamental matrix and U is as in the canonical form.

Definition 11.3. *For an absorbing Markov Chain P, the matrix $H = (1 - Q)^{-1}$ is called the fundamental matrix of P. The entry ij of H gives the expected number of steps the chain takes to reach state j if it starts in the state i. Then the expected*

time to absorption, or time to failure, is taken from the definition of the expectation of the discrete random variable T; $E(T)$, which is given by:

$$T_e = E[T_{iF}] = \sum_0^{\infty} nP(T_{iF} = n) \tag{11.4}$$

In matrix form, equation (11.4) is given by (11.5), where, H is defined above, and c is a column vector all of whose entries are 1.

$$T_e = Hc \tag{11.5}$$

11.6 Estimator for the Transition Probability Matrix

Our task is to model the marker history and lifetime for each device by a stochastic process with a countable number of states represented in a multistate model. In general, the future state transitions of a multi-state model may depend in a complicated way on past events. However, for the special case of a Markov chain the past and future are independent given its present state. Therefore the future transitions of a Markov chain depend only on its present state as described by the transition probabilities $P_{ij}(s,t) = P(Y(t) = j|X(s) = i)$; $s < t$. We will show that an estimator for the transition probability matrix can accommodate historical information, and overcome the apparent limitation inherent to Markov chain models.

Corresponding to the hazard rate for survival, we may for a Markov chain define the transition intensities

$$\alpha_{ij}(u) = lim_{\Delta u \to 0} P(Y(u+du) = j|Y(u) = i)/du \tag{11.6}$$

where $Y(t^-)$ denotes the value of the marker Y "just before" time t. Note that $\alpha_{ij}(t)dt$ is the probability that an device that has experienced i intermittent faults

(so is in state i) "just before" time t, will make a transition to state j in the small time interval $[t, t + dt)$.

Only for simple Markov chains, is it possible to give explicit expressions for the probability transition matrix in terms of the transition intensities. For example in the case of a two-state model, the components of the probability transition matrix are simply the KM survival and failure estimates. More generally we can express the $(k + 1) \times (k + 1)$ transition probability matrix $P(s, t) = \{P_{ij}(s, t)\}$ in terms of the matrix of transition intensities. To see how this is done, we partition the time interval $(s, t]$ into a number of time intervals $s = t_0 < t_1 < t_2 < \ldots < t_k = t$ and use the Markov property to write the transition probability matrix as the product:

$$P(s, t) = P(t_0, t_1) \times P(t_1, t_2) \times \ldots \times P(t_{k-1}, t_k) \tag{11.7}$$

If the number of time points increases, while the distance between them goes to zero uniformly, the matrix product approaches a limit termed a (matrix-valued) product-integral. The product-integral is written in terms of the $(k + 1) \times (k + 1)$ matrix $\alpha(\mathbf{u})$ of transition intensities, that is the matrix where the off-diagonal elements equal the transition intensities $\alpha_{hj}(u)$, $h = j$, and the diagonal elements

$$\alpha_{ii}(u) = -\sum_{j=i} \alpha_{ij}(u)$$

are chosen so that all row sums are zero. Since the transition intensities describe the instantaneous probabilities of transitions between states, $P(u, u + du)I + \alpha(u)du$, where I is the $(k + 1) \times (k + 1)$ identity matrix. This explains why we may write the limit of equation (11.7) in product-integral form as: $P(s, t) = \prod_{(s,t]} I + \alpha(u)du$. Alternatively, if we let $A(t)$ denote the cumulative transition intensity matrix with

elements $A_{ij}(t) = \int_0^t \alpha_{ij}(u)du$ we have:

$$P(s,t) = \prod_{(s,t]} \{I + dA(u)\} \tag{11.8}$$

In the discrete case, the cumulative transition intensity takes the form $\sum_{u \leq t} \Delta A(u)$, where $\Delta A(u) = A(u) - A(u^-)$. Then the product-integral in equation (11.8) becomes the ordinary matrix product $\prod_{s<u\leq t}\{I + \Delta A(u)\}$. We may derive an estimator for $P(s,t)$ using the Nelson-Aalen (NA) estimator $\hat{A}$ as a non-parametric estimate for the cumulative transition intensity matrix A. In a counting process framework, the NA estimator is constructed by counting the devices that are observed to experience a transition from state i to state j in $[0,t]$. The NA estimator is given by:

$$\hat{A}_{ij}(t) = \int_0^t \frac{dN_{ij}(s)}{R_h(s)} \tag{11.9}$$

where $dN_{ij}(t) = N_{ij}(t+dt) - N_{ij}(t)$, is the increment in the number of transitions from state i to state j observed over a small time interval $[t, t+dt)$, and $R(t) = \{\#devices : T_i > t\}$ is the risk set; the number of surviving devices at time t. Furthermore we introduce $\hat{A}_{ii}(t) = -\sum_{j \neq i} \hat{A}_{ij}$. The relation in equation (11.8) suggests that we estimate the matrix of transition probabilities by the $k \times k$ matrix $P(s,t)$, called the Aalen-Johansen (AJ) estimator, $\hat{A} = \{\hat{A}_{ij}\}$

$$\hat{P}(s,t) = \prod_{(s,t]} (I + d\hat{A}(u)) \tag{11.10}$$

The NA estimators are step-functions with a finite number of jumps on $(s,t]$. Therefore, the AJ estimator given in (11.10) is a finite product of matrices. If one or more transitions are observed at any time u, then the contribution to (11.10) from this time point is a matrix $I + \Delta\hat{A}(u)$, where $\Delta\hat{A}(u)$ is the $k \times k$ matrix with entry (i,j) equal to $\Delta N_{ij}(u)/R_i(u)$ for $h \neq j$ and entry (i,i) equal to $-\Delta N_{io}(u)/R_i(u)$

with $N_{io} = \sum_{j \neq i} N_{ij}$. The transition probability matrix defines the instantaneous probability that an intermittent fault or failure may occur and is estimated using observations on the marker variable, in this case the event of an error message.

11.6.1 Evaluating the Probability Transition Matrix

In this section we present a "worked out" example for evaluating the probability transition matrix. We assume that the life-history of the system is described by a Markov process with a finite number of states $\Omega = 1, \ldots, K$. The transition probability is denoted, as before, by $p_{ij}(s,t)$ and describes the probability that a system in state i at time s transitions into state j by time t. The $K \times K$ matrix $P(s,t)$ summarizes the transition probabilities of the Markov process. We next define the transition probability matrix in terms of the AJ estimator.

Definition 11.4. *We define the following quantities:*

1. $t_1 < t_2 < \ldots$ *are times when transitions are observed to occur*
2. d_{ijk} *is the number of items that transfer from state* i *to* j *at time* t_k
3. $d_{ik} = \sum_{i \neq j} d_{ijk}$ *is the number of items that transition out of statei at time* t_k
4. R_{ik} *is the number of devices in sate* i *just prior to time* t_k

The AJ estimator for the transition matrix $P(s,t)$ is given in equation (11.10). The interpretation of equation (11.10) is given by equation (11.11), which is the product over transition matrices between times s and t.

$$\hat{P}(s,t) = \prod_{k: s < t_k \leq t} (I + \hat{\alpha}_k) \tag{11.11}$$

Here,$\hat{\alpha}_k$ is a $K \times K$ matrix with entry (i,j) equal to $\hat{\alpha}_{ijk} = d_{ijk}/R_{ik}$, entry (i,i) equal to $\hat{\alpha}_{iik} = -d_{ik}/R_{ik}$, and all other entries are zero. I is the indentity matrix.

The product is taken over increasing t_k's. In the case of a two-state model, where state 1 represents the healthy state and state 2 the failure state, the transition and intensity matrices are two dimensional. In this case the AJ estimator can be seen as a matrix version of the KM estimator: $\hat{S}(t) = \hat{P}[1,1]$. The square brackets indicate the AJ matrix as opposed to the parenthesis which denote the time components.

11.6.2 Example of a two-dimensional multistate model

To demonstrate the computations, consider the following example:

Suppose we have a sample of 16 devices observed to failure and censoring. Out of these 16 devices 12 fail at times t_k (0.75, 0.91, 1.32, 1.7, 2.15, 2.76, 2.88, 2.98, 4.51, 6.23, 8.57, 10.23) and four are censored at times c_j (0.5, 0.8, 1.7, 2.08), where times are measured in hours. Table 11.1 summarizes the data and gives the survival probability estimates at each observation time t_k. Column 1 shows the device ID, column 2 the event time, column 3 shows the failure indicator, which is 1 if a failure is observed and 0 if the failure is censored. Column 4 shows the states that the device transitions from and column 5 the state the device transitions into. In this case, state 1 represents a healthy state and state two the failure state. Column 7 shows the at-risk population at time t_k, and the last two columns show the estimated survival and failure probabilities at that time, respectively.

In this case the AJ estimator reduces to the KM estimator as mentioned earlier. Here we show the steps involved in evaluating the AJ estimator, starting from its definition in equation (11.11) and the properties in definition 11.4

$$\hat{P}(0,t) = \prod_{k:t_k \leq t} (I + \hat{\alpha}_k) =$$

$$= \prod_{k:t_k \leq t} \left[\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} -\frac{d_{ik}}{R_{ik}} & \frac{d_{ijk}}{R_{ik}} \\ 0 & 0 \end{pmatrix} \right] =$$

Tab. 11.1: Lifetime data and survival probability estimates

ID	t_k	d_{ijk}	From	To	R_{ik}	$\hat{S}(t_k)$	$\hat{F}(t_k)$
13	0.5	0	1	1	16	1	0
1	0.75	1	1	2	15	0.933	0.066
14	0.8	0	1	1	14	0.933	0.066
2	0.91	1	1	2	13	0.861	0.138
3	1.32	1	1	2	12	0.789	0.21
4	1.7	1	1	2	11	0.717	0.282
15	1.7	0	1	1	11	0.652	0.347
16	2.08	0	1	1	9	0.652	0.347
5	2.15	1	1	2	8	0.571	0.428
6	2.76	1	1	2	7	0.489	0.51
7	2.88	1	1	2	6	0.407	0.592
8	2.98	1	1	2	5	0.326	0.673
9	4.51	1	1	2	4	0.244	0.755
10	6.23	1	1	2	3	0.163	0.836
11	8.57	1	1	2	2	0.081	0.918
12	10.23	1	1	2	1	0	1

$$= \prod_{k:t_k \le t} \begin{pmatrix} 1 - \frac{d_{ik}}{R_{ik}} & \frac{d_{ijk}}{R_{ik}} \\ 0 & 1 \end{pmatrix} =$$

$$= \begin{pmatrix} \prod_{k:t_k \le t} \left(1 - \frac{d_{ik}}{R_{ik}}\right) & 1 - \prod_{k:t_k \le t} \left(1 - \frac{d_{ijk}}{R_{ik}}\right) \\ 0 & 0 \end{pmatrix} =$$

$$= \begin{pmatrix} \hat{S}(t) & 1 - \hat{S}(t) \\ 0 & 1 \end{pmatrix}$$

Using the above data one can show that the following results hold

$$\hat{P}(0, t_1) = (I + \hat{\alpha}_1) = \begin{pmatrix} \frac{14}{15} & \frac{1}{15} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0.933 & 0.066 \\ 0 & 1 \end{pmatrix}$$

$$\hat{P}(0, t_2) = \hat{P}(0, t_1)(I + \hat{\alpha}_2) = \begin{pmatrix} \frac{14}{15} & \frac{1}{15} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{12}{13} & \frac{1}{13} \\ 0 & 1 \end{pmatrix} =$$

$$= \begin{pmatrix} \frac{56}{65} & \frac{9}{65} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0.861 & 0.138 \\ 0 & 1 \end{pmatrix}$$

$$\hat{P}(0, t_3) = \hat{P}(0, t_2)(I + \hat{\alpha}_3) = \begin{pmatrix} \frac{56}{65} & \frac{9}{65} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{11}{12} & \frac{1}{12} \\ 0 & 1 \end{pmatrix} =$$

$$= \begin{pmatrix} \frac{154}{195} & \frac{41}{195} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0.789 & 0.21 \\ 0 & 1 \end{pmatrix}$$

The AJ survival probability estimates are plotted in Fig. 11.4. The survival probability is based on the (1,1) element of the AJ estimator at each time step t_k, as seen from the results above. The resulting survival probability plot in figure (Fig. 11.4) is a step function which drops at each observed failure time. From figure 11.4 we can infer the survival probability of a new device surviving at time t^*. For example, a new device, surviving at $t^* = 4$, has a 65% chance of failing in the next time instance. To summarize, the AJ estimator, non-parametrically estimates the survival probability as a function of time by taking a product of the transition intensities of a Markov chain at each of the observed event times. Because here we only have two types of events, there is only one transition: from healthy to failure. To see the usefulness of this model, we next present a case study that generalizes these computations to higher dimensions to accommodate recurrent intermittent faults.

The expected time to failure is computed using Markov chain theory on absorbing chains as discussed in section 12.4. Specifically, we are interested in the fundamental matrix of the transition probability matrix H. As stated in 11.3, the ij entry of H gives the expected number of steps the chain takes to reach state j if it starts in the state i. Consider for example, the transition probability matrix given by the AJ estimator at time t_3, in this case the transient state matix Q is

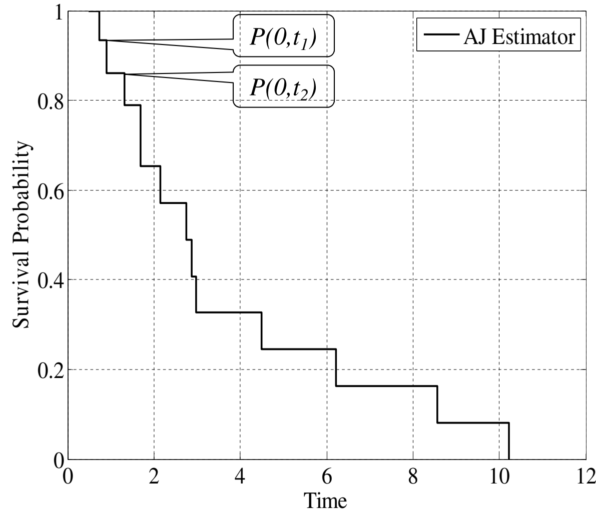


Fig. 11.4: Survival probability estimates from the AJ estimator

a scalar $Q = 0.789$, then $H = (1 - 0.789)^(-1) = 4.74$ and the expected time to absorption $T_e(t_3) = 4.74(1) = 4.74$. Repeating the above calculation at all failure and censoring times we get the following expected time to failure vector, which is plotted in figure 11.5. Again, figure 11.5 can be used to infer the expected time to failure for a new device surviving at some time t^*. Continuing the example above, for an device surviving at $t^* = 4$, its expected time to failure is just over an hour.

$$T_e = (\text{NA}, 12.75, 12.75, 7.22, 4.74, 3.54, 2.88, 2.88,$$
$$2.33, 1.96, 1.69, 1.48, 1.32, 1.19, 1.08, 1.00)$$

The first entry in the expected time to failure vector is the survival estimate at the earliest lifetime measurement $t_1 = 0.5$, which is a censoring time. Because at this time, no failures have previously been observed, this censored lifetime does not contribute to the survival probability estimate. The AJ estimate at time t_3, is the survival probability estimate at the third failure time, or including censored times, its the fifth lifetime measurement, as seen in the vector above.

In a higher dimensional multistate model, we need to take the inverse of Q

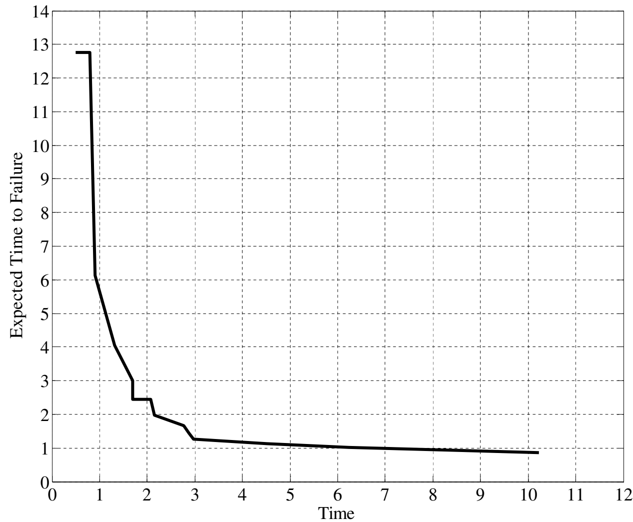


Fig. 11.5: Expected time to failure

matrices, which becomes a computational problem when there are many states to account for. Difficulties may arise in taking the inverse of anon-singular matrix. Here we assume that computational techniques for matrix inverse evaluation are available (LU decomposition, Singular Value decomposition, Gauss-Jordan elimination, etc.). In the following case study the same calculations are used but in matrix form.

11.7 Case Study

In tests, failures are repeatedly induced by stressing the computer with simulated usage. Failures are defined as unwanted/automatic restarts or shutdowns. Computer usage is simulated by running intensive programs that consume computational resources, and cause the computer to freeze or trigger a shutdown. It is believed that errors are correlated to failure and can therefore be used as precursors/markers in our multistate model. Table 11.2 shows a sample of failure times observed for a computer under test. The first column shows the start time after a restart/reboot of the computer. The second column is the time the computer was observed to fail: time to failure (TTF). The third column records the calendar time

Tab. 11.2: Sample of failure times for collected from a computer

Start	Stop	TTF	TBF
2/23/2010 7:53	2/24/10 11:05	4:33	27.19
2/24/2010 11:05	2/25/10 7:26	12:57	20.35
2/25/2010 7:26	2/25/10 8:09	6:14	0.73
2/25/2010 8:09	2/26/10 13:28	13:55	29.32
2/26/2010 13:28	2/26/10 23:16	9:07	9.79

Tab. 11.3: Life Table

ID	Time	I_f	Y
1	23.1	0	1
1	27.19	1	0
2	18.61	0	1
2	20.01	0	1
2	20.35	1	0
3	0.73	1	0

in hours from the start of the experiment (t=0). The last column shows the time in hours between each failure (TBF). This is the lifetime of an device as discussed in section 3.

In addition to failure time information, we also collect error event times. This data is collected in a similar way. Lifetimes and error event times are collected into a "life table", which contains all the data: lifetimes and error event times in one table structure. Table 11.3 gives an example of a life table. Each row in the life table represents the occurrence of an event: failure or error. In Table 11.3, the failure ID represents an device in the computer. There may be multiple error events prior to each failure; therefore, the same failure ID can appear multiple times. When a failure is observed, the failure indicator $I_f = 1$, otherwise $I_f = 0$ to indicate an error event.

We simulated lifetime and error data. The main objective of this case study is to test and validate the proposed multistate model. Our simulation design hypothesis is: 1) that with each error, the probabilities of another error or failure occurring increase and 2) with each failure the probability that an error occurs increases. In

other words, devices that experience more errors typically fail earlier, and a computer which has seen many failures is more likely to generate more errors than a computer which has not seen as many failures.

In our simulation, random lifetimes and error times are generated based on the FHT model discussed in section 12.4. Lifetimes T_q, therefore, are modeled as the first hitting time of a degradation process $X(t)$ to a fixed failure threshold a, namely, $T = inf t : X(t) = a$. We modeled the degradation process by a Wiener process with drift given by:

$$X(t) = x_0 + \nu_X t + \sigma_X W(t) \tag{11.12}$$

In equation (11.12), x_0 is the initial degradation of an device, ν_X is the degradation drift parameter, σ_X the degradation variance and $W(t)$ is a standard normal Wiener process $W(t) \sim N(0, t)$. The degradation process is conditioned on the error events represented by the marker variable $Y(t)$, which is modeled by a homogeneous Poisson process with arrival rate parameter λ, $Y(t) \sim Poisson(\lambda)$, where $Y(t) \in \mathbb{N}$ [1]. Its probability mass function is given by:

$$P(Y(t_2) - Y(t_1) = k) = \frac{\lambda exp\left(-\lambda(t_2 - t_1)\right)(t_2 - t_1)}{k!} \tag{11.13}$$

The drift of the Wiener process is parameterized as a function of the number of errors observed before failure, $\nu_X(t) = cY(t)$ where $c > 0$ is a positive constant. In other words, the drift of the Wiener process is simulated to increase each time an error occurs. An increase in the drift of the Wiener process in turn means that the degradation is more likely to reach the threshold faster. We model the effect of ageing in computers by parameterizing the rate of arrival of errors as a function of cumulative number of failures and errors up until time t, $\lambda(t) = bY(t)C(t)$ where

[1] All natural numbers 0,1,2,. . .

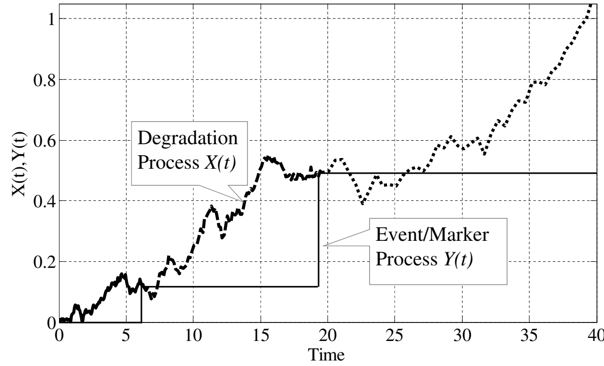
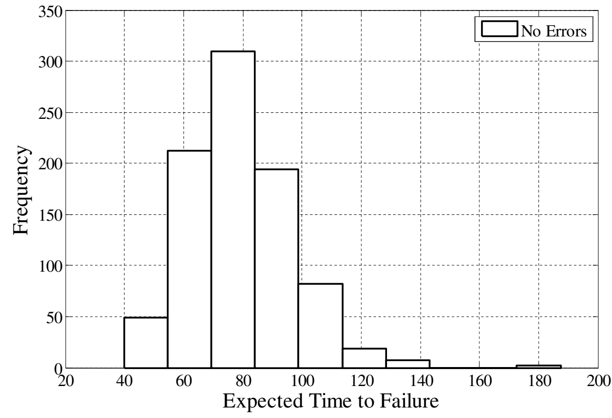


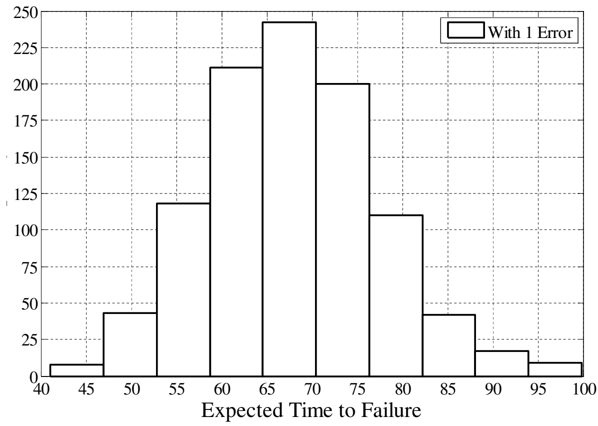
Fig. 11.6: Simulation of the degradation process conditioned on an error event process

$C(t)$ is the cumulative number of failures.

The simulated data consists of 40 simulated lifetimes and 45 event times summarized in table 11.4. The degradation and event processes are generated using starting parameter values: $(\nu_X, \sigma_X, \lambda) = (0.01,\ 0.035,\ 0.009)$ and an arbitrarily threshold level of a=1. Fig. 11.6 plots the simulated degradation process conditioned on an event process. Specifically it plots the degradation process of an device conditioned on two error events, one that occurs at 6.32 hours and the second at 19.56 hours. This device fails at 39.73 hours. In this data set (40 devices), the expected time to failure for devices that do not experience errors is 70.58 hours, for devices that experience 1 error 47.05 hours, and for devices that experience 2 errors 30.87 hours.

The simulation is repeated many times, in our case 1000, to get an empirical estimate for the convergence of the expected time to failure given a sample of 40 devices. Figure 11.7 plots the empirical distribution of the expected time to failure for 1000 experiments, each time considering only 40 devices. We can see that the expected time to failure for devices that do not experience errors is on average greater than that for devices that experience 1 or 2 errors. Similarly the expected time to failure for devices that experience 1 error is again on average greater than that





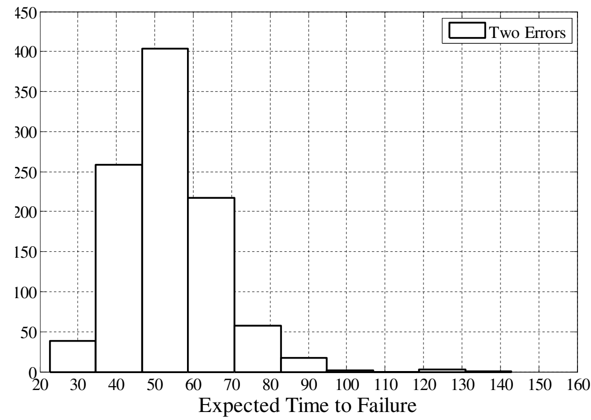


Fig. 11.7: Expected time to failure distribution for devices with 0,1 and 2 error events

for devices that experience 2 errors. Using random lifetime and error times generated from our statistical model this result validates the simulation design discussed earlier.

$$\begin{bmatrix} p_{01} & p_{01} & 0 & p_{03} \\ 0 & p_{11} & p_{12} & p_{13} \\ 0 & 0 & p_{22} & p_{23} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Fig. 11.8: Transition probability matrix used in case study

The multistate model, in this case, consists of two marker states and a failed state. In this model only two errors occurrences are considered prior to failure for each device, but this need not be the case, any discrete number of errors can be accommodated. State 0 represents the state in which the computer has not experienced any errors, state 1 the state in which the computer experiences its first error and state 2 its second and last error. States 0 through 2 represent the transient states in a Markov chain. State 3, the failure state, represents the absorbing state. Fig. 11.8 shows the structure of the transition probability matrix used in this case study for a three state multistate model.

In a three state multistate model, each device under observation has a maximum of three possible event times, two for the errors and one for the devices failure. For example, ID=4 experiences its first error at 29.89 hours, its second at 32.82 hours and fails 41.37 hours. Time is measured from t=0 for all devices. For ID=1, it does not experience any errors and it fails at 59.63 hours.

The time considered in this experiment starts at t= 0 and ends at the time of the last observed failure across the 40 lifetimes, which for this data set is t= 83.36 hours. The AJ estimator is applied over a time increment of r= 0.00018 hours which is equal to the smallest time difference between any two error events in the data set. This guarantees that at any time, at most, only 1 error event can occur. The transition probability matrix therefore is estimated at every time step, in total 4631 times, to generate a three dimensional matrix $k \times k \times m$, with $k = 4$ and $m = 4631$. Each slice of this matrix along m, gives the non-parametric estimate of the transition probability matrix at a particular time.

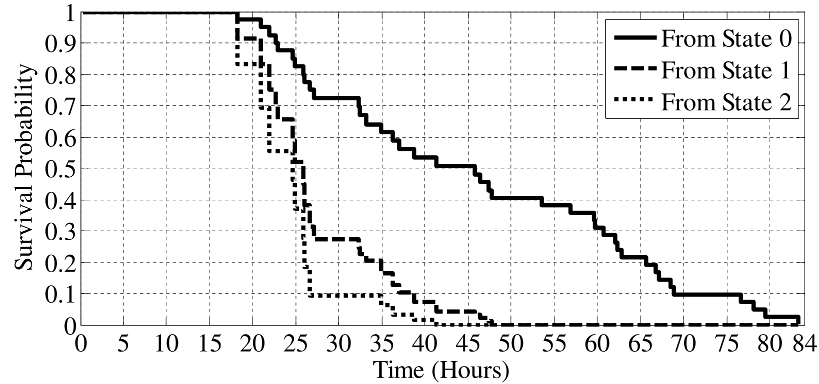


Fig. 11.9: Aalen-Johansen survival probability estimate starting from states 0, 1 and 2

Figure 11.9 shows the AJ estimator results for the survival probabilities starting from states 0, 1 and 2. These plots are generated based on a Matlab code developed at CALCE to implement the AJ estimator (equation (11.10)). The evaluation of the AJ estimator is performed at discrete times, similar to example 5.1. Survival probability from state 0 is based on element (1,4) in the transition probability matrix (see figure 11.6), from state 1, based on element (2,4), and from state 2 based on element (3,4).

From the results in figure 11.9 we observe that the AJ estimator detects the design conditions in the simulated data. We can see that devices that experience an increasing number of errors before failure have increasingly lower chances of surviving, and this result validates the simulation design discussed earlier.

Therefore, for a new computer, if we know how long it's been running since last shutdown, and we know the number or errors in any (errors of some predetermined type) it experienced, we can get point estimates of its remaining life. For example, for a test computer that is surviving at time t, its expected time to failure is plotted in Fig. 11.10, for times t ranging from 0 to 84 hours. Using equation (11.5), figure 11.10 plots the expected failure-time as the first hitting time of the

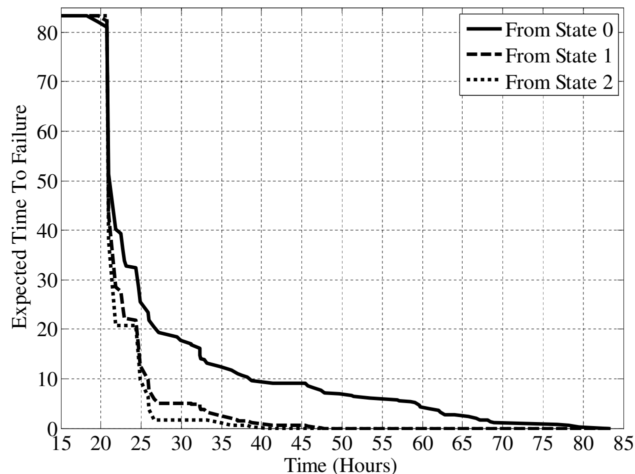


Fig. 11.10: Expected time to failure starting from state 0,1 and 2 respectively

absorbing state against time, starting in states 0, 1 and 2 respectively.

Here we see that occurrence of errors reduce the life expectancy of the devices. More specifically, for example, consider a device that is surviving at 30 hours, in other words, a computer that has not failed after 30 hours of operation. In this case, if the computer has not experienced any errors, its expected time to failure is just under 20 hours. If it experienced one error, then its expected time to failure is about 5 hours, and if it experienced two errors then its expected time to failure is about 1 hour. The exact times can be found in the data, here we are summarizing visually, based on figure 11.10.

11.8 Summary and Conclusions

Using a multistate model, we have developed a methodology to model failure time data together with event time data (errors). Event time data are used as auxiliary information to failure time data, with the aim to improve remaining life estimates. Failure times are modeled as the first hitting times to a failure state in a multistate model. The multistate model re-parameterizes the bivariate degradation/marker process to a univariate Markov chain by expressing the state space as

a list of all possible points in the bivariate state space. Each state represents the joint degradation/marker values indexed by time. The marker value is observable and represents the cumulative number of errors experienced by the device, while the degradation variable is unobservable and therefore unknown. In this approach, inference is based on the progress of the marker variable which is assumed correlated to the degradation variable.

Inference of the transition probability matrix is approached non-parametrically using the AJ estimator over small time increments spanning the life time of the device. A reducible absorbing time homogeneous Markov chain is used to compute the expected first hitting times to estimate the expected time to failure. A case study simulates a sample of 40 lifetimes and 45 error event times. The simulation generates failure times and event times according to a design hypothesis discussed in the chapter, and used to emulate experimental data. In the simulation, the marker variable is modeled as a Poisson random variable, and the degradation as a Gaussian random variable. The degradation process is conditioned on the event process; its drift parameter increases with increasing number of errors and failures in the computer. From the results we see, as anticipated, the expected time to failure decreases given an increasing number of errors experienced by an device.

This chapter contributes to the literature on degradation models on the following levels:

1. A recurrent event process is interpreted as a marker variable observed as a stochastic process and is used to relate observable events to the underlying degradation process of the system. In doing so, this work casts a degradation model as a multistate model that can be solved using Markov Chain theory, with a transition probability matrix that can be parameterized to accommodate covariates.

2. Complex multivariate relationships between covariates and the degradation

process are summarized using a multistate representation which is simple and intuitive to use for prediction.

3. Using counting process theory, time dependent covariates (in this case the occurrence of an error) influence the estimation of transition probabilities. This is perhaps the most useful aspect of the model because it can naturally accommodate a large number of covariates (in this thesis we use one) without burdening computations and complicating predictive inference equations.

4. By taking a non-parametric approach, the model avoids making some assumptions necessary in parametric modeling of degradation.

5. Connects the model with PHM, a methodology that requires real-time health assessment and predictions. Extensions to this model can include a parametric model to describe the relationship between the marker and a degradation variable. In this case, to achieve a valuable model, the degradation variable must represent the mode of a known failure mechanism, and the marker variable should be correlated to the degradation variable.

Tab. 11.4: Simulated Data for Case Study

ID	Time	FI	From	To	ID	Time	FI	From	To
1	59.63517	1	1	4	21	11.27847	0	2	3
2	3.667432	0	1	2	21	25.95885	1	3	4
2	13.66344	0	2	3	22	51.24407	0	1	2
2	26.62085	1	3	4	22	59.72076	1	2	4
3	29.51897	0	1	2	23	2.429111	0	1	2
3	37.05257	0	2	3	23	7.054876	0	2	3
3	38.69994	1	3	4	23	20.9028	1	3	4
4	29.89073	0	1	2	24	5.408587	0	1	2
4	32.82195	0	2	3	24	22.50397	0	2	3
4	41.3729	1	3	4	24	24.87877	1	3	4
5	2.82899	0	1	2	25	76.71006	1	1	4
5	22.96408	1	2	4	26	38.15745	0	1	2
6	79.5193	1	1	4	26	47.78758	1	2	4
7	32.28822	0	1	2	27	63.9671	0	1	2
7	47.3686	1	2	4	27	68.86442	1	2	4
8	49.65221	0	1	2	28	78.15233	1	1	4
8	53.49254	1	2	4	29	59.02685	0	1	2
9	67.14219	1	1	4	29	62.04048	1	2	4
10	17.60471	0	1	2	30	14.99896	0	1	2
10	31.23497	0	2	3	30	33.22369	1	2	4
10	36.27443	1	3	4	31	37.53772	0	1	2
11	2.525531	0	1	2	31	45.39904	0	2	3
11	22.69991	1	2	4	31	46.38308	1	3	4
12	65.67164	1	1	4	32	51.42378	0	1	2
13	2.554267	0	1	2	32	56.89857	1	2	4
13	17.68143	0	2	3	33	83.3691	1	1	4
13	21.86692	1	3	4	34	19.6774	0	1	2
14	4.869203	0	1	2	34	37.02043	1	2	4
14	27.16367	1	2	4	35	62.75821	1	1	4
15	62.26681	1	1	4	36	58.52609	0	1	2
16	68.14045	0	1	2	36	66.74907	1	2	4
16	68.50191	1	2	4	37	1.388588	0	1	2
17	31.1752	0	1	2	37	5.06764	0	2	3
17	44.50559	0	2	3	37	25.84599	1	3	4
17	45.78031	1	3	4	38	23.16686	0	1	2
18	57.07051	0	1	2	38	32.39662	1	2	4
18	60.75794	1	2	4	39	15.58235	0	1	2
19	15.35275	0	1	2	39	24.34467	0	2	3
19	32.28841	1	2	4	39	34.84539	1	3	4
20	0.352023	0	1	2	40	10.13809	0	1	2
20	15.44168	0	2	3	40	20.75748	0	2	3
20	18.1696	1	3	4	40	24.69688	1	3	4
21	9.604934	0	1	2					

12. SUMMARY AND CONCLUSIONS

In order to predict the remaining life of a device we need to know when other similar devices failed, as well as how they responded to stress over time. To get this information we must conduct tests, and expose a sample of devices to stress and measure their responses. Only then can we take information from a new fielded device and make any inference on its reliability. In this work the au The impetus for degradation models in PHM stems from the need to explain heterogeneous reliability qualities across a sample of devices used in dynamic stress environments. This need is further exacerbated by the requirement in PHM to predict failure-times when failure-time samples are small, and when degradation data are not predictive of failure. Small failure-time samples are common in highly reliable products or products with short product-cycles. Small failure-time samples also result from reduced accelerated test conditions, aimed to preserve the failure generating mechanisms for devices put through failure-tests.

Degradation data are collected for each device and used as auxiliary reliability information to improve reliability models. Heterogeneous degradation data, collected from degradation variables, can provide valuable insight into device-specific reliability. First hitting time models use degradation variables to define the failure-time, implicitly enforcing a causality between the underlying failure mechanism, degradation and the failure-time. Degradation variables play another important role in enhancing reliability and PHM models, because, they represent responses to stress or usage, which varies across devices. In this way degradation variables allow us to model the effect of dynamic environments on our failure-time predictions.

Central to the thesis are so called non-predictive degradation variables. These are known, typically observable, degradation variables that attain the failure - threshold level suddenly without any preceding trend, a trend useful for making predictions on lifetimes. In such a case, as we discuss in this thesis, we use latent degradation models with terminal observations on an otherwise latent *true degradation variable* and longitudinal measurements on observable marker variables. We provide justification for using bivariate latent degradation models for data collected in failure-tests, and through our model we address key data-limitations encountered in such settings.

Our baseline analytical framework is Whitmore's bivariate Wiener model for terminal degradation and marker data-observations. In chapter 1 we introduce the problem and motivate a direction of research. In chapter 2 we present the data-structures used in the thesis and examples taken from failure-tests of electronics conducted at CALCE. In chapter 3 we present main theorems and lemmas on estimation theory. In chapters 4, 5 and 8 we present extensions to Whitmore's first hitting time model. In chapter 6 we present case studies that compare the performance of our degradation model on various data-structures, the results of which form the basis of our contributions. In chapter 7 we consider the effect of covariates on estimation under the same degradation model.

The first extension and first contribution of this work is to address and provide a simple solution to the *"small failure-time sample"* problem. We developed parametric and predictive inference equations based on Whitmore's first hitting time model, for a data-structure that augments terminal degradation measurements to a terminal data-structure. With terminal degradation observations on both failed and surviving devices we were able to use, in contrast to Whitmore, terminal degradation information on surviving devices. In other words, in our approach, surviving devices become much more valuable for estimation and inference.

We compared the mean failure-time parameter of an $\mathcal{IG}$ lifetime distribution, and we showed that our approach under the TMD data-structure consistently reduces the asymptotic variance of MLEs. More importantly, our experimental results indicate that our approach performs increasingly better under smaller failure-time sample sizes. As expected, from a statistical point of view, the efficacy of the TMD over the TM data-structure is explained because ML estimation improves with enhanced data, when sample sizes are large.

The improvement in inference under the TMD data-structure also has broader commercial implications. The results show that we can "afford" to reduce designed accelerating conditions, and test durations in planned failure-tests. From an engineering perspective, reducing accelerating conditions is a welcome option, because, as we discuss in the introduction, highly accelerated test conditions may alter targeted failure-mechanisms. Typically, failure-tests are conducted over a time-period long enough to see enough devices fail. Shorter test-durations are not only less costly, they are sometimes required due to short product life-cycles. Under latent degradation conditions, our results suggest investing in failure-analysis equipment capable of measuring terminal degradation.

The second contribution of the thesis lies in our treatment of longitudinal marker observations in a latent degradation model for both parametric and predictive inference. We develop parametric inference for a general longitudinal marker data-structure and show in our results improved estimation starting from two intermediate marker observations. The efficacy of the longitudinal data-structure depends on the strength of correlation between the marker and the degradation variable. In our simulations and analysis, we use a strong correlation coefficient. Further work is needed to test estimation improvement under weaker correlations.

The third contribution of the thesis incorporates a variable failure-threshold to Whitmore's bivariate Wiener model. We develop parametric inference equations

with the failure-threshold variable modeled as a Gaussian random variable and independent of the degradation process. More realistically, as part of future work, the failure-threshold variable may also be dependent upon the degradation process, and can itself be modeled, more generally as a stochastic process. We are currently working on code to evaluate AREs under the TMD data-structure for fixed vs variable threshold models. These results are not availabe in the thesis, but are however anticipated to be part of subsequent publications.

In chapter 7, in addition to presenting covariates as part of FHT models, we present machine learning approaches as part of degradation models. Specifically, we consider Gaussian process regression and support vector machine classification as methods for including covariates. Our introduction, and justification of SVMs in this context constitutes our fourth and last contribution in part-I of the thesis.

In part-II of the thesis we investigate degradation models based on nonparametric models. In chapter 9 we analyze the reliability of a product from a health monitoring perspective in the context of PHM. In the absence of failure training-data, anomaly detection is approached through a one-class learning algorithm based on SVM classification. This is also used in a Bayesian framework to estimate the posterior class probabilities of test data with unknown class. In this work we make a contribution to the field of reliability by interpreting *health* as the outcome of a classifier. We introduce a methodology that connects machine-learning analysis to FHT models by helping determine suitable marker variables that can be used to track latent degradation. We solve a novelty detection problem with a one-class classification algorithm, and a Bayesian framework for uncertainty analysis. The results of our Bayesian classifier, we argue and show through case studies, are suited for further trending and event-time predictions.

In chapter 11 we consider degradation variables that take-on discrete values, and use computer reliability as a working example. Specifically we are interested

in modeling the occurrence of intermittent errors that lead to failure. Here we use a multistate Markov model to develop a methodology to model failure-time data together with event-time data (errors). Failure-times are modeled as the FHT to a fixed failure state in a multistate model. The marker value is observable and represents the cumulative number of error experienced by the device, while the degradation variable is unobservable and therefore unknown. In this approach, inference is based on the process of the marker variable which is assumed correlated to the degradation variable.

Inference on the transition probability matrix is approached using the non-parametric Aalen-Johansen estimator over small time increments spanning the life-time of the device. A reducible absorbing time-homogeneous Markov chain is used to compute the expected first hitting times. A simulation case-study is used to generate failure-times an event times. From the results we observe, that as anticipated, the expected failure-time decreases given an increasing number of errors experienced by a device. This part of the thesis contributes to the literature on degradation models on the following levels: (i) it casts a degradation model as a multistate model, (ii) it accommodates covariates using counting process theory, (iii) it avoids making many parametric assumptions and (iv) it connects the model with PHM.

BIBLIOGRAPHY

[1] Albert, P.S., Longitudinal data analysis (repeated measures) in clinical trials, Statistics in Medicine, Vol. 18, No. 13, 1707–1732, 1999

[2] M. Pecht,Prognostics and Health Management of Electronics, Wiley-Interscience, 2008

[3] A. Dasgupta, D. Barker, M. Pecht, Reliability prediction of electronic packages, Journal of the IES, Vol. 33, No. 3, 36–45, 1990

[4] M. Pecht, A. Dasgupta, D. Barker, C.T. Leonard, The reliability physics approach to failure prediction modelling, Quality and Reliability Engineering International, Vol. 6, No. 4, 267–273, 1990

[5] J.M. Hu, M. Pecht, A. Dasgupta,A probabilistic approach for predicting thermal fatigue life of wire bonding in microelectronics, Journal of Electronic Packaging, Vol. 113, 275–285, 1991

[6] D. Kwon, M.H. Azarian, M. PECHT, Prognostics of Interconnect Degradation using RF Impedance Monitoring and Sequential Probability Ratio Test, International Journal of Performability Engineering, Vol. 6, No. 4, 351–460, 2010

[7] A. Ramakrishnan, M. Pecht, Load characterization during transportation, Microelectronics Reliability, Vol. 44, No. 2, 333–338, 2004

[8] N. Vichare, P. Rodgers, V. Eveloy, M. Pecht, Environment and usage monitoring of electronic products for health assessment and product design, International Journal of Quality Technology and Quantitative Management, VOl. 4, No. 2, 235–250, 2007

[9] M. Pecht, J. Gu, Physics-of-failure-based prognostics for electronic products, Transactions of the Institute of Measurement and Control, Vol. 31, No. 3-4, 309–322, 2009

[10] Dasgupta A., M. Pecht, Material failure mechanisms and damage models, IEEE Transactions on Reliability, Vol. 40, No. 5, 531–536, 1991

[11] N. D. Singpurwalla,Survival in dynamic environments, Statistical Science, Vol. 10, No. 1, 86–103, 1995

[12] K. Sobczyk, Stochastic models for fatigue damage of materials, Advances in applied probability, Vol. 19, No. 3, 652–673, 1987

[13] M. Chown, G.G. Pullum, G.A. Whitmore,Reliability in Communication Technology, Chapman & Hall

[14] J. Lu, Degradation processes and related reliability models, Thesis, 1995

[15] V.N. Nair, Estimation of Reliability in Field-Performance Studies,Technometrics, Vol. 30, No. 4, 379–383, 1988

[16] A. Desmond,Stochastic models of failure in random environments,The Canadian Journal of Statistics/La Revue Canadienne de Statistique, Vol. 13, No. 3, 171–183, 1985

[17] M. Pecht, A prognostics and health management roadmap for information and electronics-rich systems, Microelectronics Reliability, Vol. 50, No. 3, 317–323, 2010

[18] C. K. Lu, W.Q. Meeker, Using degradation measures to estimate a time-to-failure distribution, Technometrics, VOl. 34, No. 2, 161–174, 1993

[19] W. Kahle, Simultaneous confidence regions for the parameters of damage processes, Statistical Papers, Vol. 35, No. 1, 27–41, 1994

[20] , G.A. Whitmore, Estimating degradation by a Wiener diffusion process subject to measurement error, Lifetime data analysis, Vol. 1, No. 3, 307–319, 1995

[21] J. Lu, Degradation processes and related reliability models, Unpublished thesis, McGill University, 1995

[22] N.D. Singpurwalla, Survival in dynamic environments, Statistical Science, Vol. 10, No. 1, 86–103, 1995

[23] K.A. Doksum, S.L.T. Normand, Gaussian models for degradation processes-Part I: Methods for the analysis of biomarker data, Lifetime Data Analysis, Vol. 1, No. 2, 131–144, 1995

[24] G. Whitmore, M. Crowder, J. Lawless, Failure inference from a marker process based on a bivariate Wiener model, Lifetime Data Analysis, Vol. 4, No. 3, 229–251, 1998

[25] L.I. Pettit, K.D.S. Young, Bayesian analysis for inverse Gaussian lifetime data with measures of degradation, Journal of statistical computation and simulation, Vol. 63, No. 3, 217–234, 1999

[26] M. Lee, V. DeGruttola, D. Schoenfeld, A model for markers and latent health status, Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol. 62, No. 4, 747–762, 2000

[27] J. Lawless, M. Crowder, Covariates and random effects in a gamma process model with application to degradation and failure, Lifetime Data Analysis, Vol. 10, No. 3, 213–227, 2004

[28] M.L.T. Lee, G.A. Whitmore, Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary, Statistical Science, Vol. 21, No. 4, 501–513, 2006

[29] J. Tang, T.S. Su Estimating failure time distribution and its parameters based on intermediate data from a Wiener degradation model, Naval Research Logistics, Vol. 55, No. 3, 265–276, 2008

[30] W. Kahle, A. Lehmann, The Wiener Process as a Degradation Model: Modeling and Parameter Estimation, Advances in Degradation Modeling, 127–146, Springer, 2010

[31] S.X.Si, W. Wang, C.H. Hu, D.H. Zhou, Remaining useful life estimation-A review on the statistical data driven approaches, European Journal of Operational Research, Vol. 213, No. 1,1–14, 2010

[32] N.D. Singpurwalla, A New Perspective on Damage Accumulation, Marker Processes, and Weibulls Distribution, Advances in Degradation Modeling, 241–249, Springer, 2010

[33] M.L.T. Lee, G.A. Whitmore, B.A. Rosner, Threshold regression for survival

data with time-varying covariates, Statistics in medicine, Vol. 29, No. 7-8, 896–905, 2010

[34] K. Sobczyk, Stochastic models for fatigue damage of materials, Mathematical and Computer Modelling, Vol. 12, No. 8, 1046–1046, 1989

[35] B. Epstein, M. Soebel Journal of the American Statistical Association, American Statistical Association, Vol. 48, No. 263, 486–502, 1953

[36] N.D. Singpurwalla, Inference from accelerated life tests when observations are obtained from censored samples, Technometrics, Vol. 13, No. 1, 161–170,1971

[37] N.D. Singpurwalla, A problem in accelerated life testing,Journal of the American Statistical Association, Vol. 66, No. 336, 841–845

[38] N.D. Singpurwalla Inference from Accelerated Life Tests Using Arrhenius Type Re-Parameterizations, Technometrics, Vol. 15, No. 2, 289–299, 1973

[39] N.R. Mann, R.E. Schafer, N.D. Singpurwalla, Methods for statistical analysis of reliability and life data, Wiley, 1974

[40] G.K. Bhattacharyya, A. Fries, Inverse Gaussian regression and accelerated life tests, Lecture Notes-Monograph Series, Vol. 2, 101–117, 1982

[41] W. Nelson, Accelerated testing: statistical models, test plans, and data analyses, Wiley, 1990

[42] M.B. Carey, R.H Koenig, Reliability assessment based on accelerated degradation: a case study, IEEE Transactions on Reliability, Vol. 40, No. 5, 499–506, 1991

[43] K.A. Doksum, A. Hóyland, Models for variable-stress accelerated life testing experiments based on Wiener processes and the inverse Gaussian distribution, Technometrics, Vol. 34, No. 1, 74–82, 1992

[44] W.Q. Meeker, L.A. Escobar, A review of recent research and current issues in accelerated testing, International Statistical Review/Revue Internationale de Statistique, Vol. 61, No. 1, 147–168, 1993

[45] G.A. Whitmore, F. Schenkelberg, Modelling accelerated degradation data using Wiener diffusion with a time scale transformation, Lifetime Data Analysis, Vol. 3, No. 1, 27–45, 1997

[46] J.C. Lu, J. Park,Q. Yang, Statistical inference of a time-to-failure distribution derived from linear degradation data, Technometrics, Vol. 39, No. 4, 391–400, 1997

[47] W.Q. Meeker, L.A. Escobar, C.J. Lu, Accelerated degradation tests: modeling and analysis, Technometrics, Vol. 40, No. 2, 89–99, 1998

[48] W.J. Owen, W.J. Padgett, Accelerated test models for system strength based on Birnbaum-Saunders distributions,Lifetime Data Analysis, Vol. 5, No. 2, 133–147, 1999

[49] A. Onar, W.J. Padgett, Accelerated test models with the inverse Gaussian distribution, Journal of statistical planning and inference, Vol. 89, No. 1-2, 119–133, 2000

[50] V. Bagdonavicius, M.S. Nikulin, Estimation in degradation models with explanatory variables, Lifetime Data Analysis, Vol. 7, No. 1, 85–103, 2001

[51] V. Bagdonavicius, O. Cheminade, M. Nikulin, Statistical planning and inference in accelerated life testing using the CHSS model, Journal of statistical planning and inference, Vol. 126, No. 2, 535–551, 2004

[52] W.J. Padgett, M.A. Tomlinson, Inference from accelerated degradation and failure data based on Gaussian process models, Lifetime Data Analysis, Vol. 10, No. 2, 191–206, 2004

[53] C. Park, W.J. Padgett, Accelerated degradation models for failure based on geometric Brownian motion and gamma processes, Lifetime Data Analysis, Vol. 11, No. 4, 511-527, 2005

[54] C. Park, W.J. Padgett,Stochastic degradation models with several accelerating variables, IEEE Transactions on Reliability, Vol. 55, No. 2, 379–390, 2006

[55] S.J. Bae, W. Kuo, P.H. Kvam, Degradation models and implied lifetime distributions,Reliability Engineering & System Safety, Vol. 92, No. 5, 601–608, 2007

[56] W.Q. Meeker, L.A. Escobar, Y. Hong, Using accelerated life tests results to predict product field reliability, Technometrics, Vol. 51, No. 2, 146–161, 2009

[57] M. Pecht, P. McCluskey, J. Evans, Failures in Electronic Assemblies and Devices, Product Integrity and Reliability in Design, Springer-Verlag 2001, 204–232

[58] D. Kwon, M. Azarian, M. Pecht, Prognostics of Interconnect Degradation using RF Impedance Monitoring and Sequential Probability Ratio Test, Vol. 6, No. 4, 351–460, 2010

[59] S. Han, M. Osterman, M. Pecht, Electrical Shorting Propensity of Tin Whiskers, IEEE Transactions on Electronics Packaging Manufacturing, Vol. 33, No. 3, 205–211

[60] N.Patil, J. Celaya, D. Das, K. Goebel, M. Pecht, Precursor parameter identification for insulated gate bipolar transistor (igbt) prognostics, IEEE Transactions on Reliability, Vol. 58, No. 2, 271–276, 2009

[61] G. Casella, R.L. Berger, Statistical Inference, Duxburry Press, 2001

[62] W.H. Press, B.P. Flannery, S.A.Teukolsky, W.T. Vetterling, Numerical Recipes, Cambridge University Press, 1992

[63] , L.B. Koralov, Y.G. Sinai, Theory of probability and random processes, Springer Verlag, 2007

[64] A.F. Karr, Probability, Springer, 1993

[65] M.L. Eaton, Multivariate statistics: a vector space approach, Wiley New York, 1983

[66] , N. V. Krylov, Introduction to the theory of diffusion processes, American Mathematical Society, 1995

[67] N. V. Krylov, Introduction to the theory of random processes, American Mathematical Society, Vol. 43, 2002

[68] , R.S. Chhikara, J.L. Folks, The inverse Gaussian distribution as a lifetime model,Technometrics, Vol. 19, No 4. 461–468, 1977

[69] N. Vichare, M. Pecht, Prognostics and Health Management of Electronics, EEE

Transactions on Components and Packaging Technologies, Vol. 29, No. 1, 1521–3331, 2006

[70] M. Bilodeau, D. Bremmer, Theory of multivariate statistics, Springer Verlag, 1999

[71] T. Amemiya, Introduction to statistics and econometrics, Harvard University Press, 1994

[72] X.K. Song, Correlated data analysis: modeling, analytics, and applications, Springer, 2007

[73] N.M. Laird, J.H. Ware, Random-effects models for longitudinal data, Biometrics Vol.38, 963–974, 1982

[74] J.P. Klein, M.L. Moeschberger, Survival analysis: techniques for censored and truncated data, Springer Verlag, 2003

[75] C.E. Rasmussen, C.K.I Williams,Gaussian processes in machine learning, THe MIT Press, 2006

[76] M. Pecht, D. Das, and A. Ramakrishnan, "The IEEE standards on reliability program and reliability prediction methods for electronic equipment", *Microelectronics Reliability*, vol. 42, no. 9 - 11, pp. 1259–1266, 2002

[77] N. Vichare and M. Pecht, "Prognostics and health management of electronics", *IEEE Transactions on Components and Packaging Technologies*, vol. 29, no. 1, pp. 222–229, 2006

[78] A. Ramakrishnan and M. Pech, "A life consumption monitoring methodol-

ogy for electronic systems", *IEEE Transactions on Components and Packaging technologies*, vol. 26, no. 3, pp. 625–634, 2003

[79] K. Feldman, T. Jazouli, and P. Sandborn, "A Methodology for Determining the Return on Investment Associated with Prognostics and Health Management", *IEEE Transactions on Reliability*, nol. 58, no. 2, pp. 305–316, 2009

[80] N. Vichare, P. Rodgers, and V. Eveloy, and M. Pecht, "In Situ Temperature Measurement of a Notebook ComputerA Case Study in Health and Usage Monitoring of Electronics", *IEEE Transactions on Device and Materials Reliability*,vol. 4, no. 4, pp. 658–663, 2004

[81] T. Stibor, P. Mohr, and J. Timmis, and C. Eckert, "Is negative selection appropriate for anomaly detection?", in *Proceedings of the 2005 conference on Genetic and evolutionary computation*, ACM, pp. 321–328, 2005

[82] S. Kumar, V. Sotiris, and M. Pecht, "Health Assessment of Electronic Products using Mahalanobis Distance and Projection Pursuit Analysis", *International Journal of Computer, Information, and Systems Science, and Engineering*, vol. 2, no. 4, pp. 242–250, 2008

[83] D. Tax and P. Juszczak, "Kernel whitening for one-class classification", *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 17, no. 3, pp. 333–347, 2003

[84] H. Wang, and Z. Song, and P. Li, "Fault detection behavior and performance analysis of principal component analysis based process monitoring methods", *Ind. Eng. Chem. Res*, vol. 41, no. 10, pp. 2455–2464, 2002

[85] J. Jackson and G. Mudholkar, "Control procedures for residuals associated with principal component analysis", *Technometrics*, vol. 21, no. 3, pp. 341–349, 1979

[86] V. Klema and A. Laub, "The singular value decomposition: Its computation and some applications", *IEEE Transactions on Automatic Control*, vol. 25, no. 2, pp. 164–176, 1980

[87] L. Ruixin, W. Dongfeng, and H. Pu *et al.*, "On the applications of SVD in fault diagnosis", in *IEEE International Conference on Systems, Man and Cybernetics*, vol 4, no. 5, pp. 3763–3768, 2003

[88] V. Vapnik, "The nature of statistical learning theory", *Springer Verlag*, 2000

[89] V. Vapnik, "Statistical learning theory", Wiley, New York, 1998

[90] C. Burges, "A tutorial on support vector machines for pattern recognition",*Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998

[91] M. Seeger, "Relationships between Gaussian processes, support vector machines and smoothing splines", *Machine Learning*, 2000

[92] K. Bennett and E. Bredensteiner, "Duality and geometry in SVM classifiers",in *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE*, Citeseer, pp. 57–64, 2000

[93] C. Hsu, C. Chang, and C. Lin *et al.*, "A practical guide to support vector classification", Citeseer, 2000

[94] Y. Grandvalet, J. Mariethoz, and S. Bengio, "A probabilistic interpretation

of SVMs with an application to unbalanced classification",*Advances in Neural Information Processing Systems*, vol. 18, pp. 467–474, 2006

[95] J. Kwok, "The evidence framework applied to support vector machines", *IEEE Transactions on Neural Networks*, vol. 11, no. 5, pp. 1162–1173, 2000

[96] D. MacKay, "The evidence framework applied to classification networks", *Neural Computation*, vol. 4, no. 5, pp. 720–736, 1992

[97] J. Kwok, "Moderating the outputs of support vector machine classifiers", *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1018–1031, 1999

[98] I. Steinwart, D. Hush, and C. Scovel, "A classification framework for anomaly detection", *Journal of Machine Learning Research*, vol. 6, no. 1, pp. 211–232, 2006

[99] W. Fan, M. Miller, S. Stolfo, W. Lee, W. and P. Chan, "Using artificial anomalies to detect unknown and known network intrusions",*Knowledge and Information Systems*, vol. 6, no. 5, pp. 507–527, 2004

[100] F. González and D. Dasgupta, "Anomaly detection using real-valued negative selection",*Genetic Programming and Evolvable Machines*,vol. 4, no. 4, pp. 383–403, 2003

[101] H. Yu, J. Han, and K. Chang, "PEBL: Web page classification without negative examples",*IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 1, pp. 70–81, 2004

[102] J. Theiler and D. Cai, "Resampling approach for anomaly detection in multispectral images",in *Proceedings of SPIE*, vol. 5093, pp. 230–240, 2003

[103] S. Marsland, ”Novelty detection in learning systems”, *Neural computing surveys*, vol. 3, pp. 157–195, 2003

[104] M. Markou and S. Singh, ”Novelty detection: a reviewpart 1: statistical approaches”,*Signal Processing*, vol. 83, no. 12, pp. 2481–2497, 2003

[105] B. Scholkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson, ”Estimating the support of a high-dimensional distribution”, *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001

[106] L. Manevitz and M. Yousef, ”One-class svms for document classification”,*The Journal of Machine Learning Research*, vol. 2, pp. 139–154, 2002

[107] A. Bánhalmi, A. Kocsor, and R. Busa-Fekete, ”Counter-Example Generation-Based One-Class Classification”, *Lecture Notes in Computer Science*, no. 4701, pp. 543–550, 2007

[108] M. Davenport, R. Baraniuk, and C. Scott, ”Learning minimum volume sets with support vector machines”, in *Proc. IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)*, Citeseer

[109] J. Nuñez Garcia, Z. Kutalik, K. Cho, and O. Wolkenhauer, ”Level sets and minimum volume sets of probability density functions”, *International journal of approximate reasoning*, vol. 34, no. 1, pp. 25–47, 2003

[110] W. Polonik and Q. Yao, ”Conditional Minimum Volume Predictive Regions for Stochastic Processes”, *Journal of the American Statistical Association*, vol. 95, no. 450, pp. 509–519, 2000

[111] P. Moreno, P. Ho, and N. Vasconcelos, ”A Kullback-Leibler divergence based

kernel for SVM classification in multimedia applications",*Advances in Neural Information Processing Systems*, vol. 16, 2004

[112] J. Gao and P. Tan, "Converting output scores from outlier detection algorithms into probability estimates", in *ICDM06: Proceedings of the Sixth International Conference on Data Mining*, pp. 212–221

[113] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods", *Advances in large margin classifiers*, pp. 61–74, 1999

[114] W. Chu, S. Keerthi, and C. Ong, "A new Bayesian design method for support vector classification", in *Special Section on Support Vector Machines of the 9th International Conf. on Neural Information Processing*, Citeseer, pp. 888–892, 2002

[115] Sotiris, V.A. and Tse, P.W. and Pecht, M.G., "Anomaly Detection Through a Bayesian Support Vector Machine",Reliability, IEEE Transactions on, Vol. 59,no. 2,pp. 277–286,2010

[116] Doksum, K.A. and Hóyland, A., "Models for variable-stress accelerated life testing experiments based on Wiener processes and the inverse Gaussian distribution", Technometrics,no. 1, 74–82,1992

[117] Doksum, K.A. and Normand, S.L.T., "Gaussian models for degradation processes-Part I: Methods for the analysis of biomarker data",Lifetime Data Analysis, Vol. 1,no. 2, pp. 131–144,1995

[118] Whitmore, GA, "Estimating degradation by a Wiener diffusion process subject to measurement error", Lifetime data analysis, Vol. 1, no. 3, pp. 307–319,1995

[119] Satten, G.A. and Longini Jr, I.M., "Markov Chains With Measurement Error: Estimating theTrue'Course of a Marker of the Progression of Human Immunodeficiency Virus Disease", Applied Statistics, vol. 45,no. 3,pp. 275–309, 1996

[120] Henderson, R. and Diggle, P. and Dobson, A., "Joint modelling of longitudinal measurements and event time data", Biostatistics, vol. 1, no. 4,pp. 465–480, 2000

[121] Meeker, W.Q. and Escobar, L.A., "Statistical methods for reliability data", Wiley New York, 1998

[122] Commenges, D., "Multi-state models in epidemiology", Lifetime data analysis, vol. 5, no. 4,pp. 315–327, 1999

[123] Commenges, D., "Inference for multi-state models from interval-censored data", Statistical Methods in Medical Research, vol. 11, no. 2, pp. 167–182, 2002

[124] Bagdonaviius, V. and Bikelis, A. and Kazakeviius, V. and Nikulin, M., "Nonparametric estimation from simultaneous degradation and failure time data", Comptes Rendus Mathematique, vol. 335, no. 2, pp. 183–188, 2002

[125] Putter, H. and van der Hage, J. and de Bock, G.H. and Elgalta, R. and van de Velde, C.J.H., "Estimation and Prediction in a Multi-State Model for Breast Cancer",Biometrical journal, vol. 48,no. 3, pp. 366–380, 2006

[126] Peña, E.A., "Dynamic modelling and statistical analysis of event times", Statistical science: a review journal of the Institute of Mathematical Statistics, vol. 21,no. 4, pp. 487–500, 2006

[127] Meira-Machado, L. and de Uña-Álvarez, J. and Cadarso-Suárez, C. and Andersen, P.K., "Multi-state models for the analysis of time-to-event data",Statistical methods in medical research, vol. 18, no. 2, pp. 1–28,2009

[128] L. M. Machado, C. C. Suarez and J. Alvarez, "Inference in the progressive three-state model", International Journal of Mathematical Models and Methods in Applied Sciences, vol. 2, No. 3, pp. 447- 454, 2008

[129] Andersen, P.K. and Pohar Perme, M., "Inference for outcome probabilities in multi-state models", Lifetime data analysis, vol. 14, no. 4, pp. 405–431, 2008

[130] Cook, R.J. and Lawless, J.F. and Lakhal-Chaieb, L. and Lee, K.A., "Robust estimation of mean functions and treatment effects for recurrent events under event-dependent censoring and termination: application to skeletal complications in cancer metastatic to bone", Journal of the American Statistical Association, vol. 104, no. 485, pp. 60–75, 2009

[131] Aalen, O.O. and Borgan, Ø. and Gjessing, H.K., "Survival and event history analysis: a process point of view", Springer Verlag, 2008

[132] R. J. Cook, J. F. Lawless, "The Statistical Analysis of Recurrent Events". Springer, New York, 2007

[133] P. K. Andersen, O. Borgan, R. D. Gill, N. Keiding, "Statistical Models Based on Counting Processes", Springer, New York, 1993

[134] Aalen, O.O. and Johansen, S., "An Empirical Matrix for Non-Homogeneous Markov Chains Based on Censored Observations", Scandinavian Journal of Statistics, Vol. 5, No. 3, pp. 141–150, 1978

[135] Brémaud, P., "Markov chains: Gibbs fields, Monte Carlo simulation, and queues", Springer, 1999

[136] Singpurwalla, N.D., "On competing risk and degradation processes", Lecture Notes-Monograph Series, vol. 49, pp. 229–240, 1996

[137] Bagdonavicius, V. and Masiulaityte, I. and Nikulin, MS, "Reliability Estimation from Failure-Degradation Data with Covariates", Advances in Degradation Modeling: Applications to Reliability, Survival Analysis, and Finance, Springer, 2009