November 2022

Statistical Methods for Reliability Test planning and Data Analysis

Oluwaseun Elizabeth Otunuga
University of South Florida

Statistical Methods for Reliability Test Planning and Data Analysis

by

Oluwaseun Elizabeth Otunuga

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
with a concentration in Statistics
Department of Mathematics & Statistics
College of Arts and Sciences
University of South Florida

Major Professor: Lu Lu, Ph.D.
Li Mingyang, Ph.D., Ph.D.
Kandethody Ramachandran
Getachew Dagne, Ph.D.

Date of Approval:
November 8, 2022

Keywords: Weibull Test Plan; Multivariate Degradation; Pareto Front; Accelerated Degradation Test, Cox Partial Log Likelihood; Variable Selection.

Dedication

This work is dedicated to God almighty and to my family; my husband, my mother, my siblings and friends who extended their help and support while doing this work.

Acknowledgments

Foremost, I would like to express my deepest gratitude to God Almighty for His love and protection over my life. Where I am today is not by my power but by His grace and mercy. Truely, all His promises are YES and AMEN. Thank you Lord.

Words cannot express my gratitude to my advisor Dr. Lu Lu for her continuous support, advise, patience, motivation, enthusiasm, and immense knowledge throughout my program. Her guidance helped me during my program and writing of this dissertation. I could not have imagined having a better advisor and mentor for my Ph.D. program. Besides my advisor, I would like to extend my sincere thanks to the chair of my committee, Dr. Mingyang Li for his invaluable patience and feedback. I also could not have undertaken this journey without my defense committee, Dr. Ramachandran Kandethody, Dr. Getachew Dagne for their encouragement, valuable comments and suggestions.

Very special thanks to my husband, Abiodun Idowu, for his support and assistance in every step of my dissertation. Thank you so much dear for showing me the beauty and opposite side of life and making my mind relax during the hard times of my dissertation. I would like to thank my family: my mother, Mrs. Esther Otunuga, for loving and supporting me spiritually through prayer throughout my life. Thank you so much mummy, I pray you will eat the fruit of your labour in peace and good health. My special thanks also goes to my siblings: Mr. and Mrs. Emmanuel Oluwatosin Otunuga, Dr. and Dr. Mrs Michael Otunuga, Mrs. Dahunsi Aanuoluwapo Otunuga (my one and only sister) and Mr. Dahunsi. Thank you all for your support, advise and care. Many thanks to my niece and nephews: Grace Otunuga, Gabriel Otunuga, Emmanuel Otunuga, Inioluwa Dahunsi,

Caleb Otunuga, and Mabel Otunuga for their love and prayer. My thanks is not complete without mentioning my husband's family. Thank you all so much for loving me like your child. May God continue to be with us all. Amen.

Table of Contents

List of Tables

List of Figures

Abstract

This dissertation develops several statistical methods to advance the techniques and applications in the fields of reliability test planning and data analysis as well as statistical modeling and analysis in survival analysis.

The first project focuses on developing new demonstration test plans for lifetime data based on considering multiple objectives. Reliability demonstration tests have been broadly used for assuring reliability performance at the desired confidence level. We consider lifetime data that follows a Weibull distribution which has been broadly used for modeling a variety of shapes of lifetime distributions. When planning a demonstration test, there are often multiple aspects to be considered including the consumer's risk, the producer's risk, the acceptance probability, and the cost. The natural trade-offs between these objectives require a careful evaluation of their interrelationship with the planning parameters and a systematic approach to making a tailored decision. We propose a Pareto front optimization approach for balancing the multiple objectives and offer a set of graphical and numerical tools for comparing solutions and selecting the best test plan to match different users' priorities.

The second project focuses on advancing the statistical modeling and analysis of accelerated degradation test (ADT) data with interdependent multiple degradation measures. In some ADTs, to assess and understand the different aspects of reliability performance, multiple characteristics of how the product degrade are measured, which are often interdependent within individual test units. A nonlinear multivariate general path model with random effects and covariates was developed to capture the variation in the indi-

vidual degradation paths from unit-to-unit while allowing to model the interdependence among the multiple degradation measurements and also capture the correlation between the initial degradation condition and the degradation rate. A full Bayesian approach for estimation and inferential analysis is demonstrated. The method is evaluated and compared to the stage-of-art practices via a simulation study and is also illustrated using synthetic optical media ADT data from ISO [3].

The third project focuses on advancing the use of penalized regression based on Cox Proportional Hazard models in survival analysis. It is a common challenge in the field of reliability and survival analysis to select a subset of key variables for accurate estimation and prediction of the reliability or survival experience when there are small data with a large number of predictor variables. Penalized Regression models work well for effective variable selection and reduce the complexity of the model. A new penalized regression model based on the Cox partial likelihood and a modified minimax concave penalty is proposed. The performance of the proposed penalized regression model compared with existing methods is demonstrated through a simulation study and its application is illustrated via two real-world examples for analyzing the heart failure data and the NKI breast cancer data.

Chapter 1: Introduction

1.1 Reliability Demonstration Test Plan

To deal with the special type of time-to-event random variables such as failure time, lifetime, survival time, etc, specialized fields of mathematical statistics which are reliability and survival analysis are developed [76]. The basis of reliability analysis is to model the lifetime by a suitable probability distribution and through the selected distribution, the life behavior will be characterized.

According to [32], with reliability analysis, we can answer questions, such as:

- What is the percentage of items that will last longer over a certain time?
- What is the probability that a test unit will fail before a given time?
- What is the expected lifetime of a component?

Reliability is the probability that a system will perform it's required performance at a certain time-point. Reliability of a product is a desirable property of great interest to both manufacturers and consumers. It is the performance of an item under specified environmental and operational conditions and over a given period of time [1]. The designing of a good product should be based on what customers want (because customers have higher expectations for more "reliable" product) and because of this, engineers have been prompted to conduct extensive testing before the release of products (or services). Indeed, testing represents a significant portion of the total product cost.

Reliability demonstration tests have been broadly used for assuring reliability performance at the desired confidence level. RDT is a test used to validate the design of a product to know whether a certain reliability requirement is met at a given time with a stated confidence level. In Reliability demonstration, products are tested under certain design to show whether their reliability reach a pre-specified threshold, and many companies and industries have used reliability demonstration tests to make decisions on the design of their products.

1.2 Access to Reliability Using Degradation Data

Another area in reliability is reliability analysis based on degradation data. Many researchers have access the reliability of a product using degradation data and this has become a significant approach to evaluate the reliability and safety of critical systems. Degradation data provide measurements on the physical degradation of the products or systems, which could offer more direct information on the failure mechanism than the lifetime data. Looking at a highly reliable product's reliability, it is often dominated by its performance degradation process in many engineering situations. Many industries and companies that produce modern products with many components (where most of these products are designed to last for a long time) are facing a lot of demand from the consumers. Some of these are the demand for excellent quality and high reliability of these products. This is a big challenge to the producer because testing these products under the normal operating environments to obtain sufficient time-to-failure data is hard [109]. To meet the consumer's demands and for high reliability products, strategies are needed to accelerate data acquisition to access the reliability of these products. So, using degradation data itself may not be sufficient for reliability analysis because some of the degradation process happen slowly over time, hence we are adding acceleration on top of degradation in this work. The extraction can be done by accelerating the extraction

of the failure information in an efficient way, for example, by exposing the test specimens to severe-than-normal conditions to accelerate the failure process [[10], [52], and [96]]. We can trace many of these failure mechanisms back to an underlying degradation process such as corrosion, crack growth, cumulative wear, fatigue, etc [43]. To access reliability for highly reliable products, Accelerated Life Tests (ALTs) and Accelerated Degradation Tests (ADTs) have been commonly used. Our work focus on using ADT. To model the degradation path and predict failure and/or assess reliability under the normal use conditions, ADTs is used to measure the degradation of products under the accelerated conditions.

1.3 Variable Selection For Survival Analysis

Looking at the second field of mathematical statistics that was mentioned above. Survival analysis is a branch of statistics which incorporate statistical methods to analyze survival data where the time that an event among living organisms occur is the outcome of the variable. The occurrence of disease, death, recovery from disease, etc can be the event. We can measure the survival time or the failure time in days, weeks etc. With survival analysis, these following questions, can be answered:

- What proportion or percentage of a population will survive longer than a certain time?
- What is the rate that those who survive will fail or die before a given time?
- Can a multiple causes of death or failure be considered?

It is a common challenge in the field of reliability and survival analysis to select a subset of key variables for accurate estimation and prediction of the reliability or survival experience when there are large number of predictor variables. Selecting a subset of key variables for estimating the response of interest is crucial for understanding the underlying input-response relationship and producing precise prediction of the response. It is necessary to improve the interpretability of a statistical model and to minimize variability

of predictions. Variable selection is vital to survival analysis and various variable selection criteria and procedures for linear regression models have been proposed by many authors [16]. [39] discussed about an overview on variable selection for survival analysis. In practice, a lot of covariates are often available as potential risk factors because a large number of predictors are usually introduced at the initial stage of modeling by the data analyst. Much attention has been received in the selection of variables for survival data analysis in the recent literature because of its complicated data structure that poses many challenges.

Chapter 2: Review

2.1 Reliability Demonstration Test Methods

Different researchers have applied reliability demonstration test methods and there are different reliability types of data. For example pass/fail data, Degradation test data, Lifetime/failure time data and Accelerated lifetime data. Because there are different reliability data, we are developing different test plan to accommodate different types of data. [44]. Binomial RDTs based on pass/fail data are more broadly studied and zero-failure tests which is a pass/fail data have been popular due to cost consideration. When planning an RDT, there are often multiple aspects to be considered including the cost, the acceptance rate of the test, the producer's risk, and the consumer's risk.

To determine the sample size and the test duration for an RDT, different method have been studied extensively. Some of which are in [5], [26], [81], [11], [22]. Degradation testing is more efficient and informative than the zero-failure testing, as shown in [82]. Many test plans have been developed by using various optimal degradation. These test plans choose measurement frequency, the sample size, and test termination time to estimate the product reliability [101], [100]. With these test plans, the statistical error of an estimate subject to a cost constraint, and the require samples to use until the best time to terminate is reached is being minimized. To assess the reliability and also to predict the remnant life of systems, more information are often provided by using degradation data than using failure time data.

With some existing methods, there are two main types of statistical test plan to help determine the sample size and the test duration for a reliability demonstration test (RDT):

- test designs based on fixed time /failure time [100], [99], [101], and
- test designs based on the number of failures [5], [26], [81], [11], [22], [74].

For some applications, it was shown that RDT designs based on the number of failures which is also called a binomial test are easy to use and this makes it popular.

On the other hand, an underlying failure time distribution and its parameters are used in the failure time method as ”planning information”. The RDT is then designed according to the precision requirements for the estimated parameters or their functions.

2.2 Types of Accelerated Tests

Accelerated Life Tests (ALTs) measure the failure time of the products by putting them under more stressful or harsher use conditions. Lifetime models are built under the accelerated conditions which are then used to extrapolate and predict failure time and reliability under the normal use conditions. The ALTs have been applied to provide timely assessment of reliability for materials, components, and subsystems [15]. Trevisanello [42] uses ALTs for high brightness light emitting diodes (LED).

The second type is the Accelerated Degradation Tests (ADTs) [64]. These tests measure the degradation of products under the accelerated conditions which are then used to model the degradation path and predict failure and/or assess reliability under the normal use conditions. To ensure an effective assessment of the reliability of a product, ADTs have been used to shorten the samples needed, reduce the duration of the test, and it provide sufficient data. This is a method used to extrapolate the lifetime of highly reliable products under normal use conditions. Also, we have seen where accelerated degrada-

tion testing (ADT) have been used as an effective tool to verify the reliability, evaluate lifetime modern products, and collect the degradation data by exposing the test specimens to severe conditions.

Many scientists have applied ADT to extrapolate the lifetime of some highly reliable products and some of which are Nelson [123], who describes the basic information on accelerated degradation models by reviewing degradation survey applications and literature. [35] applied ADT to the reliability analysis of batteries, [43] applied ADT to the super luminescent diode (SLD), and [7] applied ADT to the smart electricity meter, etc. An example of a device designed to last long is the Optical media. This is because within a test period of time and under normal use conditions, it is not possible to observe sufficient failure data. The accelerated degradation test (ADT) was used by [3] to estimate and predict the lifetime of optical media. It is also possible to apply the mixed-effects ADT approach to the failure analysis of optical disk media. Degradation models are either driven by data or derived from physical principles via stochastic processes. The data-driven models are generally used to analyze degradation data. In some ADTs, to assess and understand the different aspects of reliability performance, various characteristics of how the product degrade are measured, which are often interdependent within individual test units.

Our motivation in modeling degradation data that has many degradation characteristics (DCs) is to determine how long the information stored on a recordable optical disc is going to last and this is inspired by using the degradation data that is obtained from an ADT experiment.

2.3 Penalized Regression

In the selection of variables as discussed in the previous chapter, a major drawback of the best subset variable selection is its lack of stability as analyzed, by Breiman in [72]. Some penalized regression works by setting some coefficient to be equal to zero and then select the remaining variables e.g., least absolute shrinkage and selection operator (LASSO) penalty or regression [[61], [13], [114]]. [106] applied LASSO to cox model.

Applying some other penalties tends to result in all small but non-zero regression coefficients e.g., ridge penalty or regression. Ridge regression was proposed by Hoerl and Kennard in [31] and is an estimation procedure which is based on adding small positive quantities to the diagonal of $X^T X$ to obtain a point estimate with a smaller mean square error and also to help overcome many of the difficulties (biasness) associated with the usual least squares estimates.

In statistics, we have two critical characteristics of estimators that are to be considered. They are the variance and the bias. The bias is the difference between the expected estimator and the true population parameter. It measures the inaccuracy of the estimates. The variance measures the spread between them. If the variance and the bias are too large, i.e., if there are many predictive features in the model, it can harm the model's predictive performance.

Ridge penalty being a continuous shrinkage method achieves better predictive performance through a bias-variance trade-off. The ideal goal of ridge regression in terms of bias and variance is to obtain a low bias and a low variance. This is near difficult or impossible to achieve. Therefore, the need of the trade-off. This bias-variance trade-off favors ridge over LASSO if there is high correlation between predictors [105]. However, as ridge produces coefficient values for each of the predictor variables, it does not perform

variable selection and hence cannot produce a more parsimonious model (model with good explanatory predictive power), [115]. Ridge regression was introduced to solve the multicollinearity in multiple regression and ridge regression problem in multicollinear data was investigated by [46]. [110] applied ridge to logistic regression. Using three well-known microarray gene expression data sets, Bθvelstad [6] applied ridge to Cox's model to compare the prediction performance.

Tibshirani in [105] proposed a new method which is the LASSO for linear regression. LASSO penalty reduces the residual sum of squares subject to the condition that the sum of the absolute value of the coefficients being less than a constant. It then produces some coefficients that are 0 (this is done by forcing the sum of the absolute value of the regression coefficients to be less than or equal to a fixed value (λ) i.e., $|z_j| \leq \lambda$) and also interpretable models. LASSO has also been applied to generalized regression models [105]. The aim of LASSO regression is to identify the variables and corresponding regression coefficients that minimizes the prediction error in a model. This is done by imposing a constraint on the model parameters, which 'shrinks' the regression coefficients towards zero. In a practical sense this reduces the complexity of the model. [[117], [68]] applied LASSO to Cox's proportional hazards models commonly used for survival analysis. This proposed model minimizes the log partial likelihood subject to the constraint that sum of the absolute values of the parameters being bounded by a constant.

The disadvantage of using the LASSO approach is that it does not focuses on the accuracy of the estimation and interpretation of the contribution of individual variables but rather, it focuses on the best combined prediction. Due to this, the interpretation of the regression coefficients may not be reliable in terms of independent risk factors because if we have two or more highly collinear variables, LASSO will select them randomly and this is not good for interpretation [68]. Another disadvantage of LASSO is that with a

large regression coefficient, LASSO has large bias when shrinking the coefficient toward 0. This implies that LASSO is not a very satisfactory variable selection method if the number of variables is greater than the number of observations and this is because, at most n predictors will be picked by LASSO as non-zero even if all predictors are relevant in the model. Due to the bias-variance trade-off of ridge when correlation is low and LASSO not being a very satisfactory variable selection method if the number of variables is greater than the number of observations, elastic-net was proposed by Zou and Hastie [56].

Elastic net is a combination of ridge and LASSO penalty, and hence is also a variable selection method. Using real world dataset and through simulation, it has been shown that the elastic net often outperforms the LASSO (because it has both property of ridge and LASSO i.e., because of its effective shrinkage of coefficients like ridge and because of how it set coefficient to zero like LASSO) while enjoying a similar sparsity of representation. In addition, the elastic net encourages a grouping effect, in which strongly correlated predictors will have similar estimated coefficients. This unique effect was inherited from ridge regression. This penalty is useful when the number of predictors (p) is larger than the number of observations (n). Liu and Li [121] applied elastic net to regression analysis for spectrum data. Wu [129] developed a solution path algorithm using the least angle regression (LAR) with the elastic net penalty in Cox's proportional hazards model. This was done in two steps: In the first step, the LAR was extended to optimize the log partial likelihood plus a fixed small ridge term. After this, the path modification was defined and this leads to the solution path of the elastic net regularized log partial likelihood.

In Chapter 3 an extensive discussion about demonstration test plan will be done, and in Chapter 4, a more thorough explanation and application on how to analyze degradation data with multiple Degradation characteristics (Dc) will be done. In Chapter 5, a new penalized log partial likelihood was developed.

Chapter 3: Demonstration Test Plans For Lifetime Data Based on Considering Multiple Objectives

3.1 Introduction

Some existing method focus on zero failure test which is widely used in industry [44], [69], [122], [66], [126], [65] and the advantage of this test is that it minimizes cost. For this test, we will pass the test only if we have no failure. i.e. if the maximum allowable failure is 0. With the zero-failure test plan, there is a strong trade-off between the producer's risk and the consumer's risk because this test allow us to choose a minimum test unit in order to decrease the CR so as to ensure an acceptable CR. According to [80], the disadvantage of the zero-failure test is that while trying to improve the consumer's risk with minimal sample size, we could dramatically increase the producer's risk as well as reducing the passing rate of the demonstration test. That means it can result in unacceptable: High PR and low acceptance probability (AP): probability of passing the test. It's advantage is that it uses Minimum cost for testing.

Even if the design and structure of a system are the same, the time to failure or life-time of that system will vary from system to the next system. The Life test methods for reliability demonstration include the conventional life test and the sequential life test. The sequential life test method [20, 28, 37, 67, 88] reduces sample size at the expense of test time by varying the sample size and test duration because one sample is tested at a time [44]. This implies that it only needs fewer samples during testing and the hypothesis about the product reliability is then evaluated. This is done to find a test plan that will meet the precision requirement on estimate of an interested reliability metric. When

the samples are tested one at a time by trial-and-error efforts, the test duration of the sequential procedure may be longer. Testing for a longer duration and also testing more test units will allow more information to be gathered about the underlying failure time distribution. With this, the reliability estimate will be more precise.

The conventional life test method [14, 36, 47, 88, 118] is usually applied by the commercial industry and they tests some or all samples to failure. The sample size is determined in advance by conventional life test procedures, which generally require either many samples or a long test period. The hypothesis about the product reliability is then evaluated at the end of the test, and decisions are made about accepting or rejecting it. This test method is ineffective; It takes a long time to complete the test, or it require a large sample size, or both.

To estimate the reliability at the required time, the life data is employed, and using this estimated reliability, the confidence bounds will be estimated. After the estimation, we can then say the designed reliability is successfully demonstrated if the lower bound exceeds the reliability requirement. This test method is more informative than the zero-failure test because it can provide estimates of the reliability. In spite of this, the test may be too costly since units are run to failure at a reasonable rate.

Reliability demonstration can also be accomplished through degradation testing. Having a specified threshold value and having some products whose performance characteristics degrade over time, then these products will fail when their performance characteristic reaches the threshold value. Testing these products allows measurement of performance characteristics at different times. Measuring the product's reliability contains credible information, which is commonly used to estimate it. [44].

3.2 Weibull Demonstration Test Plan

The Weibull distribution is the most popular parametric alternative to the exponential distribution in reliability applications and it is used to model the behavior of a product with time. Binomial test considers attribute test data that capture the survival or non-survival of each of the test device, but the Weibull test focuses on lifetime data. An effective way to gain insight into your product's lifetime performance is to conduct a Weibull Analysis. For as few as two or three failures, Weibull analysis with the use of Weibull works exceptionally well and this is critical when there are severe financial consequences to failures and when prior engineering knowledge is sufficient. Weibull test plan is a plan in which assumptions about Weibull distribution is made in other to incorporate the information about the failure time. The efficiency of a Weibull demonstration test plan is completely determined by the experimental time, which depends on the unknown sample size and on the Weibull shape parameter.

As discussed earlier, with the zero-failure test plan, there is a strong trade-off between the producer's risk and the consumer's risk because it focuses more on reducing consumer's risk which results in forcing the producer to take a big risk that is unacceptable. Also, the probability of a test being successful will be small when the test plan is too rigorous and to redesign and test the product again will incur an extra costs and efforts in product development. Therefore, the probability of passing the test will be low if we focus more on the cost of the test.

Since having too large of a demonstration test will lead to an increase in the cost of implementing the test, then the best thing to do is that for different tests, the actual criteria should be evaluated quantitatively by testing at a fixed time duration. After the evaluation, we can now examine how much the test will cost, the difficulty of passing the test, the test duration, and the trade-offs between producer's and the consumer's risks. With

this information, it is now possible to balance our decision based on the user's need and the specific objectives.

According to [70], for optimization, the approach is to find a feasible solution and compare it with others until it is not possible to find a better one. In multi-objective, optimizing the set of optima in general is much larger as it allows for a flexible trade-off between the various objectives. Since there is no "most suitable" solution of achieving the desirable outcome for all criteria, then we need to know and understand the trade-offs between the objectives when we test the product at a fixed time duration. With this, we will be able to make a balanced decision to match our goal using the Pareto front approach.

When planning a demonstration test, there are often multiple aspects to be considered and this include the consumer's risk, the producer's risk, the acceptance probability, and the cost. The natural trade-offs between these objectives require a careful evaluation of their interrelationship with the planning parameters and a systematic approach to making a tailored decision. We propose a Pareto front optimization approach for balancing the multiple objectives and offer a set of graphical and numerical tools for comparing solutions and selecting the best test plan to match different users' priorities.

Given a Weibull test plan (n, t_0, c), with n number of units at a particular t_0 (test time units) value and c units of failure (these are the parameters that must be specified to determine a test plan). For us to develop this test plan, we must be able to answer the question of "How many devices are we testing?", "For how long are we testing each of the device?", or "What is the highest number of failures allowed for this test to be successful?". In other to determine the test plan to use, we need to specify the combination

of (n, t_0, c) values that will be allowed for the test to be accepted.

In this work, we will be considering the lifetime data. Having a probability that a particular unit survive (i.e. the probability of surviving), our goal is to show that for this unit to be reliable at a particular (fixed) test duration, the probability must be at or above the desired level of confidence. We assume the failure time t follows Weibull(λ, β) distribution with λ being a scale parameter and β being a shape parameter. The probability density function (PDF) is given as:

$$f(t|\lambda, \beta) = \lambda\beta t^{\beta-1} exp(-\lambda t^{\beta}), \qquad t, \lambda, \beta > 0,$$

and cumulative distribution function (CDF):

$$F(t|\lambda, \beta) = 1 - exp(-\lambda t^{\beta}), \qquad \lambda, \beta > 0.$$

Before we define the risk criteria, let us identify clearly the requirements on reliability $R(t_*)$,

where

$$R(t_*) = exp(-\lambda t_*^{\beta}).$$

According to [108] pg. 360, for a failure time distribution, the reliable life time, t_*, for specified *, is the time beyond which 100 *% of the population will survive. The most common types of risk that are used in determining the parameters of demonstration test plan are the consumer's risk and the producer's risk. Let π be the actual reliability at time $t_* = 2000$, π_0 be the minimum acceptable reliability level (smallest level at which the reliability of a product will reach before it can be accepted) and π_1 be the maximum rejectable reliability level (highest level at which the reliability of a product will reach

before it can be rejected). From [108], pg. 344, the region $\pi \in (\pi_1, \pi_0)$ is called an indifference. From the frequentist (classical) point of view, the risk from the consumer is the probability that the test is passed when the actual reliability is at the maximum rejectable reliability level ($\pi = \pi_1$) i.e $P(\textit{Test is passed}|\pi = \pi_1)$ and the frequentist producer's risk is the probability of failing the test when the actual reliability is at the minimum acceptable reliability level $\pi = \pi_0$ i.e $P(\textit{Test is failed}|\pi = \pi_0)$. For a classical risk, if we have a satisfactory device, then it will pass the test and unsatisfactory devices will fail the test. The classical risk criteria can be a better choice when we have a particular desirable or undesirable reliability value in mind. So, given a desired area of reliability values, the conditional probability of getting a desired test is being measured.

Average risk criteria are also another risk criteria that are similar to the frequentist criteria, but the only difference is that we condition on the events $\pi \geq \pi_0$ and $\pi \leq \pi_1$ for producer's risk and consumer's risk respectively. Doing this requires a suitable prior distribution for π, which is specified by $p(\pi)$. The Average Consumer's Risk is the probability of passing a test when $\pi \leq \pi_1$ i.e the actual reliability is less than or equal to the maximum rejectable reliability. We denote this as:

$$\textit{Average Consumer's Risk} = P(\textit{Test is passed}|\pi \leq \pi_1) = \frac{\int_0^{\pi_1} \sum_0^c m(y)p(\pi)d\pi}{\int_0^{\pi_1} p(\pi)d\pi},$$

where

$$m(y) = \big(1 - exp\big(-\lambda t_0^{\beta}\big)\big)^y \big(exp\big(-\lambda t_0^{\beta}\big)\big)^{(n-y)} \tag{3.1}$$

The Average Producer's Risk is the probability of not passing a test when the actual reliability π is in the acceptable region π_0.

$$\textit{Average Producer's Risk} = P(\textit{Test is failed}|\pi \geq \pi_0) = \frac{\int_{\pi_0}^1 \sum_0^c m(y)p(\pi)d\pi}{\int_{\pi_0}^1 p(\pi)d\pi},$$

The average risk can be more suitable when considering the acceptability or unacceptability of a range of value.

In reliability, Bayesian methods have been used more often. One of the reasons for using Bayesian methods is that Bayesian analysis has the strongest features of combining many sources of information together to perform inference and using expert judgment, Bayesian method for reliability develop informative prior distributions [119]. [75] proposed a method to derive the Bayesian reliability demonstration test plan for series systems with binomial subsystem data by using Mann's approximately optimum lower confidence bound model to derive the system prior based on binomial subsystem data. With this, they derived the system Bayesian reliability demonstration test plan using existing methods for meeting posterior confidence requirements. The proposed method used objective subsystem test data and this method is generally valuable for systems that as of now have substantial subsystem test data before the reliability demonstration.

From the Bayesian version of the consumer's risk, the posterior consumer's risk (PCR) is the probability that the actual reliability is at or lower than the maximum rejectable region ($\pi \leq \pi_1$) given that the test is passed while the posterior producer's risk (PPR) is the probability that the actual reliability is at or greater than the acceptable region ($\pi \geq \pi_0$) given that the test is failed. The posterior risks provide accurately the assurance that if the test is passed, then the consumer desires a maximum probability that the actual reliability is equal or less than the maximum rejectable reliability. Also, if the test is failed, then the producer desires a maximum probability that the actual reliability is equal or greater than the minimum acceptable reliability.

For Weibull testing, with t_0 (time at which we put n units to test) and t_* (the reliability time), the criteria become:

Posterior Producer's Risk

$$\begin{aligned} &= P(R(t_*) \geq \pi_0 | \textit{Test is failed}) \\ &= P\left(e^{-\lambda t_*^{\beta}} \geq \pi_0 | \textit{Test is failed} \right) \\ &= \frac{\int_0^{\infty} \int_0^{m_0} \left(1 - \sum_{y=0}^{c} m(y) \right) p(\lambda, \beta) d\lambda d\beta}{\int_0^{\infty} \int_0^{\infty} \left(1 - \sum_{y=0}^{c} m(y) \right) p(\lambda, \beta) d\lambda d\beta} \end{aligned} \tag{3.2}$$

where

$$m_0 = -\log(\pi_0) t_*^{-\beta}, \; m(y) \textit{ is given in } (3.1)$$

and

Posterior Consumer's Risk

$$\begin{aligned} &= P(R(t_*) \leq \pi_1 | \textit{Test is passed}) \\ &= P\left(e^{-\lambda t_*^{\beta}} \leq \pi_1 | \textit{Test is passed} \right) \\ &= \frac{\int_0^{\infty} \int_{m_1}^{\infty} \left(1 - \sum_{y=0}^{c} m(y) \right) p(\lambda, \beta) d\lambda d\beta}{\int_0^{\infty} \int_0^{\infty} \left(\sum_{y=0}^{c} m(y) \right) p(\lambda, \beta) d\lambda d\beta} \end{aligned} \tag{3.3}$$

with

$$m_1 = -\log(\pi_1) t_*^{-\beta}.$$

To define the risk criteria, there is no way we can specify the requirements on reliability without specifying it at a particular time which is t_*.

Because of the disadvantage of *zero-failure* test stated earlier, instead of using the *zero-failure* test, our actual criteria for different tests can be evaluated and we will then examine the trade-offs between consumer's and producer's risks, the sample size, and the probability of passing the test. After doing this and getting the result or information, we can now make our decision based on the desire goal of our test.

3.3 Pareto Front Optimization Based on Multiple Criteria

Here, we will introduce the Optimization criteria and the Pareto front approach used for multiple objective optimizations. How the multiple criteria and their trade-offs are related is also discussed. Some decisions were made by considering the multiple criteria for different user priorities.

3.3.1 Multiple Optimization

Due to the time restriction, we set $t_0 = 100$. We need to determine a test plan (n, c) with $c \in [0, 20]$, and $n \in [c + 1, 400]$. We construct a prior distribution of actual reliability $\pi = exp^{-\lambda t_0^{\beta}}$, where $\beta \sim$ *Exponential* and $\lambda \sim$ *Inverse-gamma*. M = 4000 draws of possible π was obtained and M value was chosen to obtain more precise approximations of the criteria values. We are using this prior distribution based on subject matter expert knowledge about the expectation of the performance of the system. The subject matter expert think the system reliability are likely in general to be about 0.6, so we use the prior distribution to merge the anticipated range of the product reliability. Monte Carlo integration was used to calculate the posterior consumer and producer's risks. The posterior distribution can be evaluated and then the posterior consumer's risk, the posterior producer's risk and the acceptance probability can then be estimated approximately. The

approximations are given as follows:

$$
\begin{aligned}
&PPR \\
&= P(R(t_*) \geq \pi_0 | \, Test\ is\ failed) \\
&= \frac{\sum_{j=1}^{M} \left[1 - \sum_{y=0}^{c} \binom{n}{y} (1 - e^{-\lambda t_0^{\beta}})^y * e^{-(n-y)\lambda t_0^{\beta}} \right] I \left[e^{\lambda t_*^{\beta}} \geq \pi_0 \right]}{\sum_{j=1}^{M} \left[1 - \sum_{y=0}^{c} \binom{n}{y} (1 - e^{-\lambda t_0^{\beta}})^y * e^{-(n-y)\lambda t_0^{\beta}} \right]} \\
&= \frac{\sum_{j=1}^{M} \left[1 - \sum_{y=0}^{c} \binom{n}{y} (1 - \pi^{(j)})^y (\pi^{(j)})^{(n-y)} \right] I \left[\pi^{*(j)} \geq \pi_0 \right]}{\sum_{j=1}^{M} \left[1 - \sum_{y=0}^{c} \binom{n}{y} (1 - \pi^{(j)})^y * (\pi^{(j)})^{(n-y)} \right]}
\end{aligned}
\tag{3.4}
$$

We also approximate the posterior consumer's risk (PCR) by

$$
\begin{aligned}
&PCR \\
&= P(R(t_*) \leq \pi_1 | \, Test\ is\ passed) \\
&= \frac{\sum_{j=1}^{M} \left[\sum_{y=0}^{c} \binom{n}{y} (1 - e^{-\lambda t_0^{\beta}})^y * e^{-(n-y)\lambda t_0^{\beta}} \right] I \left[e^{\lambda t_*^{\beta}} \leq \pi_1 \right]}{\sum_{j=1}^{M} \left[\sum_{y=0}^{c} \binom{n}{y} (1 - e^{-\lambda t_0^{\beta}})^y * e^{-(n-y)\lambda t_0^{\beta}} \right]} \\
&= \frac{\sum_{j=1}^{M} \left[1 - \sum_{y=0}^{c} \binom{n}{y} (1 - \pi^{(j)})^y (\pi^{(j)})^{(n-y)} \right] I \left[\pi^{*(j)} \leq \pi_1 \right]}{\sum_{j=1}^{M} \left[\sum_{y=0}^{c} \binom{n}{y} (1 - \pi^{(j)})^y * (\pi^{(j)})^{(n-y)} \right]}
\end{aligned}
\tag{3.5}
$$

The acceptance probability (AP) which is the probability of accepting the test is also approximated by:

$$
AP = P(Testispassed) = \frac{1}{M} \sum_{j=1}^{M} \left[\sum_{y=0}^{c} \binom{n}{y} (1 - \pi^{(j)})^y * (\pi^{(j)})^{(n-y)} \right] \tag{3.6}
$$

With the number of failure $y = 0, \ldots\ldots, c$, where c is the maximum failure that occur.

3.4 Case Study

For all possible inputs with c, t_0 and n, we carried out a comprehensive assessment of all test plans in other to observe how related the four criteria are with the test plan parameters (n, c, t_0). We calculated these criteria values for each of the test plan (n, c, t_0) by using the approximated PPR and PCR together with the number of test units n.

3.4.1 Trade-Offs Between Design Factors and The Criteria

For each test plan (n, c, t_0), and using formulas given in (3.4) - (3.6), we calculated the four criteria values which are the PPR, PCR, the cost (number of test unit), and AP at a fixed testing time. Using prior with Invgamma(8,0.7) and Exp(11) with the range of parameters $c \in [0, 20]$, $t_0 = 100$ and $n \in [c + 1, 400]$, we evaluated 8190 test plans $[(n, c, t_0) = (1, 0, 100)$,..., $(400, 0, 100), (2, 1, 100)$,..., $(400, 1, 100)]$. To illustrate the relationships between the four criteria at a particular test duration, Figure 3.1 is the plot that highlight the different pairs of criteria. There are 4 trade-off plots between criteria that we will be considering. Using Figure 3.1(a) which is the plot of the consumer's risk (CR) vs producer's risk (PR) for all test plan tested and reliability value of 0.8 as an example. Darker grey to light grey symbolizes smaller c value to higher c value.

From the curve, we see some specific test plans with n = 10, 20, 35, 60, and 100 to show how changing the sample size affect the test plan. If we use zero failure test i.e $c = 0$ and want to control the CR at 0.1, then the PR is close to 0.6 and this is extremely high. This implies that there is a trade-off between the CR and the PR. Also, it implies that we have more than 60% chance of rejecting something that is already meeting the requirement and this is a huge risk from the producers side. So our work is important to consider PR. As we move towards the bottom left corner in 3.1(a) by increasing the c value, we are

simultaneously reducing both the CR and PR when the number of test (n) increases. So testing more units also improve all other criteria. When we consider a zero-failure time, we observe that the minimal PR is at 0.45 while the minimal CR is at 0.29. For a fixed n, the CR decreases as the PR increases with a decrease in the maximum failures c. On the other hand, if we increase the maximum failures c, the CR increases thereby resulting in an increased in the PR. Furthermore, if we fix the maximum failures c and increase n to reduce CR, the PR increases and likewise if we reduce the test unit n to increase CR, the PR decreases. The main disadvantage of using sample size test plan is that it does not protect against producer's risk i.e., it will be difficult for producers to produce a test unit with reliability lower than the required reliability range.

From Figure 3.1(b), the plot of the AP vs. n with different c, the CR level is at 0.05, 0.10, 0.15, 0.20, 0.25 and the PR level is at 0.05, 0.10, 0.15, 0.20, 0.25. We see that as we increase the number of test units, we are also reducing the acceptance probability as we reduce the CR and this makes it harder to pass the test. We also observed that as we allow more failures, the probability of accepting the test increases for a fixed n. This happens because if we stop testing more unit and we keep allowing more failure, then the chance of passing the test will be high. On the other hand, if we reduce the maximum allowable failure, the probability of accepting the test will decrease. Also, when c is fixed, the AP decreases as n increases and vice versa.

Figure 3.1(c), is the plot of the PR vs. number of test unit for different c. From the plot, the CR level is at 0.05, 0.10, 0.15, 0.20, 0.25. As we increase the number of test unit, we are also increasing the PR very quickly. By fixing c, the PR increases and the CR decreases as we test for more units. If CR is fixed, we see that as both n and c increases, the PR reduced. Also, by fixing n, we can reduce the CR and increase the PR as c decreases.

Figure 3.1(d) is the plot of consumer's risk vs. the number of test for different c, where PR levels are controlled at 0.05, 0.1, 0.15, 0.2, and 0.25. We see that as we test for more units, the CR reduces if we try to increase the PR. If we fix the PR, then the CR will be reduced by increasing c and n. For a fixed n, as the PR increases, the CR is reduced by allowing fewer failures.

All the last three figures tells us that as we are testing more unit, we are only improving the consumers risk and we are hurting both the PR and the AP. These explain the general trade-off between CR and other two criteria.

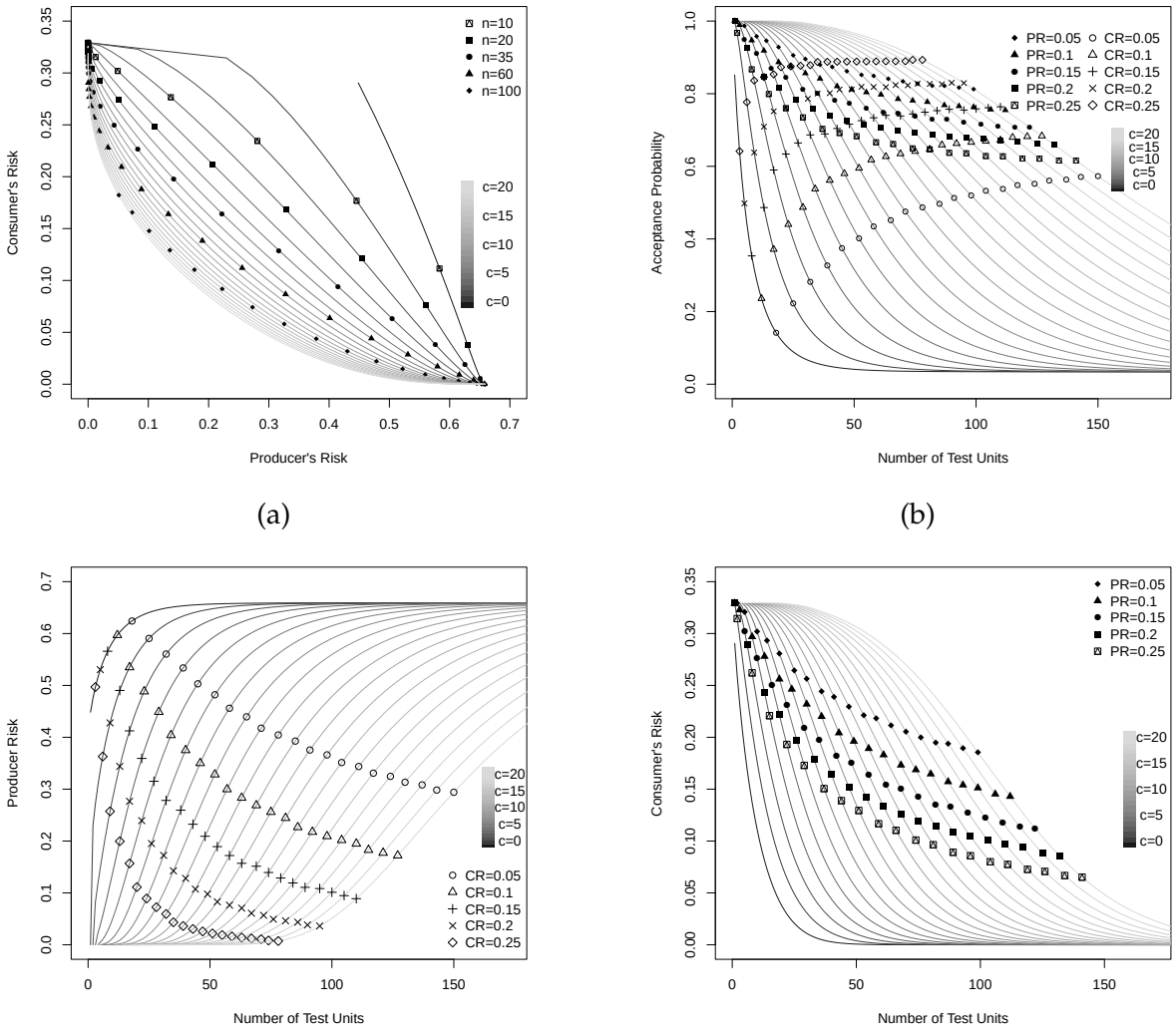


Figure 3.1: Interrelationships between the different criteria.

3.4.2 Pareto Front Optimization With Its Literature Review

Many fields have raised their attention to making the most optimal decisions based on multiple objectives or responses lately because budgets and resources have become more constrained and there is competition for resources when different objectives are considered concurrently in many applications. To make this decisions, many approach such as in [71] has been used to find a single "outstanding" solution to optimize the multiple ob-

jectives until the development of Pareto front approach [79].

In order to find competing solutions for all objectives under consideration, the Pareto front approach became increasingly popular. Pareto dominance refers to a solution that is better on at least one of the objectives and is as good as the other solution. Using the utopia point is the best performance for all the criteria but those solution on the utopia point can't be obtained in reality as a real solution. So, the solutions that are reasonable to focus on are the one on the Pareto front. The Pareto front approach finds a collection of non-dominating test plans and eliminate all the non-contenders. That is a rational set of solution to be focusing on. We then we proposed different strategies to further select test plan from the Pareto front based on the users priority. A Pareto optimal solution is one in which no other solution dominates it and whose corresponding criteria vector is undominated. i.e., whenever one objective cannot be improved without deteriorating another and a rational final choice was made from a complete set of superior solutions, then we have a Pareto optimal set.

Pareto front can be constructed based on a finite collection of solutions, and using search algorithms. The Pareto front approach was used by [79] to construct a polynomial description of the Pareto set using simulation and high-performance computing. A Pareto set member's optimality was determined by the geometric relationships between its members. Also, [113] used the Pareto-optimization to find stable folding peptides that are still not known yet. To provide sustainable land uses from global to sub-global scales, the optimization algorithms which include the Pareto Fronts and scenario analysis were used [111]. [54] also determined a set of Pareto-optimal solutions using the desirability of the objectives which reveal the preferences from an expert regarding different objective regions. By simultaneously balancing multiple criteria, [74] applied the Pareto front ap-

proach to select a test plan that is optimal. [84] developed the Pareto front approach into a two-stage to make decision.

3.4.3 Usefulness of Pareto Front in This Work

Based on taking into account the multiple criteria to remove non-contenders from making decision, we used the Pareto front approach to find a collection of non-dominating test plans. For a specified range of (n, c, t_0) values we identified the Pareto optimal solutions. After the identification, a graphical summary to determine the best demonstration test for different strategies to further select test plan from the Pareto front based on the users priority will then be highlighted.

3.4.4 Prioritizing the CR

The rest of the strategy is going to be based on which criteria we are going to consider as the most important criteria for decision making. Since the producer's aim is to always satisfy the consumers, the priority of the consumer should come first but not at the detriment of the producer. Assuming we want to focus on primarily controlling CR, then Figure 3.2 is the threshold we are going to use to reduce the solution. After applying this constraint, we will find the Pareto front of the remaining solution. Let us control the CR at 0.20, We have 21 solutions corresponding to all the 21-c value. From the figure, the cost is a square symbol (it's scale is on the right axis), AP is a triangle symbol and PR is a circle symbol with probability value on the y-axis (left axis) and the maximum allowable failures on the x-axis.

After this step, how are we going to make decision?. This will depend on the available resources and secondary criteria. Suppose we think among the remaining 3 criteria, PR is the most important and we can't accept more than 0.1 of PR, then we can see our natural solution is $n = 51$, and AP at 0.80 with $c = 10$. Suppose we think the cost (test unit) is the

most important after controlling the CR value, and we can't test more than 60 unit, then the natural design is at $PR = 0.09$, and AP is at 0.80 with $c = 12$. Suppose we can't afford more than 5 failures for our test after controlling CR, then we have $n = 27$, $PR = 0.21$, and $AP = 0.77$.

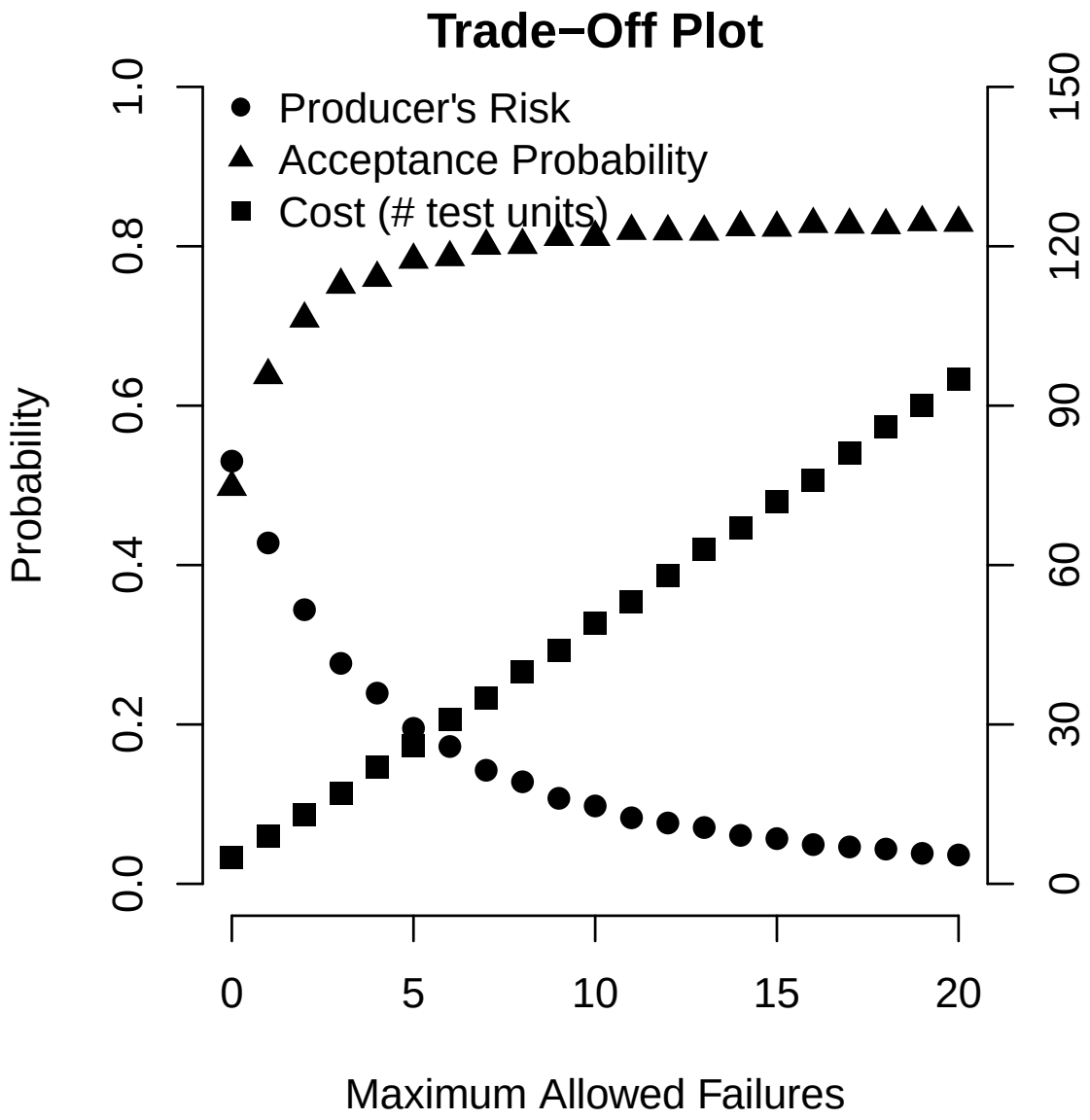


Figure 3.2: Plots of the Trade-Off for the 21 Choices on the Pareto Front Based on cost, AP and PR using CR$\leq$ 0.2

Next, we move on to prioritize the producer's risk.

3.4.5 Prioritizing the PR

If we prioritize PR, then we get a more busier Pareto front from Figure 3.3. The CR has a circle symbol while the AP has a triangle symbol. The right axis is the AP scale, and the left axis is the range of the CR on the Pareto front. The graphical summary help us to see the trend. Suppose PR is the most important criteria to control, then we reduce the solution by selecting all the subset that has PR less than or equal to 0.2, and we will find the Pareto front. After this, if we think CR is the next criteria to control, then let's control the CR to be less than or equal to 0.15. From the region where CR ≤ 0.15 to the top of the graph correspond to smaller CR value and higher AP value which is at the top right corner and the better option choice close to that region is when $c = 19$, $n = 72$, and $AP = 0.89$. Suppose the test unit is important and we can't afford more than 60 unit with low CR close to 0.13, the better option close to that line is when $c = 18$ and AP $= 0.93$.

.

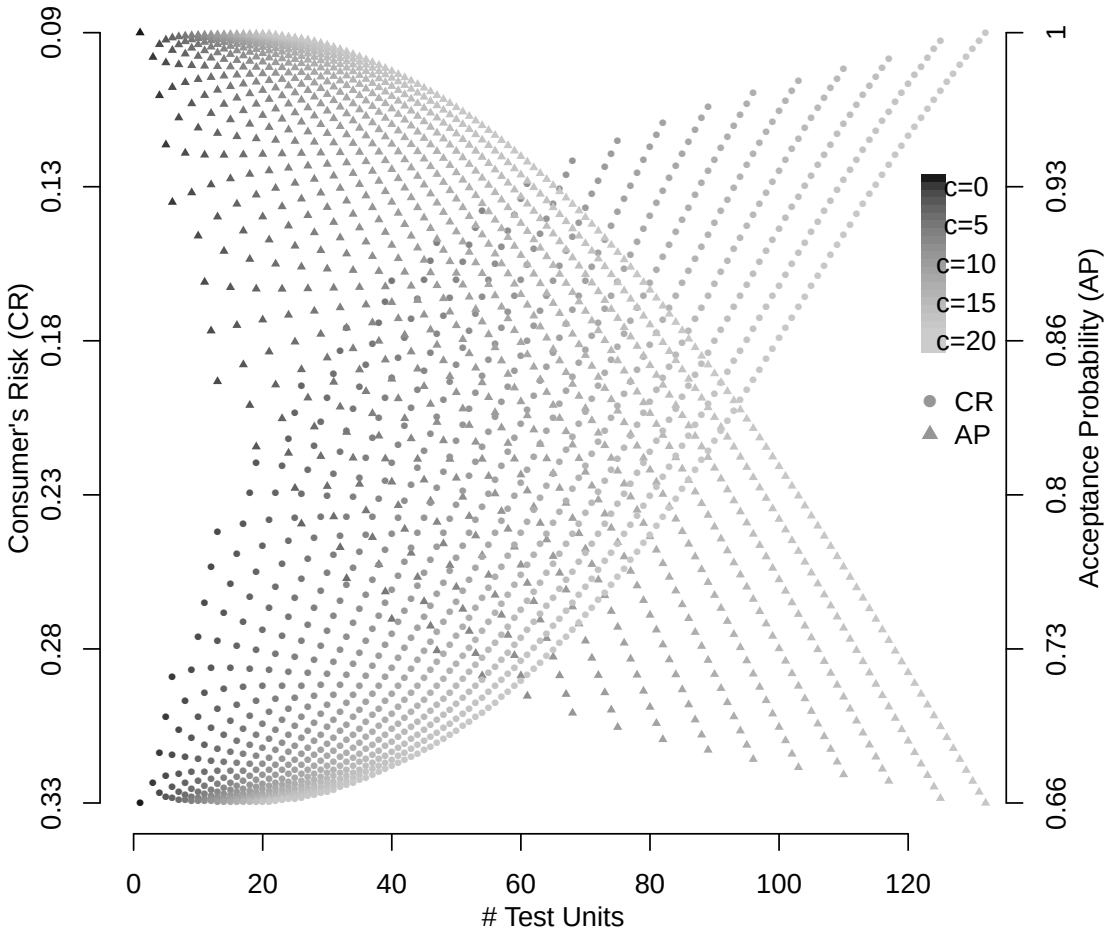


Figure 3.3: Plots of the Trade-Off for the 21 Choices on the Pareto Front Based on cost, AP and CR using PR$\leq$ 0.2

Next, we move on to prioritize the maximum allowable failures c. The number of sample to test will determine the number of failure we will have. The more expensive the test is, the lower the unit to be tested and this will affect the maximum allowable failure.

3.4.6 Prioritizing The Maximum Number of Failure "c"

For some practitioners, controlling c is the most important criteria and then based on their desire, we explore different choices of the c-value here and practitioners can pick the best plot that fit their scenario. Using a specific fixed c values of 0, 5, 10, and 15, Figure

3.4 is the trade-off plot when c is fixed. For each of this plot, the scale on the left is for PR and CR and is different because zero which is the best performance is at the top of the graph and 1 which is the worst performance is at the lowest (bottom) position. AP is on the right scale. It has zero at the bottom for the worst performance and 1 at the top is the best performance.

If we choose the zero failure test, the top left i.e $c = 0$, then basically, we can not find a solution that perform reasonably well because the region where the CR starts to increase is very small. For majority of the scenario, we have bad performance for two of the criteria using this zero failure test. For example, as we control the CR, then the AP and PR get worst using this zero failure test. So if we increase the c value, then in general, we are allowing us to have more chance to get more balanced performance.

We are actually looking for solution that are near the top where all the criteria meet for when $c > 0$ on the plot. So here are some of the choices: For $c = 15$, if the test unit is 100, the CR will be 0.1 with AP being 0.7 and PR being 0.10. From this result, as we increase the c value from 0 to 15, we are generally improving other criteria but the c to choose depend on our actual scenario of how much maximum failure we can afford.

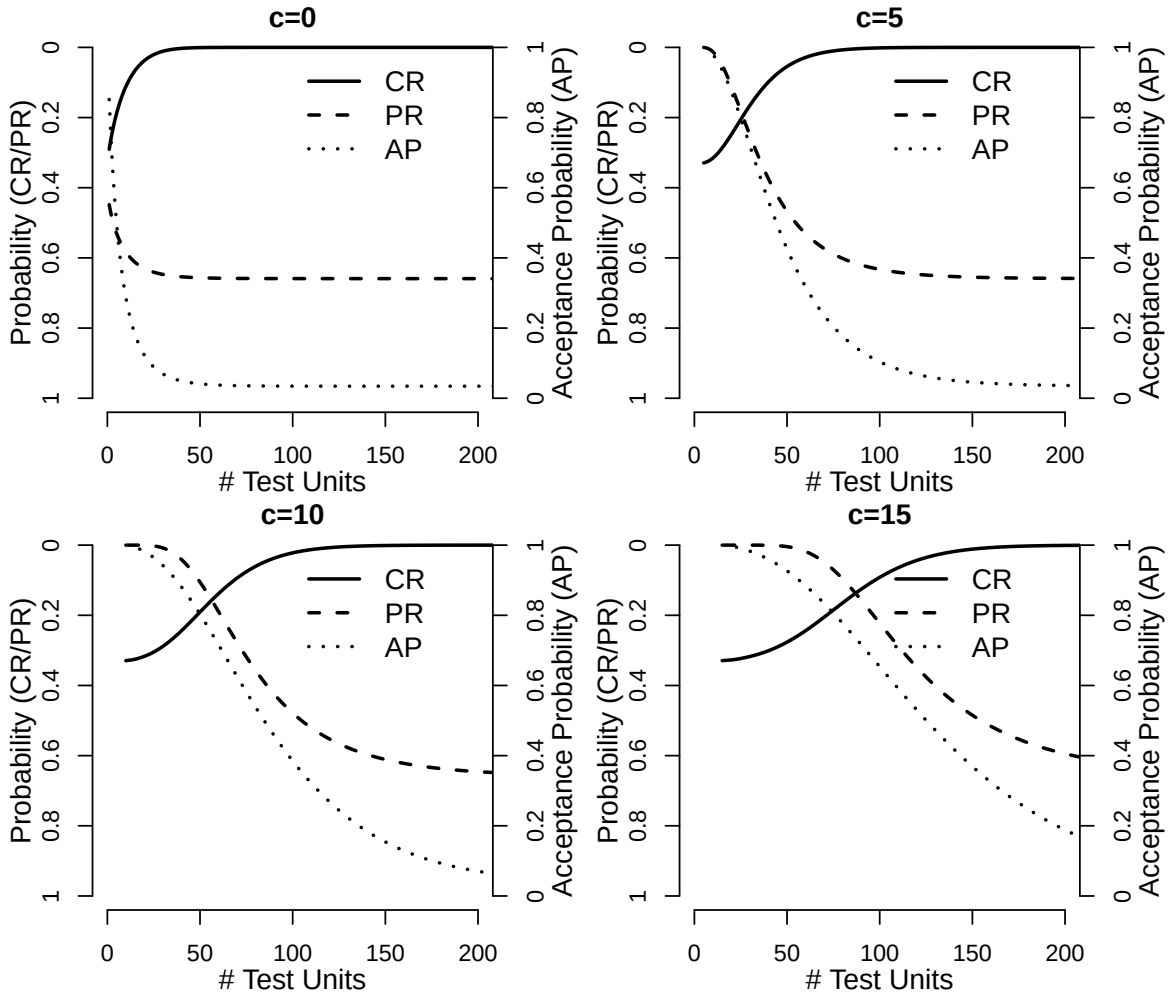


Figure 3.4: Trade-Off Plot for Fixed c = 0, 5, 10, and 15

In summary, we have shown and discussed about the three strategies of making decision in choosing a best demonstration test plan. After controlling the maximum number of failures, the CR, and the PR at a fixed testing time, it is now left to the users to decide which plan to use. This decision will be based on fund, actual testing time and resources available. Choosing a particular PR value or controlling PR based on Figure 3.3 can lead to the users choosing from an enormous rich option. There are many trade-offs between all four criteria under consideration when we focus on a particular *c* value and as a result of that, we have many competing options to choose from. However, the interconnections between the criteria are simple and clear from Figure 3.4. From Figure 3.2, we can see a clear and straight forward result to make our decision when we control the CR. The

Pareto front techniques can be effectively used to eliminate lesser solutions. So, since controlling the CR leads to the simplest choices for reaching a final decision, we need to do sensitivity analysis on different CR value.

3.5 Sensitivity Analysis

In this section, we want to discuss in detail the effect of the decision made by the user based on their choices. Figure 3.5 is a plot for using three different CR values. For each level, we have three curves that represent the criteria values for all test plans at same time duration on the Pareto front based on the PR, AP, and cost criteria. The actual criteria values are shown in Table 3.1. Based on Table 3.1 and Figure 3.5, we observe that when the threshold value for the CR decreases (i.e. 0.20, 0.1, 0.05), more units will be tested for a fixed c value. With this result, the PR increases with a decrease in the AP value. Going for smaller value of PR will leads to an increase in c and n.

For example, if the producer can accept maximum of 0.29 risk, then using 0.20 threshold for CR, we can only test for 18 units to achieve 0.73 for AP. With this, we must fail 3 units. If the user wants more units to be tested using the same PR value, then we must test 57 units with a threshold CR = 0.1 and fail 8 units with AP = 0.62 or we test n = 150 units and threshold CR = 0.05. With this, we will have to fail 20 units with AP = 0.573. From this, we see that as we increase the number of test unit and fix the PR, the CR decreases and this results in a decrease in the AP.

As c increases for a fixed CR, the PR decreases and the AP value increases. By increasing the sample size, the PR decreases and this leads to an increase in the AP. From Figure 3.5, it is impossible to improve the AP to above 0.573 and PR to below 0.2941 with CR = 0.05 regardless of the number of test units. Hence, to understand the impact of different CR value on available choices, we need to determine the performance of the test.

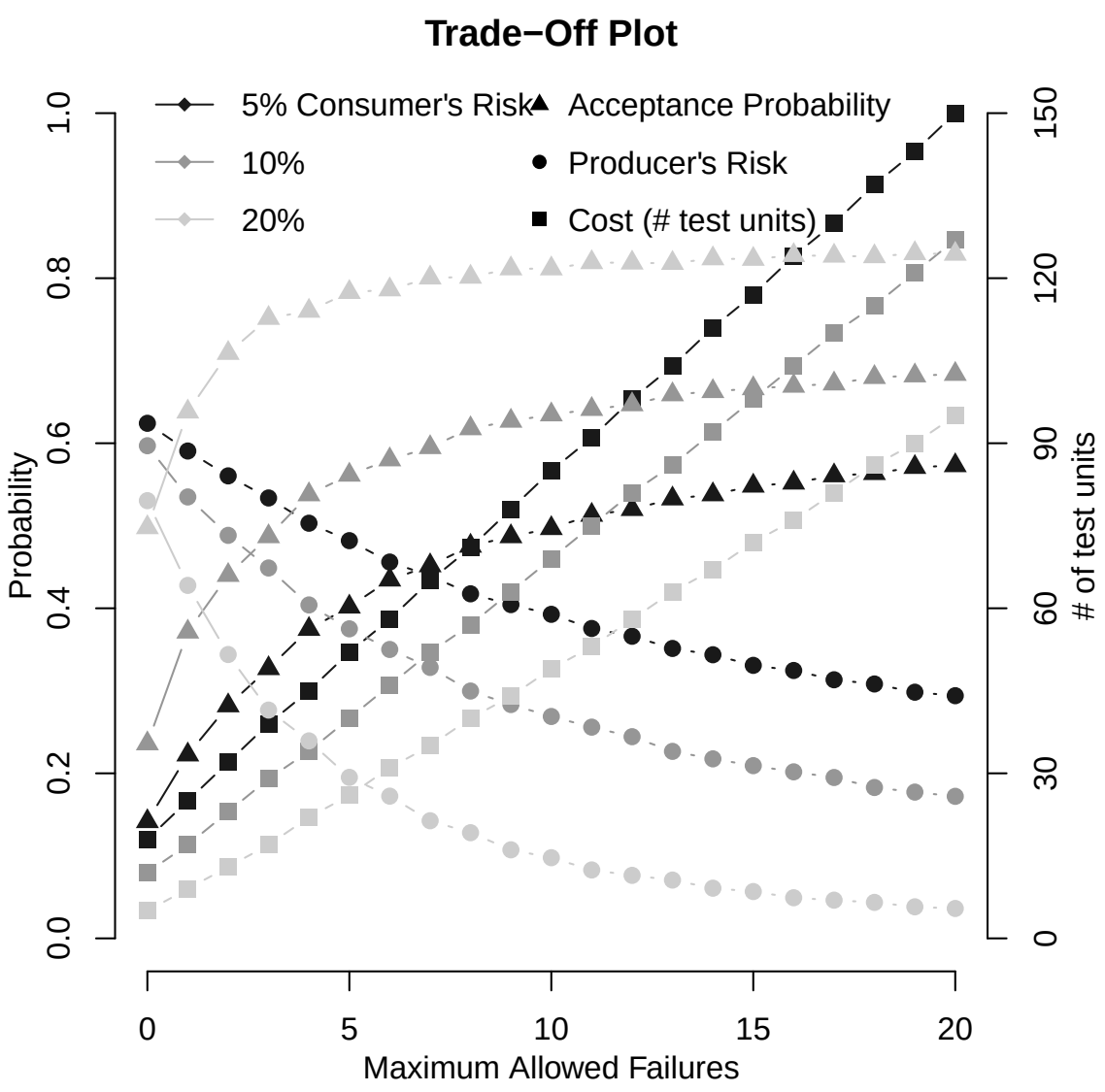


Figure 3.5: Plots for the Pareto Front Based on using different CR value

Table 3.1: Table for the Pareto Front Based on using different CR value

CR=0.20					CR=0.1					CR=0.05			
PR	AP	n	c		PR	AP	n	c		PR	AP	n	c
0.5304	0.4979	5	0		0.5970	0.2360	12	0		0.6242	0.1417	18	0
0.4455	0.5965	10	1		0.5349	0.3714	17	1		0.5906	0.2225	25	1
0.3626	0.6781	14	2		0.4884	0.4401	23	2		0.5604	0.2820	32	2
0.2944	0.7276	18	3		0.4489	0.4870	29	3		0.5337	0.3271	39	3
0.2552	0.7410	23	4		0.4040	0.5378	34	4		0.5031	0.3746	45	4
0.2093	0.7670	27	5		0.3751	0.5615	40	5		0.4820	0.4017	52	5
0.1849	0.7725	32	6		0.3501	0.5801	46	6		0.4560	0.4345	58	6
0.1647	0.7769	37	7		0.3283	0.5950	52	7		0.4394	0.4517	65	7
0.1378	0.7911	41	8		0.2997	0.6178	57	8		0.4175	0.4755	71	8
0.1245	0.7931	46	9		0.2834	0.6269	63	9		0.4045	0.4870	78	9
0.1132	0.7948	51	10		0.2689	0.6346	69	10		0.3927	0.4968	85	10
0.0964	0.8039	55	11		0.2560	0.6412	75	11		0.3755	0.5129	91	11
0.0886	0.8046	60	12		0.2444	0.6468	81	12		0.3661	0.5199	98	12
0.0817	0.8052	65	13		0.2266	0.6590	86	13		0.3513	0.5327	104	13
0.0757	0.8058	70	14		0.2175	0.6628	92	14		0.3437	0.5377	111	14
0.0657	0.8120	74	15		0.2093	0.6663	98	15		0.3308	0.5483	117	15
0.0613	0.8121	79	16		0.2018	0.6693	104	16		0.3246	0.5520	124	16
0.0574	0.8122	84	17		0.1949	0.6721	110	17		0.3134	0.5607	130	17
0.0539	0.8123	89	18		0.1828	0.6799	115	18		0.3083	0.5635	137	18
0.0474	0.8169	93	19		0.1773	0.6819	121	19		0.2983	0.5709	143	19
0.0447	0.8169	98	20		0.1721	0.6837	127	20		0.2941	0.5730	150	20

Now let's consider another prior Invgamma(9,0.7) with Exp(11) and compare it with our previous prior. Figure 3.6(a) is the plot of the consumer's risk (CR) vs producer's risk (PR) for all test plan tested at a fixed time t_0 of 100 hours and reliability value of 0.8. We

are doing sensitivity analysis on prior distribution because prior distribution are subjective choices. Bayesian method generally is subjective to the choice of prior distribution. So we want to look at choosing different prior distribution and how it affect our selective test plan. From this plot, we noticed that the range of the CR is from 0 to 0.27 and the range of the PR is from 0 to 0.72. Comparing this plot with the plot of when we use Invgamma(8,0.7) with Exp(11) in 3.1(a) (the range of the CR is from 0 to 0.37 and the range of the PR is from 0 to 0.7), we noticed that increasing the shape parameter of inverse-gamma parameter (i.e., by applying a more diffuse prior distribution. This can be seen in Figure 3.8) leads to an increase in the PR and a decrease in CR.

In figure 3.6(b), the plot of the AP vs. n with different c at a fixed test duration t_0 value, we see that when we fix the CR at 0.05 and reduce the maximum number of failure, the AP reduces from 0.68 to 0.2. Comparing this to 3.1(b), we see that if we fix the CR at 0.05 and reduce the maximum number of failure, the AP reduces from 0.56 to 0.12. This implies that the AP increases when we apply more diffuse prior distribution. Figure 3.6(c) is the plot of the PR vs. n with different c, we see that when we fix the CR at 0.1 and $c = 0$, we test for less units compared to when we use a less diffuse prior in 3.1(c).

Finally, from 3.6(d), we observed that the CR reduces compared to when less diffuse prior in 3.1(d) is used. In summary, we see that if we apply a less diffuse prior, the CR will increase and the PR will decrease. Also, the AP decreases.

For this prior, we used different value for CR, and the detail effect of the decision made by the user based on their choices is shown in the trade-off plot in Figure 3.7. For each of these c values, we have best choice to choose. Based on this plot, we observe that when the threshold value for the CR decreases (i.e. 0.20, 0.1, 0.05), more units will be tested for a fixed value of c. If the producer can accept at most 0.29 risk, then using 0.20 threshold

for CR, we can only test for 12 units to achieve 0.85 for AP. With this, we must fail 3 units. If the user wants more units to be tested, then we must test 58 units with threshold CR = 0.1. We will have to fail 9 units with AP = 0.745 or we test n = 138 units with threshold CR = 0.05 and fail 19 units. This will leave us with AP = 0.679. From this, we see that as we increase the sample size, the CR decreases, and this result to a decrease in the AP.

In summary, having a prior distribution Invgamma (9, 0.7), Exp(11) which is a more diffuse prior distribution results into testing less units with higher acceptance probability and higher PR. Having higher information from Invgamma (9, 0.7), we observe that the spread decreases compared to Invgamma (8, 0.7). Since we want higher reliability, then prior with more information i.e. Invgamma (9, 0.7) should be used. Comparing our two priors (Invgamma(8,0.7) with Exp(11) and Invgamma(9,0.8) with Exp(11)) results, we observed that to quantify the different criteria using a Bayesian approach, the calculated probability of accepting the test and the risk criteria can be affected by the prior distribution specified by the user. We observe that the acceptance probability and the risk criteria calculated can be sensitive to the prior distribution selected by the user when we use the Bayesian approach to measure the different criteria.

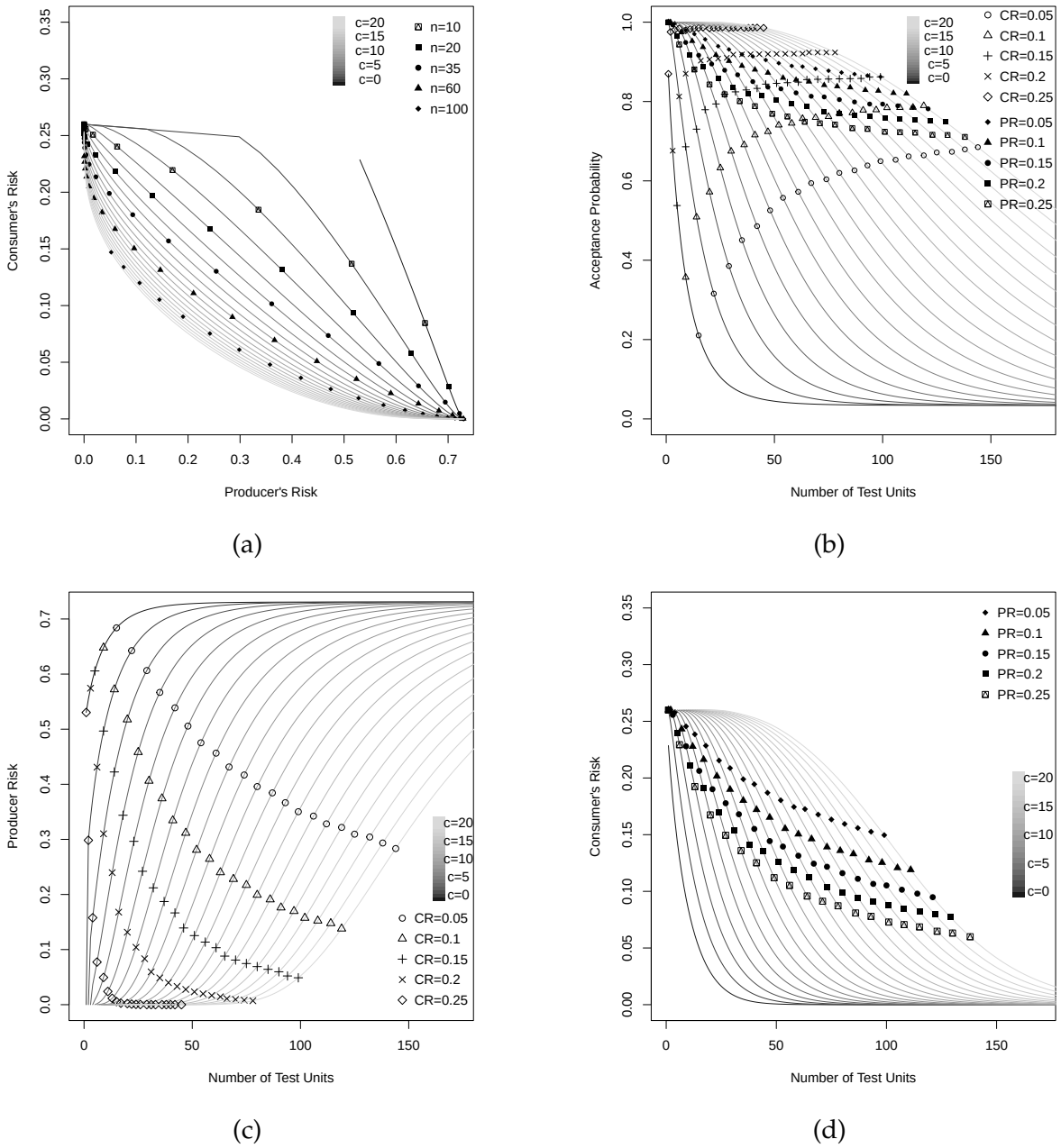
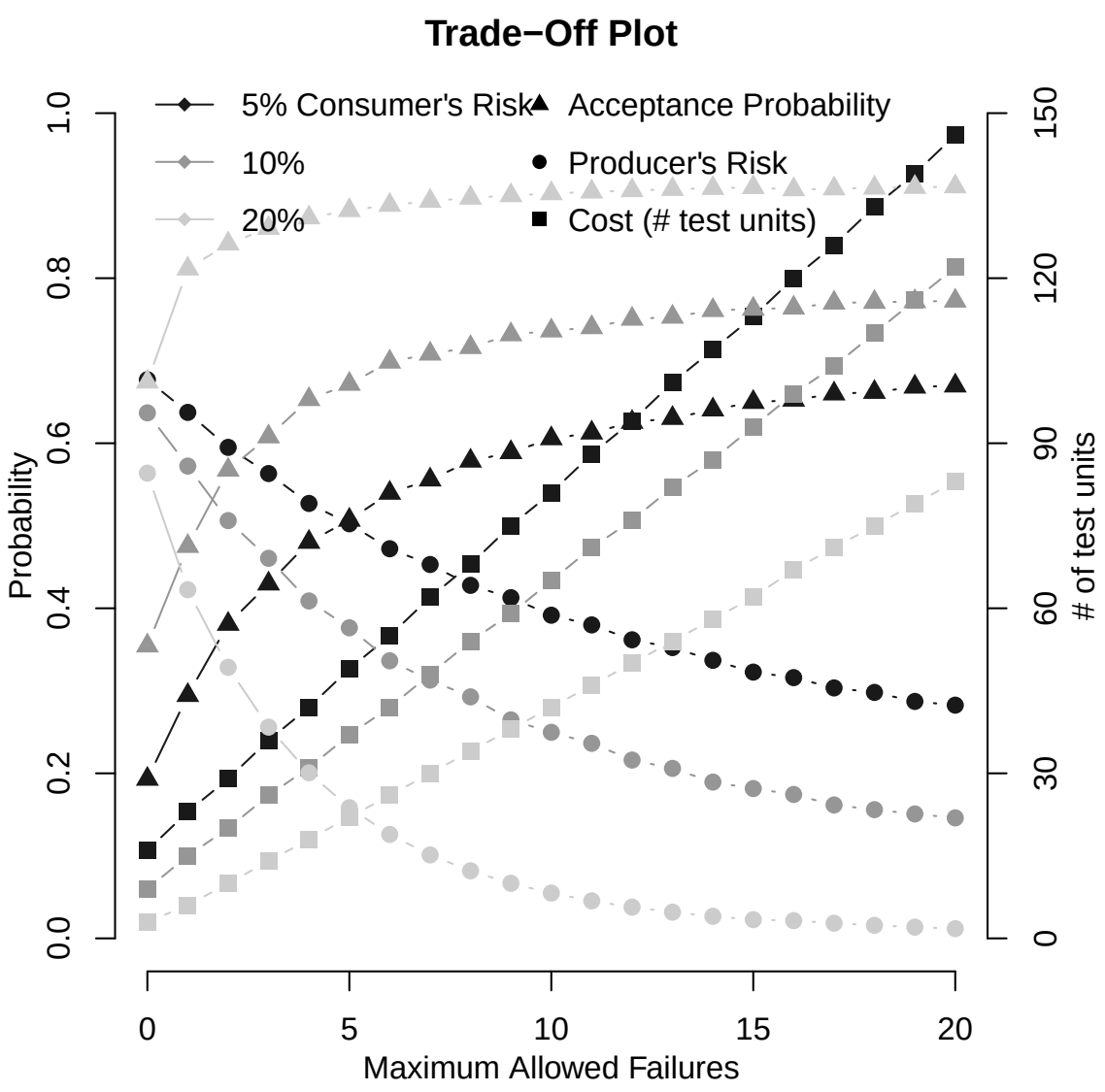


Figure 3.6: Interrelationships between the different criteria using different prior.



(a)

Figure 3.7: Plots for the Pareto Front Based on using different CR value

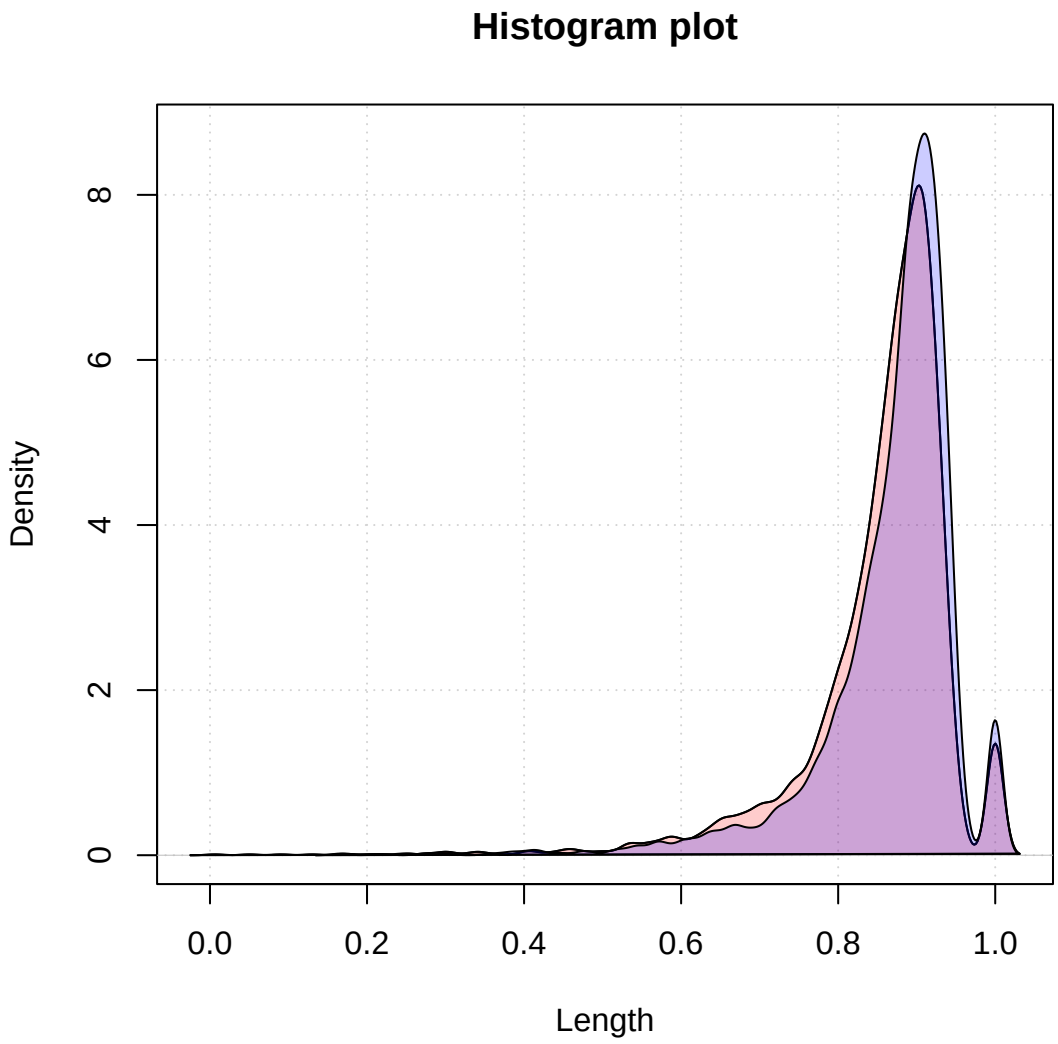


Figure 3.8: Probability Density Curves for weibull distribution using two different (Invgamma(9, 0.7)(higher density) and Invgamma(8, 0.7)) with same Exp(11).

Chapter 4: Bayesian Analysis For Accelerated Degradation Test Data With Multiple Degradation Measurements and Covariates Using the General Path Model

4.1 Introduction

We can categorize degradation models into two broad classes, which are stochastic process models and the general path models [132]. These two models were proposed to capture the three sources of variation in a degradation process. These sources of variation include variation from unit-to-unit, temporal variation, and variation based on measurement [23]. Due to time-dependent structures properties and because it captures the temporal variation within a unit, the stochastic process model is being used and this model includes the Wiener process (i.e., Brownian motion) [7], Gamma process [62], and Inverse Gaussian (IG) process ([125], [133]).

Wiener process model with random drift-volatility has been proposed by Wang [124]. Ye [132] reviewed the cases involving random effects (which modeled a unit-to-unit variation), covariates, and measurement errors. Whitmore [45] used the Wiener process with random drifts to estimate degradation. However, one of the obstacles faced is that we cannot directly use the linear drift Wiener process to describe a nonlinear degradation process and this is because of the mechanism behind the failure of the product and because of how complex the structure is. To overcome this obstacle, we need to transform the degradation data. Some of the transformation methods include log transformation [9], [103], [18] and time-scale transformation [17], [83], [90]. Si et al. [86] realized that we cannot properly transform all nonlinear degradation processes. Hence, there was a development of a Wiener process model with a nonlinear drift coefficient to characterize the dynamics and

nonlinearity of the degradation process. [40] focused on the RDT plan design problem for long life products based on degradation test data. They assume the product fails due to the degradation of some special performance and they modeled the degradation process using Wiener process with shift.

A degradation process that is always strictly increasing and positive, can be modeled using the Gamma and IG (Inverse Gaussian) processes. The Gamma process with random scales was studied by Lawless and Crowder [62]. An uncomplicated way we can model the Wiener process is by adding measurement errors, but this is not the case for the Gamma or the IG processes. Stochastic model doesn't utilize the physical knowledge of the degradation process.

General path model is being used whenever we have a nonlinear degradation path and when we have a physical understanding recommending a functional form of the degradation path. A general path model is used in this work because it allow us to leverage the physical understanding of the process based on the formulation of the degradation path model. Using a mixed-effects regression and measurement errors, it is easier to model the unit-to-unit variation using the general path models [63]. To understand different aspects of the reliability performance, some degradation tests measured the multiple characteristics of a degradation process and this type of data is called the degradation data with multiple degradation characteristics (DCs).

In general, for ADT dataset, most of the existing work consider a single characteristics but a new model that focuses on the modeling and analysis of degradation data has been found to analyze multiple Dcs. This new method model the multiple DCs interdependently. Since multivariate degradation data is more common recently, Hong et al. [78] came up with the need to develop multivariate degradation models. Huang and

Askin [120] discussed the analysis of electronic devices with multiple competing modes of catastrophic failure and degradation failure by assuming the independence of multiple degradation processes.

There are very little work that actually address correlation between the multiple degradation measurement. That is where our work fit in. Some existing work used different method such as copula method and (EM) Expectation Maximization algorithm but we used full Bayesian approach which provide a straightforward structure for inference. For copula method, adding random effects to it can be challenging and complicated when estimating the model because of how the number of model parameters increase. Some work used multivariate model by using different estimation approach such as the EM which is the Expectation Maximization algorithm. To estimate a time-to-failure distribution using degradation measures, Lu and Meeker [63] developed more general statistical models and data analysis methods. The frequentist approach was used by [30] fang et al. to analyze the ISO dataset on hierarchical model using the maximum likelihood estimation (MLE) method. Computation of frequentist approach is slow, and the parameters are not modeled probabilistically unlike the Bayesian approach that model the data and the parameter probabilistically. To integrate over high dimensions, Bayesian approach is useful.

Huibing et al. [50] proposed a Bayesian framework to integrate the population degradation information and individual degradation data using a Weiner process and used the MCMC method to estimate the unknown parameters in the model. Soliman et al. [91] used the MCMC sampling method for posterior inference of the reliability of the stress–strength model. [58] also proposed a sequential MCMC model for reliability evaluation of the Offshore Wind Farm. The full Bayesian approach is easy to implement and it provide straightforward structure for inference.

As discussed in Chapter 2, Accelerated Degradation Tests (ADTs) measure the degradation of products under the accelerated conditions. These are then used to model the degradation path and predict failure and/or assess reliability under the normal use conditions. ADTs method is used to extrapolate the lifetime of highly reliable products under normal use conditions. ADT datasets have been used by different researchers to track the degradation of products. [130] applied ADT dataset (outdoor weathering data that contain degradation measurements and environmental covariates) to model degradation paths and they proposed a class of nonlinear general path models with random effects to incorporate dynamic covariates for modeling of degradation paths. This dataset has single characteristic and the parameters were estimated using the outside iterations. [127] also applied ADT dataset (NIST coating degradation data) to track the degradation of products. This dataset has multiple DC measurements each with repeated measurements. Their proposed model incorporate nonlinearity in the degradation path, physical understanding of the degradation process, unit-to-unit variation, and covariate effects by using the Bayesian approach.

The Bayesian approach using Markov Chain Monte Carlo (MCMC) will be used in our work to integrate the multivariate random effects over multiple dimensions. A simulated and synthetic data (with multiple DC measurements) from ISO data (ADT dataset that has repeated measurements) will be used to propose our model in order to incorporate nonlinearity in the degradation path, physical understanding of the degradation process, unit-to-unit variation, and covariate effects. Our model will capture the correlation between the initial condition and the degradation rate.

The remainder of this paper is organized as follows. We will discuss the ADT dataset (ISO/IEC) and the multivariate nonlinear degradation path model to be used. We will also talk about how the model parameters and associated uncertainty can be estimated

using the MCMC method with the use of the Stan package. We will also talk about how to plot the reliability curve. To demonstrate how our proposed multivariate nonlinear degradation model performed together with when the multiple independent degradation models are being used, we will conduct a simulation study. We illustrated the implementation and performance of our method through the analysis of the synthetic ISO/IEC degradation test data.

4.2 Data and Models

4.2.1 Overview

4.2.2 ISO/IEC Data and Application

The worldwide standardization system consists of both the International Electrotechnical Commission (IEC) and the International Organization for Standardization (ISO). ISO 10995 is the international standard for the reliability testing and archival lifetime prediction of optical media. The standard from the name implies the testing conditions in terms of the combinations of stress variables—temperature and relative humidity. Fang, et al. [30] used degradation test to predict the lifetime of ISO 10995. It is assumed by the standard that the projected failure times are the actual failure times, which are then analyzed using the Eyring or Arrhenius model.

Providing guidance on current practice is ISO 10995:2011 [3] purpose, through accelerated degradation test for optical media archival life products prediction and understanding the underlying failure mechanisms. A high level of stress is thought to accelerate chemical reactions, resulting in the degradation of the material that results in the failure of disk. Different types of optical media formats such as DVD-R/-RW/-RAM and +R/+RW were tested in the ISO 10995:2011 experiment.

The operation of reading or writing data from a disk is typically accomplished by altering the transparency of an organic dye layer [49]. This brings about the degradation of the transparent portion of the dye layer over an extended period of time due to the organic nature of the dye. This process can take several years in a normal environmental condition because it has its roots in chemical kinetics, but we can speed up (accelerate) the process to a very great extent with higher temperature or humidity [34]. Using the Eyring model and some other models, we can then model the effects of these stress variables, and this is derived from the study of chemical kinetics [98]. When the information recorded on a disc cannot be recovered without significant loss, then that disc has reached the end of its useful life, and to accelerate the degradation process in order to shorten the lifetime of the disc, we will have to increase the temperature, duty cycle, voltage, relative humidity, or particle induction. ISO 10995:2011 [3] uses the accelerated degradation tests (ADTs) to estimate and predict the lifetime of optical media.

Periodically, the degradation measurements are recorded for every unit that is tested in a variety of conditions by the experimenter in ISO 10995:2011 and the observed degradation data to a regression model was fitted. The failure threshold value is at log(280). For all the test units, the predicted time-to-failure were fitted in an accelerated failure time model by assuming an exponential lifetime distribution [132]. Another problem with ISO 10995:2011 is that the model completely ignores the unit-to-unit variation among the test samples by assuming a homogeneous distribution across the population. ISO 10995 dataset has the relative humidity and temperature as the two stress variables. Each of these stress variables have four stress conditions which can be seen in Table 4.1. The dataset has $n = 90$ test unit with $r = 2$ covariates (Temperature and Humidity).

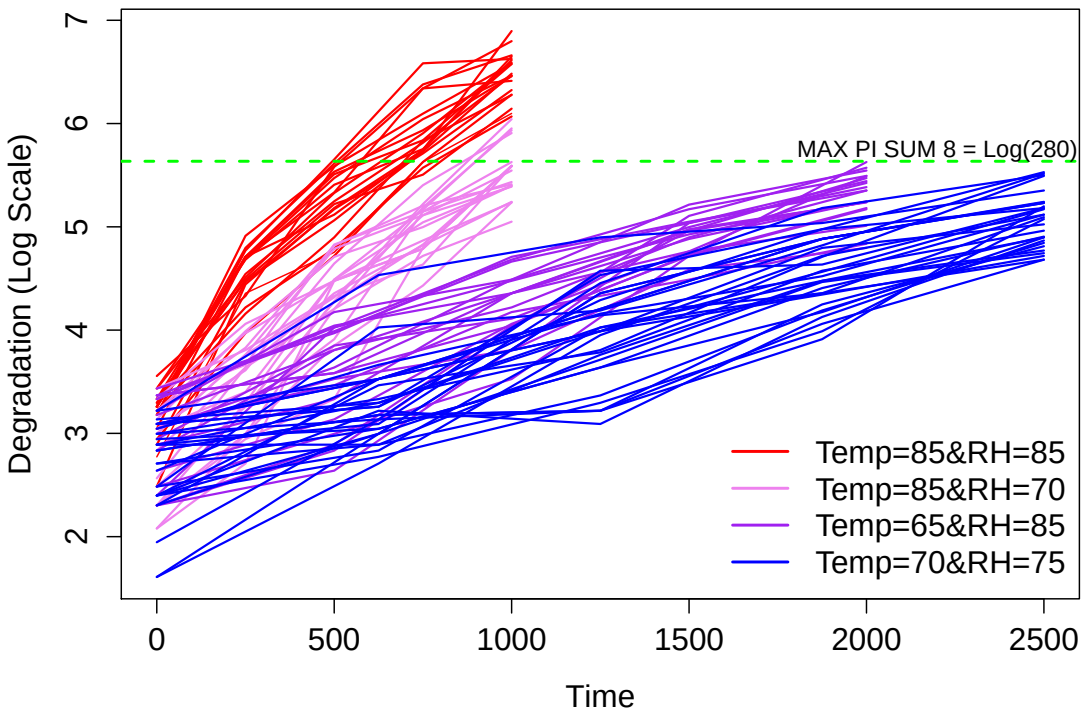


Figure 4.1: ISO 10995:2011 ADT Data Degradation Paths

Table 4.1: Stress Condition

Number	Relative Humidity in %	Temperature in ^0C
1	85	85
2	70	85
3	85	65
4	75	70

Figure 4.1 is ISO ADT Data Degradation Paths that all 90 ISO accelerated degradation test units were measured on a log scale. The test units can be recognized under four different test conditions with different colors. Different color represent different test conditions. So, degradation rate changes with different test condition. Also, we can see for unit with the same test condition (for the same color), there are differences that repre-

sent unit to unit variation. Also, the relative position of these lines on the plot are pretty consistent and are preserved across each measurement. This indicate that most of the measurement are highly correlated and that is why we need to consider correlation while modeling multiple DC measurement.

From the plot, few patterns can be observed. First, across different test conditions, different rate of degradation exist such that the units with the highest relative humidity RH (at 85%) levels and highest temperature (at 85^0C) have the highest degradation rates (the red color in the Figure). Also, units with the lowest RH (75%) and lowest temperature (70^0C) had lowest rate of degradation (blue color from the plot). Second point to consider is that degradation rates are not necessarily constant over time because a curvature can be observed on many paths, and it may not be possible to model the degradation of all test units as a linear function of time. Therefore, because of this, a polynomial regression model or nonlinear regression model is necessary to model the variation of the trend of the degradation. Third, both the initial degradation level and the rate of degradation appear to vary from unit to unit and this can be seen in Figure 4.1 in which at time 0 in the study, the degradation paths all start with different starting values (ranging from 1.609438 to 3.555348 on the log scale and correspondingly 5 to 35 on the raw degradation measures), which indicates a different starting condition for the test units at the beginning of the study. Also, the degradation rates and initial degradation conditions show a positive correlation, and this implies that the temporal degradation rate and the initial condition are positively correlated and have a unit-to-unit variation. We can see that the red paths from the plot have the highest initial degradation values and because of this, it has higher degradation rates over time.

The temporal degradation rate, as well as the individual initial condition, needs to be accounted for by multivariate random effects. From ISO 10995:2011, the performance of

the threshold is log(280) [38] for Optical Media. The green horizontal line in Figure 4.1 is the threshold for soft failure for the media used. If the level of degradation crosses this value of threshold, then the test unit is considered to fail. From the 4 test conditions given in Figure 4.1, we can see that harshest test condition, which is the test at 85^0C for temperature and 85% for the RH level failed because they pass across the threshold value. Also, at 85^0C for temperature and 70% for the RH level, we see that part of this level also failed by crossing the threshold value. With the remaining two levels, which are the one with 65^0C for temperature, 85% for the RH levels and the one with 70^0C for temperature, 75% for the RH levels, no failure was observed while doing the test. Based on this observation, we can justify that both the RH and temperature are acceleration factors for this optical media dataset and when it comes to studying high reliability products, the ADT test is critical because it helps speed up or accelerate degradation process and also helps to collect data quickly.

4.2.3 Degradation Data

In ISO 10995:2011 standard, they conducted the experiment in 4 testing conditions with a combination of temperature and relative humidity and with different specimens. There are 20 specimens each in condition 1 which is (85^0C, 85%), condition 2: (85^0C, 70%), and condition 3: (65^0C, 85%). For condition 4: (70^0C, 75%), we have 30 specimens. The specimen for the first two conditions are recorded from time 0 to 1000 hours with 250 interval; for third condition, the specimens are recorded from 0 to 2000 hours with interval of 500; and the fourth conditions are recorded from time 0 to 2500 hours with interval of 625.

In this experiment, a longer time period under less severe stress condition is used. Using this dataset, we created two synthetic data and part of the data are shown in Table 4.2 and 4.3. During this test period, not all the test units in this experiment have failed, and

the degradation measurements was fitted into a nonlinear regression so as to calculate the projected failure. We will propose a multivariate degradation path model that will capture the correlation between multivariate measurements in the next subsection.

Table 4.2: 1ST D.C. FOR ISO DATA

Temp = 85^0C, $RH = 85\%$						Temp = 85^0C, $RH = 70\%$					
Hours						Hours					
Disk	0	250	500	750	1000	Disk	0	250	500	750	1000
A3	26	94	190	335	642	B1	10	20	67	112	156
A4	26	111	247	343	718	B4	20	43	120	166	219
A9	24	118	285	723	754	B6	21	37	104	222	368
A10	12	85	178	312	988	B7	21	30	89	155	221
A12	24	136	267	444	719	B11	28	58	88	120	268

Table 4.3: 2ND D.C. FOR ISO DATA

Temp = 85^0C, $RH = 85\%$						Temp = 85^0C, $RH = 70\%$					
Hours						Hours					
Disk	0	250	500	750	1000	Disk	0	250	500	750	1000
A1	16	78	116	278	445	B2	8	20	47	84	188
A2	25	64	134	342	532	B3	12	26	72	185	421
A5	27	89	185	246	466	B5	32	45	76	103	267
A6	21	111	207	567	896	B8	22	26	72	125	267
A7	26	121	274	589	781	B9	25	46	124	182	224

4.2.4 Hierarchical Degradation Path Model

Z_{ipgq} which is response variable in ISO data has a length of 450 because, each of the 90 test units repeated measures have 5 different time points. Using the original ADT dataset,

we created two synthetic data with two DC measurements (i.e., $P = 2$) of $n = 45$ test unit each in other to achieve a multivariate model. In developing this model, we fitted a non-linear regression model of log-transformed error rate, and the time-to-failure is assumed to follow a lognormal distribution. The stress level will only affect the location parameter of the lognormal distribution.

Using the hierarchical degradation model, we want to get the correlation among the multivariate measurements on the two DCs. In developing this model, we used two levels structure to build the model. Having $i = 1, ..., n$, n is the number of test unit , the first level $log(y_{ipgq})$ is the log scale of the response variable and this is defined as the sum of degradation level for unit i on the pth DC under stress condition g at time t_{iq}, where $g = 1, ..., G$, $p = 1, 2$, $q = 1, ..., w_i$, $G = 4$ and $w_i = 5$ is the total number of measured time points for unit i. This first level is assumed to be linear in measurement error and in the transformed time variable $t_{ipgq}^{\gamma_i p}$. The hierarchical degradation model is given as:

$$logy_{ipgq} = Z_{ipgq} + \epsilon_{ipgq} \tag{4.1}$$

where the degradation path is given as

$$Z_{ipgq} = \beta_{0ip} + \beta_{1ipg} t_{ipgq}^{\gamma_i p} \tag{4.2}$$

The second equation is a nonlinear model of time. The coefficient is affected by the accelerating factor like RH, Temperature. We are allowing the intercept which represent the initial error rate and our degradation rate to be correlated together. β_{0ip} (Initial error rate) was measured prior to accelerated aging, β_{1ipg} is the degradation rate and Z_{ipgq} is the actual degradation path of unit i on the pth DC under stress condition g at time t_{iq}. The measurement error ϵ_{ipgq} are i.i.d normally distributed with zero mean and unknown variance i.e., ϵ_{ipgq} $N(0, \sigma_{rp}^2)$. σ_{rp}^2 is the common variance for measurements on the pth DC

with $r = 2$ random effects. Meeker and Escobar [64] talked about ADT modeling when stress factors, such as temperature, humidity, and voltage are presence. The impact of environmental stress on material properties are incorporated into the general path model above (4.1) and Eyring model (which is given below) in (4.3) models the temperature and another factor, such as humidity.

$$Acceleration\ \ Factor = AT^{\alpha} exp\left[\frac{\Delta H}{kT} + \left(B + \frac{C}{T}\right) logRH\right] \tag{4.3}$$

where

T is the Temperature measured in degrees Kelvin,

ΔH is the activation energy per molecule,

A is the pre-exponential time constant,

T^{α} is the pre-exponential temperature factor,

RH is the relative humidity,

B and C are the RH exponential constants and

k is Boltzmann's constant.

$k = 1.38071.3807 * 10^{-23}(J/moleculedegreeK)$.

In other to measure unit to unit variation, we will let β_{0ip} and $log\mathrm{A}_{ip}$ be our random effects that are associated with unit i for P DCs and using a reduced Eyring function, we define β_{0ip}, β_{1ipg} and $logA_{ip}$ as follow:

$$\beta_{0ip} = \mu_{1ip} + \epsilon_{1ip}$$

$$logA_{ip} = \mu_{2ip} + \epsilon_{2ip}$$

$$\beta_{1ipg} = exp(logA_{ip} + B_p logRH_{gp} + \Delta H_p \frac{11605}{T_{gp} + 273.15}) \tag{4.4}$$

where μ_{1p}, μ_{2p}, B_p, ΔH_p and γ_p are the fixed effects.

We will denote $Q_p = (\mu_{1p},\ \mu_{2p},\ B_p,\ \Delta H_p \text{ and } \gamma_p)^{'}$ to be a vector of fixed effect parameters and $\alpha_{ip} = (\beta_{0ip}, logA_{ip})^{'}$ denotes the vector of a random effect related to P DCs. The vector comprising of the random effect and the fixed effect related to the *pth* DC is given to be $\Psi_p = (Q_p^{'}, \alpha_{ip}^{'})^{'}$. Since for unit i, there are two random effects for P DCs, then those two random terms ϵ_{1ip} and ϵ_{2ip}, follow a multivariate normal (MVN) distribution with mean 0 and variance Σ, i.e., $MVN(0, \Sigma)$ where Σ is a P x P dimensional variance-covariance matrix for the random effects on P DCs measurements and each P has two random effects. We are looking at the correlation between the two random effect.

$$\begin{bmatrix} \epsilon_{1ip} \\ \epsilon_{2ip} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma \right)$$

and from this equation, we have

$$\begin{bmatrix} \beta_{0ip} \\ logA_{ip} \end{bmatrix} \sim \begin{bmatrix} \mu_{11} \\ \mu 12 \\ \mu 21 \\ \mu 22 \end{bmatrix} + MVN\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma \right) \tag{4.5}$$

where

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{1112} & \sigma_{1121} & \sigma_{1122} \\ \sigma_{1211} & \sigma_{12}^2 & \sigma_{1221} & \sigma_{1222} \\ \sigma_{2111} & \sigma_{2112} & \sigma_{21}^2 & \sigma_{2122} \\ \sigma_{2211} & \sigma_{2212} & \sigma_{2221} & \sigma_{22}^2 \end{bmatrix} \tag{4.6}$$

and $\sigma_{11}^2, \sigma_{1112}, ..., \sigma_{22}^2$ are the hyper-parameters from the var-covariance matrix.

4.2.5 Reliability Model

Let d_p be the threshold value of the degradation for the *pth* DC, then we define a soft failure as a failure that occurs when the degradation reaches a threshold value and denote the associated soft-failure time as T_p, where $p = 1, 2$. The reliability of the system at time

t is defined as

$$R(t) = Pr\{T_1 \geq t, T_2 \geq t\} = Pr\{Z_{i1gq} \geq d_1, \ldots\ldots, Z_{ipgq} \geq d_p\} \tag{4.7}$$

Where Z_{ipgq} is given in (4.2), $d_p > 0$.

The above equation (4.7) depends on the joint failure time distribution model for all the P DC measurements. Calculating a closed-form expression for $F(t) = 1 - R(t)$ might be easy in some cases, but in general, it is hard to get such a closed-form expression especially in some practical path models due to how hard it is to get the integration for random effect terms. Therefore, when more than one of the parameters is random, we have to use simulation method to evaluate $F(t)$.

4.2.6 Stan

Stan is a programming language based on C++, and it is used for Bayesian modeling and inference. It is used for statistical modeling and for fitting models that are efficient, robust and scalable. Given a user-specified model and data, Stan uses the No-U-Turn sampler (NUTS) [77] to obtain posterior simulations. For the estimation of model parameters when random effect is present, we describe the Hamiltonian Monte Carlo (HMC) which is a Markov chain Monte Carlo (MCMC) method. The techniques of the Stan's Markov chain Monte Carlo (MCMC) are based on the Hamiltonian Monte Carlo (HMC) and the HMC is a robust and more efficient sampler. Stan provides interfaces to other programming languages.

Stan can be used through R interface, through Python and command-line shell. In this paper, the R interface was used to run our model. Stan has the advantage of extensibility and flexibility, and the arbitrary target functions is supported. The compilation

of the program is written by Stan, and it then uses the available dataset to run the model thereby producing the posterior simulation process of the model parameters automatically. Advantage of Stan is its ability to analyze large dataset and complex model fast. From the Hamiltonian Monte Carlo (HMC) algorithm, the no-u-turn (NUTS) which is a Stan's built-in sampler is derived and it was first proposed by Hoffman and Gelman (2014) [12].

4.3 Statistical Inference

As a random process, Markov chains from Markov Chain Monte Carlo (MCMC) have the memoryless property which implies that it can only be influenced by the process current state and not by its past. The Monte Carlo part from MCMC is a computational algorithm that rely on repeated random samples to estimates the properties of a distribution to obtain numerical results. The motivation for using MCMC simulation is that the MCMC simulation approximate an uncompromising posterior distribution. With Bayesian approach using MCMC, the likelihood function conditional on the unobserved variables is only considered, which makes the computation faster. The MCMC method uses the derivatives of the sampled density function for the purpose of generating transitions efficiently across the posterior. Based on numerical integration, it uses an approximate Hamiltonian dynamics simulation, and this is corrected by performing a Metropolis acceptance step. Stan doesn't need more samples to get it's posterior result. We used the Stan package in R for the MCMC algorithm and we will describe the MCMC algorithm for the estimation of the parameter.

4.3.1 Stan Fundamental Parts

Stan code will be used to fit our model in (4.1) using RStan. RStan has a structure that consists of three fundamental parts. These three parts are the data block, the parameters block and the model block. Using our model in (5.1), we can implement this and write

this as follow in Stan:

$$y_{ipgq} \sim lognormal(Z_{ipgq}, \sigma_{rp}^2) \tag{4.8}$$

where Z_{ipgq} is defined in (4.2). The degradation model can then be rewritten as

$$y_{ipgq} \sim lognormal(\beta_{0ip} + \beta_{1ipg} t_{ipgq}^{\gamma p}, \sigma_{rp}^2) \tag{4.9}$$

The above model (4.9) can be written in the model block using Stan. Also, equation (4.5) and (4.6), will be written in the transformed parameters block before entering the model block.

4.3.2 Parameter Estimation

4.3.2.1 Estimation of Parameters

Here we want to discuss about the estimation of the model parameters. To estimate these parameters, we will need the degradation data together with the hierarchical degradation model discussed using (4.1). The vector of the model parameters Ψ_p which was discussed in the previous chapter will be estimated and this vector includes the fixed effects and the random effects. Each of the parameters will be assigned a prior distribution to obtain an estimate and this is according to the Bayesian approach. The approach used by the Bayesian is as follows: Given a likelihood function of some data, Bayesian method will introduce a probability distribution to an uncertain parameter and update the parameter estimate based on the newly introduced information. For the frequentist approach, probabilities will not be assigned to any parameter value instead, it uses the point estimates of unknown parameter to predict the new data point.

4.3.2.2 MCMC Sampling Technique

Let $y_{ipg} = (y_{ipg1}, ..., y_{ipgw_i})^{'}$, $y_{ip} = (y_{ip1}^{'}, ..., y_{ipG}^{'})^{'}$, and $y_i = (y_{i1}^{'}, ..., y_{iP}^{'})^{'}$. Then $y = (y_1^{'}, ..., y_n^{'})^{'}$ serves as the vector of degradation measurements for all n test units. We denote $\Gamma = (\Psi_1^{'}; \sigma_{11}^2, ..., \Psi_P^{'}; \sigma_{2P}^2)^{'}$ as the model parameters' vector. The probability density function (PDF) of a standard normal distribution is given as:

$$f(\Gamma) = \Phi_{normal}\left[\frac{logy_{ipgq} - \beta_{0ip} - \beta_{1ipg}t_{ipgq}^{\gamma_p}}{\sigma}\right]$$

For the multivariate normal distribution, the Probability Density Function (PDF) is given as

$$f(\beta_{0ip}, logA_{ip}) = \frac{1}{2\pi|\Sigma|^{1/2}}exp\bigg(-\frac{1}{2}(\beta-\mu)^T\Sigma^{-1}(\beta-\mu)\bigg)$$

where

$$\beta = \begin{bmatrix}\beta_{0ip}\\ logA_{ip}\end{bmatrix}$$

and

$$\mu = \begin{bmatrix}\mu_{1p}\\ \mu_{2p}\end{bmatrix}$$

with Σ given in (4.6)

In other to get the marginal likelihood, we develop the degradation path, Z_{ipgq}, as a function of random effect coefficients, β_{0ip} and $\log A_{ip}$, to be

$$f(y_i|\beta_{0ip}, logA_{ip}; \Gamma) = \frac{1}{\sqrt{2\pi}\sigma}exp\bigg(-\frac{1}{2\sigma^2}\bigg(logy_{ipgq} - \beta_{0ip} - \beta_{1ipg}t_{ipgq}^{\gamma_p}\bigg)^2\bigg)$$

If we integrate the two random effects $(\beta_{0ip}, logA_{ip})$ out, we will have the marginal likelihood to be

$$L_1(\Gamma, \Sigma|y; \beta_{0ip}, logA_{ip}) = \int \int f(y_i|\beta_{0ip}, logA_{ip}; \Gamma) * f(\beta_{0ip}, logA_{ip}) \ d\beta_{0ip} \ dlogA_{ip} \quad (4.10)$$

Drawing from a density $f(\Gamma)$ for parameters Γ is simply the goal of sampling and this is typically a Bayesian posterior $f(\Gamma|y)$ given data y.

4.3.2.3 The Hamiltonian and The Auxiliary Momentum

Hamiltonian Monte Carlo (HMC) is a method of MCMC and it was applied to lattice field theory simulations of quantum chromodynamics by Duane, et al [59]. HMC create efficient developments spanning the posterior by using the derivatives of the sampled density function. The application of HMC in statistics started in 1996 by Neal and it was applied to neural network models [85]. There have been other applications of HMC to statistical problems (e.g., [51], [89]).

As stated earlier, Stan was used for the MCMC algorithm and like most other HMC implementations, Stan uses the leapfrog integrator. In other to give a stable result for Hamiltonian systems of equations, leapfrog integrator which is a numerical integration algorithm was used. The leapfrog algorithm takes discrete steps of some small-time interval ϵ and using the leapfrog integrator with number of steps L and discretization time ϵ for a given number of iterations, the HMC algorithm starts by sampling a new momentum and the new parameter value Γ is updated according to Hamiltonian dynamics.

In Stan, some interfaces set the step size ϵ and an approximate integration time t. Using the no-U-turn sampling (NUTS) algorithm [12]. Stan is able to automatically optimize ϵ to match an acceptance-rate target using the dual averaging [128]. This warmup opti-

mization procedure is extremely flexible and using the notation of Hoffman and Gelman in [12], Stan exposes each tuning option for dual averaging for completeness.

4.3.3 Posterior Analysis

For full Bayesian inference, Stan uses Markov chain Monte Carlo (MCMC) methods to generate samples from the posterior distribution. At different positions in the chain, the transition probabilities of the Markov chain do not change so that for u, $u^{'} \geq 0$, the probability function $f(\Gamma^{(u+1)}|\Gamma^{(u)})$ is the same as $f(\Gamma^{(u^{'}+1)}|\Gamma^{(u^{'})})$. The target density $f(\Gamma)$ defined by a Stan program is the equilibrium distribution $f(\Gamma^{(u)})$ and this is typically a proper Bayesian posterior density $f(\Gamma|y)$ defined on the log scale up to a constant.

4.3.4 Prior Distribution Used

To fit degradation model for this analysis in Stan, scaling of the dataset should be done because the features vary in units, magnitudes, and range. This should be followed by pre-processing to scale the dataset because Euclidean distance between two data points is being used by most MCMC algorithms to bring all features to the same level of magnitude. With the synthetic ISO data, the temperatures and relative humidity were scaled using Min-Max Scaling [3] so that the value will be between 0 and 1. With Stan, we can generate MCMC posteriors draws for each parameter.

For the model parameters, denoted as Q_p, each of these parameters $\mu_{1p}, \mu_{2p}, B_p, H_p$, and γ_p follows Normal (0,10). From (4.6), we see that Σ is a covariance matrix and since it has a scale and a location, prior distribution will be assigned to it. The only way to do the assigning is to assign the prior distribution to a correlation matrix instead. This is because the correlation matrix is a standardized version of a covariance matrix. Lewandowski, Kurowicka, and Joe [27] which is shorten to LKJ, developed a correlation called LKJ prior (LKJcorr) and this can be assigned to the correlation matrix. Having Σ to be a symmetric

matrix where the unit is diagonal and positive-definite, $(\Sigma|\omega) \propto det(\Sigma)^{(\omega-1)}$, where ω is the shape parameter and ω can be 1 ($\omega = 1$), or $\omega > 1$, or $0 < \omega < 1$. For the LKJ prior, values closer to 1 are less skeptical of strong correlations (-1, +1), and higher values (e.g., 2, 4) are more skeptical of strong correlation coefficients. An implicit parameterization of the LKJ correlation matrix density was provided by Stan in terms of its Cholesky factor and this is denoted as *lkj_corr_cholesky* in Stan. For example, given $S \sim lkj\_corr\_cholesky(1.0)$, this implies that $S * S^{'} \sim lkj\_corr(1.0)$.

For the random effects parameters, the prior distribution is assigned as follows:

$$\Sigma_* = \begin{bmatrix} \sigma_{11}^2 & 0 & 0 & 0 \\ 0 & \sigma_{12}^2 & 0 & 0 \\ 0 & 0 & \sigma_{21}^2 & 0 \\ 0 & 0 & 0 & \sigma_{22}^2 \end{bmatrix} \Omega \begin{bmatrix} \sigma_{11}^2 & 0 & 0 & 0 \\ 0 & \sigma_{12}^2 & 0 & 0 \\ 0 & 0 & \sigma_{21}^2 & 0 \\ 0 & 0 & 0 & \sigma_{22}^2 \end{bmatrix} \tag{4.11}$$

where $\sigma_{11}^2, \sigma_{12}^2, \sigma_{21}^2, \sigma_{22}^2$ follows Exponential(1.0) and Ω follows $lkj\_corr\_cholesky(2.0)$.

This sign Ω which follows $lkj\_corr\_cholesky(2.0)$ are copies of the same diagonal matrix containing variances on the diagonal. If we multiply eqn (4.11) together, we will get a covariance matrix.

4.3.5 Stan Convergence

In Stan, to know if your MCMC chain converges or if the posterior draws are stationary distributed, you must do a diagnostic test. This diagnostic test includes checking the trace plot of the MCMC chain, checking the density plot, checking for the Rhat value, and checking the Autocorrelation Function Plots (ACF). We will talk about each of these diagnostic tests as follows.

MCMC Trace plot is a time series plot of the Markov chains. The x-axis of the trace plot represents time, and the y-axis represents the posterior values of the draws. If the sample drawn from MCMC chains becomes stationary, then we say the trace plot is well mixed and there are no apparent anomalies. MCMC chains can be stationary if there is an increase in the number of warm-up period (burn-in period). MCMC chains can be iterated more often by increasing the number of iterations to improve the trace plot. By doing this, you are increasing the sample size drawn from the MCMC run and this will allow the chains to explore the sample space many times. Figure 4.2 shows the trace plot of some of the parameters.

Another way to know whether the MCMC simulation chains have been reached for a stationary state and have already converged, is by plotting the density Plot. This plot has a bell shape among all the parameters. Figure 4.3 is the density plot for some parameters.

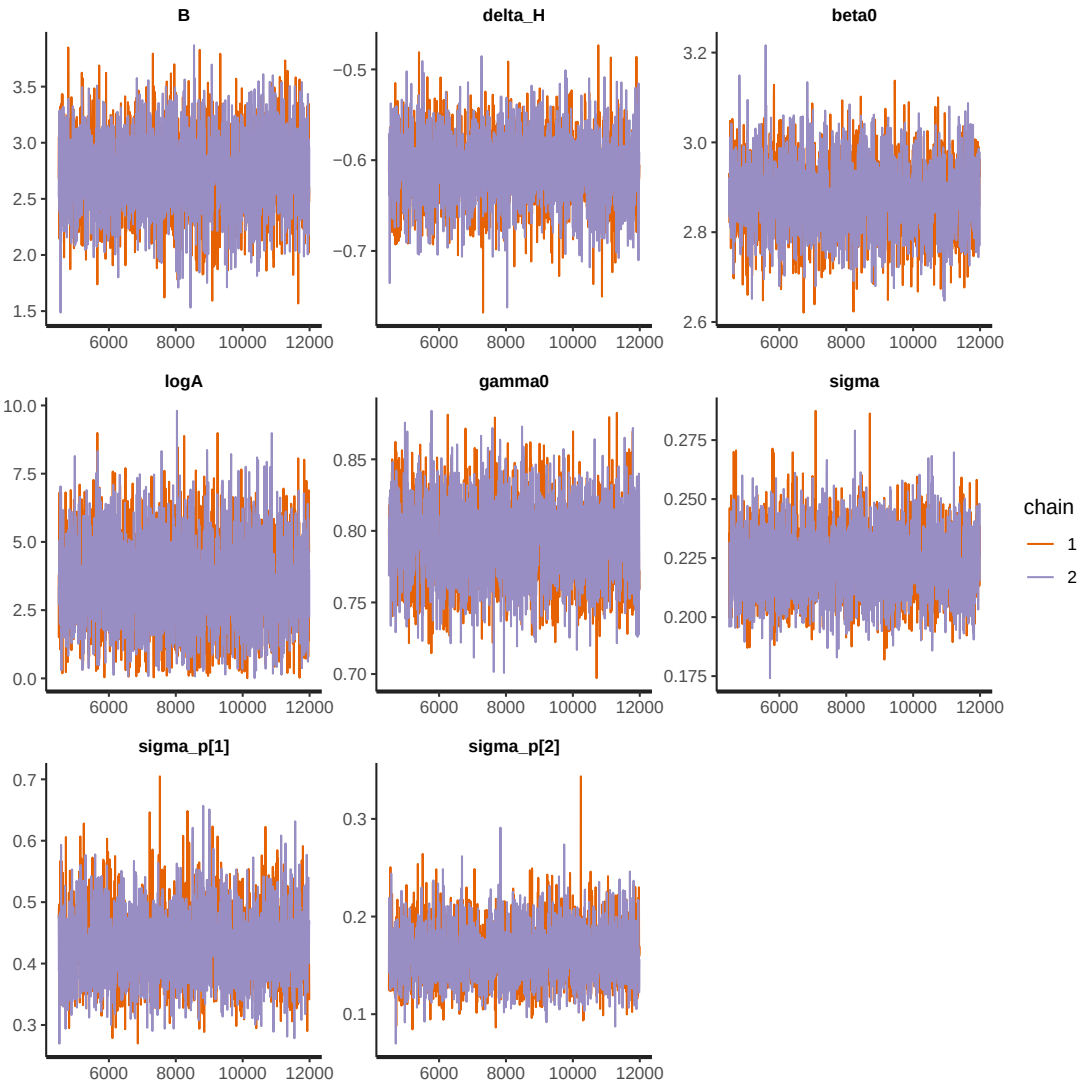


Figure 4.2: Trace Plots for Parameters that Converged

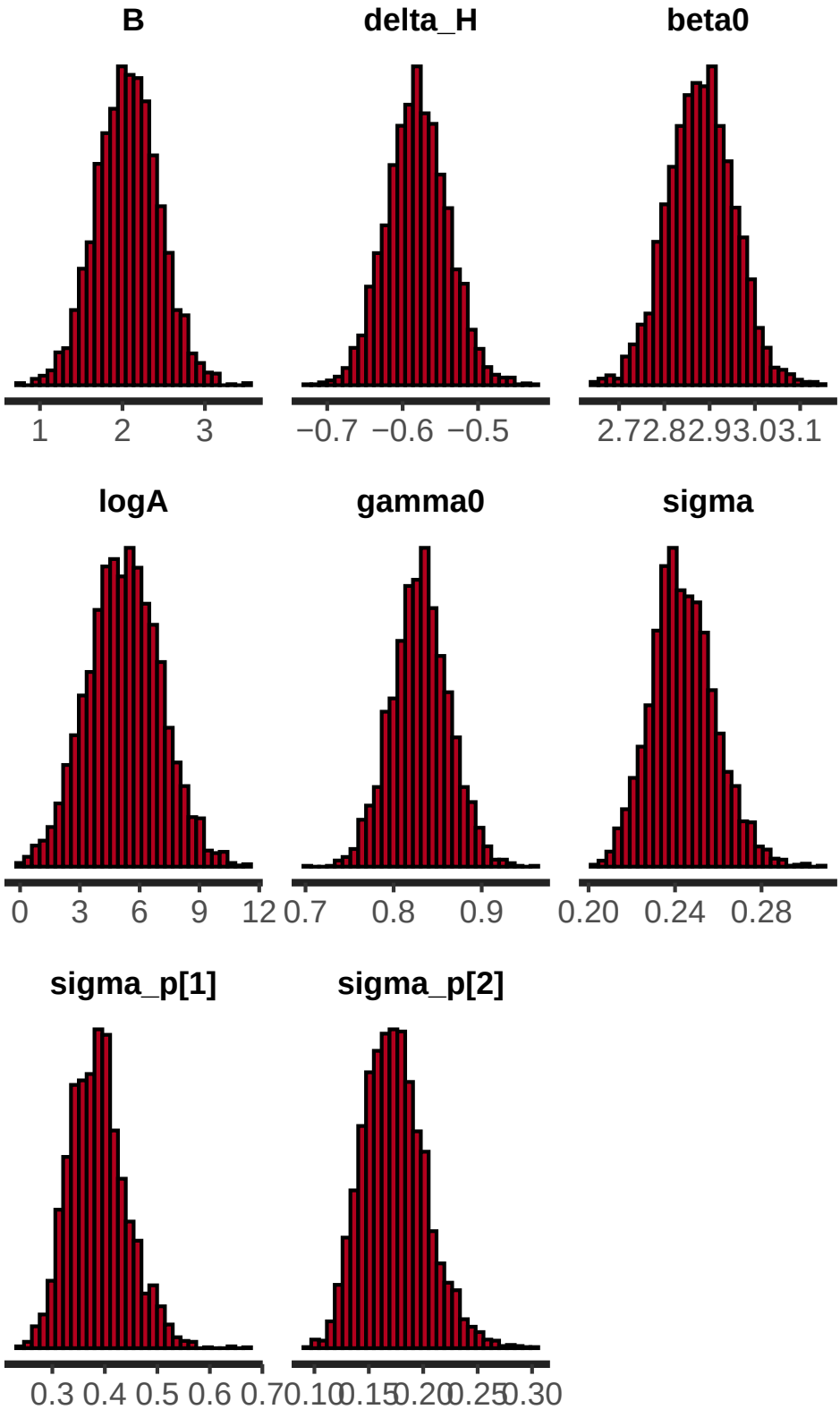


Figure 4.3: Density Plots for Parameters that Converged

Other way to show the convergence of the MCMC chain (diagnostic) is by looking at the R-hat value. R-hat compares the between and within chain estimates for model parameters and other univariate quantities of interest. If R-hat is greater than 1, then the MCMC chains have not mixed well. It is recommended to use R-hat that is less than 1.05.

To explain the relationship between lags and autocorrelations, we will need the Autocorrelation Function Plots (ACF). Using ACF plot is also another way to check for convergence. Using this plot, if the autocorrelation reduced quickly from lag 1, then the MCMC chain has converged. A good ACF plots should show that at large autocorrelation, there is a short lag, but the autocorrelation goes to zero quickly as the lag increases. Thinning the MCMC chain will help to improve the ACF plots. Figure 4.4 is the ACF plot for some parameters.

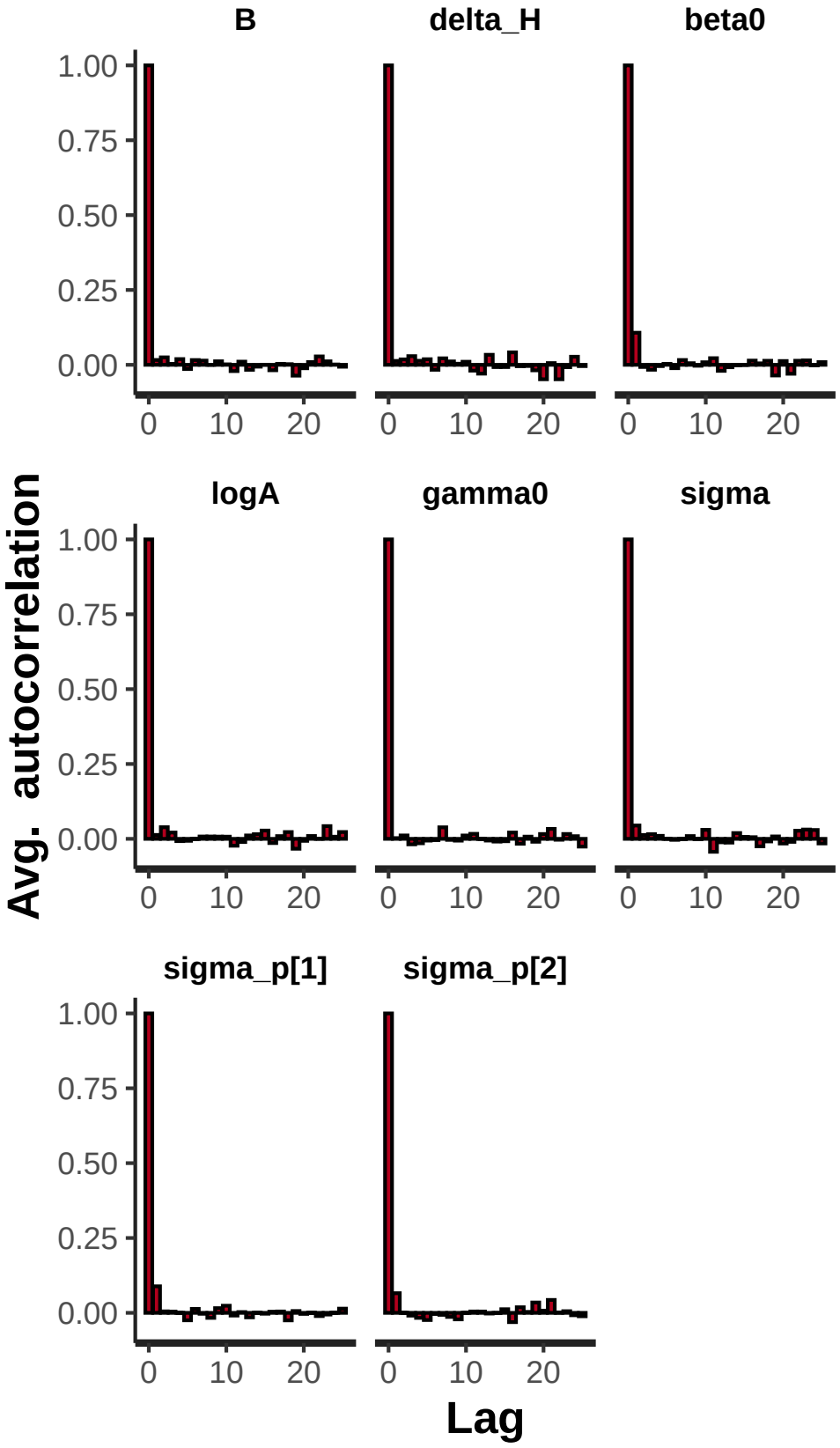


Figure 4.4: ACF Plots for Parameters that Converged

4.3.5.1 Monte Carlo Simulation to Draw Degradation Paths

F(t) can be evaluated by using Monte Carlo simulation in most practical reliability cases and the idea is this, from an assumed degradation path model, a large number of degradation paths are generated after which we use proportion of the path which crosses the threshold d_p at each time t to evaluate R(t). The algorithm to do this is as follow:

1. From Stan, given a large number of A, generate A simulated realizations of the parameters.

2. For each path, find the number of crossing times at all time points based on the computed failure time for A realization of the parameters based on the simulated failure time

3. Whatever t value we desired, use

   $F(t) \approx \frac{Number\ of\ First\ Crossing\ Times\ of\ the\ Simulation \leq t}{A}$ to evaluate F(t). For each time point t, we can generate $i = A$ sample draws with 500 degradation paths using the above method.

4.3.6 Procedure For Plotting the Reliability Curve

The term reliability is the probability that an object of interest or a system will work under some operating conditions for any specified time. Using the optical media dataset, this dataset has short lifespan of 30 years but from previous study, this can be extended to between 30 years and 300 years. To fit the reliability model of the Optical media dataset, the parameter estimation derived using Stan will be used. In this work, a more detailed prediction will be done to provide summaries of the reliability together with its uncertainty bounds as this will help companies and manufacturers. The detailed procedures on how to plot reliability plot is given below: Using the results obtained from Stan and using $j = 500$ degradation paths for each $i = 3000$ MCMC draws with $k = 259$ time

interval, we will summarize over 500 paths within each draw over time. After this, the proportion of the number of degradation measurements that are greater than $log(280)$ (threshold value) at each time point will be taken as the proportion of the failure rate at each time point. With the result, the failure rate curve for each draw will be plotted, and we will have 3000 different curves for failure rate in total. Since $R(t)$ reliability function has a relationship with $F(t)$ failure rate function i.e., $R(t) = 1 - F(t)$, we can then have the projected reliability curve with its 95% credible interval.

4.4 Simulation Settings

In this section, we are going to simulate a new dataset under the same data structure of ISO and under the same problem setting so as to assess the performance of the proposed methodology. Having a normal used condition at 50% for RH and 25^0C for temperature, we aim to predict reliability at this normal used condition. The ISO data have just one DC measurement. Two DC measurements was created synthetically to create a multivariate model. Two random effects are used in our model and since we have two datasets, then the correlation between these two datasets will be 4 x 4 matrix. For each of the two degradation measurements, if any of these two-fail using the degradation thresholds of $log(280)$, then the material will fail. In our simulation study, as given in the ISO data, for the simulation, we choose this setting of an equal sample size of ten (i.e., $n = 10$) with four ($b = 4$) test conditions. This implies that we have 2n test units in total at each Temperature and relative humidity level. For each sample size, we simulated two degradation measurements with the two datasets we have, using the multivariate degradation model given in (4.1). Also in our study, we used $c = 5$ repeated measurements as seen in the ISO data with time points that were measured and equally spread throughout the duration of the test. A varying correlation level at $Corr(v_p; v_p^{'}) = r = 0, 0.9$ for all $p \neq p^{'} \epsilon\ 1, 2$ was used to understand the effect of using low and high correlation, and we allowed the same correlation coefficient between the two DCs. We did this in other to

assess the performance across the two levels.

By using just one sample size level of ten ($n = 10$) given in the ISO data, one repeated measurement level of five ($c = 5$), one ($b = 4$) test conditions level of four and two ($r = 0, 0.9$) correlation levels that we are using for comparison, we evaluated 2 different scenarios. For each of these scenarios, we simulated $A = 100$ data sets from the model (4.1). To estimate the model parameters, the MCMC approach using the Stan package was used for each dataset that we simulated. In addition, at the normal use condition of 50% RH and $25^0 C$ Temperature, we estimated the reliability as a function of time by applying the method explained in section 4.3 and for each simulation scenario, we summarized the estimates across $A = 100$ simulated data sets. This helps to evaluate the performance based on accuracy and precision. With the two simulation settings, we will know the performance of the method used in this study across the two different choices in the correlation level. The independent degradation model which has no correlation between the two multiple DCs will be compared with the multivariate degradation model which uses the correlation of the two DCs.

4.5 The Performance of the Model Estimation

For all the model parameter estimates, for all $A = 100$ simulations, the root of mean squared error was calculated to measure how precise the estimated model parameters are across multiple DCs. Let β_p be the coefficient parameter of Relative Humidity (RH) for pth DC measurement, the root mean square error for $\hat{\beta}_p$ is calculated as, RMSE($\hat{\beta}_p$) = $\sqrt{\sum_{a=1}^{A} \frac{(\hat{\beta}_{pa} - \beta_p)^2}{A}}$ where $\hat{\beta}_{pa}$ is the estimate of $\hat{\beta}_p$ using the MCMC algorithm (using Stan) from the ath simulation. We use the average root of mean square error $Ave.RMSE(\hat{\beta}_p) = \frac{\sum_{p=1}^{P} RMSE(\hat{\beta}_p)}{P}$ to quantify the average precision across the P DC measurements. Figure 4.5 below is the Ave.RMSE error of $\hat{\beta}_p$'s for the coefficient of delta_H and the measurement error. In this figure, we compare both the independent degradation model which has no

correlation between the two multiple DCs and which also estimates parameters by separate models with the multivariate degradation model given in (4.1). In this figure, the left plot uses the multivariate model with $r = 0$ and 0.9 and the right plot uses the independent model with $r = 0$ and 0.9. The first two points on the left panel plot are the points of the Ave.RMSE of delta_H for $r = 0$ and 0.9 respectfully and the last two points on the left panel plot are the points of the Ave.RMSE of sigma (σ) for $r = 0$ and 0.9 respectfully using multivariate model.

With these plots, we observed some few patterns. First, for different correlation levels, the average RMSE of each estimated $\beta(delta\_H, \sigma)$ does not change, and this implies that the correlation among the two DCs with the estimated model coefficient has not much significant impact on its accuracy. Also, using both the independent and multivariate models, there is similarity between the Ave.RMSE of estimated β and this implies that if correlation is not present among the two DC measurements, the bias of the coefficient of the parameter estimate will not increase. This same pattern was observed for other parameters such as the logA, B and β_0 except for Σ.

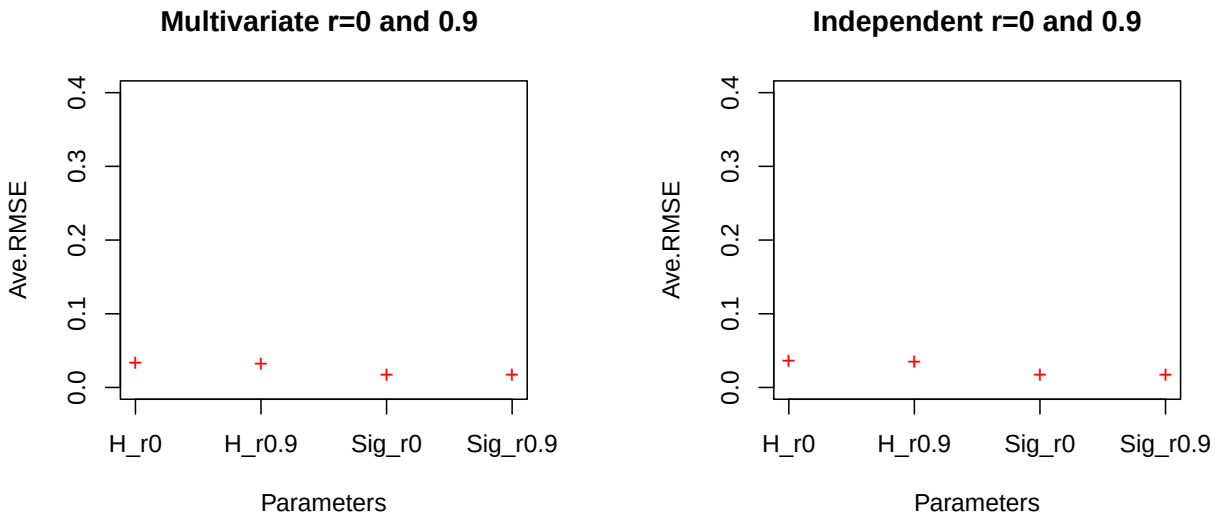


Figure 4.5: RMSE plot for the average of model parameter β_p across the two degradation characteristics using the independent together with the multivariate degradation models.

Figure 4.6 is the RMSE plot for the average of all the units (i.e. the variance with the covariance) in Σ over the two degradation characteristics for different correlation choices when the independent together with the multivariate degradation models is being used. We noticed that when we use a multivariate degradation model with different correlation levels, the average RMSE of that estimate remain unchanged. This implies that the accuracy for the estimated variance-covariance of the random effects will remain the same even if we use the right model that captures the correlation structure regardless of the size of the correlation. On the other hand, in contrast to other parameters, when we use the independent degradation model for degradation characteristic measurements that are correlated, there is an increase in the average RMSE of the estimated Σ as the correlation level increases.

Furthermore, compare to the multivariate model with $r = 0$ correlation among multiple DC measures. For independent model, we see that the Ave. RMSE is smaller. This implies that the independent model for when $r = 0$ produces little more accurate esti-

mates of *Sigma* compared to the multivariate model for when $r = 0$. On the other hand, when the correlation increases i.e., $r = 0.9$, we notice that the average RMSE of the multivariate model produces more accurate estimate of Σ compared to the independent model. For example, we can see that for multivariate model, when correlation is 0.9 ($r = 0.9$), the Ave.RMSE is around 0.1 and when correlation is 0.9 ($r = 0.9$) for independent model, the Ave.RMSE is around 0.38. Hence, we can see in the estimated correlation parameter that there is huge difference between the multivariate and the independent model when there is high correlation. Independent model is 4 times bigger in the estimated correlation coefficient than the Multivariate model. Considering the multivariate structure, it will allow us to get much more precise estimation of the correlation.

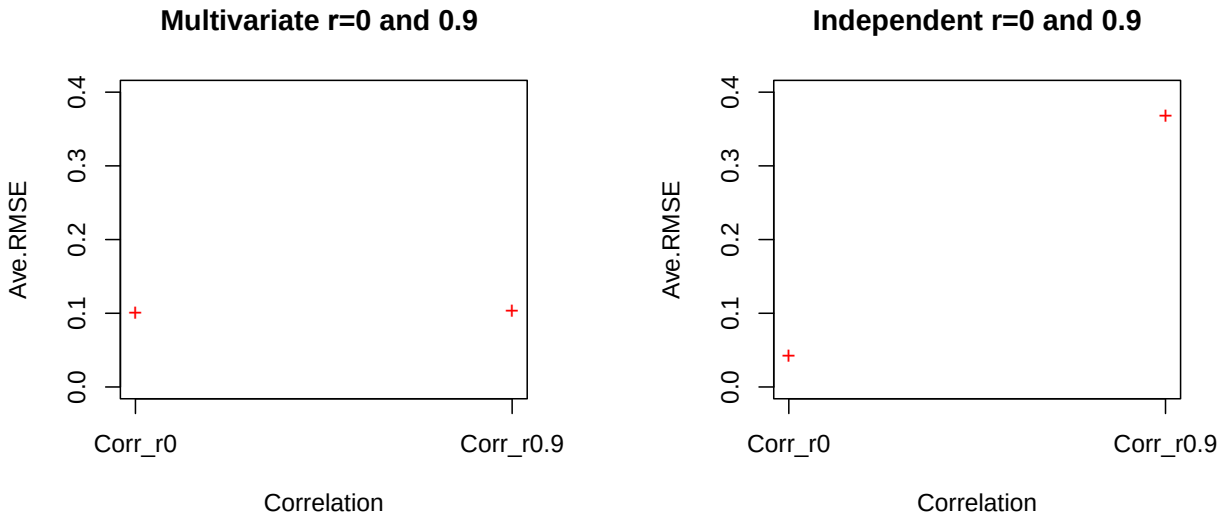


Figure 4.6: RMSE plot for average of all the units in Σ over the two degradation characteristics when the independent together with the multivariate degradation models is being used.

4.6 Reliability Estimation Comparison

In this section, at 50% RH and 25^0C Temp. normal use condition, we calculate the predicted system reliability based on the method described in the previous section and we compared the results between the independent degradation and multivariate models

using different correlation values. Figure 4.7 shows the predicted reliability for both the multivariate and independent models and for individual DC measurements using the two simulated datasets with $n = 10$ test units, $m = 4$ repeated measurements for each test unit. This plot is to show how related the estimated system reliability of the independent model is with the multivariate model. From this plot, the system reliability that was predicted using the multivariate model is the olympic curve that is solid, the apricot and violet dash curves correspond to reliability using the first data and second data respectively and the independent model is the solid red curve.

From Figure 4.7, the time point is in log scale, and this means we need to convert it back to hours. By comparing reliability functions of the two datasets, the reliability using the 1st data is higher than the reliability using the 2nd data before time point 12.40 which is 242,801 hours (30 years) and then becomes lower than the other 2nd data after 242,801 hours. Given the two DC measurements, the DC measurements that fail first among the two DC measurements will jointly determine the reliability of the system. The reliability curve then observes the two DC measurements with the lowest reliability function. From the plot, we see that the independent model from the predicted system reliability was lower than each of the individual reliability functions. This is because the independent model was derived from the product of the individual reliability functions. Looking at the independent model from our plot, we noticed that the reliability of the system is underestimated because the interdependence of the two DC measurements is not accounted for.

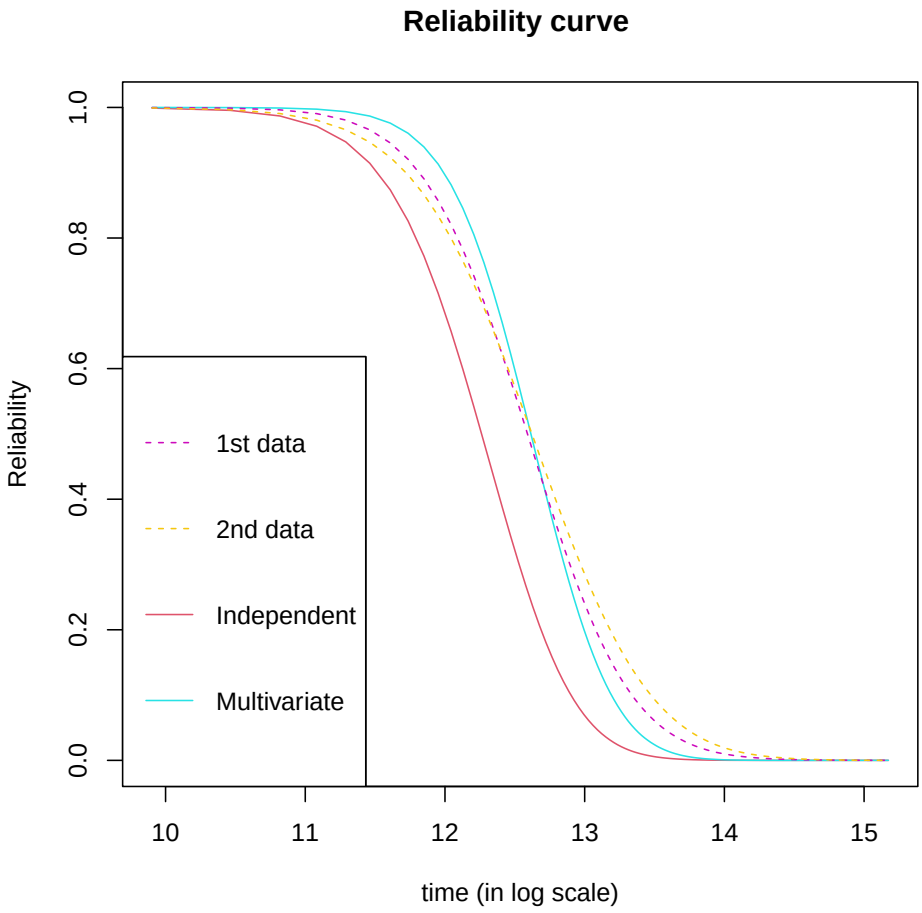


Figure 4.7: The reliability plot functions predicted at normal use condition using a single simulation for individual degradation characteristics measurements and systems utilizing an independent together with a multivariate degradation models at n = 10, r=0.9.

We want to predict the reliability of the system at the normal use condition at 50% RH, and 25^0 C Temp. Figure 4.8 is the plot for comparing the reliability estimate based on a single simulation scenario. Using the multivariate model (point estimate is the solid blue curve and it's 95% CI is the dash blue curves) together with the independent model (solid red curve and dash red curves for it's 95% CI), we compared the predicted reliability. This plot is showing the range where the reliability drops from 1 at 2042 days to 0 at 82,614 days and it's 95% pointwise CI.

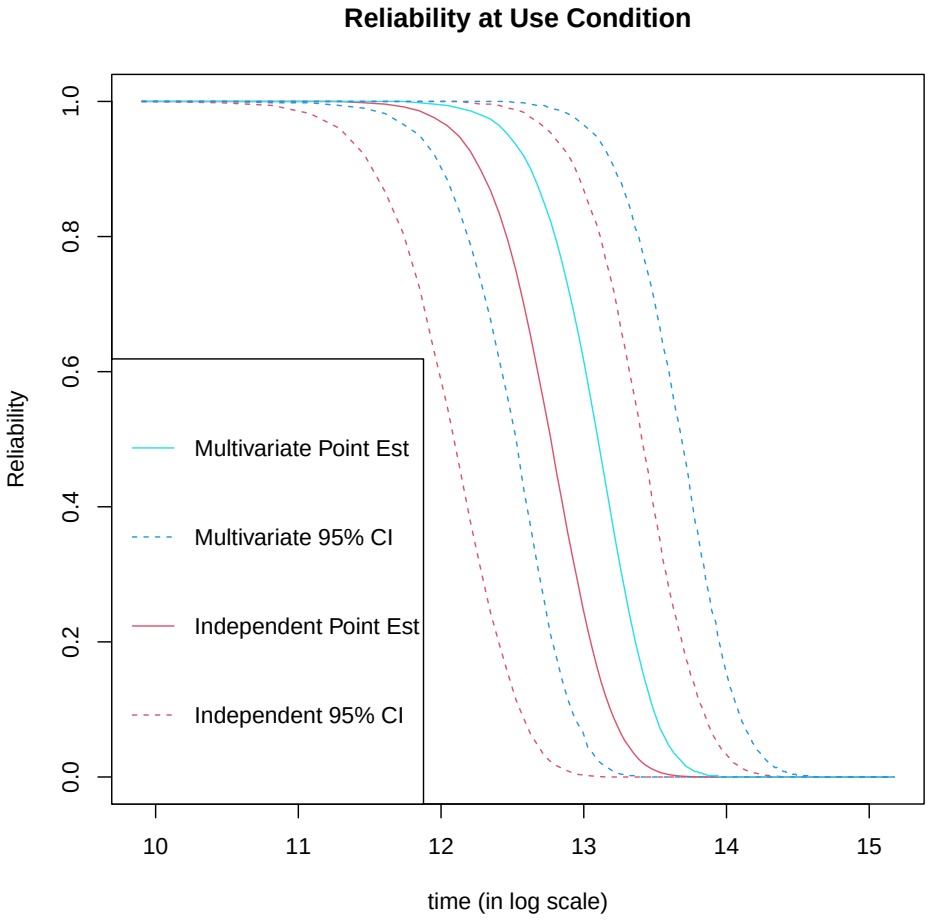


Figure 4.8: Comparing the reliability between an independent and multivariate degradation model at a normal condition of 50% RH, and 25^0 C Temp using the reliability curves together with their 95% confidence intervals based a single simulation.

The general pattern here is that the red is to the left. So that means that the independent model are under-estimating the reliability i.e. the predicted reliability of the independent model is smaller than that of the multivariate model. There is similarity in the precision of the predicted reliability when the two models are used. Therefore, the independent model predict the earlier failure of the system than when we use the multivariate model. Due to this, a lot of resources can be saved using the multivariate degradation model because it avoids taking unnecessary early action when this information is used for system management.

Figure 4.9 is the RMSE plot of the reliability using multivariate and the independent models respectively for r (two) correlation levels. The first two plots at the top are the plots of when there is no correlation ($r = 0$) for both multivariate and independent model while the lower two plots are when correlation is high ($r = 0.9$) for the two model. From the two plots at the top panel, we can see that when there is no correlation, the two methods perform very similarly. Also, for $r = 0.9$, from the two plots at the bottom panel, we see that the multivariate model is lower than the independent model. So, ignoring the correlation will lead to severely inflated RMSE.

For example, we see that the RMSE for the independent model for $r = 0.9$, is 0.3 while the RMSE for the multivariate model for $r = 0.9$ is 0.18. This implies that the multivariate degradation model predicted more accurate reliability using the RMSE than the independent model for when $r = 0.9$. Also, as the correlation value i.e. r increases for independent model, we observe an increase in the predicted reliability RMSE. The plots show that there is an increase in the RMSE value for when $r = 0.9$ for the independent model compared to when $r = 0$. From this plot, we also noticed that for the first 30,000 hours, the predicted reliability RMSE is close to 0 and as it reaches its maximum at 370,000 hours, it increases when multivariate degradation model is used. After this, it starts dropping to around 1,200,000 hours. This looks like what we observed in. Figure 4.7 where the disparity between the multivariate (olympic) and independent (red) models continues to increase between 30000 [time point 10.30 in log scale] and 370000 hours [time point 12.8 in log scale] then after that it starts to decrease until it reaches 0 at 1,200,000 hours [time point 14 in log scale]. In summary, when high correlation is being used, the predicted reliability accuracy is improved by the multivariate model than when we use the independent model.

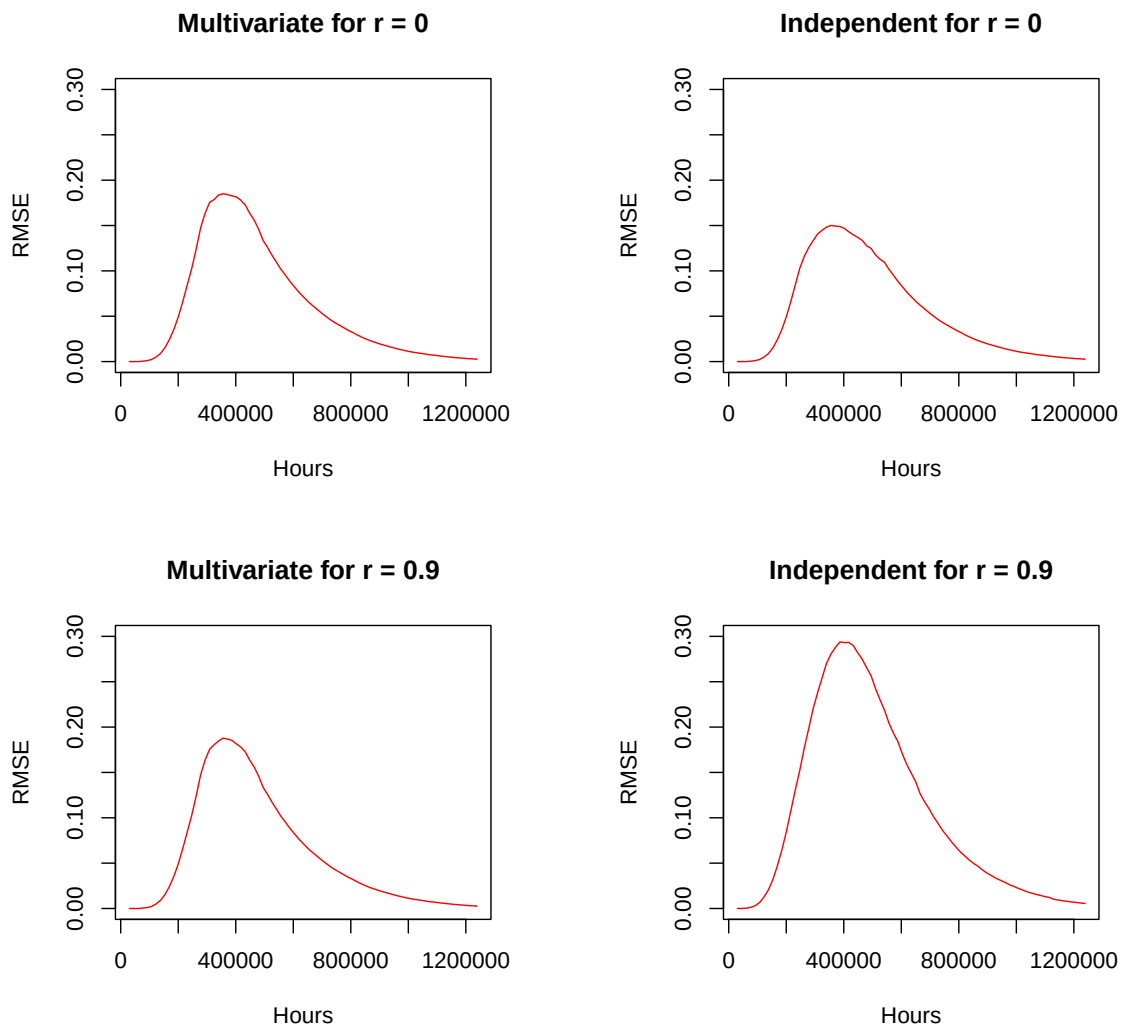
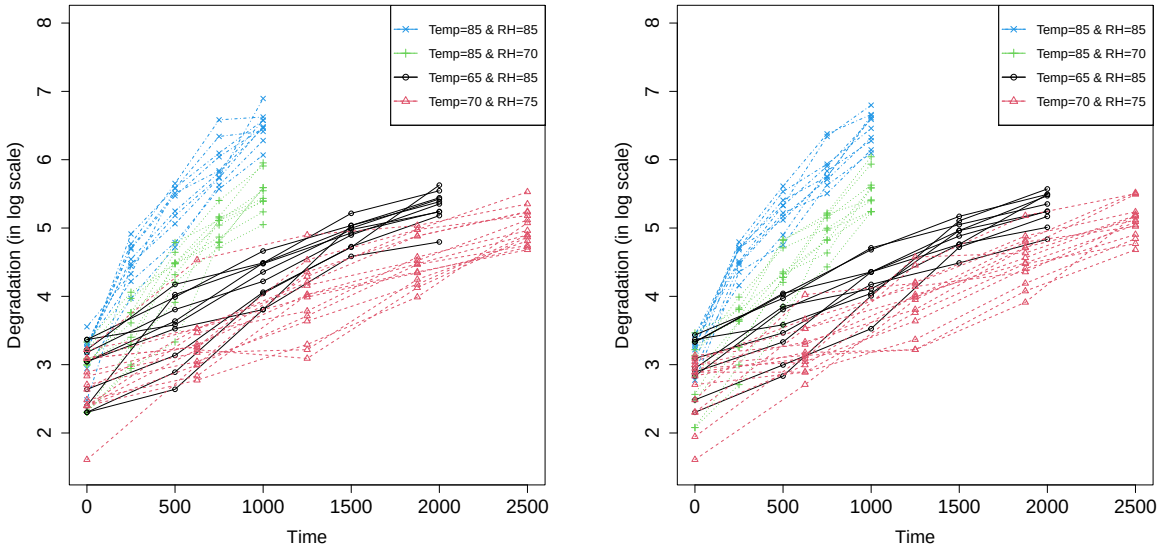


Figure 4.9: A plot that describes the RMSE under normal use condition for the predicted reliability when using an Independent with multivariate degradation models.

4.6.1 Application to ISO 10995:2011 Dataset

In this chapter, we used the ISO 10995:2011 dataset shown in Figure 4.2 and 4.3 to illustrate our proposed method by fitting the multivariate nonlinear degradation model given in (4.2). The MCMC method (The Hamiltonian Markov chain (HMC)) described in section 4.3 from Stan package was used to estimate the model parameters. The original ISO data have only one DC measurement. By grouping measurements from the independent sample units, we generate the two DC measurements and because of this, the correlation

between the two DCs using the multivariate and the independent models should be small thereby producing the same results. The threshold of the failure is log (280) for $p = 1, 2$, and for this device, the normal use condition is at 50% for RH and 25^0C for temperature. The summary of the multivariate and independent parameter estimates and their standard errors can be seen in Table 4.4.



(a) DC 1 (b) DC 2

Figure 4.10: ISO degradation data plot with artificial DCs for 45 units.

Due to grouping method which was done manually, we can see that the estimate covariance and the parameter values that we got using the independent model are similar to the multivariate model. Using two DCs given in Figure 4.11, the curves of the 95% pointwise CIs of the estimated reliability is given. There is high reliability before 345,000 hours which drops gradually to 0 at around 420,000 hours. The multivariate and independent model of the predicted reliability are similar to each other by looking at the plot in Figure 4.11.

Table 4.4: Estimated parameters and their standard errors (in parentheses) when multivariate and independent models for ISO data are being used

Multivariate Model					Independent Model			
DC	1st dataset		2nd Dataset		1st dataset		2nd Dataset	
beta0	2.859 (0.0010)		2.876 (0.0010)		2.857 (0.0011)		2.871 (0.0011)	
delta_H	-0.613 (0.0006)		-0.617 (0.0005)		-0.614 (0.0006)		-0.625 (0.0005))	
logA	3.571 (0.0202)		4.245 (0.0198)		3.573 (0.0206)		4.260 (0.0222)	
B	2.749 (0.0054)		2.623 (0.0054)		2.748 (0.0054)		2.627 (0.0058)	
gamma0	0.801 (0.0004)		0.824 (0.0004)		0.797 (0.0004)		0.822 (0.0004)	
σ	0.245 (0.0002)		0.243 (0.0002)		0.245 (0.0002)		0.244 (0.0002)	
Σ	beta0[1]	beta0[2]	logA[1]	logA[2]	beta0[1]	logA[1]	beta0[2]	logA[2]
beta0[1]	0.193 (0.0008)	0.051 (0.0019)	-0.050 (0.0018)	-0.003 (0.0025)	0.196 (0.0008)	-0.055 (0.0012)	0.201 (0.0008)	-0.060 (0.0017)
beta0[2]	0.051 (0.0019)	0.198 (0.0007)	-0.013 (0.0025)	-0.053 (0.0020)				
logA[1]	-0.050 (0.0018)	-0.013 (0.0025)	0.036 (0.0004)	0.017 (0.0022)	-0.055 (0.0012)	0.038 (0.0005)	-0.060 (0.0017)	0.039 (0.0005)
logA[2]	-0.003 (0.0025)	-0.053 (0.0020)	0.017 (0.0022)	0.036 (0.0004)				

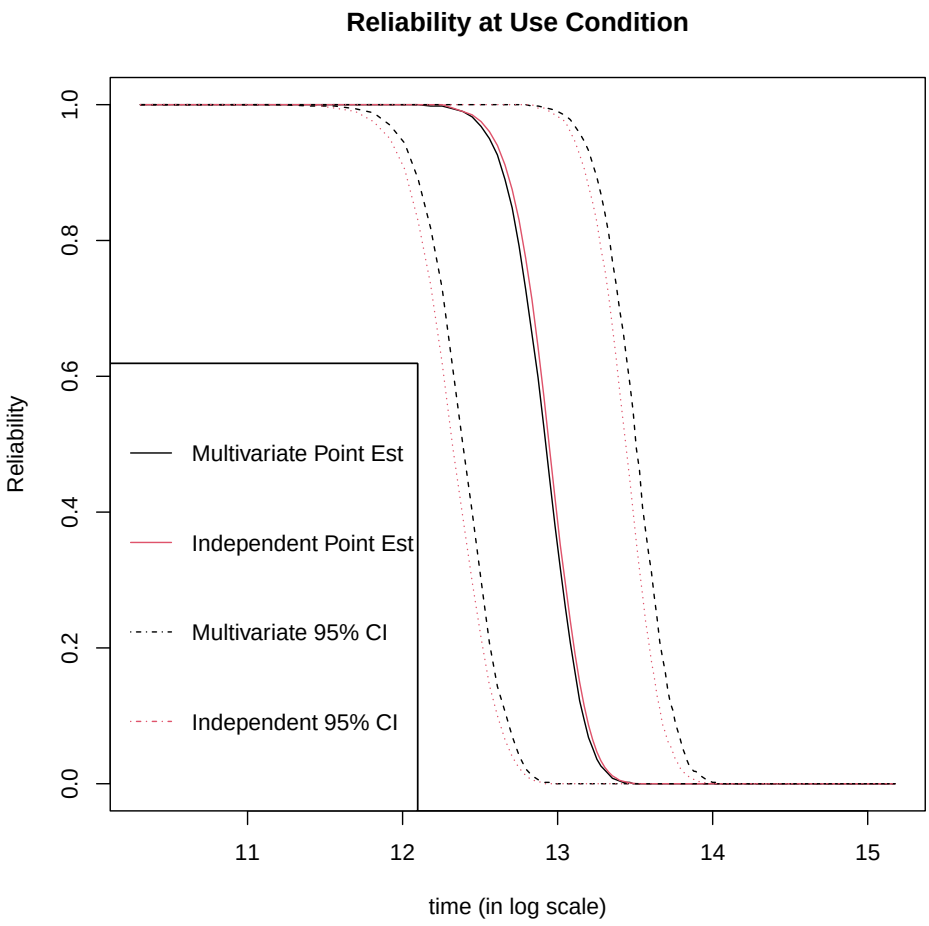


Figure 4.11: Curves of the predicted reliability and their 95 percent confidence intervals of Multivariate and Independent degradation models at $Temp = 25^0 C$, RH=50% normal use condition of two DC's.

Chapter 5: Penalized Regression for Survival Analysis

Stepwise regression [8] and subset selection [4] have been broadly used to choose important variables and improve predictability. [114] carried out a procedure for stepwise regression analysis and [13] discussed about variable and subset selection. Although they are practically useful in many applications, these selection procedures ignore inherited stochastic errors, and their theoretical properties are not fully understood. To achieve an adequate smaller subset of important variables will require searching through subsets of potential predictors and doing this can be unstable (Having q which is a minimum least-squares predictor in a collection of predictors. If a small change in the data used to derive the sequence of q can cause large changes in q, then we say the procedure is unstable) [72]. To avoid this drawback, penalized regression methods have been developed in recent years that perform subset selection.

Based on the dimension of microarray data (high dimension), partial least square was introduced by some researchers to reduce the dimension [117]. To utilize penalized regression, an optimized set of guidelines has been published to deal with gene expression data [21]. Some authors have proposed different penalties to remove the biasness in the selection of features in their model. To get less biased regression coefficients in sparse model (some coefficients being exactly zero) and to have a consistent variable selection by reducing bias in LASSO, Minimax Concave Penalty (MCP) penalty was proposed by Zhang [24].

Also, an alternate to the LASSO penalty with less biased estimates for nonzero regression coefficients was proposed by Jianqing and Runze in [60] and this is called Smoothly Clipped Absolute Deviation (SCAD) penalty. These two penalties are nonconvex, and they have the oracle property i.e. the penalized estimator is asymptotically equivalent to the oracle estimator on only the true support. This implies that the penalty will perform well as if the true underlying model were given or known in advance. According to [24], MC+ penalty which include MCP and PLUS (penalized linear unbiased selection) was derived. Focusing on the MCP, by minimizing the maximum concavity provides sparse convexity to a large extent. According to [24], a larger γ values affords less unbiasedness and more concavity.

To select the exact variables to be included in the model as we increase γ, a new penalized regression model based on the Cox partial likelihood and a modified minimax convex penalty is proposed. This model will identify the important variables by retaining the good features. The performance of the proposed penalized regression model compared with existing methods will be demonstrated through a simulation study and its application is illustrated via two real-world dataset examples for analyzing the heart failure data and the NKI breast cancer data.

Consider the linear regression model

$$y = X\beta + \epsilon \tag{5.1}$$

Where y which is an n x 1 vector is the response variable that depends on predictors X which is an n x j matrix with β being an j x 1 matrix and $\epsilon \sim N(0, \sigma^2 I_n)$.

If we have a small number of covariates i.e., m is small, then it will be easy to use the forward and backward selection to select the best variables to use. Using the above equation in (5.1), we see that $y - X\beta = \epsilon$, and letting $X\beta = \delta$, we have the penalized least square function to be,

$$W(\boldsymbol{\beta}; \lambda) \equiv \frac{1}{2n} \sum_{i=1}^{n} (y_i - \delta_i)^2 + \sum_{j=1}^{m} \rho(|\beta_j|; \lambda) \tag{5.2}$$

where $\rho(|\beta_j|; \lambda)$ is the penalty function (it can be SCAD, MCP or any penalty function) with a parameter indexed by $\lambda \geq 0$, and λ controls the tradeoff between the loss function and the penalty function.

Let's look at ridge regression. To avoid over-fitting issue, ridge techniques works well because if λ value in ridge is very large, it will add too much weight (and lead to under-fitting). Also, if λ value in ridge is 0, this will lead to ordinary least square regression. When we have a large multivariate data with larger number of predictors (p) than the number of observations (n), the ridge regression performs better than the ordinary least square method. Ridge penalty is given as:

$$\rho(\beta; \lambda)) = \lambda\beta^2$$

and its derivative is

$$\rho'(\beta; \lambda) = 2\lambda\beta$$

As discussed earlier in Chapter 2, the purpose of ridge which is also known as shrinkage or regularization methods is to shrink the coefficient values towards zero and this shrinking requires the selection of λ value (which is a tuning parameter that determines the amount of shrinkage). The main purpose of this shrinkage is to let the less contributive

variables have a coefficient close to zero or equal zero but while doing this, it will include all the predictors in the final model. This is a disadvantage of the ridge regression.

Talking about LASSO penalty, as stated earlier in Chapter 2, the disadvantage of using LASSO is the biasedness when the number of variables n is lower than the number of predictors p ($p > n$) and this biasedness interferes with how variables are accurately selected. Unlike ridge, LASSO forces some of the coefficient estimates to zero in order to reduce the complexity of the model. It penalizes the regression model with a penalty term called L1-norm (the sum of the absolute coefficients) by shrinking the regression coefficients toward zero. The advantage of ridge over LASSO is that ridge regression perform better than LASSO when we have many predictors than number of variables.

LASSO penalty is given as

$$\rho(\beta;\lambda)) = \lambda|\beta| \tag{5.3}$$

and its derivative is given as

$$\rho'(\beta;\lambda) = \lambda$$

What the derivative does is to show us how the algorithm is being penalized to avoid overfitting. Penalizing implies that you do not want your algorithm to be overfitted to your dataset.

Elastic-net is the combination of LASSO and ridge regression. It effectively shrink coefficients like the way ridge regression does and set some coefficients to zero as in LASSO.

Elastic net penalty is given as

$$\rho(\beta;\lambda)) = \lambda_1|\beta| + \lambda_2|\beta|^2 \tag{5.4}$$

and its derivative is given as

$$\rho'(\beta;\lambda) = \lambda + 2\lambda\beta$$

SCAD penalty which is a nonconvex penalty was proposed to reduce the large bias in LASSO penalty towards 0 when we have large regression coefficient. The SCAD penalty corresponds to a quadratic spline (piecewise polynomials) function with knots (where the splines are joined together) at λ and $\gamma\lambda$ (we can see this in the derivative equation below). SCAD penalty is given as:

$$\begin{aligned} \rho(\beta;\lambda)) &= \lambda \int_0^{\beta} min\left\{1, \frac{(\gamma - \frac{t}{\lambda})_+}{\gamma - 1}\right\} dt; \quad \gamma > 2, t > 0. \\ &= \begin{cases} \lambda\beta & |\beta| \leq \lambda \\ \frac{2\gamma\lambda|\beta| - \beta^2 - \lambda^2}{2(\gamma-1)} & \lambda < |\beta| < \gamma\lambda \\ \frac{\lambda^2(\gamma+1)}{2} & |\beta| \geq \gamma\lambda \end{cases} \end{aligned} \tag{5.5}$$

In the above equation, the subscript + implies that all quantities that are not positive will be equal to zero. The derivative is given below:

$$\rho'(\beta;\lambda)) = \begin{cases} \lambda & |\beta| \leq \lambda \\ \frac{\gamma\lambda - |\beta|}{(\gamma-1)} & \lambda < |\beta| < \gamma\lambda \\ 0 & |\beta| \geq \gamma\lambda \end{cases} \tag{5.6}$$

At a universal penalty level $\lambda_{universal} = \sigma\sqrt{\frac{2}{n} logp}$, the probability of selecting the right variable is high for MCP without us requiring that the $min_{\beta_j \neq 0}|\beta_j|/\lambda_{universal}$ must be greater than a quantity of the order $\sqrt{r^0}$, where r^0 is the rank i.e., $r^0 \equiv$ number $j : \beta_j \neq 0$ [73]. Using the conditions which are unbiasedness and selection features placed on MCP, the MCP penalty provides the best convexity for the penalized loss in sparse regions by minimizing the maximum concavity. The MCP penalty selection consistency

applies to the case of $p > n$. The MCP penalty is given as

$$\begin{aligned} \rho(\beta;\lambda)) &= \lambda \int_0^{\beta} \left(1 - \frac{t}{\lambda\gamma}\right)_+ dt; \quad \gamma > 1, t > 0. \\ &= \begin{cases} \lambda\beta - \frac{\beta^2}{2\gamma} & |\beta| \leq \lambda\gamma \\ \frac{\lambda^2\gamma}{2} & \text{otherwise} \end{cases} \end{aligned} \tag{5.7}$$

The derivative of MCP penalty is:

$$\rho'(\beta;\lambda)) = \begin{cases} (\lambda - \frac{|\beta|}{\gamma})sign(\beta) & |\beta| \leq \lambda\gamma \\ 0 & \text{otherwise} \end{cases} \tag{5.8}$$

where sign(β) denotes the sign of the coefficients.

The role of parameter γ in MCP is to control how fast the penalization rate goes to zero. This applies to SCAD also. Below is the plot of the derivative of LASSO, SCAD and MCP.

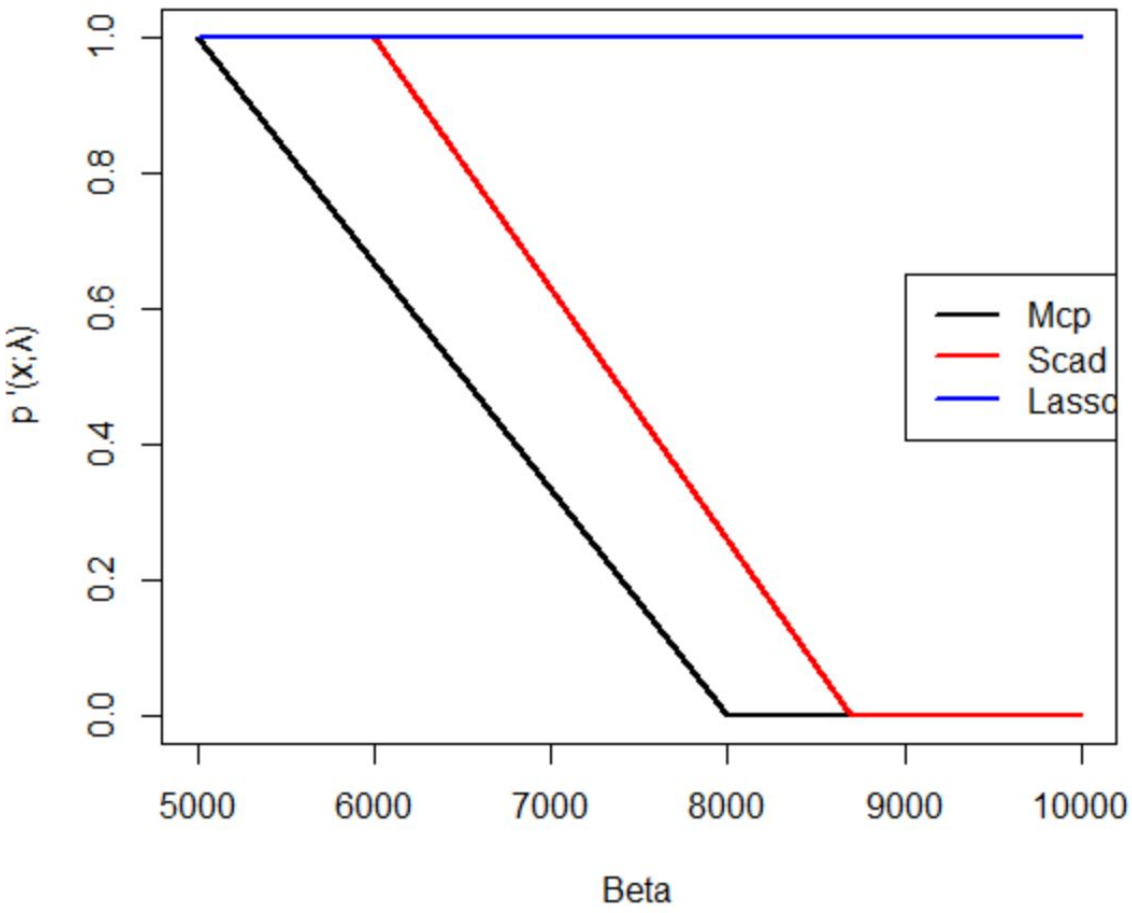


Figure 5.1: Derivative plot of MCP ($\gamma = 3$), SCAD ($\gamma = 3.7$) and LASSO penalty.

Rate of penalization is the rate of penalizing your model to prevent overfitting and relaxing the rate of penalization means your model is being penalized. SCAD first applied LASSO penalization rate and then starts to smoothly relaxes the rate of penalization as the absolute value of the coefficient is increasing. We can observe this through the derivative of SCAD penalty (5.6) and Figure 5.1. Compared to the SCAD penalty, as the absolute value of the coefficient increases (as $|\beta|$ increases), MCP immediately relaxes the rate of penalization down to zero

5.0.1 Survival Analysis

Survival analysis is a branch of statistics that analyzes time-to-event-data to learn and estimate the survival experience of objects of interest. The time-to-event data measures the length of time from a time origin to the occurrence of an event of interest (e.g., failure of a product or death of a patient). We can analyze survival data using different methods. The parametric approaches rely on assumptions of a certain lifetime distribution (e.g., Weibull, Gamma, Lognormal, etc.). The non-parametric methods which include the commonly used Kaplan-Meier product limit estimator [95], are used to estimate the survival function based on the time to the occurrence of the event. There are three assumptions of Kaplan Meier. Kaplan Meier assumed that the survival probabilities are the same for subjects recruited early and late in the study.

Also, for Kaplan Meier, every patient who are censored are assumed to have the same survival prospects as those who continue to be followed. Lastly, using Kaplan Meier, events are assumed to happen at the specified time. The semi-parametric methods are the method in which the distribution of the outcome remain unknown even if the regression parameters (β) are known (e.g., the Cox proportional hazards model). This paper considers the Cox proportional hazards model which does not rely on an assumption of the lifetime distribution and allows us to leverage the covariates for regression analysis. However, the proposed method is also applicable to parametric analysis.

5.0.1.1 Cox Proportional Hazards Model

Cox proposed the Cox proportional hazards model which has been widely used in the analysis of survival data [102]. In Cox regression model, the hazard function, which is the risk of death at time t for an individual is given by

$$h(t|X) = h_0(t)exp(\beta^{'}X) \tag{5.9}$$

where $h_0(t)$ is called the baseline hazard, X is the vector of covariates and β is the vector of coefficients of the covariates.

Let $N(t) = \{i : Z_i \geq t\}$ denote the set of individuals who are "at risk" for failure at time t, (risk set). The partial likelihood is defined as the product of the conditional probabilities of seeing the observed deaths over the set of observed death times, given the set of individuals at risk at those times. At each failure time Z_t, a risk set $N(Z_t)$ usually consists of individuals who have been followed up till that particular time and have not yet experienced the event of interest just before that time point [104]. Using (5.9), and at each failure time Z_n, the contribution to the likelihood is:

$$\begin{aligned} L_n(\beta) &= P(\textit{individual n fails} \mid \textit{one failure from } N(Z_n)) \\ &= \frac{P(\textit{individual n fails} | \textit{at risk at } Z_n)}{\Sigma_{l \in N(Z_n)} P(\textit{individual l fails} \mid \textit{at risk at } Z_n)} L_n(\beta) \\ &= \frac{h(Z_n|X_n)}{\Sigma_{l \in N(Z_n)} h(Z_n|X_l)} \end{aligned}$$

In cox regression model, there is a log-linear relationship between the covariates and the hazard function. Also, the Cox proportional hazards model can be considered as a modified "simple" linear regression model [102].

Under the proportional hazard assumption and using (5.9), the Cox partial likelihood is given as

$$\begin{aligned} L(\beta) &= \prod_{n=1}^{w} \frac{h(Z_n|X_n)}{\Sigma_{l \in N(Z_n)} h(Z_n|X_l)} \\ &= \prod_{n=1}^{w} \frac{h_0(Z_n) exp(\beta^{'} X_n)}{\Sigma_{l \in N(Z_n)} h_0(Z_n) exp(\beta^{'} X_n)} \end{aligned}$$

Taking the logarithm of the Cox partial likelihood gives

$$l(\beta) = \sum_{n=1}^{w} \left(\beta^{'} X_n - log \sum_{l \in N(Z_n)} exp(\beta^{'} X_n)\right)$$

In other to estimate the regression coefficients β, we will need to maximize the log partial likelihood function over β. Due to high dimensional space of some dataset especially gene dataset, [61], [53] indicated that we cannot apply Cox proportional hazards model directly to predict survival time because it was designed for small datasets and does not scale well to high dimensions. To overcome this challenge of high dimensionality, different type of penalized regression model such as SCAD, MCP, LASSO should be applied to Cox partial likelihood function so as to control over-fitting. These three above penalties set some coefficients estimates to zero in order to reduce the complexity of the model. By doing this, the function becomes a penalized log partial likelihood function. We will then estimate β by maximizing the penalized log partial likelihood function. When we apply MCP penalty to this Cox partial likelihood function, the MCP penalized log partial likelihood function will be written as

$$W(\beta) = l(\beta) + \sum_{j=1}^{m} \rho(|\beta_j|; \lambda)$$

Where $\rho(|\beta_j|; \lambda)$ is given in eqn (5.7), λ is the tuning parameter of the penalty and m is the number of covariates. For other different penalties, we will apply the Cox partial likelihood function to them before minimizing the objective function to obtain β. To maximize the objective function, we take the derivative of the objective function with respect to β and set it to zero. Since LASSO, SCAD and MCP have absolute value, then the derivative of the vector norm will be applied to obtain β.

Using LASSO as an example, and using the penalized least square function in (5.2), the objective function is given as,

$$W(\beta) = \frac{1}{2}(y - X\beta)^{'}(y - X\beta) + \lambda \sum_{j=1}^{m} |\beta_j| \tag{5.10}$$

Let X_{-j} denote the matrix consisting of columns of X after removing the jth column and X_j denote the jth columns of X where $X_j^{'} X_j = I$

Setting the derivative of (5.10) to zero with respect to β_j gives us

$$\begin{aligned} \frac{dW(\beta)}{d\beta_j} &= -X_j^{'}(y - X_{-j}\beta_{-j}) + X_j^{'}(X_j\beta_j) + \lambda\delta(|\beta_j|) \\ &= -z_j + I\beta_j + \lambda\delta(|\beta_j|) = 0 \end{aligned} \tag{5.11}$$

Where $z_j = X_j^{'}(y - X_{-j}\beta_{-j})$ and it doesn't depend on β_j.

Using the vector norm,

Case 1: When $\beta_j \neq 0$,

$$\lambda\delta(|\beta_j|) = \lambda\frac{\beta_j}{|\beta_j|}$$

and substituting this into (5.11) to solve for β_j gives us

$$\beta_j = z_j\left(1 + \frac{\lambda}{|\beta_j|}\right)^{-1} \tag{5.12}$$

We will need to replace $|_j|$ that is on the right-hand side and note that

$$|\beta_j| = |z_j|\left(1 + \frac{\lambda}{|\beta_j|}\right)^{-1} \tag{5.13}$$

Substituting eqn (5.13) into eqn (5.12) gives

$$\beta_j = \frac{z_j}{|z_j|}(|z_j| - \lambda) \tag{5.14}$$

Case 2: When $\beta_j = 0$,

Since the vector norm is not differentiable at zero, the subdifferential and the Karush-Kuhn-Tucker (KKT) conditions states that

$0 \in -z_j + I\beta_j + \lambda v$ (Stationarity condition)

Since $|v| \leq 1$, where v is any vector, this implies that $|z_j| \leq \lambda$.

Hence,

$$\beta_j = \frac{z_j}{|z_j|} f_{LASSO}(z, \lambda),$$

where

$$f_{LASSO}(z, \lambda) = \begin{cases} S(z, \lambda); & |z| > \lambda \\ 0; & |z| \leq \lambda \end{cases} \tag{5.15}$$

and

$$S(z, \lambda) = \begin{cases} z - \lambda; & |z| > \lambda \\ 0; & |z| \leq \lambda \\ z + \lambda; & |z| < -\lambda \end{cases} \tag{5.16}$$

$S(z, \lambda)$ is the soft-thresholding operator defined for $\lambda \geq 0$ (5.14).

From (5.15), whenever $|z_j| \leq \lambda$ that is whenever $|X_j^{'}(y - X_{-j}\beta_{-j}))| \leq \lambda$, β will be zero and when $|z_j| > \lambda$, β will not be zero.

The lasso univariate solution in (5.15) is closely related to wavelet soft-thresholding method. Soft thresholding works by first setting the coefficients whose absolute values are lower than the threshold λ to zero and then shrink the nonzero coefficients toward zero. With this, it provides smoother results [131]. We can understand the rationale behind the SCAD and MCP, by looking at their univariate solution which is a solution of β_j. SCAD univariate solution can be solved as follow.

Using the first formula of SCAD in (5.5), we see that the univariate solution will be the same as the univariate solution of LASSO in (5.15). Using the second formula of SCAD in (5.5) and using the penalized least square function in (5.2), the objective function is given as,

$$W(\beta) = \frac{1}{2}(y - X\beta)^{'}(y - X\beta) + \sum_{j=1}^{m} \frac{2\gamma\lambda|\beta| - \beta^2 - \lambda^2}{2(\gamma - 1)} \tag{5.17}$$

Setting the derivative of (5.17) to zero with respect to β_j gives us

$$\begin{aligned} \frac{dW(\beta)}{d\beta_j} &= -X_j^{'}(y - X_{-j}\beta_{-j}) + X_j^{'}(X_j\beta_j) + \delta\left(\frac{2\gamma\lambda|\beta| - \beta^2 - \lambda^2}{2(\gamma - 1)}\right) \\ &= -z_j + I\beta_j + \left(\frac{\gamma\lambda\frac{\beta_j}{|\beta_j|} - \beta_j}{\gamma - 1}\right) = 0 \end{aligned} \tag{5.18}$$

Where $z_j = X_j^{'}(y - X_{-j}\beta_{-j})$ and it doesn't depend on β_j.

Using the vector norm,

Case 1: When $\beta_j \neq 0$,

Solving for β_j gives us

$$\beta_j = z_j\left(1 - \frac{1 - \gamma\lambda\frac{1}{|\beta_j|}}{\gamma - 1}\right)^{-1}$$

$$
\begin{aligned}
|\beta_j| &= |z_j|\left(1 - \frac{1 - \gamma\lambda\frac{1}{|\beta_j|}}{\gamma - 1}\right)^{-1} \\
&= |z_j|\left(1 - \frac{1}{\gamma - 1} + \frac{\gamma\lambda}{|\beta_j|(\gamma - 1)}\right)^{-1} \qquad (5.19)
\end{aligned}
$$

We will need to solve for $|\beta_j|$

$$
|\beta_j| = \frac{|z_j| - \frac{\gamma\lambda}{(\gamma-1)}}{1 - \frac{1}{\gamma-1}} \qquad (5.20)
$$

Substituting eqn (5.20) into eqn (5.19) gives

$$
|\beta_j| = \frac{\frac{z_j}{|z_j|}\left(|z_j| - \frac{\gamma\lambda}{(\gamma-1)}\right)}{1 - \frac{1}{\gamma-1}} \qquad (5.21)
$$

Case 2: When $\beta_j = 0$,

Using the subdifferential with $|v| \leq 1$, where v is any vector and the Karush-Kuhn-Tucker (KKT) condition $0 \in -z_j + I\beta_j + v\gamma\lambda(\gamma - 1)$, we have

$$
|z_j| \leq \frac{\gamma\lambda}{(\gamma - 1)} \qquad (5.22)
$$

Joining (5.22) and (5.21) together gives us

$$
\beta_j = \begin{cases} \frac{\frac{z_j}{|z_j|}\left(|z_j| - \frac{\gamma\lambda}{(\gamma-1)}\right)}{1 - \frac{1}{\gamma-1}} & |z_j| > \frac{\gamma\lambda}{(\gamma-1)} \\ 0 & |z_j| \leq \frac{\gamma\lambda}{(\gamma-1)} \end{cases}
$$

Using the third formula of SCAD in (5.5) and using the penalized least square function in (5.2), the objective function is given as,

$$
W(\beta) = \frac{1}{2}(y - X\beta)'(y - X\beta) + \sum_{j=1}^{m} \frac{\lambda^2(\gamma + 1)}{2} \qquad (5.23)
$$

Setting the derivative of (5.23) to zero with respect to β_j gives us

$$\begin{aligned}\frac{dW(\beta)}{d\beta_j} &= -X_j^{'}(y - X_{-j}\beta_{-j}) + X_j^{'}(X_j\beta_j) + 0 \\ &= -z_j + I\beta_j + 0 = 0\end{aligned}$$

This implies that $z_j = \beta_j$

The final solution of SCAD is then given as

$$\beta_j = \frac{z_j}{|z_j|} f_{SCAD}(z, \lambda, \gamma),$$

where,

$$f_{SCAD}(z, \lambda, \gamma) = \begin{cases} S(z, \lambda); & |z| \leq 2\lambda \\ \frac{S(z, \frac{\gamma\lambda}{\gamma-1})}{1-\frac{1}{\gamma-1}}; & 2\lambda < |z| \leq \gamma\lambda \\ z; & |z| > \gamma\lambda \end{cases} \tag{5.24}$$

where and $S(z, \lambda)$ is given in (5.16) and

$$S(z, \frac{\gamma\lambda}{\gamma-1}) = \begin{cases} z - \frac{\gamma\lambda}{\gamma-1}; & |z| > \lambda \\ 0; & |z| \leq \lambda \\ z + \frac{\gamma\lambda}{\gamma-1}; & |z| < -\lambda \end{cases}$$

Based on the definition of soft thresholding given above, hard thresholding works by setting the coefficients whose absolute values are lower than the threshold λ to zero. Compared to the soft thresholding, hard thresholding provides better edge preservation than the soft threshold. The SCAD univariate solution given in (5.24) converges to soft thresholding as $\lambda \to \infty$. However, as $\lambda \to 2$, (5.24) does not converge to hard thresholding; but it converges to

$$\begin{cases} S(z,\lambda); & |z| \leq 2\lambda \\ z; & |z| > 2\lambda \end{cases}$$

Hence, as γ approaches its minimum value, f_{SCAD} converges to discontinuous functions and the solution moves from λ to 2λ as z exceeds 2λ.

MCP univariate solution is given below. Using the first formula of MCP in (5.7) and using the penalized least square function in (5.2), the objective function is given as,

$$W(\beta) = \frac{1}{2}(y - X\beta)^{'}(y - X\beta) + \sum_{j=1}^{m} \lambda|\beta| - \frac{\beta^2}{2\gamma} \tag{5.25}$$

Setting the derivative of (5.25) to zero with respect to β_j gives us

$$\begin{aligned} \frac{dW(\beta)}{d\beta_j} &= -X_j^{'}(y - X_{-j}\beta_{-j}) + X_j^{'}(X_j\beta_j) + \delta\left(\lambda|\beta| - \frac{\beta^2}{2\gamma}\right) \\ &= -z_j + I\beta_j + \left(\lambda\frac{\beta_j}{|\beta_j|} - \frac{\beta_j}{\gamma}\right) = 0 \end{aligned} \tag{5.26}$$

Where z_j doesn't depend on β_j.

Using the vector norm,

Case 1: When $\beta_j \neq 0$ and solving for β_j gives us

$$\beta_j = z_j\left(1 - \frac{1}{\gamma} + \lambda\frac{1}{|\beta_j|}\right)^{-1}$$

$$|\beta_j| = |z_j|\left(1 - \frac{1}{\gamma} + \lambda\frac{1}{|\beta_j|}\right)^{-1} \tag{5.27}$$

Solving for $|\beta_j|$ gives us

$$|\beta_j| = \frac{|z_j| - \lambda}{1 - \frac{1}{\gamma}} \tag{5.28}$$

Substituting eqn (5.28) into eqn (5.27) gives

$$\beta_j = \frac{\frac{z_j}{|z_j|}\left(|z_j| - \lambda\right)}{1 - \frac{1}{\gamma}} \tag{5.29}$$

Case 2: When $\beta_j = 0$,

Using the subdifferential with $|v| \leq 1$, where v is any vector and the Karush-Kuhn-Tucker (KKT) condition $0 \in -z_j + I\beta_j + v\lambda$, we have

$$|z_j| \leq \lambda \tag{5.30}$$

Joining (5.30) and (5.29) together gives us

$$\beta_j = \begin{cases} \frac{\frac{z_j}{|z_j|}\left(|z_j| - \lambda\right)}{1 - \frac{1}{\gamma}}; & |z_j| > \lambda \\ 0; & |z_j| \leq \lambda \end{cases} \tag{5.31}$$

Using the second formula of MCP in (5.7) and using the penalized least square function in (5.2), the objective function is given as,

$$W(\beta) = \frac{1}{2}(y - X\beta)'(y - X\beta) + \sum_{j=1}^{m} \frac{\lambda^2\gamma}{2} \tag{5.32}$$

Setting the derivative of (5.32) to zero with respect to β_j gives us

$$\begin{aligned} \frac{dW(\beta)}{d\beta_j} &= -X_j'(y - X_{-j}\beta_{-j}) + X_j'(X_j\beta_j) + 0 \\ &= -z_j + I\beta_j + 0 = 0 \end{aligned}$$

The final solution of MCP is then given as

$$\beta_j = \frac{z_j}{|z_j|} f_{MCP}(z, \lambda, \gamma)$$

$$f_{MCP}(z,\lambda,\gamma) = \begin{cases} \frac{S(z,\lambda)}{1-\frac{1}{\gamma}}; & |z| \leq \lambda\gamma \\ z; & |z| > \gamma\lambda \end{cases} \tag{5.33}$$

where $S(z,\lambda)$ is same as (5.16).

The MCP univariate solution given in (5.33) turn to a firm threshold as we change γ value. It (5.33) converges to soft thresholding as $\gamma \to \infty$ and as $\gamma \to 1$, it becomes equivalent to hard thresholding. The name "firm thresholding" comes from the solution bridging the gap between soft thresholding and hard thresholding as we change γ. As γ approaches its minimum value, f_{MCP} also converges to discontinuous functions and as z exceeds λ, the solution jumps from 0 to λ.

5.0.2 MMCP Penalty

Here, we propose a penalty function by modifying the MCP penalty. We will give a brief description, conditions and some concepts of MMCP method alongside with our main results.

The MMCP penalty is given as

$$\begin{aligned} \rho(\beta;\lambda) &= \lambda \int_0 \beta \left(1 - \frac{t}{\gamma\lambda\alpha}\right)_+ dt \\ &= \begin{cases} \lambda|\beta| - \frac{\beta^2}{2\alpha\gamma}; & \beta \leq \lambda\alpha\gamma \\ \frac{\lambda^2\gamma\alpha}{2}; & otherwise \end{cases} \end{aligned} \tag{5.34}$$

The derivative of (5.34) is given as

$$\rho^{'}(\beta;\lambda) = \begin{cases} \lambda - \frac{|\beta|}{\alpha\gamma}; & \beta \leq \lambda\alpha\gamma \\ 0; & otherwise \end{cases} \tag{5.35}$$

where $\lambda \geq 0$ is the penalty term, $0 < \alpha \leq 1$, and $\gamma > 0$ is a parameter that controls how fast the penalization rate goes to zero. Its value needs to be specified to get the best λ value. According to [24], a larger γ values affords less unbiasedness and more concavity. So, if we introduce α to the equation, it will help to reduce the rate at which γ is increasing so as to afford more unbiasedness and less concavity i.e. more sparse convexity to the broadest extent. Since MCP is nonconvex, there will be numerical challenges in fitting the model but because we are adding α to the model to make it more sparse convex, then there won't be a lot of numerical challenge in fitting the model compare to using MCP. Also, this will help our model to select the exact variables to be included in the model as we increase γ.

For the MMCP penalty at each λ level, when $\alpha = 1$, the penalty changes to MCP penalty, as $\alpha = 1$ and $\gamma \to \infty$, the penalty changes to the "$\mathcal{L}1$ penalty". We will assume that $\rho(\beta; \lambda)$ is non-decreasing in t and its derivative given in (5.35) is continuous. Also, as $\alpha = 1$ and $\gamma \to 0+$, the penalty changes to the "$\mathcal{L}0$ penalty

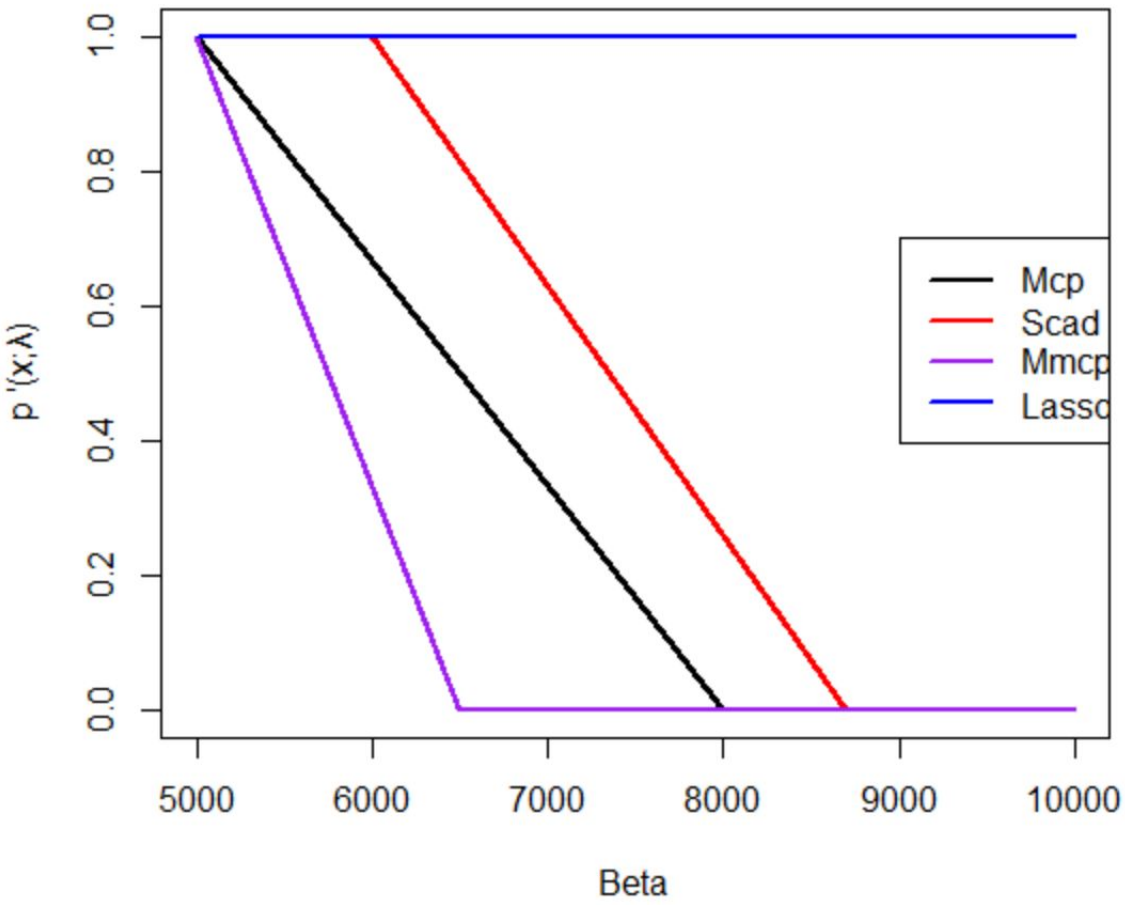


Figure 5.2: Derivative plot of MCP ($\gamma = 3$), SCAD ($\gamma = 3.7$), MMCP ($\gamma = 3, \alpha = 0.5$), and LASSO penalty.

For the minimizers of (5.1) to have variable selection features with zero components [29], we will assume that in t, the penalty $\rho(t;\lambda)$ has a continuous derivative in $(0,\infty)$ represented as $\rho^{'}(t;\lambda)$ and is nondecreasing. We will also assume that $\rho^{'}(0+;\lambda) > 0$ so that the minimizers of (5.10) will have features of selecting variables with zero components [29].

Also, whenever $\rho^{'}(0+;\lambda) < \infty$, we will assume $\rho^{'}(0+;\lambda) = \lambda$ so that there is an interpretation for λ as the threshold level for individual coefficient β_j under the standardization $|x_j|^2 = n$

The penalty given in (5.10) minimizes the maximum concavity which is:

$$\eta(\rho) \equiv \eta(\rho;\lambda) \equiv sup_{0<x_1<x_2} \frac{\rho^{'}(x_1;\lambda) - \rho^{'}(x_2;\lambda)}{x_2 - x_1} \tag{5.36}$$

This is subject to the following unbiasedness in selection and selection features of variables:

$$\rho^{'}(x;\lambda) = 0 \quad \forall\, x \geq \lambda\alpha\gamma, \quad \rho^{'}(0+;\lambda) = \lambda. \tag{5.37}$$

Where $\rho^{'}(x;\lambda) = 0$ *for all* $x \geq \lambda\alpha\gamma$ determine the unbiasedness and $\rho^{'}(0+;\lambda) = \lambda$ determines the selection of variables.

For the MCP, $\eta(\rho;\lambda) = \frac{1}{\gamma}$, for the SCAD, $\eta(\rho;\lambda) = \frac{1}{(\gamma-1}$), and for the MMCP,

$$\eta(\rho;\lambda) = \frac{1}{\alpha\gamma} \tag{5.38}$$

Looking at the univariate solution of MMCP, let us consider a simple linear regression of y upon x. The rationale behind the MMCP can also be understood by considering its univariate solution. For this simple linear regression problem, the MMCP estimator has the following closed form:

Using the first formula of MMCP in (5.34) and using the penalized least square function in (5.2), the objective function is given as,

$$W(\beta) = \frac{1}{2}(y - X\beta)^{'}(y - X\beta) + \sum_{j=1}^{m} \lambda|\beta| - \frac{\beta^2}{2\gamma\alpha} \tag{5.39}$$

Setting the derivative of (5.39) to zero with respect to β_j gives us

$$\begin{aligned}\frac{dW(\beta)}{d\beta_j} &= -X_j'(y - X_{-j}\beta_{-j}) + X_j'(X_j\beta_j) + \delta\left(\lambda|\beta| - \frac{\beta^2}{2\gamma\alpha}\right) \\ &= -z_j + I\beta_j + \left(\lambda\frac{\beta_j}{|\beta_j|} - \frac{\beta_j}{\gamma\alpha}\right) = 0 \end{aligned} \tag{5.40}$$

z_j doesn't depend on β_j.

Using the vector norm,

Case 1: When $\beta_j \neq 0$ and solving for β_j gives us

$$\beta_j = z_j\left(1 - \frac{1}{\gamma\alpha} + \lambda\frac{1}{|\beta_j|}\right)^{-1}$$

$$|\beta_j| = |z_j|\left(1 - \frac{1}{\gamma\alpha} + \lambda\frac{1}{|\beta_j|}\right)^{-1} \tag{5.41}$$

Solving for $|\beta_j|$ gives us

$$|\beta_j| = \frac{|z_j| - \lambda}{1 - \frac{1}{\gamma\alpha}} \tag{5.42}$$

Substituting eqn (5.42) into eqn (5.41) gives

$$\beta_j = \frac{\frac{z_j}{|z_j|}(|z_j| - \lambda)}{1 - \frac{1}{\gamma}} \tag{5.43}$$

Case 2: When $\beta_j = 0$,

Using the subdifferential with $|v| \leq 1$, where v is any vector and the Karush-Kuhn-Tucker (KKT) condition $0 \in -z_j + I\beta_j + v\lambda$, we have (5.30).

Joining (5.43) and (5.30) together gives us

$$\beta_j = \begin{cases} \frac{\frac{z_j}{|z_j|}\left(|z_j|-\lambda\right)}{1-\frac{1}{\gamma\alpha}}; & |z_j| > \lambda \\ 0; & |z_j| \leq \lambda \end{cases} \tag{5.44}$$

Using the second formula of MMCP in (5.34) and using the penalized least square function in (5.2), the objective function is given as,

$$W(\beta) = \frac{1}{2}(y - X\beta)^{'}(y - X\beta) + \sum_{j=1}^{m} \frac{\lambda^2\gamma\alpha}{2} \tag{5.45}$$

Setting the derivative of (5.45) to zero with respect to β_j gives us

$$\begin{aligned} \frac{dW(\beta)}{d\beta_j} &= -X_j^{'}(y - X_{-j}\beta_{-j}) + X_j^{'}(X_j\beta_j) + 0 \\ &= -z_j + I\beta_j + 0 = 0 \end{aligned}$$

This implies that $z_j = \beta_j$. The final solution of MMCP is then given as

$$\beta_j = \frac{z_j}{|z_j|} f_{MMCP}(z, \lambda, \gamma, \alpha)$$

$$f_{MMCP}(z, \lambda, \gamma, \alpha) = \begin{cases} \frac{S(z,\lambda)}{1-\frac{1}{\gamma\alpha}}; & |z| \leq \lambda\gamma \\ z; & |z| > \gamma\lambda \end{cases} \tag{5.46}$$

For $\lambda \geq 0$ and $0 < \alpha \leq 1$, where $S(z, \lambda)$ is same as (5.16). Note that $S(z, \lambda)$ is the univariate solution in (5.16).

The MMCP univariate solution given in (5.46) turn to a firm threshold as we change γ and α value. It converges to soft thresholding as $\gamma \to \infty$, and $\alpha \to \infty$. As $\gamma \to 1$ and $\alpha \to 1$, it becomes equivalent to hard thresholding. As γ approaches its minimum value

and α approaches 1, f_{MMCP} also converges to discontinuous functions and as z exceeds λ, the solution jumps from 0 to λ.

Due to MCP, MMCP and SCAD being nonconvex, although our new penalty provides more sparse convexity that the rest, we will demonstrate the potential of coordinate descent algorithms for fitting MCP model and other penalty. It will also be demonstrated that this approach is faster than other approach based on theoretical convergence properties. Optimizing a function with coordinate descent algorithms involves optimizing each of the parameters one at a time, cycling through them until convergence is achieved. According to [93], using convexity diagnostics, we will also determine areas of the parameter space where the objective function is locally convex, even when the penalty is not convex. SCAD and MCP regression models with nonconvex penalties were investigated by [93] using coordinate descent algorithms. Next, we will describe the algorithms for fitting linear regression models penalized by MMCP, as well as its convergence.

5.0.3 Coordinate Descent Algorithms

In this section we will describe coordinate descent algorithms for least squares regression penalized by MMCP, and also investigate the convergence of this algorithm. Using equation (5.2) , to find the value β that optimizes this equation, the local linear approximation (LLA) algorithm [55] (which was proposed for nonconvex penalty) makes a linear approximation to the penalty. The solution will then be computed by using the least angle regression [57] (LARS) algorithm. LARS is efficient for computing the entire path of a convex penalty solutions. For each value of λ, the process is repeated iteratively until convergence occurs over a grid. We can see its implementation in [55].

The idea of coordinate descent algorithms is simple and efficient. To pass over each parameter only requires O(np) operations. So, by reducing the number of iterations to

less than p, the computation burden can be reduced by getting the solution faster even further than the np^2 operations needed to solve a linear regression problem. Coordinate descent algorithms also prove useful when dealing with problems of extremely high dimensionality because the computational burden increases only linearly with p. The coordinate descent algorithm uses the univariate solution of a penalty function to obtain the coordinate-wise minimizer of the objective function [24].

For j in 1, ..., m, we will partially optimize the penalty function W in (5.2) with respect to β_j while fixing the rest of β (i.e., β_k) at its most recently updated values. Using LASSO as an example, the algorithm of the coordinate descent is as follows:

We will minimize the penalty function W in (5.2) with respect to β_j while fixing the rest of β_k

$$W(\beta_j|\beta_k;\lambda) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \sum_{j\neq k} x_{ij}\beta_j - x_{ik}\beta_k)^2 + \lambda|\beta_j| \ + \ constant$$

Then taking the derivative with respect to β_k gives

$$\begin{aligned} dW(\beta_j|\beta_k;\lambda) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \sum_{j\neq k} x_{ij}\beta_j - x_{ik}\beta_k)x_{ik} + \lambda d|\beta_j| \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \sum_{j\neq k} x_{ij}\beta_j)x_{ik} - x_{ik}^2\beta_k + \lambda d|\beta_j| \end{aligned}$$

Using

$$\hat{l}_{ij} = y_i - \sum_{j\neq k} x_{ij}\hat{\beta}_j$$

Here, $\hat{l}_{ij}$ $(i = 1 : n)$ are the partial residuals with respect to the jth predictor.

From equation (5.11), we see that

$$\hat{z}_j = \frac{1}{n}\sum_{i=1}^{n} \hat{l}_{ik} x_{ik}$$

Where $\hat{z}_j$ is the ordinary least square estimator.

Since the soft threshold is given in (5.16) as $S(z, \lambda)$, if we let $\hat{\beta}_j$ to denote the minimizer of $W(\beta_j|\hat{\beta}_k; \lambda)$, then

$$\hat{\beta}_j = S(\hat{z}_j, \lambda)$$

Hence, we have the following algorithm.

At step j of iteration c,

repeat

for $j = 1, 2, \ldots, p$

$\hat{z}_j = \frac{1}{n}\sum_{i=1}^{n} l_i x_{ij} + \hat{\beta}_j^c$

After the calculation, $\hat{\beta}_j j = S(\hat{z}_j, \lambda)$ will be updated so as to obtain $\hat{\beta}_j^{c+1}$

For all i, we will obtain $l_i - (\hat{\beta}_j^{c+1} - \hat{\beta}_j^c)x_{ij}$ in other to obtain l_i

Do this until **convergence**.

Using this update, the coordinate descent algorithms will then iterate until convergence is reached. With this, a path of solutions is produced by repeating this process across a grid of values for λ, where λ determines the selection of variables. [93]. In essence, minimizing the penalty function W using the coordinate descent algorithms will produce a path of solution when we repeat the above process across a grid of values for λ thereby selecting or minimizing the features or variables (β).

Having a target function, what a coordinate descent algorithm does is that it optimizes this function with respect to a single parameter at a time. Iteratively, it then cycles

through all parameters until there is convergence (when the cost function cannot be decreased anymore when we apply the coordinate descent, then we say it has converged).

This algorithm is efficient because if a continuous path solution is to be computed, few iterations will be needed for the solution to be close to our initial values. Also, the update will be obtained quickly due to the minimization of W with respect to β_j being obtained from the univariate regression of the current residuals $\epsilon = y - X\beta$ on x_j at a cost of O(n) operations.

According to the lemma proposed by [93], MMCP is not convex, neither the proposed LLA algorithm nor the coordinate descent algorithms are guaranteed to converge to a global minimum in general. While W may contain a nonconvex penalty component, it can still be convex with respect to β. [24] stated that if a_* represent the minimum eigenvalue of $\frac{X'X}{n}$, then from (5.14), the MCP objective function is convex if $\gamma > \frac{1}{a_*}$. Applying this to MMCP, the MMCP objective function is convex if $\gamma > \frac{1}{\alpha a_*}$. In this case, the coordinate descent algorithms converge to the global minimum. Our interest is to obtain the estimate of β i.e., $\hat{\beta}$ for a range of values of λ starting from when λ is maximum (when the penalized coefficients are zero) to when λ is zero (when the model is excessively large). A path of solutions regularized by λ is produced because the estimated coefficients vary continuously with $\lambda \in [\lambda_{min}, \lambda_{max}]$ based on the convexity of the objective function. Due to the continuous nature of the coefficient paths, one can make a reasonable decision to pick initial values by starting from one extreme of the path and using the estimate β from the previous value of λ as the initial value for the next value of λ. From (5.16), the $\lambda_{max} = z_{max}$, where $z_{max} = max_j\{\frac{x_j'y}{n}\}$. Starting from the maximum λ value with $\beta^0 = 0$ and move to the minimum λ value, the initial values will never be far from the solution. The solutions along a grid of 100 λ values will be computed.

5.0.4 Diagnostics

5.0.4.1 Diagnostic of Local Convexity.

When $p > n$, global convexity is neither possible nor relevant in high-dimensional setting. We can still obtain stable estimates and smooth coefficient paths in the parameter space if the objective function is convex in the local region that contains these sparse solutions. It is desirable to use sparse solutions for which the number of nonzero coefficients is much lower than p in high-dimensional setting.

Diagnostics would determine which regions of MMCP penalty are locally convex and which region are not. This cutoff $\gamma > \frac{\lambda}{\alpha a_*}$ will be modified so that only active covariates i.e. the covariates with nonzero coefficients are included in the calculation of a_*. These active covariates will increase as λ decreases, hence for large value of λ, there will be no problem for local convexity of the objective function.

According to [93], for a certain value of λ, let $\hat{\beta}(\lambda)$ denote the minimizer of (5.2), $A^0(\lambda)$ = $\{j : \hat{\beta}_j(\lambda) \neq 0\}$ be the set of nonzero coefficient for a certain value of λ, and $A^0(\lambda_-)$ be the set of coefficient that are zero but will be nonzero after reducing λ value to a very small amount, then $A(\lambda) = A^0(\lambda) \bigcup A^0(\lambda_-)$ and $a_*(\lambda)$ is the minimum eigenvalue of $\frac{x'_A(\lambda)x_A(\lambda)}{n}$, where $x_A(\lambda)$ is the design matrix obtained from only the covariates for which $j \in A(\lambda)$. The λ interval over which the objective function is "locally convex" is defined to be (λ^*, ∞). We obtain this by letting

$$\lambda^* = inf\{\lambda : \gamma > \frac{1}{\alpha a_*(\lambda)}\} \tag{5.47}$$

The objective function is locally nonconvex in the region $[0, \lambda^*]$. λ^* must be a value of λ for which $A^0(\lambda) \neq A^0(\lambda_-)$ because $a_*(\lambda)$ changes only when the composition of $A(\lambda)$ changes.

5.0.4.2 *How to Select γ, α and λ.*

Choosing the best tuning parameters γ, α and λ will help in the estimation of MMCP model and we can achieve this by using an information criterion such as the Akaike Information criteria (AIC) [48], the Bayesian Information Criteria (BIC) [41], The risk inflation criterion (RIC) [94], Cp [87] or by using cross-validation. Each of this method has its disadvantage. The Akaike Information criteria (AIC) [48], the Bayesian Information Criteria (BIC) [41] can be used to select the best lambda value to use after applying a penalty function. So, in the nonconvex region of the objective function, the AIC and BIC have a chance to select local minima in some settings. Continuous penalized methods are commonly used because subset selection is not computationally feasible for large covariates.

For cross-validation, there is a considerable amount of computation involved, particularly when it is applied to a three-dimensional grid of values for γ, α *and* λ, among which some may lack convex objective functions, causing a slow convergence. This prevents practitioners from fully evaluating the choice of γ. Thus, the combination of BIC/AIC, cross-validation and convexity diagnostics could be used. If γ is given for a path of solutions, then the BIC/AIC method should be used to select the best λ and α value and convexity diagnostics should be used next to determine the locally convex regions of the solution path. If γ is not given for a path of solutions, then the BIC/AIC method should be used to select the best γ and the cross-validation should be used to select the best λ and α value and convexity diagnostics should be used next to determine the locally convex regions of the solution path. To make the penalty more convex, γ should be increased if the solution chosen lies in the region below λ^* and if it lies above λ^*, we can reduce γ. We can now use cross-validation to choose the best λ and to know the best α for this value of γ chosen after iterating the process of finding the best value of γ for some time.

5.0.5 Package

The package (ncvreg) [92] which is the regularization paths for SCAD and MCP penalized regression models [93] was adopted and we added the code of the solution path of MMCP to the package. Under this package, ncvsurv (Fit an MCP- or SCAD- penalized survival model) and cv-ncvsurv (Cross validation to fit an MCP- or SCAD- penalized survival model) was used in this project. Before using the cross-validation to choose the best λ, we first used the AIC/BIC to select the best γ by using different α value. In the package, the summary of the output has the Marginal false discovery rate (mFDR), Average mFDR, the Expected nonzero coefficients, and the Nonzero coefficients. In using Penalized regression for variable selection, we have to be confident about the selections of the variables. There has been difficulty in quantifying how confident we are with the variables selected due to the complexity of defining a "false discovery" in the penalized regression setting. In the orthogonal cases, the mFDR is calculated thus:

From equation (5.15), we see that $|(X_j)^{'}(y - X_{-j}\beta_{-j})| = z_j > \lambda$, where

$$z_j = (X_j)^{'}(y - X_{-j}\beta_{-j}) = -(X_j)^{'}(y - X\beta) + \beta_j^{(m)}$$

Thus, the probability that the jth variable is selected is given as

$$P\left(\frac{1}{n}|(X_j)^{'}(y - X_j\beta_j)| > \lambda\right)$$

This implies that we can estimate the number of selections that are false in the model if we can characterize the distribution of $\frac{1}{n}(X_j)^{'}(y - X_j\beta_j)$ under the null. This is easy to do in the case of orthonormal design:

$$E(\hat{\cup} \cap \zeta) = 2|\zeta|\Phi(-\lambda\sqrt{n}/\sigma),$$

Where $\hat{\cup}$ is the set of the selected variables and σ^2 can be estimated by $\frac{(y-X\beta)^T(y-X\beta)}{(n-|\hat{\cup}|)}$. Also, $|\zeta|$ is the set of null variables which can be replaced by q using the total number of variables as the upper bound for the null variables. The false discovery is then given as

$$\hat{FD} = 2q\Phi(-\lambda\sqrt{n}/\hat{\sigma}).$$

Also, the false discovery rate is given as

$$\hat{FDR} = \frac{\hat{FD}}{|\hat{\cup}|}$$

According to the path wise approaches definition, let us denote $\varkappa_j$ as the set of variables with non-zero coefficients in the model at the point in the path where feature j is selected, in these approaches a feature j is considered a false discovery if $X_j \perp Y | \{X_k : k \in \varkappa_j\}$. According to [97], False discovery is one that is independent of the outcome, and he considers a marginal perspective in which a selected feature j is false if it is marginally independent of the outcome.

This definition used in single-feature testing: $X_j \perp Y$. Using a simpler definition makes it possible to estimate the expected number of false discoveries as well as their rate, which is called the marginal false discovery rate (mFDR).

The mFDR estimates the marginal false discovery rate of a penalized regression model. With highly correlated predictor, the estimate tends to be slightly conservative, but it is accurate in most settings. This implies that it is much more powerful when two variables are correlated because it is challenging to distinguish between which of the variable (or none, or both) is driving changes in Y and which is merely correlated with Y. Miller, [97], [33] talked more about mFDR. Average mFDR is the mean of all the mFDR from each variable. If the Average mFDR (Ave. mFDR) is small, then it implies that the

nonzero variable(s) selected are infact the right variable(s) to be used in the model. For the cross-validation method, one of the summaries of the output from the package is cross-validation error (c.v.e.). Cross-validation error is the deviance obtained while running the cross-validation. Intuitively, the validation error estimates test error by checking the model's performance on a dataset not used for training.

5.0.6 Simulation Setting

Here, we simulated dataset on four scenarios to show the performance of the new penalty function compared to other existing penalty functions. To simulate cox proportional hazards models, we generated a survival time. According to [112], using the inverse probability method, we can generate event times from the proportional hazards model. Having a conditional hazard function given in (5.9),

We obtain a conditional survival function:

$$S(t|x) = exp(h(t|x)) \tag{5.48}$$

Where $(h(t|x))$ is given in (5.9).

Using Weibull baseline hazard built-in R function and using equation (5.25), the conditional hazard function with shape parameter τ and scale parameter ω is given as

$$h(t|x) = \left(\frac{t}{\omega}\right)^{\tau}; \quad \tau > 0, \omega > 0$$

Using the above equation, we obtain the conditional survival function to be

$$S(t|x, \omega) = exp\left(\frac{t}{\omega}\right)^{\tau}$$

Using rweibull with a scaled scale $\omega^{'}(\beta)$, where

$$\omega^{'} = \frac{\omega}{exp(\frac{\beta^{'}x}{\tau})^{'}},$$

its conditional hazard function will be given as

$$\begin{aligned} h(t|x,\omega^{'}) &= \left(\frac{t}{\omega^{'}}\right)^{\tau} \\ &= \left(\frac{t}{\frac{\omega}{exp(\frac{\beta^{'}x}{\tau})}}\right)^{\tau} \end{aligned} \tag{5.49}$$

Using (5.49), the conditional survival function will be

$$S(t|x,\omega^{'}) = exp\left(\frac{t}{\frac{\omega}{exp(\frac{\beta^{'}x}{\tau})}}\right)^{\tau} \tag{5.50}$$

Using equation (5.50) as the scale parameter of the Weibull distribution in r-package and using different τ as shape parameter, we simulated dataset of 200 observations with right censoring time (with fixed censoring rate of 200). To know the effect of the number of variables and how changing the shape parameter will affect our result, we simulated the dataset using 17 variables and 27 variables (including the event indicator and the censoring time for each). For each of the 17 variables used, we simulated the variables from different distributions. Five active variables that affect the survival function were simulated from different distributions (we can see the summary statistics of each of these distributions in Table A.1 and Table A.2 of the appendix) and additional 12 variables were included as noise variables (we used different distribution to simulate these noise variables). Also, the 27 variables include 5 active variables and 22 noise variables. Five active covariates from each scenario are used in the simulation study using Weibull distribution, lognormal distribution, Exponential distribution, and Logistic distribution. Differ-

ent numbers of noise variables (about 70% or 80% of total variables) are explored in the simulation.

The shape parameter of Weibull distribution has effect on the failure rate. Weibull distribution with $\tau < 1$ has a failure rate that decreases with time (early-life failure) and Weibull distribution with $\tau > 1$ have a failure rate that increases with time (wear-out failure). In our simulation studies, we used three different number for the shape parameter i.e., $\tau = 0.8, 1, 1.3$. From MMCP penalty, different α parameter value between 0 and 1 will be used to show the effect of α value in the penalty.

Using the coordinate descent algorithms technique from (ncvreg package [92]) as used by [93] and using our simulated dataset, we will compare our derived penalty with different existing penalties. For each of this penalty, we will use different γ value.

5.0.7 The Performance of the Penalty Estimation

5.0.7.1 Performance of the Penalty Estimation Using Simulated Dataset

We compared the existing penalties with our proposed penalty based on the simulated dataset. Increasing the shape parameter from Weibull distribution bring about an increase in the number of events in the dataset. From Table 5.1 with the BIC method, using different shape parameter value from Weibull distribution and using the BIC to select the γ value for the penalty, we observed that when the shape parameter from Weibull distribution increases to 1 (i.e. when τ increases from 0.8 to 1) and using $\alpha = 0.5$, the Cox partial likelihood function with MMCP, MCP and SCAD penalty selected the correct nonzero coefficients with 0 Ave. mFDR (they selected the 5 active variables from the total of 27 covariates). Still using the penalties mentioned above, when $\tau = 1.3$, all three methods (MMCP, MCP and SCAD) picked 7 nonzero coefficients which include all the 5 true

active variables, but SCAD has the worst Ave. mFDR among the three penalties (Ave. mFDR of 0.255). For each of the three shape parameters used, LASSO picked the greatest number of nonzero coefficients and the highest Ave. mFDR value among all the penalties, which indicates LASSO tends to overfit and selects more variables than necessary. In summary, using BIC method to select γ variable, as the number of events increases (by changing the shape parameter of Weibull distribution), the number of nonzero variables selected at each penalty increases. This implies that higher failure rate suffers more from over-fitting. Comparing all the penalties, Figure 5.3 is a plot showing the convexity of each of the four penalties. This shows whether the objective function is convex or not. The shaded region (except for LASSO) is where the objective function is convex and the solution in this shaded region may only be local optima and not global optimal of the objective function. The top left panel corresponds to the MMCP penalty, and the top right plot is based on the MCP penalty. The bottom left plot utilizes the SCAD penalty, and the bottom right plot is based on the LASSO penalty.

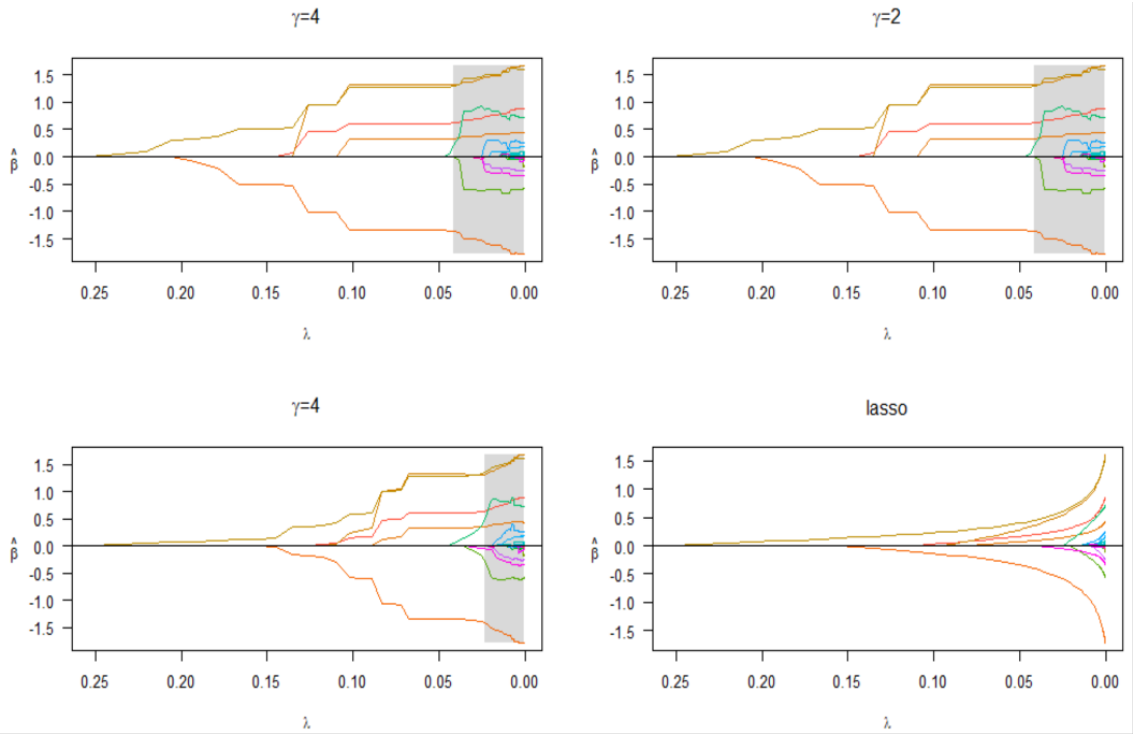


Figure 5.3: Plot to check the local convexity diagnostic for all penalties using the BIC method.

Based on the plot in Figure 5.3, for each penalty when Shape $\tau = 1.3$, we can see that each λ value chosen falls under the convex region. To use the cross validation (C.V) method, γ value selected from the first method (using BIC), will be used using the ncvsurv package. The cross-validation method will then select the best c.v.e. using the best λ value. Table 5.1 with the cross-validation method compared each penalty using 10-fold cross validation, where one-fold serves as test dataset and the remaining 9 folds serve as training set. Different λ values were applied to the loss function with different penalties. The best λ value is the one that offers the lowest c.v.e. As the shape parameter from Weibull distribution increases, both MMCP and MCP selected 5 nonzero coefficients for all the shape parameters (τ) used. SCAD selected 5 nonzero variables when $\tau = 0.8$, 6 variables when $\tau = 1$, and 7 variables when $\tau = 1.3$. LASSO has the worst selection among the penalties. It selected 11 variables (which includes the 5 exact variables) when

$\tau = 0.8$, 16 variables when $\tau = 1$, and 19 variables when $\tau = 1.3$.

For dataset with 17 variables (as seen in Section 2, Table A.3 of appendix, with the BIC method), only Cox partial likelihood function with MMCP penalty selected the correct nonzero coefficient (5 variables) for all the shape parameter used when BIC method is applied. Using the cross-validation method (as seen in the appendix, Table A.3 with the cross-validation method), the Cox partial likelihood function with MMCP, MCP, and SCAD selected 5 nonzero coefficients with the same c.v.e., when $\tau = 1.3$, and when $\tau = 0.8$, these three penalties selected 4 nonzero variables. Comparing this with Table 5.1, using the C.V method with the 27 variables dataset, the Cox partial likelihood function with MMCP and MCP selected the exact number of nonzero coefficients (5 variables) for all the shape (τ) parameters. These penalties also have the same c.v.e. This implies that if we increase the number of variable (p), the Cox partial likelihood function with MMCP and MCP will give a better selection but when we compared these two tables together using the BIC method and focus on MMCP and MCP, MMCP performed better than MCP.

Table 5.1: Comparison of different penalties with BIC and cross validation method using three different shape values

Shape (τ) $= 0.8$ *has* 40 *events*					Shape (τ) $= 1$ *has* 53 *events*				Shape (τ) $= 1.3$ *has* 61 *events*			
BIC method using three different shape values												
Penalty	MMCP	MCP	SCAD	LASSO	MMCP	MCP	SCAD	LASSO	MMCP	MCP	SCAD	LASSO
λ	0.0475	0.0475	0.0475	0.0179	0.0635	0.0635	0.0635	0.0194	0.0389	0.0389	0.0447	0.0363
Nonzero Coef	5	5	5	8	5	5	5	13	7	7	7	9
γ	3	2	4	-	3	2	4	-	4	2	4	-
Avg.mFDR	0	0	0	0.365	0	0	0	0.602	0.204	0.204	0.255	0.431
Cross Validation method using three different shape values												
λ	0.0774	0.0774	0.0475	0.0135	0.0965	0.0965	0.0592	0.0128	0.1034	0.1034	0.0592	0.0168
Nonzero Coef	5	5	5	11	5	5	6	16	5	5	7	19
γ	4	2	4	-	3	2	4	-	4	2	4	-
C.V.E	4.84	4.84	4.87	4.96	4.78	4.78	4.79	4.91	4.99	4.99	5	5.24

Since including α in our model should help our model to select the exact variables to be included in the model as we increase γ, we compared MCP with MMCP penalty by using different γ value. Table 5.2 with BIC method is the table showing the difference between when we apply Cox partial likelihood function with MMCP and MCP penalty as γ increases with a fixed $\alpha = 0.1$ value and $\tau = 0.8$ using BIC method. Based on our previous result with dataset of 27 variables being better than 17 variables, we will only focus on dataset with 27 variables and $\tau = 0.8$ *and* 1. We observed that as γ increases, the Cox partial likelihood function with MMCP penalty chose only the correct 5 nonzero coefficients with zero Ave. mFDR for each γ value while the number of nonzero coefficients selected by MCP increases from 5 to 8 as γ increases with increase in the Ave. mFDR from 0 to 0.362. Based on this result, we can see that as γ increases, the number of nonzero coefficient chosen by MCP increases compared to when we use MMCP penalty. From our derived penalty, as γ increases, α counter the increase in γ since $0 < \alpha < 1$ thereby selecting the right variables. We also used different alpha value to know the effect of α as we reduce it to a number close to zero.

Still on Table 5.2 with cross-validation method when we use $\tau = 0.8$ and $\alpha = 0.1$ value. This plot shows the difference between when Cox partial likelihood function is used with MMCP and MCP as γ increases with cross-validation method. From the table, Cox partial likelihood function with MMCP penalty selected the correct number of nonzero coefficient (5 variables) as γ increases with the lowest c.v.e as compared with Cox partial likelihood function with MCP penalty. For instance, when $\gamma = 30$, MMCP penalty selected 5 nonzero coefficients with c.v.e of 4.89 which is lower than the c.v.e of MCP (5.01) when $\gamma = 30$ and 8 nonzero coefficient were selected. Also, we can see that the C.V.E. of our new model increased at $\gamma = 40$, hence $\gamma = 30$ should be the maximum γ to consider here. We also observed that increasing γ value also increases the c.v.e. We compared when $\alpha = 0.1$ with when $\alpha = 0.2$ with same τ value using the C.V. method

and we noticed that increasing α value will also increase the c.v.e. at each γ value. Since smallest c.v.e gives a better result (better variable selection), then smaller α value will result in having a good variable selection. In order to show the implementation of Cox partial likelihood function with MMCP using real dataset, we applied it to two different data sets.

Table 5.2: Comparison of MMCP and MCP penalties with BIC and cross-validation method using four different gamma values with $\alpha = 0.1$ and 0.2

$\alpha = 0.1, \tau = 0.8$	$\gamma = 10$		$\gamma = 20$		$\gamma = 30$		$\gamma = 40$	
Penalty	MMCP	MCP	MMCP	MCP	MMCP	MCP	MMCP	MCP
BIC method								
λ	0.0475	0.0475	0.0475	0.0335	0.0475	0.0220	0.0475	0.0206
Nonzero Coef	5	5	5	6	5	8	5	8
Avg.mFDR	0	0	0	0.160	0	0.361	0	0.362
Cross Validation method								
λ	0.0830	0.0335	0.0774	0.0385	0.0628	0.0272	0.0475	0.0291
Nonzero Coef	5	7	5	6	5	8	5	7
Cross validation error	4.83	5.06	4.83	5.10	4.83	5.01	4.89	4.92
$\alpha = 0.2, \tau = 0.8$	$\gamma = 10$		$\gamma = 20$		$\gamma = 30$		$\gamma = 40$	
Penalty	MMCP	MCP	MMCP	MCP	MMCP	MCP	MMCP	MCP
BIC method								
λ	0.0475	0.0475	0.0475	0.0335	0.0475	0.0220	0.0475	0.0206
Nonzero Coef	5	5	5	6	5	8	5	8
Avg.mFDR	0	0	0	0.160	0	0.361	0	0.362
Cross Validation method								
λ	0.0475	0.0335	0.0475	0.0385	0.0359	0.0272	0.0335	0.0291
Nonzero Coef	5	7	5	6	7	8	7	7
Cross validation error	4.89	5.06	4.89	5.10	4.89	5.01	5.04	4.92

5.0.7.2 *Performance of the Penalty Estimation Using Heart Failure Dataset*

We applied the Cox partial likelihood function with each of the four penalties to the heart failure dataset. One of the subgroups of all cardiovascular diseases (CVDs) that comprehend cerebrovascular diseases (which is stroke), coronary heart attacks and other pathologies are heart failure. These altogether kill approximately 17 million people every year. This accounts for 31% of all the deaths worldwide. This dataset contains 13 features which can be used to predict death rate by heart failure. We can prevent most of these CVDs by addressing the behavioral risk factors such as unhealthy diet, smoking, and so on. In other to detect which factor is contributing more to the failure of the heart early and how to manage these factors, we will apply the Cox partial likelihood function with different penalties. The reason for using different penalties is to know the best penalty that will detect the right factor contributing to heart failure. The dataset includes the age of the heart failure patient, their sex, whether they have diabetes, anaemia, high blood pressure and so on.

Table 5.3 with BIC method is a table showing the result obtained when Cox partial likelihood function with different penalties were being used with heart failure dataset. Small $\alpha = 0.1$ was used with BIC method. From the table, both MMCP and MCP selected 3 nonzero coefficients which are age, ejection-fraction and serum-creatinine with Ave. mFDR of 0 while SCAD selected 4 nonzero coefficients (age, ejection-fraction, high-blood pressure and serum-creatinine) with 0.173 Ave. mfdr and LASSO selected 7 nonzero coefficients (age, ejection-fraction, high-blood pressure, serum-sodium, anaemia, creatinine-phosphokinase and serum-creatinine) with the highest (worst) Ave. mfdr of 0.406. The nonzero coefficients chosen by MMCP and MCP looks like the best variables based on the Ave. mFDR. To confirm this selection, we applied the C.V. method with the same γ value used for the BIC method.

In Table 5.3 with cross-validation method, the C.V. method was used with a fix $\alpha = 0.1$ and the γ value obtained using BIC to compare all the four penalties. From the result, both Cox partial likelihood function with MMCP and MCP selected the same number of nonzero coefficients that BIC method selected. Comparing these two penalties, the c.v.e. for MMCP penalty (10.03) is smaller than that of MCP which is 10.04. Also, the R-square value obtained using MMCP penalty is higher (0.52) than the R-square for MCP penalty (0.44). SCAD and LASSO penalties selected 5 and 7 nonzero coefficients respectively. From this result, we can conclude that MMCP with the minimum c.v.e and highest R-square value performed better although MCP also selected the same number of non-zero coefficient as MMCP but higher c.v.e. So, age, ejection-fraction and serum-creatinine are the right factors contributing to heart failure in this analysis.

Figure 5.4 is the cross-validation plot for each of the penalties using the C.V method. It shows how each objective function chose their nonzero coefficient. The top left plot is for the MMCP objective function plot, the top right plot is the MCP objective plot, the bottom left plot is the SCAD objective function plot, and the bottom right plot is the LASSO objective function plot. Also, Figure 5.5 is the R-square plot for each of the penalties. The top left plot is for the MMCP objective function plot, the top right plot is the MCP objective plot, the bottom left plot is the SCAD objective function plot, and the bottom right plot is the LASSO objective function plot.

Table 5.3: Comparison of different penalties with BIC and cross validation method using heart failure dataset

Using $\alpha = 0.1$				
Penalty	MMCP	MCP	SCAD	LASSO
BIC method				
λ	0.0701	0.0701	0.0569	0.0200
Nonzero Coef	3	3	4	7
Avg.mFDR	0	0	0.173	0.406
γ	16	2	3	
Cross Validation method				
λ	0.0864	0.0752	0.0495	0.0230
Nonzero Coef	3	3	5	7
γ	16	2	3	
Cross validation error	10.03	10.04	10.06	10.05
R-Square	0.52	0.44	0.42	0.42

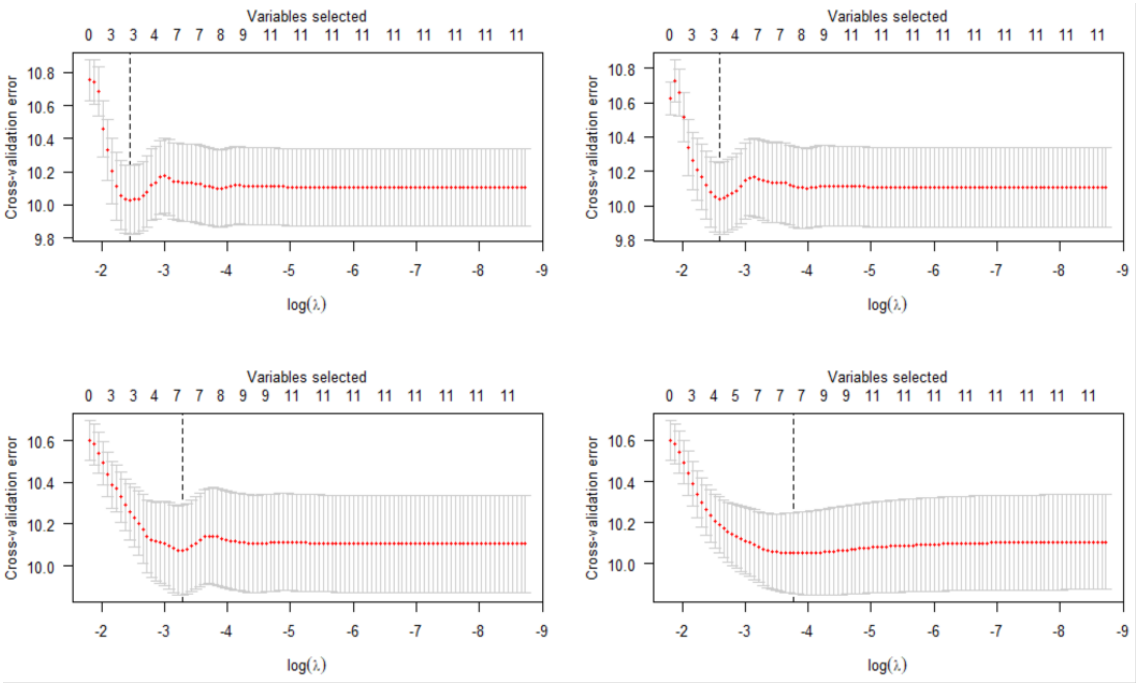


Figure 5.4: Cross validation error plot for each penalties using the C.V method

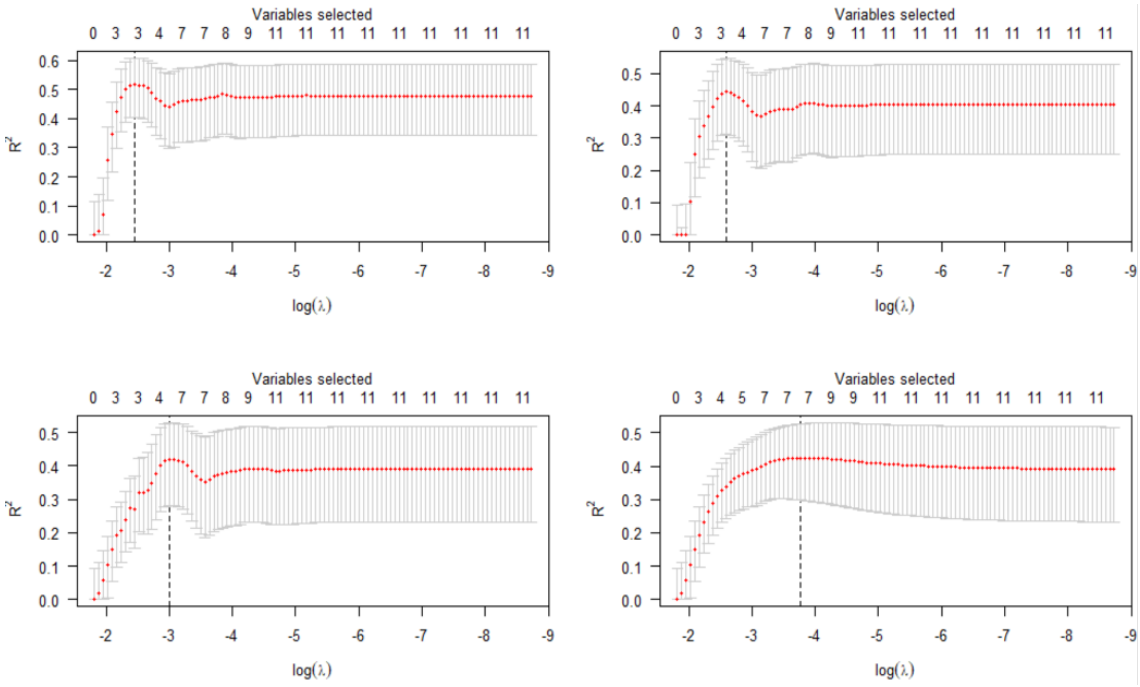


Figure 5.5: R-square plot for each penalties using the C.V method

Table 5.4: Comparison of MMCP and MCP penalties with BIC and cross validation method using two different gamma values for heart failure dataset

Using $\alpha = 0.1$				
	$\gamma = 10$		$\gamma = 20$	
Penalty	MMCP	MCP	MMCP	MCP
BIC method				
λ	0.0701	0.0230	0.0701	0.0214
Nonzero Coef	3	7	3	7
Avg.mFDR	0	0.421	0	0.414
Cross Validation method				
λ	0.0927	0.0325	0.0752	0.0303
Nonzero Coef	3	7	3	7
Cross validation error	10.03	10.09	10.04	10.07
R-Square	0.52	0.40	0.44	0.41

5.0.7.3 *Performance of the Penalty Estimation Using NKI Breast Cancer Dataset*

The relationship between gene expression profiles and survival data has so far been explored such as the breast cancer therapy [19], prediction of outcome of the origins of cancer in lung [116], and so on. Many articles have used Cox regression model with hierarchical clustering, to categorize patients according to their risk levels [107] and genes that are significant are selected using this regression model [2]. The nonlinear relationships between the gene expression level and survival time need to be accounted for in a flexible way and Bayesian approach was applied by some researchers. [25].

We will analyze a disparate dataset of gene expression profiling data (breast cancer gene expression data set) by applying the Cox partial likelihood function with MMCP penalty and other penalties discussed before. Breast cancer data sets, NKI will be used to know how effective the new penalty (MMCP) works. The NKI dataset, consists of gene expression levels extracted from 272 tumors (breast cancer patients) and is analyzed using about 1570 most varying genes. Using this dataset, since the number of variables is higher than the number of observations $(p > n)$, then, we expect MMCP, MCP and SCAD to perform better than LASSO since LASSO works best when $p < n$. The Cox partial likelihood function with each of the four penalties was applied to the NKI dataset.

From Table 5.5 with BIC method, the Cox partial likelihood function with different penalties was used with the NKI dataset and using the BIC method, with $\alpha = 0.3$ we were able to get our result. From the table, both MMCP and MCP selected 11 nonzero coefficients which include the gene expression and one other variable (grade). These variables are NM-016359, grade, NM-003430, NM-001333, NM-006096, NM-000926, Contig23211-RC, NM-003981, Contig56390-RC, NM-003258, NM-016109 with Avg. mFDR of 0.098 for MMCP and 0.097 for MCP while SCAD and LASSO have no nonzero coefficients. To confirm this selection, we applied the C.V. method with the same γ value used for the BIC

method.

In Table 5.5 with cross validation method, the C.V. method was used with a fix $\alpha = 0.3$ to compare all the four penalties. From the result, Cox partial likelihood function with MMCP selected the smallest number of non-zero coefficients compared to the remaining penalties. Also, MMCP has c.v.e. of 9.93 and this is the smallest c.v.e. when we compare this with the c.v.e. obtained from other penalties. Although MMCP selected the smallest number of coefficients which is 14, (NM-016359, grade, AF201951, AL117638, NM-003430, NM-001333, NM-006096, NM-000926, Contig23211-RC, Contig42011-RC, Contig56390-RC, NM-003258, NM-016109, L27560) and also has the smallest c.v.e., it has the same R-square value of 0.43 as the one obtained when MCP was used. SCAD and LASSO penalties selected the same number of nonzero coefficients (16 coefficients). From this result, we can conclude that MMCP with the minimum c.v.e value performed better. So, NM-016359, grade, AF201951, AL117638, NM-003430, NM-001333, NM-006096, NM-000926, Contig23211-RC, Contig42011-RC, Contig56390-RC, NM-003258, NM-016109, L27560 are the factor contributing to NKI breast cancer in this analysis.

Table 5.5: Comparison of different penalties with BIC and cross validation method using NKI breast cancer dataset

Using $\alpha = 0.3$				
Penalty	MMCP	MCP	SCAD	LASSO
BIC method				
λ	0.0948	0.09481	-	-
Nonzero Coef	11	11	-	-
Avg.mFDR	0.098	0.097	-	-
Cross Validation method				
λ	0.0840	0.0840	0.0840	0.0840
Nonzero Coef	14	15	16	16
γ	350	250	150	-
Cross validation error	9.93	9.94	9.94	9.94
R-Square	0.43	0.43	0.42	0.42

We reduced the NKI dataset by removing all the gene expressions and we analyzed the remaining 12 variables (age, chemo, hormonal, amputation, histtype, diam, posnodes, grade, angioinv, lymphinfil) using the Cox partial likelihood function with the four penalties. Table 5.6 is the table showing the comparison between different penalties using the reduced NKI dataset by using both BIC and C.V method with BIC and cross validation method. From Table 5.6 with BIC method, Cox partial likelihood function with MMCP and MCP selected 1 non-zero coefficient (grade) with zero Average mFDR. This shows that the performance of all the penalties is similar because $p < n$. From Table 5.6 with cross validation method, MMCP and MCP selected 1 non-zero coefficient with the smallest c.v.e and the highest R-square value. SCAD on the other hand, selected 2 nonzero coefficients while LASSO selected 4 nonzero coefficients with the highest c.v.e. From our observation, both MMCP and MCP performed better than other penalties for the reduced NKI dataset.

Table 5.6: Comparison of different penalties with BIC and cross validation method using reduced NKI breast cancer dataset

Using $\alpha = 0.3$				
Penalty	MMCP	MCP	SCAD	LASSO
BIC method				
λ	0.0942	0.0665	0.0665	0.0764
Nonzero Coef	1	1	1	1
Avg.mFDR	0	0	0	0
γ	3	2	3	-
Cross Validation method				
λ	0.0879	0.0879	0.0620	0.0437
Nonzero Coef	1	1	2	4
γ	3	2	3	-
Cross validation error	10.03	10.03	10.05	10.09
R-Square	0.36	0.36	0.35	0.32

Table 5.7 is the table showing the comparison between MMCP and MCP as we change γ value using the reduced NKI dataset with BIC and cross validation method. From this table, as γ increases for the BIC method, both MMCP and MCP selected 1 non-zero coefficient (grade) with zero Average mFDR. From Table 5.7 with cross validation method, for $\gamma = 10$, MMCP has the smallest c.v.e of 10.06 and the highest R-square value of 0.34. With this, Cox partial likelihood function with MMCP selected the smallest non-zero coefficient which is 1. As γ increases to 30, both MMCP and MCP selected 4 nonzero coefficients with the same c.v.e. and same R-square value. This shows that when γ is at 10, MMCP performed better than MCP and as we increase γ to 30, they both performed similar. In this case, The α value and the γ value with the best c.v.e and the smallest Avg. mFDR was used for the final model selection.

Table 5.7: Comparison of MMCP and MCP penalties with BIC and cross validation method using two different gamma values for reduced NKI dataset.t

Using $\alpha = 0.2$				
	$\gamma = 10$		$\gamma = 30$	
Penalty	MMCP	MCP	MMCP	MCP
BIC method				
λ	0.0665	0.0713	0.0713	0.0764
Nonzero Coef	1	1	1	1
Avg.mFDR	0	0	0	0
Cross Validation method				
λ	0.0927	0.0325	0.0752	0.0303
Nonzero Coef	1	4	4	4
Cross validation error	10.06	10.09	10.09	10.09
R-Square	0.34	0.32	0.32	0.32

Chapter 6: Conclusion and Contribution

6.1 Demonstration Test Plans For Lifetime Data Based on Considering Multiple Objectives

For conducting a demonstration test, it is important to choose a plan that is the best and using a zero-failure test with low test units can cause a trade-off between the CR and PR. When we control the CR, it leads to a decrease in the AP. We focused on the producer's risks, the consumer's risk, the acceptance probability, the testing time and the test unit for a successful test. Having given the contending goals for advancing a demonstration test plan, we recommend utilizing a Pareto front to remove choices that are non-contending so as to guide us in making justifiable decisions. With the three scenarios that we explored, if we first control the consumer's risk using prior Invgamma(8,0.7), it will result in a simple set of excellent solutions with a great plan to simultaneously improve the rest of the criteria for each possible c value. After finding the best solutions with the trade-offs summarized in Table 3.1, the user can then make a decision by using their AP and PR requirement, their requirement on the sample test and how long they want to test the units for. Finally, the choice of the prior distribution and the threshold of the user in a Bayesian analysis can substantially affect our ultimate choice.

We explored the impact of t_0 and found out that it has small impact on the selective test plan. For our future work, we will try and explore different test unit to see the effect of the cost (test unit).

6.2 Bayesian Analysis For Accelerated Degradation Test Data With Multiple Degradation Measurements and Covariates Using the General Path Model

In this chapter, the multivariate degradation path model with two random effect was proposed. Based on the correlation between the two random effects, this model captures the unit-to-unit variation between them. The MCMC algorithm from Stan package was used to obtain the model parameters and system reliability estimate and we demonstrated this by analyzing the synthetic ISO dataset. The ISO dataset was divided into equal size of two groups and using the measurements from the two units from both groups, a multivariate model was created. The result obtained shows similarity in the multivariate and the independent model due to low correlation between the degradation measurement. The product reliability was estimated by simulating new data with two correlation level using ISO dataset with our degradation model. From these two applications, the multivariate degradation path models of same structures can be estimated using the developed model (MCMC). Advantage of using the MCMC is that samples from some probability distributions can be drawn with MCMC and also computation of the inference is fast when estimating the MCMC parameter. Stan was used for the MCMC algorithm and like most other HMC implementations, Stan uses the leapfrog integrator which is a numerical integration algorithm. Leapfrog was used by Stan to give a stable result for Hamiltonian systems of equations. Based on our more detailed prediction, it would help manufacturers to have a better understanding of their products' reliability performance for a whole service life.

6.3 Penalized Regression For Survival Analysis

In this work, in order to solve the challenges faced in selecting variables in survival analysis when there is a large number of predictor variables relative to the number of observed data, the Modified Minimax Concave penalty (MMCP) was introduced and dis-

cussed extensively. We modified the Minimax Concave penalty (MCP) to achieve this our newly derived penalty by applying α variable to the penalty to slow down the rate at which unbiasedness reduced. We demonstrated the applicability of Cox partial likelihood function with MMCP penalty by using simulated dataset with 200 observations and two real life data sets, namely, the heart failure dataset and the NKI breast cancer dataset. We also fitted these datasets to other existing penalties through cox loss function. The Bayesian Information Criteria and the cross-validation (using cross-validation error with R-square) were used to select the best penalty (penalty that choose the right non-zero coefficient). We also compared our derived penalty (MMCP) with MCP penalty as γ increases. The result obtained in Table 5.4 through Table 5.7 shows that our new penalty performs better than other existing penalties when $p < n$ and when $p > n$. Our believe is that this new penalty will be used to select many significant variables present in many real-life data. For future work, we will try and apply the full Bayesian approach as a criterion-based method for the selection of variables.

References

[1] Iso 8402. *Quality vocabulary*, 50:63–68, 1986.

[2] Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature medicine*, 14(8):822–827, 2008.

[3] Test method for the estimation of the archival lifetime of optical media. *Organization International Standards*, 2011.

[4] Bertsimas, D., King, A. and Mazumder, R. Best subset selection via a modern optimization lens. *The annals of statistics*, 44(2):813–852, 2016.

[5] Blumenthal, S., Greenwood, J. A., and Leon, H. H. Series systems and reliability demonstration tests. *Oper. Res.*, 32(3):641–648, 1984.

[6] Bθvelstad, H. M., Nygård, S., Stθrvold, H. L., Aldrin, M., Borgan, Ø., Frigessi, A., and Lingjærde, O. C. Predicting survival from microarray data—a comparative study. *Bioinformatics*, 23(16):2080–2087, 2007.

[7] Doksum, K. A., and Hbyland, A. Models for variable-stress accelerated life testing experiments based on wener processes and the inverse gaussian distribution. *Technometrics*, 34(1):74–82, 1992.

[8] Efroymson, M. A. Multiple regression analysis. *Mathematical Methods for Digital Computers*, 1960.

[9] Elwany, A., and Gebraeel, N. Real-time estimation of mean remaining life using sensor-based degradation models. 2009.

[10] Escobar, L. A., and Meeker, W. Q. A review of accelerated test models. *Statistical science*, pages 552–577, 2006.

[11] Guo, H., Honecker, S., Mettas, A., and Ogden, D. Reliability estimation for one-shot systems with zero component test failures," in proceedings. 2010.

[12] Hoffman, M. D., Gelman, A., et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[13] Höskuldsson, A. Variable and subset selection in pls regression. *Chemometrics and intelligent laboratory systems*, 55(1-2):23–38, 2001.

[14] Kleyner, A. *Reliability demonstration in product validation testing*. in Handbook of Performability Engineering. Berlin, Germany: Springer-Verlag, 2008.

[15] Meeker, W. Q., Escobar, L. A. and Hong, Y. Using accelerated life tests results to predict product field reliability. *Technometrics*, 51(2):146–161, 2009.

[16] Miller, A. *Subset selection in regression*. chapman and hall/CRC, 2002.

[17] Whitmore, G. A., and Schenkelberg, F. Modelling accelerated degradation data using wiener diffusion with a time scale transformation. *Lifetime data analysis*, 3(1):27–45, 1997.

[18] Park, C., and Padgett, W. J. Stochastic degradation models with several accelerating variables. *IEEE Transactions on Reliability*, 55(2):379–390, 2006.

[19] Sotiriou, C., and L. Pusztai. Gene-expression signatures in breast cancer. *New England Journal of Medicine*, 360(8):790–800, 2009.

[20] Tal, O., McCollin, C., and Bendell, T. Reliability demonstration for safety-critical systems. *in IEEE Transactions on Reliability*, 50(2):194–203, 2001.

[21] Waldron, L., Pintilie, M., Tsao, M.-S., Shepherd, F. A., Huttenhower, C., and Jurisica, I. Optimized application of penalized regression methods to diverse genomic data. *Bioinformatics*, 27(24):3399–3406, 2011.

[22] Willits, C. J., Dietz, D. C., and Moore, A. H. Series-system reliability estimation using very small binomial samples. *IEEE Trans. Rel.*, 46:296–302, 1997.

[23] Si, X.-S., Wang, W., Hu, C.-H., and Zhou, D.-H. Estimating remaining useful life with three-source variability in degradation modeling. *IEEE Transactions on Reliability*, 63(1):167–190, 2014.

[24] Zhang, C.H. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

[25] Bonato, V., Baladandayuthapani, V., Broom, B. M., Sulman, E. P., Aldape, K. D., and Do, K.-A. Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics*, 27(3):359–367, 2011.

[26] Kececioglu, D. Reliability life testing handbook. *Englewood Cliffs, NJ: Prentice-Hall, Inc.*, 1994.

[27] Lewandowski, D., Kurowicka, D., and Joe, H. Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, 100(9):1989–2001, 2009.

[28] Sun, D., and Berger, J. O. Bayesian sequential reliability for weibull and related distributions. *Ann. Inst. Statist. Math.*, 46(2):221–249, 1994.

[29] Hoch, J. C. Donoho, D. L., Johnstone, I. M. and Stern, A. S. Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(1):41–67, 1992.

[30] Fang, G., Rigdon, S. E., and Pan, R. Predicting lifetime by degradation tests: A case study of iso 10995. *Quality and Reliability Engineering International*, 34(6):1228–1237, 2018.

[31] Hoerl, A. E., and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[32] Kenett, R.S., Baker, E. Process improvement and cmmi for systems and software. *CRC Press*, 2010.

[33] Miller, R. E., and P. Breheny. Marginal false discovery rate control for likelihood-based penalized regression models. *Biometrical Journal*, 61(4):889–901, 2019.

[34] Nikles, D. E., and Wiest, J. M. Accelerated aging studies and the prediction of the archival lifetime of optical disk media. In *Recent Advances in Metrology, Characterization, and Standards for Optical Digital Data Disks*, volume 3806, pages 30–34. SPIE, 1999.

[35] Cheng, S., Li, B., Yuan, Z., Zhang, F., and Liu, J. Development of a lifetime prediction model for lithium thionyl chloride batteries based on an accelerated degradation test. *Microelectronics Reliability*, 65:274–279, 2016.

[36] Lawless, J. F. *Statistical Models and Methods for Lifetime Data, 2nd ed. Hoboken, NJ, USA*. Wiley, 2003.

[37] Liu, J. F. On time-sequential test for a class of distributions. *Statistica Sinica*, 5(1):251–260, 1995.

[38] Slattery, O., Lu, R., Zheng, J., Byers, F., and Tang, X. Stability comparison of recordable optical discs—a study of error rates in harsh conditions. *Journal of Research of the National Institute of Standards and Technology*, 109(5):517, 2004.

[39] Fan, J., Li, G., and R. Li. An overview on variable selection for survival analysis. *Contemporary Multivariate Analysis And Design Of Experiments: In Celebration of Professor Kai-Tai Fang's 65th Birthday*, pages 315–336, 2005.

[40] Jin, G., and Matthews, D. Reliability demonstration for long-life products based on degradation testing and a Wiener process model. *in IEEE Transactions on Reliability*, 63(3):781–797, 2014.

[41] Schwarz, G. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[42] Trevisanello, L., Meneghini, M., Mura, G., Vanzi, M., Pavesi, M., Meneghesso, G., and Zanoni, E. Accelerated life test of high brightness light emitting diodes. *IEEE Transactions on Device and Materials Reliability*, 8(2):304–311, 2008.

[43] Yang, G. Environmental-stress-screening using degradation measurements. *IEEE Transactions on Reliability*, 51(3):288–293, 2002.

[44] Yang, G. *Life cycle reliability engineering*. John Wiley & Sons, 2007.

[45] Whitmore, G.A. Estimating degradation by a wiener diffusion process subject to measurement error. *Lifetime data analysis*, 1(3):307–319, 1995.

[46] Pasha, G.R., and Shah, M.A. Application of ridge regression to multicollinear data. *Journal of research (Science)*, 15(1):97–106, 2004.

[47] Ahmed, H., and Chateauneuf, A. How few tests can demonstrate the operational reliability of products. *Quality Technol. Quant. Manag.*, 8(4):411–428, 2011.

[48] Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

[49] Bennett, H. Understanding cd-r and cd-rw. 2003.

[50] Hao, H., and Su, C. A bayesian framework for reliability assessment via wiener process and mcmc. *Mathematical Problems in Engineering*, 2014, 2014.

[51] Ishwaran, H. Applications of hybrid monte carlo to bayesian generalized linear models: Quasicomplete separation and neural networks. *Journal of Computational and Graphical Statistics*, 8(4):779–799, 1999.

[52] Liao, H., and Tian, Z. A framework for predicting the remaining useful life of a single unit under time-varying operating conditions. *Iie Transactions*, 45(9):964–980, 2013.

[53] Schumacher, M., Binder, H., and T. Gerds. Assessment of survival prediction models based on microarray data. *Bioinformatics*, 23(14):1768–1774, 2007.

[54] Trautmann, H., and Mehnen, J. Preference-based pareto optimization in certain and noisy environments. *Engineering Optimization*, 41:23–38, 2009.

[55] Zou, H., and Li, R. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of statistics*, 36(4):1509–1533, 2008.

[56] Zou, H., and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

[57] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[58] Chao, H., Hu, B., Xie, K., Tai, H.-M., Yan, J., and Li, Y. A sequential mcmc model for reliability evaluation of offshore wind farms considering severe weather conditions. *IEEE Access*, 7:132552–132562, 2019.

[59] Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

[60] Fan, J., and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[61] Gui, J., and H. Li. Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, 21(13):3001–3008, 2005.

[62] Lawless, J., and Crowder, M. Covariates and random effects in a gamma process model with application to degradation and failure. *Lifetime data analysis*, 10(3):213–227, 2004.

[63] Lu, C. J., and Meeker, W. O. Using degradation measures to estimate a time-to-failure distribution. *Technometrics*, 35(2):161–174, 1993.

[64] Meeker, W. Q., Escobar, L. A., Lu, C. J. Accelerated degradation tests: modeling and analysis. *Technometrics*, 40(2):89–99, 1998.

[65] Meeker, W. Q., Hahn, G. J., and Doganaksoy, N. Reliability verification testing, and analysis in engineering design. new york, ny, usa. *Marcel Dekker*, 2003.

[66] Meeker, W. Q., Hahn, G. J., and Doganaksoy, N. Planning life tests for reliability demonstration. *Quality Progress*, page 80–82, August 2004.

[67] Padgett, W. J., and Wei, L. J. A sequential test and interval estimation in time truncated life testing. *India J. Statist.*, 44(2):242–250, 1982.

[68] Ranstam, J., and Cook, J.A. Lasso regression. *Journal of British Surgery*, 105(10):1348–1348, 2018.

[69] Sun, Q., Zhang, Z., Feng, J., and Pan, Z. A zero-failure reliability demonstration approach based on degradation data. In *2012 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering*, pages 947–952, 2012.

[70] Deb, K. *Multi-Objective Optimization using Evolutionary Algorithms*. Wiley-Interscience Series in Systems and Optimization, Baffins Lane, Chichester, West Sussex, England: John Wiley and Sons Ltd, 2001.

[71] Lange , K. *Optimization, 2nd edition. New York, NY:*. Springer, 2013.

[72] Breiman, L. Heuristics of instability and stabilization in model selection. *The annals of statistics*, 24(6):2350–2383, 1996.

[73] Donoho, D. L. and Johnstone, J.M. Ideal spatial adaptation by wavelet shrinkage. *biometrika*, 81(3):425–455, 1994.

[74] Lu, L., Mingyang, L., and Anderson-Cook, C. M. Multiple objective optimization in reliability demonstration tests. *Journal of quality Technology*, 48(4), 2016.

[75] Ten, L., and Xie, M. Bayes reliability demonstration test plan for series-systems with binomial subsystem data. *Annual Reliability and Maintainability Symposium. Proceedings. International Symposium on Product Quality and Integrity*, pages 241–246, 1998.

[76] Tietjen, G. L. Reliability and survival analysis. In *A Topical Dictionary of Statistics*, pages 79–87. Springer, 1986.

[77] Carpenter, B., Hoffman, D. M., and A. Gelman. Stan, scalable software for bayesian modeling. in proceedings of the nips workshop on probabilistic programming. In *In Proceedings of the NIPS Workshop on Probabilistic Programming*, 2012.

[78] Hong, Y., Zhang, M., and Meeker, W. Q. Big data and reliability applications: The complexity dimension. *Journal of Quality Technology*, 50(2):135–149, 2018.

[79] Kasprzak, E. M., and Lewis, K. E. Pareto analysis in multiple optimization using the collinearity theorem and scaling method.

[80] Krasich, M. Accelerated testing for demonstration of product lifetime reliability. *Annual Reliability and Maintainability Symposium*, pages 117–123, 2003.

[81] Leemis, L. M. Lower system reliability bounds from binary failure data using bootstrapping. *Q. Quality Technol.*, 38:2–13, 2006.

[82] Leemis, L. M., and Trivedi, K. S. A comparison of approximated interval estimators for the bernoulli parameter. *The American Statistician*, 50:63–68, 1996.

[83] Liu, L., Li, X. Y., Jiang, T. M., and Sun, F. Q. Utilizing accelerated degradation and field data for life prediction of highly reliable products. *Quality and Reliability Engineering International*, 32(7):2281–2297, 2016.

[84] Lu, L., Anderson-Cook, C. M., and Robinson, T. J. Optimization of designed experiments based on multiple criteria utilizing a pareto frontier. *Technometrics*, 53:353–365, 2011.

[85] Neal, R. M., et al. Bayesian learning for neural networks. *Lecture Notes in Statistics*, (118), 1996a.

[86] Si, X.-S., Hu, C.-H., Chen, M.-Y., and Wang, W. An adaptive and nonlinear drift-based wiener process for remaining useful life estimation. In *2011 Prognostics and System Health Managment Confernece*, pages 1–5. IEEE, 2011.

[87] Colin L Mallows. Some comments on cp. *Technometrics*, 42(1):87–94, 2000.

[88] MIL-HDBK-781A. *Reliability Test Methods; Plans, and Environments for Engineering Development, Qualification, and Production*. Department of Defense, Washington, DC, USA, 1996.

[89] Schmidt, M. N. Function factorization using warped gaussian processes. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 921–928, 2009.

[90] Ye, Z.S., Chen, N., and Shen, Y. A new class of wiener process models for degradation analysis. *Reliability Engineering & System Safety*, 139:58–67, 2015.

[91] Soliman, A. A., Abd-Ellah, A.H., Abou-Elheggag, N.A., and Ahmed, E. A. Reliability estimation in stress–strength models: an mcmc approach. *Statistics*, 47(4):715–728, 2013.

[92] Breheny, P. Ncvreg package. *CRAN repository*, 2021.

[93] Breheny, P., and Huang, J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The annals of applied statistics*, 5(1):232, 2011.

[94] Foster, D. P., and George, E. I. The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22(4):1947–1975, 1994.

[95] Goel, M. K., Khanna, P., and Kishore, J. Understanding survival analysis: Kaplan-meier estimate. *International journal of Ayurveda research*, 1(4):274, 2010.

[96] Limon, S., Yadav, O. P., and Liao, H. A literature review on planning and analysis of accelerated testing for reliability assessment. *Quality and Reliability Engineering International*, 33(8):2361–2383, 2017.

[97] Breheny, P.J. Marginal false discovery rates for penalized regression models. *Biostatistics*, 20(2):299–314, 2019.

[98] Kahan, P.T. A study of the eyring model and its application to component degradation: A conceptual introduction to hamiltonian monte carlo. *Journal of Computational and Graphical Statistics*, 8(4):779–799, 1999.

[99] McKane, S. W., Escobar, L. A. , Meeker, W. Q. Theory for optimum accelerated censored life tests for weibull and extreme value distribution. *Technometrics*, 47:182–190, 2005.

[100] Meeker, W. Q., and Nelson, W. Weibull variances and confidence limits by maximum likelihood for singly censored data. *Technometrics*, 19:473–476, 1977.

[101] Nelson, W., Meeker, W. Q. Theory for optimum accelerated censored life tests for weibull and extreme value distribution. *Technometrics*, 20:171–177, 1978.

[102] Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[103] Gebraeel, N. Z., Lawley, M. A., Li, R., and Ryan, J. K. Residual-life distributions from component degradation signals: A bayesian approach. *IiE Transactions*, 37(6):543–557, 2005.

[104] Lau, B., Cole, S. R., and S. J. Gange. Competing risk regression models for epidemiologic data. *American journal of epidemiology*, 170(2):244–256, 2009.

[105] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[106] Tibshirani, R. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

[107] Vasselli, J. R., Shih, J. H., Iyengar, S. R., Maranchie, J., Riss, J., Worrell, R., Torres-Cabala, C., Tabios, R., Mariotti, A., Stearman, R., et al. Predicting survival in patients with metastatic kidney cancer by gene-expression profiling in the primary tumor. *Proceedings of the National Academy of Sciences*, 100(12):6958–6963, 2003.

[108] Hamada, M. S., Wilson, A. G., Reese, C. S., and Martz, H. F. *Bayesian Reliability*. Springer, 2008.

[109] Kahng, A. B., Mantik, S., and Markov, I. L. Min-max placement for large-scale timing optimization. In *Proceedings of the 2002 international symposium on Physical design*, pages 143–148, 2002.

[110] Le Cessie, S. and Van Houwelingen, J. C. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(1):191–201, 1992.

[111] Seppelt, R., Lautenbach, S. and Volk, M. Identifying trade-offs between ecosystem services, land use, and biodiversity: a plea for combining scenario analysis and optimization on different spatial scales. *Current Opinion in Environmental Sustainability*, 5:458–463, 2013.

[112] Bender, R., Augustin, T., and Blettner, M. Generating survival times to simulate cox proportional hazards models. *Statistics in medicine*, 24(11):1713–1723, 2005.

[113] Gronwald, W., Hohm, T., and Hofmann, D. Transmuted Pareto distribution. *ProbStat Forum*, 07:1–11, 2014.

[114] Johnsson, T. A procedure for stepwise regression analysis. *Statistical Papers*, 33(1):21–29, 1992.

[115] Utazirubanda, J. C., M. León, T., and P. Ngom. Variable selection with group lasso approach: Application to cox regression with frailty model. *Communications in Statistics-Simulation and Computation*, 50(3):881–901, 2021.

[116] Herbst, R. S., Heymach, J. V., and Lippman, S. M. Molecular origins of cancer: Lung cancer. *New England Journal of Medicine*, 359(13):1367–1380, sep 2008.

[117] Nguyen, D. V., and D. M. Rocke. Partial least squares proportional hazard regression for application to dna microarray survival data. *Bioinformatics*, 18(12):1625–1632, 2002.

[118] Abou-Amara, I., Evans, G. W., and Usher, J. S. A multi-objective approach to reliability demonstration testing. *Qualification, and Production" Department of Defense, Washington, DC, USA*, pages MIL–HDBK–781A, 1996.

[119] Alyson, G. W., and Kassandra, M. F. Bayesian reliability: Combining information. *Journal of Quality Engineering*, 29, 2017.

[120] Huang, W., and Askin, R. G. Reliability analysis of electronic devices with multiple competing failure modes involving performance aging degradation. *Quality and Reliability Engineering International*, 19(3):241–254, 2003.

[121] Liu, W., and Li, Q. An efficient elastic net with regression coefficients method for variable selection of spectrum data. *PloS one*, 12(2):e0171122, 2017.

[122] Lu, M. W., and Ruddy, R. J. Laboratory reliability demonstration test considerations. *IEEE Trans. Rel.*, 50(1):12–16, Mar. 2001.

[123] Nelson, W. Analysis of performance-degradation data from accelerated tests. *IEEE Transactions on Reliability*, 30(2):149–155, 1981.

[124] Wang, X. Wiener processes with random effects for degradation data. *Journal of Multivariate Analysis*, 101(2):340–351, 2010.

[125] Wang, X., and Xu, D. An inverse gaussian process model for degradation data. *Technometrics*, 52(2):188–197, 2010.

[126] Ke, H. Y. Sampling plans for vehicle component reliability verification. *Quality Rel. Eng. Int.*, 15(5):363–368, 1999.

[127] Lu, L., Wang, B., Hong, Y., and Ye, Z. General path models for degradation data with multiple characteristics and covariates. *Technometrics*, 63(3):354–369, 2021.

[128] Nesterov, Y. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

[129] Wu, Y. Elastic net for cox's proportional hazards model with a solution path algorithm. *Statistica Sinica*, 22:27, 2012.

[130] Xu, Z., Hong, Y., and Jin, R. Nonlinear general path models for degradation data with dynamic covariates. *Applied Stochastic Models in Business and Industry*, 32(2):153–167, 2016.

[131] German-Sallo, Z. Nonlinear wavelet denoising of data signals. *UbiCC J*, 6:895–900, 2011.

[132] Ye, Z.S., and Xie, M. Stochastic modelling and analysis of degradation for highly reliable products. *Applied Stochastic Models in Business and Industry*, 31(1):16–32, 2015.

[133] Ye, Z.S., and Chen, N. The inverse gaussian process as a degradation model. *Technometrics*, 56(3):302–311, 2014.

Appendix A: Supplementary Materials

A.1 Penalized Regression For Survival Analysis

A.1.1 Additional Results For the Simulation Study

Table A.1 and A.2 are the tables showing the summary statistics of each of the distributions used in the simulation study. Table A.1 is the table for the summary statistics of when 27 variables are used while Table A.2 is the table for the 17 variables used in the simulation study. Table A.3 is the table showing the Comparison between different penalties with BIC and cross validation method when we have 17 variables by using three different shape values of 0.8, 1.1 and 1.3.

Table A.4 (27 variables) is the table showing the Comparison between MMCP and MCP penalties with BIC and cross validation method using four different gamma values for when alpha=0.2 and $\tau = 0.8$, and $\alpha = 0.2$ and $\tau = 1.1$ respectively. While table A.5 shows the difference between MMCP and MCP penalties with BIC and cross validation method using four different gamma values for when $\alpha = 0.1$ and $\tau = 1.1$.

A.2 Additional Results For the NKI Breast Cancer Data Example

Table A.6 compares the MMCP and MCP penalties using the BIC and cross validation method by applying four different gamma values of 10, 20, 30 and 40 with $\alpha = 0.3$ while Table A.7 compares the MMCP and MCP penalties with cross validation method using four different gamma values with $\alpha = 0.3$.

Table A.1: Summary statistics for each of the 27 variables used for the simulation study.

The 5 active variables from the 22 variables	Summary Statistics		
Distributions	minimum	mean	maximum
Weibul (1,2.9)	0.0037	2.9572	13.5050
Normal (1,3.2)	-9.3461	1.3134	10.7401
Lognormal (0.4,1)	0.0500	2.6074	36.4485
Exponential (0.7)	0.0021	1.3910	8.6044
Logistics (3,2.9)	-13.0807	3.1644	19.7402
Noise from the remaining 27 variables			
Negativebinomial (2,0.4)	0.0000	3.0300	14.0000
Beta (0.03,1,4)	0.0000	0.6057	0.9970
Binomial (2,0.7)	0.0000	1.365	2.0000
Cauchy (3, 1.5)	-19.1060	3.8710	156.4460
Gamma (2, 0.3)	0.2529	6.4927	33.2214
Geometric (0.3)	0.0000	2.2300	17.0000
Hypergeometric (2, 12,8)	0.0000	1.1250	2.0000
Negativebinomial (5,0.4)	0.0000	7.2800	19.0000
Uniform (-1, 7)	-0.9906	2.9662	6.9085
Poisson (9)	2.0000	8.7500	18.0000
Negativebinomial (3,0.4)	0.0000	4.4650	21.0000
Uniform (-1, 5)	-0.9064	2.1536	4.9533
Binomial (5, 0.7)	1.0000	3.6800	5.0000
Cauchy (4, 1.5)	-144.4820	3.0330	101.2290
Gamma (3, 0.3)	1.3970	10.2130	30.3590
Geometric (0.5)	0.0000	0.9850	5.0000
Hypergeometric (12,12,8)	1.0000	3.9900	7.0000
Negativebinomial (6,0.4)	1.0000	9.5100	29.0000
Hypergeometric (11,11,8)	1.0000	4.0450	7.0000
Poisson (10)	3.0000	9.8100	19.0000
Negativebinomial (5,0.4)	0.0000	7.6050	27.0000
Beta (0.07, 3, 4)	0.0000	0.3550	0.9244

Table A.2: Summary statistics for each of the 17 variables used for the simulation study.

The 5 active variables from the 27 variables	Summary Statistics		
Distributions	minimum	mean	maximum
Weibul (1,2.9)	0.0037	2.9572	13.5050
Normal (1,3.2)	-9.3461	1.3134	10.7401
Lognormal (0.4,1)	0.0500	2.6074	36.4485
Exponential (0.7)	0.0021	1.3910	8.6044
Logistics (3,2.9)	-13.0807	3.1644	19.7402
Noise from the remaining 17 variables			
Negativebinomial (2,0.4)	0.0835	2.4205	33.3964
Beta (0.03,1,4)	0.0000	0.5947	0.9961
Cauchy (7, 1.5)	-166.4080	5.8520	81.9100
Cauchy (3, 1.5)	-69.2060	2.9700	77.3120
Gamma (5, 0.3)	2.6390	16.7360	49.4790
Geometric (0.1)	0.0000	8.0450	57.0000
Hypergeometric (10, 12,8)	1.0000	3.7150	7.0000
Negativebinomial (5,0.4)	0.0000	7.2850	24.0000
Uniform (-1, 7)	-0.9822	3.0411	6.9240
Poisson (9)	2.0000	9.1400	17.0000
Negativebinomial (3,0.4)	0.0000	4.2200	21.0000
Uniform (-1, 5)	-0.9811	1.9902	4.9923

Table A.3: Comparison of Different Penalties with BIC and Cross Validation Method using 17 variables with Three Different Shape Value

$\alpha = 0.2$, $Shape(\tau) = 0.8$ has 38 events					Shape $(\tau) = 1.1$ has 50 events				Shape $(\tau) = 1.3$ has 63 events			
BIC method using three different shape values												
Penalty	mmcp	mcp	scad	LASSO	mmcp	mcp	scad	LASSO	mmcp	mcp	scad	LASSO
λ	0.0413	0.0385	0.0335		0.0728	0.0728	0.0550	0.0193	0.0757	0.0757	0.0757	0.0163
Nonzero Coef	5	6	7	-	5	5	6	8	5	5	5	8
γ	4	4	4	-	5	3	3	-	5	3	3	-
Avg.mFDR	0	0.161	0.277	-	0	0	0.155	0.360	0	0	0	0.366
cross validation method using three different shape values												
λ	0.1262	0.0890	0.0774	-	0.0479	0.0479	0.0479	0.0090	0.1073	0.1000	0.0811	0.0115
Nonzero Coef	4	4	4	-	7	7	7	13	5	5	5	11
g	4	4	4	-	5	3	3	-	5	3	3	-
Cross validation error	4.74	4.74	4.74		4.76	4.76	4.76	4.81	4.90	4.90	4.91	5.07

Table A.4: Comparison of MMCP and MCP penalties (27 variables) with BIC and cross-validation method using four different gamma values with $\alpha = 0.2$ and, $\tau = 1.1$ and 0.2

$\alpha = 0.2, \tau = 0.8$	$\gamma = 10$		$\gamma = 20$		$\gamma = 30$		$\gamma = 40$	
Penalty	MMCP	MCP	MMCP	MCP	MMCP	MCP	MMCP	MCP
BIC method								
λ	0.0475	0.0475	0.0475	0.0335	0.0475	0.0220	0.0475	0.0206
Nonzero Coef	5	5	5	6	5	8	5	8
Avg.mFDR	0	0	0	0.160	0	0.361	0	0.362
Cross Validation method								
λ	0.0830	0.0335	0.0774	0.0385	0.0628	0.0272	0.0475	0.0291
Nonzero Coef	5	7	5	6	5	8	5	7
Cross validation error	4.89	5.06	4.89	5.10	4.98	5.01	5.04	4.92
$\alpha = 0.2, \tau = 1.1$	$\gamma = 10$		$\gamma = 20$		$\gamma = 30$		$\gamma = 40$	
Penalty	MMCP	MCP	MMCP	MCP	MMCP	MCP	MMCP	MCP
BIC method								
λ	0.0635	0.0316	0.0635	0.0448	0.0635	0.0316	0.0635	0.0256
Nonzero Coef	5	5	5	7	5	9	5	10
Avg.mFDR	0	0	0	0.272	0	0.427	0	0.481
Cross Validation method								
λ	0.0965	0.0363	0.0635	0.0339	0.0448	0.0339	0.0418	0.0275
Nonzero Coef	5	9	5	9	9	9	9	10
Cross validation error	4.78	4.82	4.80	4.92	4.81	4.94	4.80	4.95

Table A.5: Comparison of MMCP and MCP penalties (27 variables) with BIC and cross validation method using four different gamma values with $\alpha = 0.1$ and $\tau = 1.1$

$\alpha = 0.1, \tau = 1.1$	$\gamma = 10$		$\gamma = 20$		$\gamma = 30$		$\gamma = 40$	
Penalty	MMCP	MCP	MMCP	MCP	MMCP	MCP	MMCP	MCP
BIC method								
λ	0.0635	0.0316	0.0635	0.0448	0.0635	0.0316	0.0635	0.0256
Nonzero Coef	5	5	5	7	5	9	5	10
Avg.mFDR	0	0	0	0.272	0	0.427	0	0.481
Cross Validation method								
λ	0.1035	0.0363	0.0965	0.0339	0.0783	0.0339	0.0635	0.0275
Nonzero Coef	5	9	5	9	5	9	5	10
Cross validation error	4.78	4.82	4.78	4.92	4.79	4.94	4.80	4.95

Table A.6: Comparison of MMCP and MCP penalties with BIC and cross validation method using four different gamma values with $\alpha = 0.3$ for NKI dataset

$\alpha = 0.3$	$\gamma = 350$		$\gamma = 450$		$\gamma = 550$		$\gamma = 650$	
Penalty	MMCP	MCP	MMCP	MCP	MMCP	MCP	MMCP	MCP
BIC method								
λ	0.0948	0.0840	0.0948	0.0948	0.0948	-	0.0948	-
Nonzero Coef	11	11	11	11	11	-	11	-
Avg.mFDR	0.098	0.096	0.098	0.096	0.097	-	0.097	-
Cross Validation method								
λ	0.0840	0.0840	0.0840	0.0840	0.0840	0.0840	0.0840	0.0840
Nonzero Coef	14	16	15	16	15	16	15	16
Cross validation error	9.93	9.94	9.94	9.94	9.94	9.94	9.94	9.94
R-square	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.42

Table A.7: Comparison of MMCP and MCP penalties with cross validation method using four different gamma values with $\alpha = 0.3$ for NKI dataset

$\alpha = 0.1, \tau = 0.8$	$\gamma = 350$		$\gamma = 450$		$\gamma = 550$		$\gamma = 650$	
Penalty	MMCP	MCP	MMCP	MCP	MMCP	MCP	MMCP	MCP
BIC method								
λ	0.0840	0.0840	0.0840	0.0840	0.0840	0.0840	0.0840	0.0840
Nonzero Coef	14	16	15	16	15	16	15	16
Cross validation error	9.93	9.94	9.94	9.94	9.94	9.94	9.94	9.94
R-square	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.42