Simulation-based

Bayesian Optimal Accelerated Life Test

Design and Model Discrimination

by

Ehab Nasir

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2014 by the
Graduate Supervisory Committee:

Rong Pan, Chair
George Runger
Esma Gel
Ming-Hung Kao
Douglas Montgomery

ARIZONA STATE UNIVERSITY

August 2014

ABSTRACT

Accelerated life testing (ALT) is the process of subjecting a product to stress conditions (temperatures, voltage, pressure etc.) in excess of its normal operating levels to accelerate failures. Product failure typically results from multiple stresses acting on it simultaneously. Multi-stress factor ALTs are challenging as they increase the number of experiments due to the stress factor-level combinations resulting from the increased number of factors. Chapter 2 provides an approach for designing ALT plans with multiple stresses utilizing Latin hypercube designs that reduces the simulation cost without loss of statistical efficiency. A comparison to full grid and large-sample approximation methods illustrates the approach computational cost gain and flexibility in determining optimal stress settings with less assumptions and more intuitive unit allocations.

Implicit in the design criteria of current ALT designs is the assumption that the form of the acceleration model is correct. This is unrealistic assumption in many real-world problems. Chapter 3 provides an approach for ALT optimum design for model discrimination. We utilize the Hellinger distance measure between predictive distributions. The optimal ALT plan at three stress levels was determined and its performance was compared to good compromise plan, best traditional plan and well-known 4:2:1 compromise test plans. In the case of linear versus quadratic ALT models, the proposed method increased the test plan's ability to distinguish among competing models and provided better guidance as to which model is appropriate for the experiment.

Chapter 4 extends the approach of Chapter 3 to ALT sequential model discrimination. An initial experiment is conducted to provide maximum possible information with respect to model discrimination. The follow-on experiment is planned by leveraging the most current information to allow for Bayesian model comparison through posterior model probability ratios. Results showed that performance of plan is adversely impacted by the amount of censoring in the data, in the case of linear vs. quadratic model form at three levels of constant stress, sequential testing can improve model recovery rate by approximately 8% when data is complete, but no apparent

advantage in adopting sequential testing was found in the case of right-censored data when censoring is in excess of a certain amount.

DEDICATION

To my mother Safia Saad and late father Dr. Awad Nasir.

ACKNOWLEDGMENTS

This Dissertation would not have been possible if not for the encouragement, support and patience of my advisor, Dr. Rong Pan. I'm very grateful to his guidance and continuous help during my study period. I'm also sincerely thankful to Dr. George Runger, Dr. Douglas Montgomery, Dr. Esma Gel, and Dr. Ming-Hung (Jason) Kao for serving on my graduate advisory committee and for providing helpful comments and suggestions especially during my qualifying exam that helped improve this thesis.

I would also like to thank many faculty members in the department of Industrial Engineering Program for their excellent teaching and course offerings throughout the years of my Masters and PhD studies. Among them, I would like to specially acknowledge Dr. George Runger and Dr. Douglas Montgomery as what they have taught me in classes proved extremely valuable and constituted the corner stone in my professional career as Quality and Reliability Engineer in the Semi-conductor industry.

Finally, I would like to thank my wife, Nancy Yousif, for her unconditional support, patience and encouragement at all times throughout my graduate study while I'm working a full time job in a highly demanding industry. My love goes to her and to our kids, Danya, Amro and Kareem, whom they have been as patient and supportive as their mother.

TABLE OF CONTENTS

Page

CHAPTER Page

CHAPTER Page

CHAPTER Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

CHAPTER 1

GENERAL INTRODUCTION

1.1. Introduction

In this dissertation, we are interested in Accelerated life test planning methods under practical constraints, with particular focus on Bayesian methods to incorporate existing engineering knowledge in the planning phase of the experiment.

Accelerated life tests (ALTs) are widely used throughout industry, primarily to estimate lifetime performance of products at field or use conditions. Testing at use conditions is impractical due to the length of testing time required to produce fails. Therefore, reliability engineers would instead perform testing at increased levels of applied stresses (for example, temperature, temperature amplitude (ΔT), humidity, voltage/bias, or pressure) to produce fails within reasonable time durations. Obtained data at test levels would then be used to make predictions of product performance at use conditions through extrapolation using an appropriate ALT model. This will enable timely decision making satisfying a business need.

Accelerated life models have two parts: a life distribution which is a statistical model for the time to failure data at each stress level, and a physical model or relationship which is a mathematical model that links the parameters of life distribution to stress levels. The intent of ALT testing is to accelerate a given physical mechanism without introducing new fail mechanism that do not exist in the use environment. Therefore, overstressing should be avoided.

Testing resources such as time, test units and test equipment (e.g. stress chambers) are usually very limited. Therefore, careful test planning is critical for the efficient use of such resources while extracting maximum possible information. Typically, ALT plans specify the levels of accelerating variable/s and the allocation of available test units to these levels. Optimum test plans are obtained given different criteria of interest; for example, the estimation precision of a life distribution quantile at use conditions. Optimum test plans serve as the base for obtaining good practical compromise test plans. ALT planning methods can be grouped into constant-stress, step-stress, and ramp-stress ALT methods. For the constant-stress methods see, e.g., Meeker and Han [11], Nelson [13], Nelson and Kielpinski [14], and Nelson and Meeker [15]. Majority of

available research has focused on optimal plans for the single stress-factor as in Nelson and Meeker [15]. However, in reality there are increased numbers of failure mechanisms that are driven by a combined effect of multiple stress factors affecting the product simultaneously. Hence our interest focuses on ALT plans for the constant-stress multi-stress factor case, as shown in Chapter 2 of this dissertation, and we tackle the design problem from a Bayesian perspective, as we are motivated by the fact that engineers acquire wealth of knowledge during the course of their experimentation that should be utilized in planning of future ALTs.

Majority of today's Bayesian experimental design work follows a decision theoretic approach to experimental design as outlined in Lindley [9], and Raiffa and Schlaiffer [17]. This approach suggests that a good way to design experiments is to specify a utility function reflecting the purpose of the experiment, to regard the design choice as a decision problem, and to select a design that maximizes the expected utility. However, the difficulty with this approach is that the exact utility function is often a complicated integral of high dimension, and as such approximation or simulation methods must be used for its evaluation, (Chaloner and Verdinelli [4]). With advances in computing power nowadays, the Simulation-based strategies have gained solid grounds and can be used to approximate the expected utility function necessary for the evaluation of the Bayesian optimal design. To avoid intensive sampling and computation as a result of increased number of stress factor-level combinations, we explore the use of a Latin hypercube designs for sampling the design space, alongside non-parametric surface smoothing techniques to approximate the pre-posterior variance of a quantity of interest at use condition to arrive at our optimal design.

Implicit in the design criteria used in current ALTs is the assumption that the form of the acceleration model is correct. In many real-world problems this assumption could be unrealistic. A more realistic goal of an initial stage of ALT experimentation is to find an optimal design that helps in selecting a model among rival or competing model forms; i.e., a design that could assist in model discrimination. The ability to choose between competing model forms in an early experimentation stage has an important impact on the effective design of subsequent

experimentation phases. A considerable work has been done in the development of experimental designs for discrimination among regression models. See, for example, Hunter and Reiner [7], Box and Hill [3], Hill et al. [6], Atkinson and Cox [1]. More recently, many authors focused on the development of T-optimum criterion for model discrimination; see, for example, Atkinson and Donev [2], Ponce de Leon and Atkinson [8]. However, all of the above attempts at model discrimination work has been in the context of traditional experimental design, i.e., standard experimental designs such as factorial, fractional factorial, Box central composite, etc. None to our knowledge has been explicitly targeting model discrimination in planning of accelerated life tests experiments, where failure time censoring is commonly expected. Nelson [13] (p. 350) has cautioned that the statistical theory for traditional experimental design is correct only for complete data (all units fail in a test), and one should not assume that properties of standard experimental designs hold for censored and interval-censored data as they usually do not hold. Therefore, our work in Chapter 3 draws its importance from its attempt at contributing to model discrimination literature in accelerated life test planning when censoring is inevitable due to practical test constraints.

Motivated by the sequential nature of the learning process in general, we investigate the effect of sequential experimentation on ALT model discrimination in Chapter 4 as compared to the non-sequential method proposed in Chapter 3. The Bayesian theory allows us to implement a sequential scheme and make use of the most recent information obtained from data. In addition to that, any available prior information about the models under investigation can be considered in the analysis. The MCMC-based methods have the advantage of its capacity to handle both linear and nonlinear models. Hence, our proposed methodology is utilizing Bayesian MCMC methods.

Sequential testing and design of experiments have been studied previously, one may refer to Chernoff [5], Pilz [16], Wetherill and Glazebrook [19] and Michlin et at. [12]. Sequential testing in the context of ALT planning and inference has been studied by Liu and Tang [10], and Tang and Liu [18] for the single-variable constant-stress accelerated test. However, we know of no previous work that aimed at sequential testing in the context of ALT model discrimination. That

is an additional motivation to the work presented in Chapter 4.

1.2. Dissertation Organization

This dissertation consists of three papers corresponding to Chapters 2, 3, and 4 respectively. Chapter 2 provides a simulation-based method for the design of multi-stress factor accelerated life tests in a Bayesian decision theoretic framework. Multi-stress factor ALTs are challenging due to the increased number of experiments required as a result of stress factor-level combinations resulting from the increased number of factors to be studied. The methodology introduces the use of Latin hypercube sampling scheme to meet that challenge and reduce the simulation cost without loss of statistical efficiency. Optimization of expected utility function; the posterior variance of a life distribution quantile of interest at use condition, is carried out by a developed algorithm that utilizes Markov chain Monte Carlo methods (Gibbs sampler) and nonparametric smoothing techniques applied to the Latin hypercube design space in which each stress factor has as many levels as there are runs in the design with levels chosen to maximize the minimum distance between design points. The approach illustrated with an application to an ALT planning problem with practical constraints. A comparison of proposed approach to the full blown grid simulation is provided to illustrate computational cost gain, and a comparison to the large-sample approximation method reveals the flexibility of our approach in determining optimal stress settings with less assumptions being made and more intuitive unit allocations.

Chapter 3 discusses Bayesian accelerated life test planning with a focus on differentiating among competing acceleration models, when there is uncertainty as to whether the relationship between log mean (life) and the stress (possibly transformed) is linear or exhibits some curvature. The proposed criterion is based on the Hellinger distance measure between predictive distributions. The optimal stress-factor setup and unit allocation are determined at three stress levels subject to test-lab equipment and test-duration constraints. Optimal designs are validated by their recovery rates, where the true, data-generating, model is selected under the DIC (Deviance Information Criterion) model selection rule. We also compare performance of

obtained plans with other test plans including the typically used three stress-levels good compromise plan, the best traditional plan and the well-known 4:2:1 compromise ALT test plans.

Chapter 4 extends the approach of Chapter 3 to sequential model discrimination in accelerated life test planning. Comparison of model recovery rates under the two approaches are made and the possibility of identifying a “winning” model form at a much earlier stage than would be possible with non-sequential testing, at consequently lower experimental cost is investigated.

Chapter 5 provides overall discussion and conclusions based upon the results obtained in Chapters 2, 3, and 4 respectively.

References

[1] A. C. Atkinson, and D. R. Cox, Planning experiments for discriminating between models, J. Roy. Statist. Soc. Ser. B **36** (1974), 321-348.

[2] A. C. Atkinson, and A. N. Donev, Optimum experimental designs, Oxford, England: Oxford University Press.

[3] G. E. P. Box, W. J. Hill, Discrimination among mechanistic models, Technometrics **9** (1967), 57-71.

[4] K. Chaloner, and I. Verdinelli. Bayesian Experimental Design: A Review. Statistical Science, **10** (1995), 273-304.

[5] H. Chernoff, Sequential Analysis and Optimum Design, Society for Industrial and Applied Mathematics (SIAM), Philadelphia (1972).

[6] W. J. Hill, W. G. Hunter, D. W. Wichern, A joint design criterion for the dual problem of model discrimination and parameter estimation, Technometrics **10** (1968), 145-160.

[7] W. G. Hunter, A. M. Reiner, Designs for discriminating between two rival models, Technometrics **7** (1965), 307-323.

[8] P. D. Leon and A. C. Atkinson, Optimum experimental design for discriminating between two rival models in the presence of prior information, Biometrika **78** (1991), 601-608.

[9] D. V. Lindley, Bayesian Statistics – A review. SIAM, Philadelphia (1972).

[10] X. Liu and L. C. Tang, A Sequential Constant-Stress Accelerated Life Test Scheme and its Bayesian Inference, Quality and Reliability Engineering International **25** (2009), 91-109.

[11] W. Q. Meeker, and G. J. Hahn, Asymptotically optimum over-stress tests to estimate the survival probability at a condition with a low expected failure probability, Technometrics **19** (1977), 381–399.

[12] Y. H. Michlin, L. Meshkov and I. Grunin, Improvement on Sequential Testing in MIL-HDBK-781A and IEC 61124, IEEE Transactions on Reliability **57** (2008), 379-387.

[13] W. Nelson, Accelerated Testing, John Wiley & Sons: New York, (1990).

[14] W. Nelson, and T. Kielpinski, Theory for Optimum Censored Accelerated Life Tests for Normal and Lognormal Distributions, Technometrics **18** (1976), 105–114.

[15] W. Nelson, and W. Q. Meeker, Theory for optimum accelerated Censored Life Tests for Weibull and Extreme Value Distributions, Technometrics **20** (1978), 171-177.

[16] J. Pilz, Bayesian Estimation and Experimental Design in Linear Regression Models, John Wiley & Sons, Germany (1991).

[17] H. Raiffa, R. Schlaiffer, Applied Statistical Decision Theory. Division of Research, Harvard Business School, Boston (1961).

[18] L. C. Tang and X. Liu, Planning and Inference for a Sequential Accelerated Life Test, Journal of Quality Technology **42**, No.1 (2010).

[19] G. B. Wetherill and K. D. Glazebrook, Sequential Methods in Statistics, 3rd edition. London, UK: Chapman and Hall (1986).

CHAPTER 2

SIMULATION-BASED BAYESIAN OPTIMAL DESIGN FOR MULTI-FACTOR ACCELERATED LIFE TESTS

Ehab Nasir[1], Rong Pan[2]

[1] Industrial Engineer, Intel Corporation, Chandler, Arizona.

[2] Associate Professor, School of Computing, Informatics, and Decision Systems Engineering, Arizona State University

Abstract

We consider simulation-based methods for the design of multi-stress factor accelerated life tests ALTs in a Bayesian decision theoretic framework. The Bayesian methodology is an attractive alternative to the maximum likelihood MLE approach when considerable uncertainty exists in the planning values of the model parameters. Multi-stress factor ALTs are challenging due to the increased number of experiments required as a result of stress factor-level combinations resulting from the increased number of factors to be studied. This negatively impacts the overall cost of the experiment and its practical feasibility. We propose the use of Latin hypercube sampling scheme to meet that challenge and reduce the simulation cost without loss of statistical efficiency. Exploration and optimization of expected utility function, the posterior variance of a life distribution quantile of interest at use condition (design stress) in our case, is carried out by a developed algorithm that utilizes Markov chain Monte Carlo methods (Gibbs sampler) and nonparametric smoothing techniques applied to the Latin hypercube design space in which each stress factor has as many levels as there are runs in the design with levels chosen to maximize the minimum distance between design points. We illustrate our approach with an application to an ALT planning problem with practical constraints when the underlying life model has a Weibull distribution with type-I censored data. A comparison of proposed approach to a full blown grid simulation is provided to illustrate computational cost gain. We also provide a comparison to the large-sample approximation method that reveals the flexibility of our approach

in determining optimal stress settings with the advantage of less assumptions being made and more intuitive unit allocations.

Key Words - Bayesian inference, Monte Carlo simulation, Gibbs sampler, Latin hypercube sampling, nonparametric smoothing.

2.1. Introduction

Accelerated life tests (ALTs) are widely used in reliability studies. Data from ALTs usually involve type-I censoring (fixed time tests; more often used in practice) and/or type-II censoring (fixed number of failures tests; less practical). Testing resources such as time, test units and test equipment (e.g. stress chambers) are usually very limited. Therefore, careful test planning is critical for efficient use of such resources while extracting maximum possible information. Typically, ALT plans specify the levels of the accelerating variable/s and the allocation of available test units to these levels. One can find an optimum test plan for a given criterion, such as the estimation precision of a life distribution quantile at use conditions. Optimum test plans serve as the base for obtaining good practical test plans (compromise plans).

Accelerated life tests (ALTs) methods and results have been studied by other researchers, and can be generally categorized into methods for test planning and methods for analysis of/ inference from test data. Test planning methods can be further subdivided into constant-stress, step-stress, and ramp-stress ALT methods. For the constant-stress methods see, e.g., Meeker and Han [22], Nelson [30], Nelson and Kielpinski [31], and Nelson and Meeker [32]. Majority of available research has focused on optimal plans for the single stress-factor as in Nelson and Meeker [32]. However, in reality more and more failure mechanisms are driven by a combined effect of multiple stress factors affecting the product simultaneously (e.g. temperature, humidity and voltage or current stresses). In this study we focus on Bayesian ALT plans for the constant-stress multi-stress factor case.

Most of today's Bayesian experimental design work follows a decision theoretic approach

to experimental design as outlined in Lindley [19], and Raiffa and Schlaiffer [38]. This approach suggests that a good way to design experiments is to specify a utility function reflecting the purpose of the experiment, to regard the design choice as a decision problem, and to select a design that maximizes the expected utility. Selecting a utility function that appropriately describes the goal of the experiment is very important. A design that is optimal for estimation may not necessarily be optimal for prediction. However, the exact utility function is often a complicated integral of high dimension, and as such approximation or simulation methods must be used for its evaluation, Chaloner and Verdinelli [8]. Most approximations to the expected utility function involve using a normal approximation to the posterior distribution. Several normal approximations are possible as outlined in Berger [5], and involve either the expected Fisher information matrix or the matrix of second derivatives of the logarithm of either the likelihood or the posterior density, Bai and Kim [3]. Analytic and approximation-based strategies can be found in Behnken and Watts [4], Chaloner and Larntz [6, 7], Polson [37], Verdinelli et al. [47], Clyde et al. [11], and Zhang and Meeker [51]. With the help of modern advances in computing power, the Simulation-based strategies have gained solid grounds and can be used to approximate the expected utility function necessary for the evaluation of the Bayesian optimal design. Simulation based methods, such as Monte Carlo simulation are available, provided that the prior distribution of the model parameters and sampling distribution of the data are available for efficient random variable generation, and the utility function can be evaluated for any given realization of the experiment. Simulation-based methods have been used in Erkanli and Soyer [12], Sun et al. [42], Hamada et al. [16], and Xiao and Loon-Ching [20].

In this paper we develop a simulation-based Bayesian method for planning accelerated life tests for the constant-stress case in the presence of multiple factors affecting the response of interest. We outline the methodology for a case where the fail mechanism is activated by three stress factors, temperature, relative humidity and voltage bias, under an assumed Weibull life distribution and Type-I censoring. To avoid intensive sampling and computation as a result of increased stress factor-level combinations, a Latin hypercube designs for sampling the design

space are utilized, alongside non-parametric surface smoothing techniques to approximate the pre-posterior variance of a quantity of interest at use condition (design stress).

In the remainder of this paper, we present the motivation for Latin hypercube designs LHDs in Section 2.2, the Bayesian ALT design problem in Section 2.3, model description in Section 2.3.1, and the optimization criterion in Section 2.3.2. Section 2.4 outlines the optimization algorithm, and in Section 2.5 we provide a case study that demonstrates the approach and discuss results. Section 6 provides a comparison of our proposed method to the full grid method, while in Section 2.7 we show how it compares to the large-sample approximation method with a reference solution point using the full grid method as well. Section 2.8 concludes the study and provides future research directions.

2.2. Motivation for Latin Hypercube Designs (LHD)

Majority of available ALT planning studies have focused on optimal plans for the single stress-factor. However, in reality fail mechanisms that are driven by a combination of different stress factors are common and should be further investigated in methods for ALT planning. Products in the field typically operate under several simultaneous stresses that contribute to their overall fail rate. Multi-stress factor ALTs are challenging due to the increased number of experiments required as a result of stress factor-level combinations resulting from the increased number of factors to be studied. This negatively impacts the overall cost of the experiment and its practical feasibility. We propose the use of Latin hypercube designs to meet that challenge and reduce the simulation cost of the proposed methodology. A factorial design, full or fractional could have been used instead of a LHD, but it will fail to address the challenge in the multi-stress factor ALT planning. A full factorial design would require l^k experiments for a test with k factors and l levels for each factor, compared to l experiments only as required by an LHD (l,k) design. Fractional factorials, on the other hand, can help in reducing the number of experiments, but add the challenge of how to decide on which fraction to use and allocation of test units. Additional advantages of LHD designs are the fact that they are computationally cheap to generate and can

cope with many input variables, Sacks and Welch [39], and they allow the user to tailor the number of samples to the available computational budge. In the following sections we give formal definition of a Latin hypercube design LHD, its draw backs as related to design space coverage and their identified remedies.

2.2.1. Latin Hypercube Designs (LHD) – Definition

Latin hypercube sampling, McKay, et al. [21], is a method of sampling that can be used to generate input values for estimation of expected value of functions of output variables. It was originally developed as an alternative to pseudo Monte Carlo sampling. Latin hypercube design LHD follows the idea of a Latin square design where there is only one sample in each row and each column. LHD generalizes this concept to an arbitrary number of dimensions. In LHD of a multivariate distribution, a sample size l from multiple factors is drawn such that for each factor the sample is marginally maximally stratified. A sample is maximally stratified when the number of strata equals the sample size l and when the probability of failing in each of the strata is l^{-1}. Fore an example, given k factors $X_1, \dots, X_k$ the range of each factor X is divided into l equally probable intervals (strata), then for each factor a random sample is taken at each interval (stratum). The l values obtained for each of the factors are then paired with each other in a random way or based on some rules. Now we have l samples, where the samples cover the l intervals for all factors. Thus, the sampling scheme does not require more samples for more factors (dimensions). This method insures that each of the factors in X is represented in a fully stratified manner.

2.2.2. Latin Hypercube Designs (LHD) – Drawback, and Remedy

The application of the random generated LHD does not come without shortcomings as in some cases it show undesired properties and may act poorly in prediction and/or estimation. Some extreme arrangement, for example, when all samples happen to fall a long a diagonal, result in an LHD sample that performs poorly with respect to design space coverage (poor predictions in unexplored areas), and high spatial correlation (co-linearity). Several LHD design criteria have been proposed in literature to overcome these drawbacks and optimize the space-

filling properties of an LHD. Interested readers may refer to the following sources: Tang [44], Owen [33], and Tang [45] for discussion on Orthogonal Array-based LHD (OALHD) which extends to low-dimensional projections the uni-variate stratification properties of the random LHD, and provides better balance for larger experiments, more uniform designs, and no correlation among estimation of linear effects. Park [36] for Optimal LHD design, where both optimal design and random LHD designs are combined to provide smaller prediction error as compared to the random LHD. Morris et al. [27] for Maximin LHD design that has good symmetric properties obtained by maximizing the minimum inter-site distance, and provides a compromise between entropy/maximin criterion and projective properties of LHD. Keramat et al. [18] for the Modified LHD Monte Carlo (MLHDMC) characterized with faster sample generation, particularly in an optimization procedure, smaller estimation variance with same computational time, and more accurate results for Average Quality Index (AQI) or parametric yield estimation than standard LHD. Ye [49] for the Orthogonal LHD (OLHD) where there is no correlation among estimation of linear effects. Ye et al. [50] for the Optimal symmetric LHD (SLHD), that provides compromise between computing effort and design optimality with better/ maximum entropy and minimum inter-site distance criteria as compared to random LHD design. It also provides orthogonal properties, that is the estimation of the linear effect of each variable is uncorrelated with all quadratic effects, and bi-linear interactions, generalization of OLHD, flexibility in run size, yet retains some of the orthogonality of an OLHD. Fang et al. [14] for the Uniform LHD, which is a space filling design that minimizes the discrepancy between design points. Palmer et al. [34] for the Minimum Bias LHD, that provides compromise between empirical model bias reduction and dispersion of points within design space. Steinberg et al. [41] for the Orthogonal LHD (OLHD), where all main effects are orthogonal, no correlation among estimation of linear effects, and can be used to construct LHD designs with low correlation of first-order and second-order terms, the method generates orthogonality when many more factors are included as compared to Ye's OLHD. Cioppa et al. [10] for the Efficient Nearly Orthogonal LHD (EN-OLHD) which is characterized by near orthogonal properties in higher dimension space, provides flexibility in fitting models when exploring high-dimensional computer simulations where there is considerable a priori uncertainty

about the forms of the response surfaces. Joseph et al. [17] for the Orthogonal-Maximin LHD, where its fast algorithm provides optimal designs that are optimized for both of pairwise correlation and distance criteria. In our proposed algorithm, we chose to utilize the Uniform LHD of Fang et al. [14] as our base design for its space filling properties that minimizes the discrepancy between design points (i.e. requiring even spacing of design points), however modified by maximizing the minimum distance between pairs of design points (sphere-filling). As the number of variables increase, it becomes harder to fill the design space, and as optimization is pushing design points further apart, the sample tends to create a vacuum in the center of the design space. Thus, further enhancement to prediction coverage could be obtained by forcing inclusion of center points and/or corner points of the design space.

2.3. The Bayesian ALT Design Problem

Bayesian experimental design approach suggests that a good way to design experiments is to specify a utility function that reflects the purpose of the experiment and to select the design that maximizes the expected utility of the experiment. Assuming that the goal of an experiment can be formally expressed through a utility function of the general form $u(\eta, y, \theta)$, the Bayesian solution is to find the best design that maximize the expected utility $U(\eta)$ with respect to the joint density of (y, θ). Formally stated as

$$\eta^* = \arg\max_{\eta \in H} U(\eta)$$

where

$$U(\eta) = \int u(\eta, y, \theta).p_\eta(\theta, y)\, d\theta\, dy \tag{1}$$

$p_\eta(\theta, y)$ is a probability distribution of parameter θ and response y that is possibly influenced by the selected design η. It can also be expressed as $p_\eta(\theta, y) = p(\theta).p_\eta(y|\theta)$, where $p(\theta)$ is the prior distribution of parameter θ.

2.3.1. Acceleration Model Description

In this study we consider practical ALT testing constraints in a semi-conductor industry. We assume pre-specified values for the total number of units available for stress testing (N) as dictated by available budget for testing, maximum testing time (t_c) as driven by stress chambers availability and/or time by which a data-driven decision is to be made. The feasible ranges for stress factors at both high and low levels are functions of the capability of the stress lab equipment. The optimal test plan specifies the high and low stress settings for stress factors given their allowable ranges, and unit allocation at each stress level while optimizing a criterion on interest at the design stress (use level). We outline the methodology with a case in which two stress factors are involved (i.e. temperature and relative humidity). Extending the methodology to more than two stress factors should be straightforward.

Based on past experience with similar fail mechanism, the reliability engineer believes that the Weibull distribution would adequately describe C4 bump life (First level interconnect Cu bump in a semiconductor package), which implies a smallest extreme value (SEV) distribution for the log-life. That is, if T is assumed to have a Weibull distribution, $\mathrm{T}\sim\mathrm{WEIB}(\alpha,\beta)$, then $\log(t)\sim SEV(\mu,\sigma)$, where $\sigma=\frac{1}{\beta}$ is the scale parameter and $\mu=\log(\alpha)$ is the location parameter. The Weibull CDF and PDF can be written as

$$F(t|\alpha,\beta)=\Phi_{SEV}\left(\frac{\log(t)-\mu}{\sigma}\right)=1-exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right] \quad (2)$$

$$f(t|\alpha,\beta)=\frac{1}{\sigma t}\phi_{SEV}\left(\frac{\log(t)-\mu}{\sigma}\right)=\frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1}exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right],\ t>0 \quad (3)$$

In above parameterization, $\beta>0$ is the shape parameter and $\alpha>0$ is the scale parameter as well as the 0.632 quantile.

Also, the Peck's model (a special case of the more general Eyring model) was expected to describe the temperature $(Temp)$ and relative humidity (RH) acceleration. So the acceleration model is represented by

$$t(RH, Temp) = A(RH)^{-m} . \exp\left(\frac{E_a}{K \times Temp}\right), \quad (4)$$

where,

- $t(RH, Temp)$ is the life characteristic related to temperature and relative humidity.
- A is a constant, $(-m)$ is the inverse power law parameter of RH, and (E_a) is the activation energy of the chemical reaction in electron volts.
- $Temp$ is temperature in Kelvin $(c^o + 273.15)$.
- K is Boltzmann's constant $(8.617385\ E^{-5}\ eV/K)$

This model can be expressed in linearized form by taking the logarithmic of both sides as

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \quad (5)$$

where μ denotes $\log_2 t(RH, Temp)$, x_1 and x_2 denote $\log_2(RH)$ and $\left(\frac{1}{K \times Temp}\right)$ respectively. It is easy to see that β_0 is $\log_2(A)$, where A is constant. $\beta_1$is $-m$, the inverse power law parameter of RH, and β_2, is E_a, the activation energy of the chemical reaction in electron volts.

We standardize the accelerating variables for simplicity and to maintain comparable relative scales for the variables. Therefore, above model (4) can be expressed as

$$\mu = \gamma_0 + \gamma_1 \xi_1 + \gamma_2 \xi_2 , \quad (6)$$

Where the standardized variables are expressed as

$$\xi_1 = \frac{(x_1 - x_{1.low})}{(x_{1.high} - x_{1.low})} , \qquad \xi_1 \in [0, 1] \quad (7)$$

$$\xi_2 = \frac{(x_2 - x_{2.low})}{(x_{2.high} - x_{2.low})} , \qquad \xi_2 \in [0, 1] \quad (8)$$

New coefficients are related to previous ones through

$$\begin{cases} \gamma_0 = \beta_0 + \beta_1 x_{1.low} + \beta_2 x_{2.low} \\ \gamma_1 = \beta_1 (x_{1.high} - x_{1.low}) \\ \gamma_2 = \beta_2 (x_{2.high} - x_{2.low}) \end{cases} \quad (9)$$

Thus, at $x_1 = x_{1.low}, \xi_1 = 0$, and at $x_1 = x_{1.high}, \xi_1 = 1$. The same applies to x_2.

Thus, $$\begin{cases} \mu_{low} = \gamma_0 \\ \mu_{high} = \gamma_0 + \gamma_1 + \gamma_2 \end{cases} \quad (10)$$

For Type-I censored data; the probability of obtaining a censored observation at time t_c is given by

$$\Pr(\mathrm{t} > \mathrm{t}_c) = exp\left[-\left(\frac{t_c}{\alpha}\right)^{\beta}\right], \qquad t_c > 0 \quad (11)$$

2.3.2. Optimization Criterion

We consider a reliability goal of estimating an early 100 p^{th} quantile of life distribution, $\tau_p(x_U)$, at the use condition or design stress , x_U with as much precision as possible. Because $\tau_p(x_U)$, is positive, it is reasonable to use an ALT criterion based on the estimation precision of $\log[\tau_p(x_U)]$. Similar criterion has been suggested in Zhang and Meeker [51].

$$\begin{aligned} \log[\tau_p(x_U)] &= \mu_U + \frac{1}{\beta}.\log[-\log(1-p)] \\ &= \gamma_0 + \gamma_1\xi_{1U} + \gamma_2\xi_{2U} + \sigma.\log[-\log(1-p)] \\ &= a^T\theta \end{aligned} \quad (12)$$

where $a = \left[1, \xi_{1U}, \xi_{2U}, log[-\log(1-p)]\right]^T$, and $\theta = (\gamma_0, \gamma_1, \gamma_2, \sigma)$ are the ALT model parameters with two accelerating variables.

A utility function can then be defined as the posterior variance of quantity in (12) and an optimum plan is obtained by minimizing it. The posterior variance for a given design η depends on the unobserved data y. Therefore, a pre-posterior expectation of the posterior variance over the marginal distribution of data y is used to average over the unobserved data y, and the following Bayesian planning criterion is obtained.

$$\begin{aligned} \psi(\eta) &= E_{y|\eta}\left[var_{\theta|y,\eta}\left(\log\left[\tau_p(x_U)\right]\right)\right] \\ &= E_{y|\eta}\left[var_{\theta|y,\eta}(a^T\theta)\right] \\ &= E_{y|\eta}[a^T\, var_{\theta|y,\eta}(\theta)a] \end{aligned} \quad (13)$$

The posterior variance of the model parameters, $Var_{\theta|y,\eta}(\theta)$, is obtained from the joint posterior distribution of the parameters which can be obtained from Bayes' theorem

$$p_{\theta|y,\eta}(\theta) = \frac{p_{y|\theta,\eta}(y).p(\theta)}{\int p_{y|\theta,\eta}(y).p(\theta)}, \quad (14)$$

where $p_{y|\theta,\eta}(y)$ is the likelihood of the data y under plan η, $p(\theta)$ is the joint prior distribution of the model parameters θ. Criterion (13) that calculates the posterior variance and its marginal expectation over all possible data y has no closed form solution and exact numerical solution is intractable. Thus approximation or simulation techniques need to be used.

2.4. Optimization Algorithm

Our optimization algorithm is Monte Carlo simulation-based in which the optimal design η^* is arrived at by evaluating the design criterion in (13) for each of the candidate designs, and selecting the design with highest utility. We summarize the algorithm steps as outlined in subsections below.

2.4.1. Identify Potential Set of ALT Designs

Typically an engineer will have an idea of the ALT test he/she is interested in running. For example, due to budget and/or test time constraints, the engineer would prefer to run a two stress-level test or could add a third level for robustness and run the test for a specified length of time. However, the setup of these stress levels and unit allocation to each level would need to be optimized according to a design criterion of interest for the experiment to yield the most benefit. Hence, in this step the engineer will list down potential designs to investigate η_i $(i = 1, \dots, d)$ and practical constraints imposed on test, for example test budget N and test/censoring time t_C.

2.4.2. Construction of ALT Design Grid

Equally spaced design grid of all combinations of the allowed ranges of stresses of interest, for example temperature and relative humidity can be used to simulate experiments. However, doing so is computationally inefficient. We, instead propose the use of Latin hypercube experimental designs as discussed in Section 2.2. These LHD designs are constructed at each level of the test. For example, in a two stress-level test, there would be a LHD at level 1 and another LHD at level 2. Since ALT data are used simultaneously from all stress levels to predict a quantity of interest at design stress (use condition), a combined LHD design grid for all test levels in generated from individual LHDs. The combined LHD is further augmented with corner and/or center points into an mLHD "modified Lantin hypercube". ALT data are simulated and utility function is calculated over the finalized mLHD design grid. See case study for demonstration of steps.

2.4.3. Evaluation Steps Over the ALT Design Grid

1. Over the design grid; mLHD, randomly simulate fail data from the joint density of parameters and data (θ, y)

$$(\theta, y) \sim p(\theta, Y) \sim p(\theta).p(y|\theta) \sim prior(\theta).likelihood(y|\theta) \quad (15)$$

That is, independently generate random fail data using the acceleration model as in Equation (6).

1.1. Simulated failure times are compared against a predetermined test time t_c to determine the censoring time for each test unit. Those units failing before or at t_c are considered exact failures, while others exceeding t_c are considered right censored. If testing time availability of each stress chamber is different (additional stress lab constraint), then censoring scheme can easily be modified to account for that by comparing fail times at different stress levels against the individual

level test time t_{c_i} where $i = 1, \dots, l$ and l is the number of stress levels in the experiment, for example $l = 2$ in this case study.

1.2. The number of simulated fail times at each cell of the mLHD design grid is determined by unit allocation for the design being considered. For example, in demonstrate case of the two-stress level testing, we loop over all possible combination of unit allocation to the two-stress levels specified. For example, if we consider corner cell (S_{11}, S_{21}) of Figure 2.4, for a total $N = 12$ units, with a design that allocates $n_1 = 8$ units to lower stress level S_1 and $n_2 = 4$ units to higher stress level S_2, there will be a total of 12 observations generated in corner cell (S_{11}, S_{21}), i.e. 8 under stress level S_{11}, and 4 under stress level S_{21}. This will carry on for the other cells identified by the $mLHD$ to complete fail data simulation for the identified design allocation $(n_1 = 8, n_2 = 4)$. The same will be repeated for other design allocations.

2. For each simulated experiment (fail times at a particular grid cell), use Gibbs sampler to evaluate the observed posterior utility

$$u_i = var_{\theta|y,\eta}\left(\log\left[\tau_p(x_{UC})\right]\right) \quad (16)$$

Where $\tau_p(x_{UC})$ is the 100 p^{th} quantile of the lifetime distribution at use condition (design stress) and $(i = 1, \dots, K)$, where K is the number of active cells with observations as identified by the grid of the $mLHD$.

2.1. While performing this step, the likelihood function must be adjusted to account for the censoring structure in the data. When using WinBUGS MCMC simulation through Gibbs sampler, this can be accomplished using the function $I(t_c, t)$ for the right censored observations (data).

3. Since this is a planning phase of the test, the decision of design selection is based on the values of the pre-posterior expected utility. Thus, we do the following:

3.1. Approximate pre-posterior utility function by fitting a smooth surface to the Monte Carlo sample generated in step (2) as a function of evaluated designs.

3.2. We used the local regression model LOESS in R “modreg” library with its default settings for the smooth surface fitting, details are given in Section 2.4.4. This step in effect is equivalent to evaluating the integration in the solution to the optimal design problem in Equation (1).

4. The optimal design is found by maximizing the fitted surface (minimum pre-posterior variance).

The expected utility $U(\eta)$ surface is generally of continuous nature. However, the direct application of the Monte Carlo simulation, in step (3.1) of the algorithm will require large scale simulations to be applied, and will only be computationally inefficient due to the large number of iterations needed and duplication of effort in neglecting valuable information already generated at a nearby design point. That is, repeated simulations at close by points on the design grid. Therefore, to reduce computational cost, step (3.1) of the algorithm utilizes the non-parametric surface fitting approach proposed by Muller and Parmigiani [28, 29] for finding optimal designs. The use of surface smoothing for finding optimal Bayesian designs has been previously considered by Erkanli and Soyer [12] in planning a constant-stress ALT, and Liu and Tang [20] in planning accelerated degradation tests.

2.4.4. LOESS in R, Smooth Surface Fitting

In section below, we give a brief description, usage and arguments used in the R-function LOESS that we used in creating the smooth surface in step (3.1) of the algorithm which is equivalent to evaluating the integration in the solution to the optimal design problem in Equation (1). We also show an example of the R-code we used in our program as pertains to step (3.1) of the algorithm.

1) Description: fits a polynomial surface determined by one or more numerical predictors using local fitting.

2) Usage:

```
loess(formula, data, weights, subset, na.action, model = FALSE,
    span = 0.75, enp.target, degree = 2,
    parametric = FALSE, drop.square = FALSE, normalize = TRUE,
    family = c("gaussian", "symmetric"),
    method = c("loess", "model.frame"),
    control = loess.control(...), ...)
```

3) Arguments description:

formula	A formula specifying the numeric response and one to four numeric predictors (best specified via an interaction, but can also be specified additively). Will be coerced to a formula if necessary.
data	An optional data frame, list or environment containing the variables in the model.
weights	Optional weights for each case.
subset	An optional specification of a subset of the data to be used.
na.action	The action to be taken with missing values in the response or predictors. The default is given by getOption ("na.action").
model	Should the model frame be returned?
span	The parameter α which controls the degree of smoothing.
enp.target	An alternative way to specify span, as the approximate equivalent number of parameters to be used.
degree	The degree of the polynomials to be used, normally 1 or 2.
parametric	Should any terms be fitted globally rather than locally? Terms can be specified by name, number or as a logical vector of the same length as the number of predictors.
drop.square	For fits with more than one predictor and degree=2, should the quadratic term be dropped for particular predictors? Terms are specified in the same way as for parametric.
normalize	Should the predictors be normalized to a common scale if there is more than one? The normalization used is to set the 10% trimmed standard deviation to one. Set to false for spatial coordinate predictors and others know to be a common scale.
family	If "Gaussian" fitting is by least-squares, and if "symmetric" a re-descending M estimator is used with Tukey's biweight function.
method	Fit the model or just extract the model frame.
control	Control parameters
...	Control parameters can also be supplied directly.

4) Example: our R-code for the portion using LOESS regression

```
stress1.1<- c(data.use$V7) # low rh
stress1.2<- c(data.use$V8) # low temp
stress2.1<- c(data.use$V9) # high rh
stress2.2<- c(data.use$V10) # high temp
tau.p.01.sd <- c(data.use$V11) # posterior std of 1st   percentile estimate at Use
Condition
mu.1<- c(data.use$V5) # low rh, low temp
mu.2<- c(data.use$V6) # high rh, high temp
s1<-stress1.1
s2<-stress1.2
s3<-stress2.1
s4<-stress2.2
sd<-tau.p.01.sd
tau.p.01.var<-tau.p.01.sd^2
model.lo <- loess(sd ~ mu.1 + mu.2, span=.5, degree=2)
span<-50
st1 <- seq(min(mu.1), max(mu.1), len=span)
st2 <- seq(min(mu.2), max(mu.2), len=span)
newdata <- expand.grid(mu.1=st1,mu.2=st2)
fit.sd <- matrix(predict(model.lo, newdata), span, span)
fit.var<-fit.sd^2
```

2.5. Application: Industrial Case Study

2.5.1. Description of Design Problem

Reliability of first level C4 lead-free interconnect (First Level Interconnect Copper Bump) in a semi-conductor assembly is under evaluation. It is desired to conduct a two-stress factor accelerated life test at two levels in order to estimate the device lifetime at which no more than 1% of the total population is likely to fail with high estimation precision. Fail mode of interest is C4 bump electrical short failure, and fail mechanism is Cu migration in B-HAST testing (biased-highly accelerated stress test). Temperature and relative humidity are believed to accelerate the fail mechanism of interest. The use conditions are:

- Temperature target of 40°C, however it can range from a min of 30°C to a max of 50°C.
- Relative humidity target of 20%, however it can range from a min of 15% to a max of 30%.

Stress lab constraints and assumptions are:

- B-HAST stress chambers are available for 21 days (504 hours maximum test time).
- Temperature (T), and relative humidity (RH) setting are as follows:
  - Lower stress chamber can be set to run temperature range from 70°C to 95°C, and relative humidity range from 30% to 55%.
  - Higher stress chamber can be set to run temperature range from 105°C to 130°C, and relative humidity range from 60% to 85%.
  - Equipment's tolerance allows both temperature and relative humidity to be varied in increments of ∓ 5 units on their respective scales.
- Due to sample cost, a run budget of $N = 12$ are available for stress testing.

Our objective is to determine optimal unit allocation and stress level settings so as to maximize the prediction precision of the first percentile of reliability at use condition subject to allowable testing budget.

2.5.2. Acceleration Model

Acceleration model details have been described in Section 2.3.1.

2.5.3. Prior Distribution Elicitation

From Equation (6) the parameter vector θ is $(\gamma_0, \gamma_1, \gamma_2, \sigma)^T$, and we would need to specify a prior distribution for each of the parameters or $p(\theta)$. We would initially use the parameters in their original units (before transformation) to relate to the engineer's prior knowledge. Standardization is applied once prior distributions in original units have been effectively solicited from engineers.

Given historical learning and previous experience with similar fail mechanism, the reliability engineer believes that the appropriate independent prior distributions on the parameters can be specified as follows: For the activation energy, a uniform distribution that gives an equal likelihood for values that range from 1.8 to 2.1 would be appropriate to use, i.e., in the statistical software

R, $runif(SS, 1.8, 2.1)$. However, a wider range may need to be considered for the RH inverse power law parameter, which is $runif(SS, 0.2, 2.5)$. Not much was known about the intercept so it was given a vague (diffuse) normal distribution with mean of 0.0 and very low precision of $1.0E^{-6}$ ($\sigma = 1000\ or\ \sigma^2 = 1E^{+6}$). A positive density support was assumed for the Weibull shape parameter as: $rgamma(SS, 2, scale = 1)$. The parameter SS in above independent prior distributions reflects the sample size of random samples desired to be generated in simulation. Distribution notations used are in accordance with R-language syntax.

2.5.4. Simulation Search for an Optimal Design

Equally spaced design grid of all combinations of the allowed ranges for temperature (*Temp*) and relative humidity (*RH*) stresses can be used to simulate experiments. However, doing so is computationally inefficient and in a comparison study we contrast that with our proposed sampling scheme that aims at reducing computational burden without sacrificing statistical efficiency. We, instead, consider a set of testing conditions, in which each stress factor has as many levels as there are runs in the experiment and levels that are chosen to maximize the minimum distance between design points while requiring even spacing of design points. This is accomplished through a modified Latin hypercube experimental design as discussed in Section 2.2.

2.5.4.1. Construction of ALT Design Grid

Since it is desired to conduct a two-stress factor accelerated life test at two levels in order to estimate the device lifetime at use condition, the construction of the ALT design grid following our proposed methodology will entitle three steps: 1) LHD design at the lower two-stress factor level, 2) LHD design at the higher two-stress factor level, and 3) an augmented “modified” mLHD incorporating both the high and low stress levels as the finalized ALT design grid to simulate fail data and evaluate utility function.

1) LHD for the Lower Two-Stress Factor Level

Lower stress chamber can be set to run temperatures in the range from 70°C to 95°C, and relative humidity in the range from 30% to 55%. Note that fail mechanism is driven by the combined effect of temperature and relative humidity simultaneously, so lower stress, S_1 is in the form of $Temp.low$ ℃/$RH.low$ %. A full grid for all combinations of temperature and relative humidity can be used, however it will be at high computational cost, we instead use a more efficient representation through an LHD. Taking into account equipment capability:

- Lower stress chamber can be set to run temperature range from 70°C to 95°C, and relative humidity range from 30% to 55%.
- Equipment's tolerance allows both temperature and relative humidity to be varied in increments of ∓ 5 units on their respective scales.

One can use equipment tolerance to equally divide both ranges for $Temp.low$ ℃ and $RH.low$ % into six intervals, resulting in $LHD_{S_1}(6,2)$ as shown in Figure 2.1 and summarized in Table 2.1.

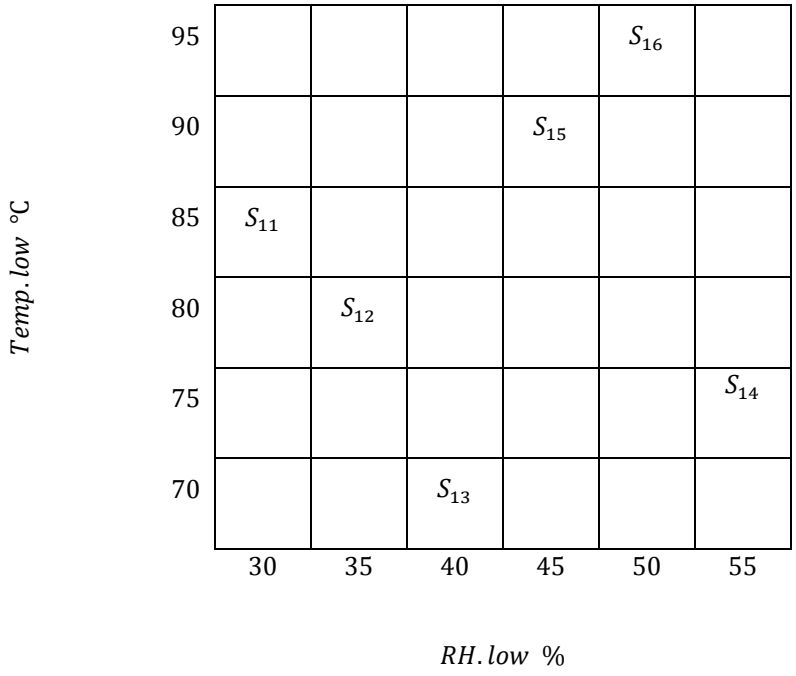


Figure 2.1: Latin Hypercube Grid for Lower Stress Chamber, $LHD_{S_1}(6,2)$

Lower Stress Chamber Setup		
Stress	RH% (Low)	Temp°C (Low)
S_{14}	55	75
S_{11}	30	85
S_{15}	45	90
S_{12}	35	80
S_{13}	40	70
S_{16}	50	95

Table 2.1: Latin Hypercube Grid for Lower Stress Chamber, $\boldsymbol{LHD_{S_1}(6,2)}$

2) LHD for the Higher Two-Stress Factor Level

Higher stress chamber can be set to run temperatures in the range from 105°C to 130°C, and relative humidity in the range from 60% to 85%. Higher stress, S_2 is in the form of $Temp.high℃/RH.high\%$. Similar to the lower stress chamber, one can use equipment tolerance to equally divide both ranges for $Temp.high$ ℃ and $RH.high$ % into six intervals, resulting in $LHD_{S_2}(6,2)$ as shown in Figure 2.2 and summarized in Table 2.2.

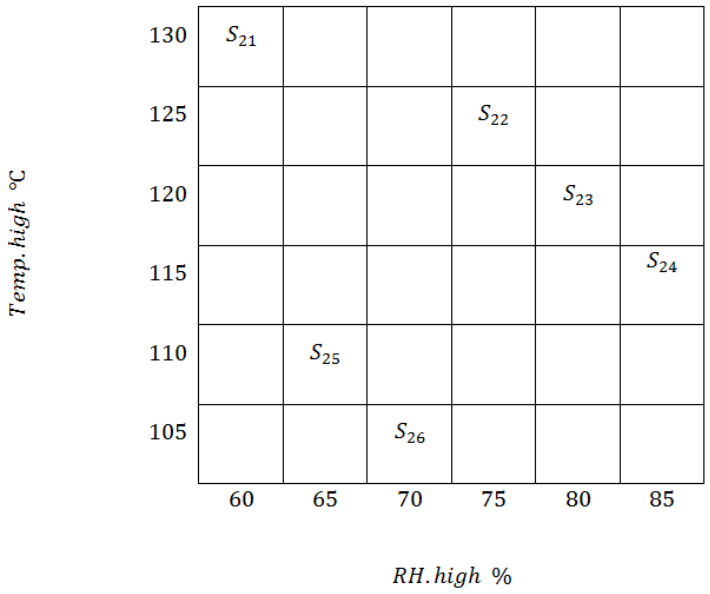


Figure 2.2: Latin Hypercube Grid for Lower Stress Chamber, $\boldsymbol{LHD_{S_2}(6,2)}$

Higher Stress Chamber Setup		
Stress	RH% (high)	Temp°C (high)
S_{22}	75	125
S_{23}	80	120
S_{25}	65	110
S_{21}	60	130
S_{26}	70	105
S_{24}	85	115

Table 2.2: Latin Hypercube Grid for Higher Stress Chamber, $\boldsymbol{LHD_{S_2}(6,2)}$

3) mLHD for the ALT Design Grid

In accelerated life testing (ALT), data obtained from all stress levels are used to predict quantity of interest at use condition through acceleration factors AF. The design grid at witch fail data are simulated and point-wise local utility function are calculated is an $mLHD$ constructed from both $LHD_{S_1}(6,2)$ and $LHD_{S_2}(6,2)$ along with corner and/or center point augmentation. Figure 2.3 shows the resulting pre-augmented $LHD_{(S_1,S_2)}(6,2)$. Figure 2.4 shows the augmented design grid, augmentation points are represented by letter "O". Finalized design grid is summarized in Table 2.3. Figure 2.5 displays the progression of the construction of the Latin hypercube design grid for our ALT problem.

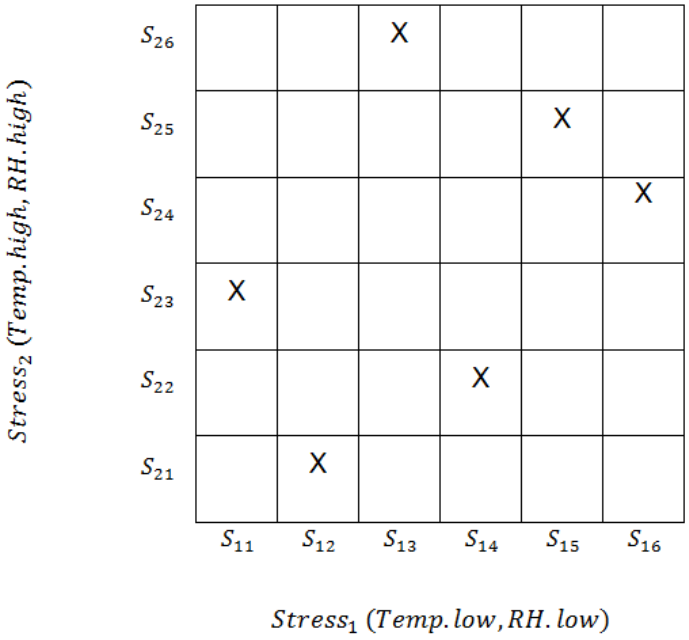


Figure 2.3: Pre-augmented Latin Hypercube Grid for ALT, $LHD_{(S_1,S_2)}(6,2)$

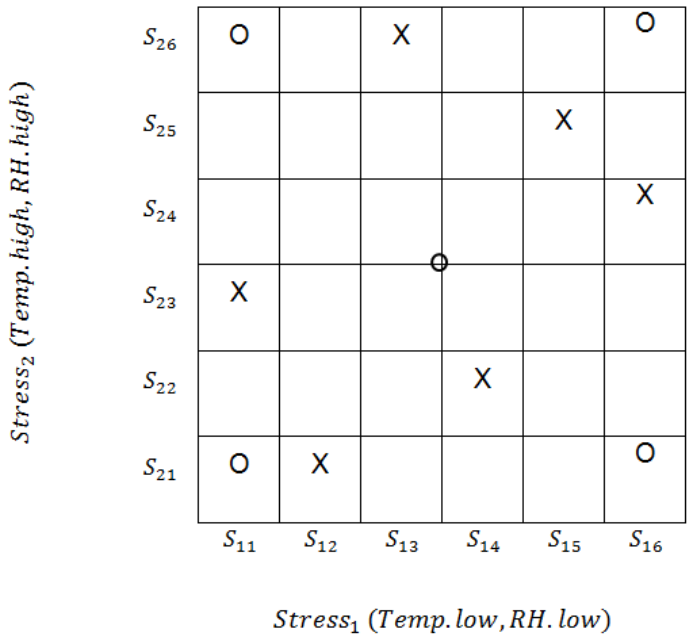


Figure 2.4: Augmented Latin Hypercube Grid for ALT, $mLHD_{(S_1,S_2)}(6,2)$

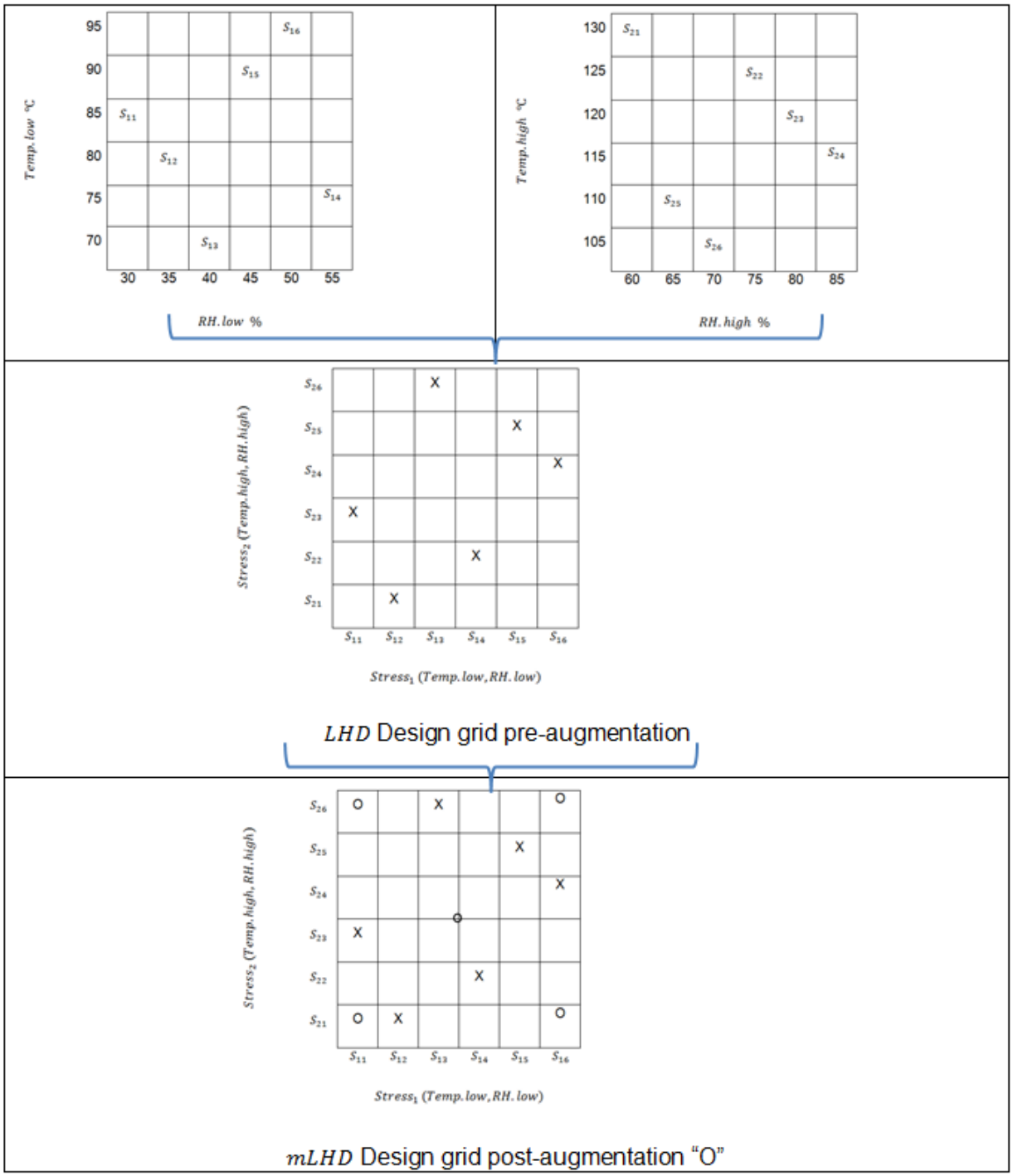


Figure 2.5: Progression of Construction for the Latin Hypercube Design Grid for ALT Problem

Run #	Lower Stress Chamber Setup		Higher Stress Chamber Setup		Run Source
	RH% (Low)	Temp°C (low)	RH% (High)	Temp°C (High)	
1	55	75	75	125	mLHD
2	30	85	80	120	mLHD
3	45	90	65	110	mLHD
4	35	80	60	130	mLHD
5	40	70	70	105	mLHD
6	50	95	85	115	mLHD
7	30	70	60	105	AUG-C1
8	55	95	85	130	AUG-C2
9	30	70	85	130	AUG-C3
10	55	95	60	105	AUG-C4
11	42.5	82.5	72.5	117.5	AUG-CN
LHD #	(1)		(2)		
	(3)				

Table 2.3: Modified Latin Hypercube Grid with 6 Runs and Corner/Center Augmentations

2.5.4.2. Decision Variables for the ALT Design

Now that we have constructed the design grid for the two-stress factor ALT planning problem, we list the decision variables we would like to solve for to determine the optimal plan setup. Typical ALT plan would require specification of the following:

1) Number of stress levels the test will be run at. In this case study, the number of stress levels has been already fixed at two levels given the engineer's desire to run a two- stress level test.
2) Magnitude of applied stress at each of the two stress levels? The engineer's desire is to find this out given test constraints at hand.
3) Allocation of units to each of the two stress levels? The engineer's desire is to find this out given test constraints at hand.

So given (1), the optimal design will determine the optimal set up for (2) and (3) under our proposed design optimization criterion discussed in Section 2.3.2.

2.5.4.3. Evaluation Steps Over the ALT Design Grid

1. Over the design grid, mLHD, we randomly simulate fail data from the joint density of parameters and data (θ, y)

$$(\theta, y) \sim p(\theta, Y) \sim p(\theta).p(y|\theta) \sim prior(\theta).likelihood(y|\theta) \quad (17)$$

That is, independently generate random fail data using the acceleration model as in Equation (6).

1.1. Simulated failure times are compared against a predetermined test time t_c to determine the censoring time for each test unit. Those units failing before or at t_c are considered exact failures, while others exceeding t_c are considered right censored. If testing time availability of each stress chamber is different (additional stress lab constraint), then censoring scheme can easily be modified to account for that by comparing fail times at different stress levels against the individual level test time t_{c_i} where $i = 1, \dots, l$ and l is the number of stress levels in the experiment, for example $l = 2$ in this case study.

1.2. The number of simulated fail times at each cell of the design grid is determined by unit allocation for the design being considered. We loop over all possible combination of unit allocation to the two stress levels specified in this study. For example, if we consider corner cell (S_{11}, S_{21}) of Figure 2.4, for a total $N = 12$ units, with a design that allocates $n_1 = 8$ units to lower stress level S_{1i} and $n_2 = 4$ units to higher stress level S_{2i}, there will be a total of 12 observations generated in corner cell (S_{11}, S_{21}), i.e. 8 under stress level S_{11}, and 4 under stress level S_{21}. This will carry on for the other cells identified by the $mLHD$ to complete fail data simulation for the identified design allocation $(n_1 = 8, n_2 = 4)$. The same will be repeated for other design allocations.

2. For each simulated experiment (fail times at a particular grid cell), use Gibbs sampler to evaluate the posterior utility

$$u_i = var_{\theta|y,\eta}\left(\log\left[\tau_p(x_U)\right]\right) \quad (18)$$

Where $\tau_p(x_{UC})$ is the 100 p^{th} quantile of the lifetime distribution at use condition (design stress) and $(i = 1, \dots, K)$, where K is the number of active cells with observations as identified by the grid of the $mLHD$. In this example, the engineer has chosen $p = 0.01$ (the 1st percentile).

2.1. While performing this step, the likelihood function must be adjusted to account for the censoring structure in the data. When using WinBUGS MCMC simulation through Gibbs sampler, this can be accomplished using the function $I(t_c, t)$ for the right censored observations (data).

3. Since we are in the planning phase of the test, we base our decision of design selection based on the values of the pre-posterior expected utility. Thus, we do the following:

3.1. Approximate pre-posterior utility function by fitting a smooth surface to the Monte Carlo sample generated in step (2) as a function of evaluated designs. An example for a design would be $[(S_1: 70℃/30\%, n_1: 4),\ (S_2: 130℃/85\%, n_2: 8),\ N = 12]$

3.2. We used the local regression model LOESS in R "modreg" library with its default settings for the smooth surface fitting, details are given in Section 2.4.4. This step in effect is equivalent to evaluating the integration in the solution to the optimal design problem in Equation (1).

4. The optimal design is found by maximizing the fitted surface (minimum pre-posterior variance), and can be read off as the mode. Since the engineer has started off with a fixed number of stress levels to be used (i.e. two), the optimal design will answer questions (2) and (3) of Section 2.5.4.2, where the magnitude of stress and unit allocation at the two stress levels are determined.

2.5.5. Optimal Design: Simulation Search Results

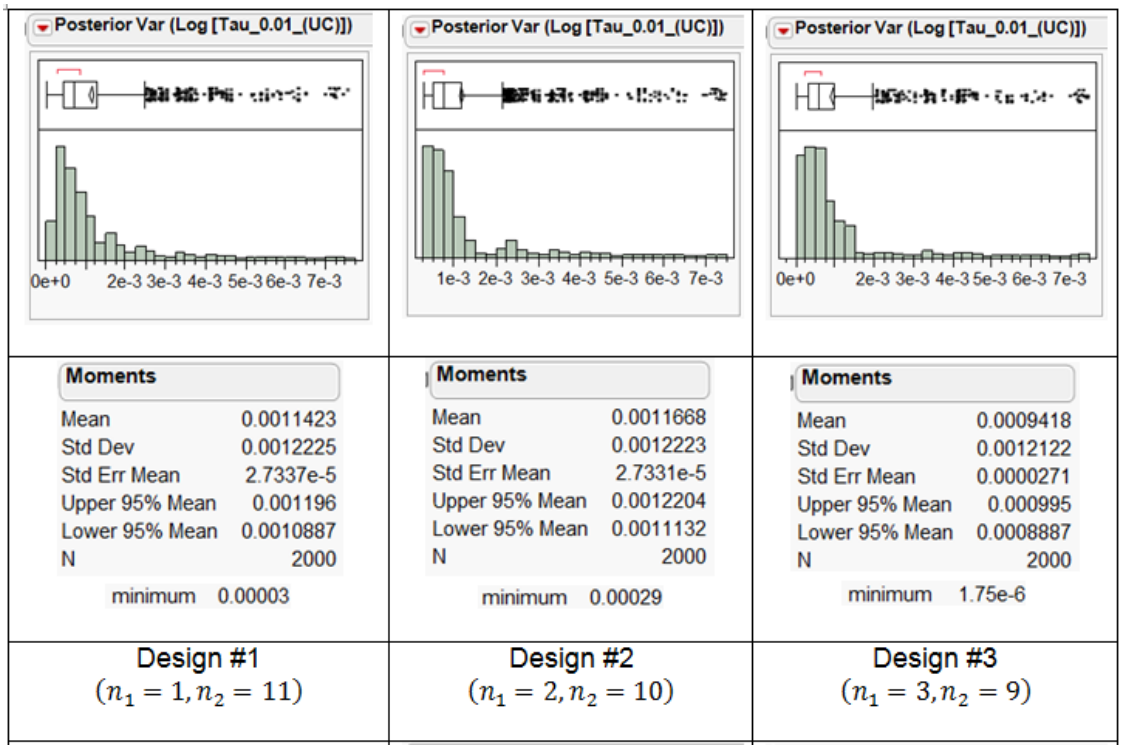
Table 2.4 summarizes the optimal test plans for the different designs of interest (i.e. test unit allocation and stress magnitude at the two stress levels). The table lists the minimum pre-posterior variance of the logarithm of 1st percentile estimate of life distribution at use condition

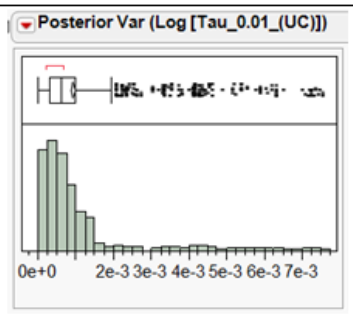
($\mathrm{var}_{\theta|y,\eta}(\log[\tau_{0.01}(x_{UC})])$), its mean pre-posterior value and the standard error mean of a sample of 2000 simulations per design. Each row in Table 2.4 represents an optimal design among designs sharing the same unit allocation (n_1, n_2) according to the criterion of lower mean pre-posterior variance of $\log[\tau_{0.01}(x_{UC})]$. Then among all optimal designs having different allocations, we select an overall optimal design that minimizes the same criterion across obtained optimal designs, this is, design #10 in table 2.4 with optimality criterion value of [0.0004137]. Figure 2.6 provides summary statistics for the different optimal designs per 2000 simulated runs of each design. In Section 2.5.6, we provide an evaluation study of the recommended optimal design in comparison with few others designs.

Given our prior knowledge regarding model parameters and the simulated fail data that are constraint by test units availability, stress lab capability and available testing time, the optimal design is given by S_1^* $(RH_1 = 30\%, Temp_1 = 70°\mathrm{C})$ and S_2^* $(RH_2 = 60\%, Temp_2 = 105°\mathrm{C})$ with the optimal unit allocation that allocates majority of test units to the lower stress condition S_1^* (less extrapolation to the use condition will drive less variability in prediction) such as $n_1 = 10, n_2 = 2$. This, in effect, may be approaching the one-level design where all units are allocated to the lower stress level. It is in part due to the ranges of model parameters that the reliability engineer has specified as prior distributions. For example, the range for the activation energy was set to run uniformly from 1.8 to 2.1, and this is considered as a precise estimate of this parameter to begin with; therefore, there would be no need for a larger sample size to be allocated at the higher stress condition to better estimate this parameter.

Design			Optimal Stress Levels				Minimum Pre-posterior variance of log(1st %tile) at UC	Mean Pre-posterior variance of log(1st %tile) at UC	Std Err mean of Pre-posterior variance of log(1st %tile) at UC
#	N =12 units		S_1^*		S_2^*				
	n_1	n_2	$RH_1\%$	$T_1°C$	$RH_2\%$	$T_2°C$			
1	1	11	45	90	85	115	2.94E-05	0.0011423	2.734E-05
2	2	10	40	70	85	130	2.93E-04	0.0011668	2.733E-05
3	3	9	30	70	85	130	1.76E-06	0.0009418	2.710E-05
4	4	8	30	70	85	130	9.96E-09	0.0009229	2.636E-05
5	5	7	55	75	60	105	9.57E-08	0.0008114	2.396E-05
6	6	6	30	70	60	105	2.15E-04	0.0008557	2.367E-05
7	7	5	30	70	60	105	1.16E-07	0.0005819	1.650E-05
8	8	4	30	70	60	105	3.73E-06	0.0005064	1.606E-05
9	9	3	30	70	60	105	2.13E-05	0.0005124	1.482E-05
10*	10	2	30	70	60	105	1.37E-11	**[0.0004137]**	9.565E-06
11	11	1	30	70	60	105	2.40E-09	0.0004983	1.463E-05

Table 2.4: Two-level Optimal Designs for (N=12) Test Units

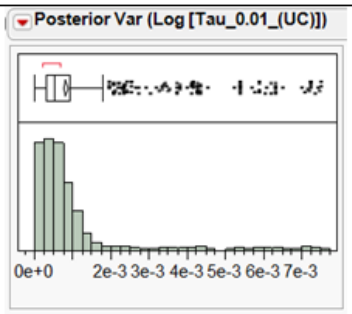




Moments	
Mean	0.0009229
Std Dev	0.0011792
Std Err Mean	2.6367e-5
Upper 95% Mean	0.0009746
Lower 95% Mean	0.0008712
N	2000

minimum 9.95e-9

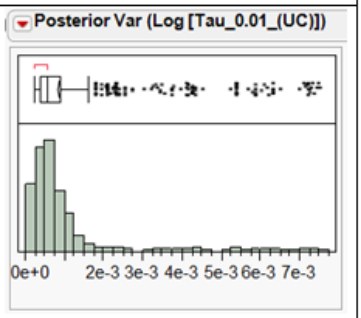
Design #4
$(n_1 = 4, n_2 = 8)$



Moments	
Mean	0.0008114
Std Dev	0.0010718
Std Err Mean	2.3966e-5
Upper 95% Mean	0.0008584
Lower 95% Mean	0.0007644
N	2000

minimum 9.55e-8

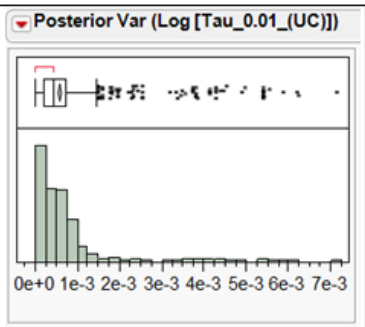
Design #5
$(n_1 = 5, n_2 = 7)$



Moments	
Mean	0.0008557
Std Dev	0.0010587
Std Err Mean	2.3674e-5
Upper 95% Mean	0.0009021
Lower 95% Mean	0.0008092
N	2000

minimum 0.00021

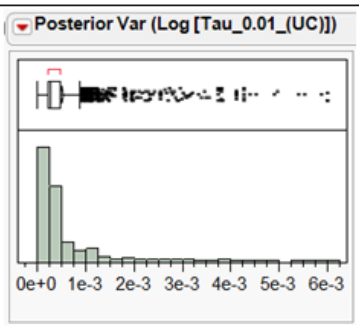
Design #6
$(n_1 = 6, n_2 = 6)$



Moments	
Mean	0.0005819
Std Dev	0.0007383
Std Err Mean	0.0000165
Upper 95% Mean	0.0006143
Lower 95% Mean	0.0005495
N	2000

minimum 1.17e-7

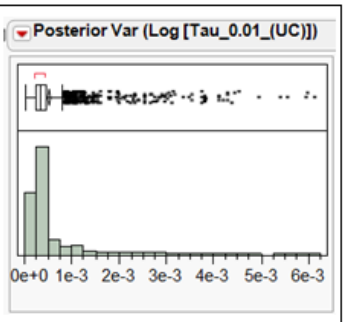
Design #7
$(n_1 = 7, n_2 = 5)$



Moments	
Mean	0.0005064
Std Dev	0.0007183
Std Err Mean	1.6062e-5
Upper 95% Mean	0.0005379
Lower 95% Mean	0.0004749
N	2000

minimum 3.73e-6

Design #8
$(n_1 = 8, n_2 = 4)$



Moments	
Mean	0.0005124
Std Dev	0.0006626
Std Err Mean	1.4816e-5
Upper 95% Mean	0.0005414
Lower 95% Mean	0.0004833
N	2000

minimum 2.13e-5

Design #9
$(n_1 = 9, n_2 = 3)$

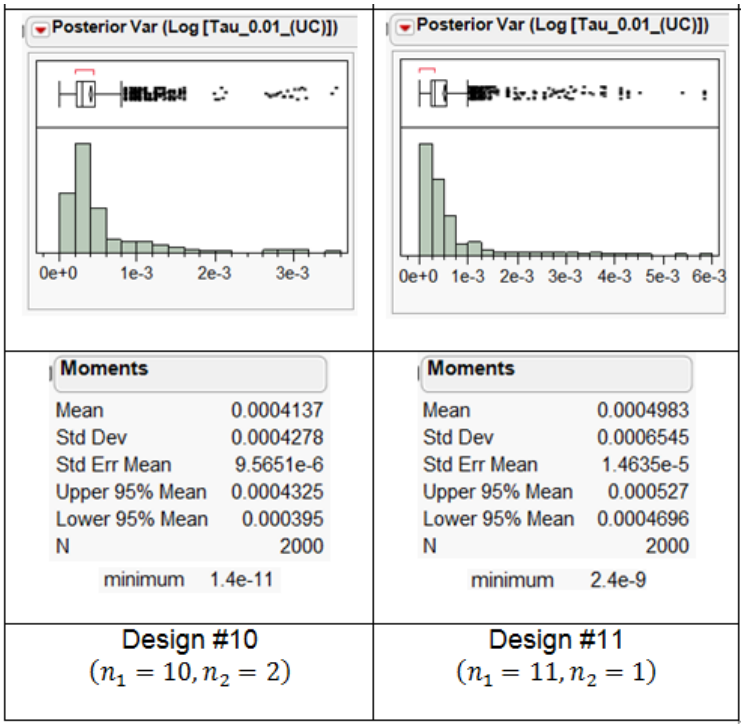


Figure 2.6: Summary Statistics for the Different Optimal Designs per 1000 Simulated Runs of Each Design

2.5.6. Evaluation of Recommended Optimal Design

In this section, we conduct an evaluation study of the obtained optimal design relative to three other designs that were not selected as optimal given our design criteria. We select comparison designs in terms of allocation. Thus, in Table 2.5, optimal design is numbered as # 3, comparison design # 4 is close to optimal in having more units allocated toward n_1, comparison design # 2 recommends equal unit allocation between n_1 and n_2 and comparison design # 1 recommends opposite allocation with more units allocated toward n_2.

In each trial a total budget of 12 units were assumed and data simulated according to the test levels and allocation as recommended by optimal designs as in Table 2.4. True model parameters are considered fixed and were taken as the means of the prior distributions assumed in previous case study except for activation energy that is assumed at the lower range of its

uniform distribution and set at 0.226, acceleration model and form of likelihood all maintained the same as in previous design problem. A total of 2500 data sets per design were simulated. For each of the simulated dataset, we computed the posterior mean and standard deviation of $\log[\tau_{0.5}(x_{UC})]$, where $\tau_{0.5}$ is the 50th percentile of life distribution. Note that $\log(\tau_{0.5})$ is a function of parameters θ and the posterior pdf of θ is proportional to likelihood (data|θ) x prior(θ). Table 2.5 summarizes the results. Table 2.6 interprets the 50th percentile at use condition across the different designs in comparison with its true value.

Design			Optimal Stress Levels				$\log(\tau_{0.5})$ Prediction at Use Condition (UC)	
			S_1^*		S_2^*		$x_{UC}: (RH = 20\%, T = 40°\text{C})$	
#	n_1	n_2	$RH_1\%$	T_1°C	$RH_2\%$	T_2°C	$\log(\hat{\tau}_{0.5})$	$Std.Dev[\log(\hat{\tau}_{0.5})]$
1	2	10	40	70	85	130	9.712	0.3205
2	6	6	30	70	60	105	10.061	0.3420
3*	10	2	30	70	60	105	10.531	0.2909
4	11	1	30	70	60	105	10.463	0.3034
							True $\log(\tau_{0.5}) = 10.561$	

Table 2.5: Comparison across Optimal ALT Designs

Design #	Use condition(UC)		Percentile Estimate at (UC)
	RH_{UC} %	T_{UC} °C	$\hat{\tau}_{0.5}$ hrs
1	20	40	16,514.59
2	20	40	23,411.90
3*	20	40	37,458.91
4	20	40	34,996.38
$True\ \tau_{0.5}(UC)$	38,599.71 hrs		

Table 2.6: UC 50th Percentile Interpretation across Optimal ALT Designs

Note that the true value of $\tau_{0.5}(UC)$ is calculated through the linearized acceleration model at use condition. That is,

$$\log[\tau_{0.5}(UC)] = \beta_0 + \beta_1 * \left(\frac{1}{K * Temp_{UC}}\right) + \beta_2 * RH_{UC}$$

- With parameter values fixed at the prior distribution means for β_0 and β_2 and activation energy $\beta_1 = 0.226$. $Temp$ is temperature in Kelvin $(c^o + 273.15)$ and K is Boltzmann's constant $\left(8.617385\ E^{-5}\ \frac{eV}{K}\right)$.

2.6. Comparison to Full Grid Method

Equally spaced design grid of all combinations of the allowed ranges for temperature (*Temp*) and relative humidity (*RH*) stresses (Full grid) will be used in this section to simulate experiment in place of our modified mLHD. Stress equipment tolerances for both of temperature and relative humidity were considered as grid increments. R-function $\boldsymbol{expand.grid\{base\}}$ was used to create the design grid from all combinations of stress factors; {in R: expand.grid(temp.low = seq(70, 95, 5), rh.low = seq(30, 55, 5), temp.high = seq(105, 130, 5), rh.high = seq(60, 85, 5))}. That resulted in a total of 1,296 combinations to consider as opposed to 11 combinations as in our proposed algorithm.

Run #	Low Stress Chamber Setup		High Stress Chamber Setup		Run Source
	RH% (Low)	Temp°C (low)	RH% (High)	Temp°C (High)	
1	70	30	105	60	Full Grid
2	75	30	105	60	Full Grid
3	80	30	105	60	Full Grid
⋮	⋮	⋮	⋮	⋮	Full Grid
939	80	30	115	80	Full Grid
⋮	⋮	⋮	⋮	⋮	Full Grid
1296	95	55	130	85	Full Grid

Table 2.7: Equally Spaced Grid Design

We applied the same settings, and followed the same simulation steps as described in section 2.4 with only exception of using a full grid instead of our mLHD design. Total sample size $= n_1 + n_2$, where n_1 is the unit allocation at stress level S_1 and n_2 is the unit allocation at stress

level S_2. We applied 2500 simulated runs per design to arrive at an optimal design setting. Given similar amount of model parameters prior knowledge as in section 2.5.3, the resulting simulated fail data that were constraint by test units availability, stress lab equipment capability and available testing time, resulted in optimal design given by: S_1^* $(RH_1 = 30\%, Temp_1 = 75°\mathrm{C})$ and S_2^* $(RH_2 = 65\%, Temp_2 = 110°\mathrm{C})$ with optimal unit allocation that allocates majority of test units to the lower stress condition S_1^* in which $n_1 = 10\ and\ n_2 = 2$. This unit allocation is in agreement with results from our method. This optimal unit allocation could also be explained relative to the ranges of model parameters that the reliability engineer has specified as prior distributions. For example, the range for the activation energy was set to run uniformly from 1.8 to 2.1, which reflects strong knowledge with less variability around the estimate of this parameter to begin with; therefore, there would be a lesser need for a larger sample size allocated at the higher stress condition to better estimate this parameter. As far as stress magnitudes are concerned, the two approaches slightly differ in the Temperature setup, but share the same Temperature delta ($\Delta T_{mLHD} = \Delta T_{Full\ Grid} = 35°\mathrm{C}$). They slightly differ in Relative humidity delta ($\Delta RH_{mLHD} = 30\%, \Delta RH_{Full\ Grid} = 35\%$).

Tables 2.8a and 2.8b show a summary comparison of the optimal design between full grid approach and our proposed modified LHD design. By using our approach the gain is tremendous in the cost of computations as reflected by the run time with minimal loss of accuracy in the evaluated utility function of interest.

Design	Low Stress Chamber Setup			High Stress Chamber Setup			Minimum Pre-posterior variance of log(1st %tile) at UC	Mean Pre-posterior variance of log(1st %tile) at UC	Std Err mean of Pre-posterior variance of log(1st %tile) at UC
	RH % Low	Temp °C Low	Unit allocation (n_1)	RH % High	Temp °C High	Unit allocation (n_2)			
mLHD	30	70	10	60	105	2	1.370E-11	0.0004137	9.565E-06
Full Grid	30	75	10	65	110	2	1.364E-11	0.0004129	9.561E-06

Table 2.8a: Equally Spaced Grid vs. mLHD Optimal Design Comparison

Design	% Gain in Estimate Accuracy	% Reduction in Number of Runs	% Reduction in Total Run Time*
mLHD	--	99	96.7
Full Grid	0.4	--	--
* HP EliteBook 8560w Workstation, Intel Core i5, 4 GB RAM, 2.6 GHz			

Table 2.8b: Equally Spaced Grid vs. mLHD Optimal Design Comparison

2.7. Comparison to Large-Sample Approximate & Full Grid Methods

In this section we compare designs obtained using our proposed simulation-based approach to its counterpart that is based on asymptotic or large-sample approximation of the Bayesian design criterion. For basis of comparison, we use the design problem described in Zhang and Meeker [51], where the recommended design was obtained using an asymptotic approximation method. Although we have shown the advantages of our proposed approach to the

full grid one in section 6, we do also provide the solution point obtained by the full grid as an additional comparison.

2.7.1. Problem Description

The reliability engineer is to investigate a new type of adhesive bond by estimating the $p = 0.1$ quantile of the lifetime distribution at 50^oC. A sample of $n = 300$ units is available, and the testing time is restricted to 6 months (t_c = 183 day). No failures would be expected for testing at 50^oC. Thus a high-temperature ALT was proposed.

Based on past experience with similar adhesive bonds, the engineer thought that the Weibull distribution would adequately describe adhesive bond life (implying an SEV distribution for the log-life). Also, the Arrhenius relationship was expected to describe the temperature acceleration up to 120^oC. The acceleration model can be expressed as $\mu = \gamma_0 + \gamma_1 x$ and $x = -11{,}605/(temp^oC + 273.15)$ with the experimental region between $x_{Use} = -35.9$ and $x_{High} = -29.5$. γ_1 is negative, implying more failures at higher temperatures. $|\gamma_1|$ is interpreted as the effective activation energy of the chemical reaction in units of electron volts. Section 4.2 of Zhang and Meeker [51] discusses prior distribution specification for the different parameters.

In their numerical search for the two-point optimum plan, Zhang and Meeker [51] assumed that one of the optimum test points must be the highest allowable variable level. Therefore, they fixed $\xi_{high} = 1$ to reduce the dimension of the optimization space. They argued that in most ALTs, censoring at the use condition is heavy which generally implies that one of the optimum test points must be the highest allowable variable level. We did not place the same assumption/restriction (i.e. the highest stress level being one of the test points), but rather allowed our optimization procedure to determine optimal stress levels freely within their allowable ranges. We considered two ranges (high & low) for the acceleration factor (temperature) where optimization is done without restriction. Although temperature is one variable, this split in range reflects the fact that two stress ovens are available in the stress lab that differ in the temperature

range each is capable of. High temperature ovens are typically more expensive to buy due to material of construction of high tolerance to heat. The ranges considered were as follows:

- $Temp_{low}$ (Stress 1) has the range of 85 to 100^oC, and
- $Temp_{high}$ (Stress 2) has the range of 105 to 120^oC.

2.7.2. Designs Comparison Result

From Table 2.5, there seems to be an agreement in the setup of the high stress level between the three approaches. Although, we did not force the highest stress level to be one of the design points (less assumptions is preferable), our ALT design returns the value of 120^oC (rounded up from 118.5^oC) as highest temperature stress setup which is in agreement with the fixed value of 120^oC forced by the large-sample approximation method. It also returns the value of 85^oC (rounded up from 84.73^oC) as the lower stress setup which is 9^oC lower than the value returned by the large-sample approximation method of 94^oC. The Full Grid method returns the value of 120^oC (rounded up from 119.2^oC) as highest temperature stress setup which is in agreement with our method and the fixed value of 120^oC forced by the large-sample approximation method. It also returns the value of 90^oC (rounded up from 87.62^oC) as the lower stress setup which 5^oC higher than our proposed method of 85^oC and is 4^oC lower than the value returned by the large-sample approximation method of 94^oC. Thus, in effect both simulation methods stretch the low and high stress levels as far apart as possible which is known to be a good design practice that reduces the variance of the estimated quantities. The two simulation methods, ours and the full grid provide roughly close proportions for unit allocation that differ from the large-sample approximate method. Simulation-based designs tend to allocate more units toward the lower stress level (65%-35%, for our proposed method) and (63%-37%, for the full grid method) as compared to the approximation method which splits the units almost (50%-50%). More units towards the lower stress level is preferable as the expected number of fails decreases with the decrease in stress, so adding more units aims at obtaining more fails at the lower stress level. This will improve predictions at use level. The large-sample approximation method reports

an average minimum posterior variance estimate of the 1st percentile of life distribution at use condition as 0.0897, while the simulation methods yields a variance estimate of 0.07698 (our proposed method), and 0.07667 (full grid method) for the same quantity of interest at use condition.

Method			
Large-Sample Approximation			
Condition	Temperature (oC)	Proportion (π_i)	Unit Allocation (n_i)
Use	50	-	-
Low	94	0.501	150
High	120	0.499	150
Simulation-Based Method (mLHD)			
Condition	Temperature (oC)	Proportion (π_i)	Unit Allocation (n_i)
Use	50	-	-
Low	85	0.65	195
High	120	0.35	105
Simulation-Based Method (Full Grid)			
Condition	Temperature (oC)	Proportion (π_i)	Unit Allocation (n_i)
Use	50	-	-
Low	90	0.63	189
High	120	0.37	111

Table 2.9: Two-point Optimal ALT to Estimate $\boldsymbol{\tau_{0.1}(x_{UC})}$

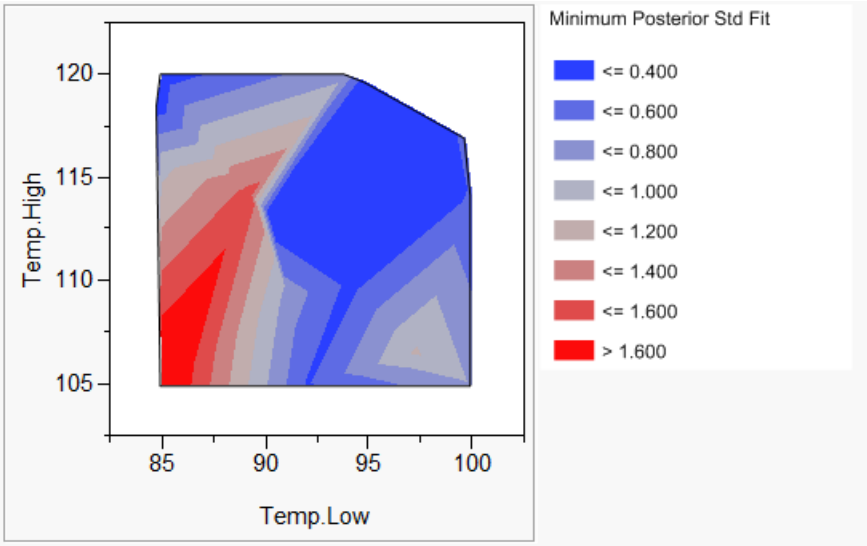


Figure 2.7: Contour Plot of Minimum Posterior Standard Deviation of 1st Percentile Estimate at UC from Our Proposed Simulation-based Design

2.8. Conclusion and Areas for Future Research

We have presented a simulation-based Latin Hypercube method for the planning of accelerated life tests in a Bayesian decision theoretic framework. In this context the design problem is thought as a decision problem and optimal design is obtained by maximizing an expected utility function which reflects the objective of the experiment. We illustrated the proposed approach through a problem with practical constraints when the underlying life model has a Weibull distribution with type-I censoring and failure mechanism that is driven by two accelerating variables (relative humidity and temperature). The applied approach was able to incorporate available prior information or knowledge on model parameters along with simulated future data from an appropriate probability model and use them to revise our knowledge according to Bayes' Theorem. Comparison of proposed approach to the full grid method demonstrated a tremendous saving in the cost of computations as reflected by the run time with minimal loss of accuracy in the evaluated utility function of interest. A second comparison of proposed approach to the large-sample approximation method revealed our approach's flexibility in determining optimal stress setting with less assumptions and more intuitive unit allocations.

As with many of the complex MCMC problems, the main limitation is typically the computationally intensive calculations and the need for point-wise evaluation of utility function. We helped remedy it by using a modified Latin Hypercube sampling scheme reinforced with design point augmentation followed by the application of curve-fitting optimization approach. The simulation-based Bayesian approach described in this paper could be extended to the ALT planning problems with more than two accelerating variables and more complicated models, such as non-linear acceleration models and accelerated lifetime models with non-constant scale parameters.

References

[1] D. S. Bai, and S. W. Chung, An optimal design of accelerated life test for exponential distribution, Reliability Engineering and System Safety **31** (1991), 57-64.

[2] D. S. Bai, and S. W. Chung, Optimal design of partially accelerated life test for exponential distribution under Type-I censoring, IEEE Transactions on Reliability **41** (1992), 400-406.

[3] D. S. Bai, and M. S. Kim, Optimum simple step-stress accelerated life tests for the Weibull distribution and type-I censoring, Naval Research Logistics **40** (1993), 193-210.

[4] D. W. Behnken, and D. G. Watts, Bayesian estimation and design of experiments for growth rates when sampling from the Poisson distribution, Biometrics **28** (1972)**,** 999-1009.

[5] J. O. Berger, Statistical Decision Theory and Bayesian Analysis, 2nd ed. N.Y.: Springer-Verlag, (1985).

[6] K. Chaloner, and K. Larntz, Bayesian Design for Accelerated Life Testing, Journal of Statistical Planning and Inference **33** (1992), 245–259.

[7] K. Chaloner, and K. Larntz, Optimal Bayesian Design Applied to Logistic Regression Experiments, Journal of Statistical Planning and Inference **21** (1989), 191–208.

[8] K. Chaloner and I. Verdinelli. Bayesian Experimental Design: A Review. Statistical Science, **10** (1995), 273-304.

[9] P. S. Chan, H. Y. So, H. K. T. Ng, and W. Gao, Optimal Sample Size Allocation for Multi-level Stress Testing with Extreme Value Regression Under Type-I Censoring, Naval Research Logistics Manuscript, (2012).

[10] T. Cioppa, and T. Lucas. Efficient Nearly Orthogonal and Space-Filling Latin Hypercubes. Technometrics **49** (2007), 1, pg. 45.

[11] M. Clyde, P. Müller, and G. Parmigiani, Optimal Design for Heart Defibrillators, in Case Studies in Bayesian Statistics 2, eds. C. Gatsonis, J. Hodges, R. E. Kass, and N. Singpurwalla, New York: Springer-Verlag, (1995) pp. 278–292.

[12] A. Erkanli, and R. Soyer, Simulation-Based Designs for Accelerated Life Tests, Journal of Statistical Planning and Inference **90** (2000), 335–348.

[13] L. A. Escobar, and W. Q. Meeker, Planning accelerated life tests with two or more experimental factors, Technometrics Vol. **37** (4) (1995), 411-427.

[14] K. Fang, C. Ma, and P. Winker. Centered L_2-Discrepency of Random Sampling and Latin Hypercube Design, and Construction of Uniform Designs. Mathematics of Computation Vol. **71**(2000), 237, 275-296.

[15] A. A. Giunta, S. F. Wojtkiewicz Jr., and M. S. Eldred, Overview of Modern Design of Experiments Methods for Computational Simulations, AIAA (2003)-0649.

[16] M. Hamada, H. F. Martz, C. S. Reese, and A. G. Wilson, Finding Near-Optimal Bayesian Experimental Designs via Genetic Algorithms, The American Statistician **55** (2001), 175–181.

[17] V. Joseph, and Y. Hung. Orthogonal-Maximin Latin Hypercube Designs. Statistica Sinica **18**, (2008), 171-186.

[18] M. Keramat, and R. Kielbasa. Modified Latin Hypercube Sampling Monte Carlo (MLHSMC) Estimation for Average Quality Index. Analog Integrated Circuits and Signal Processing, **19** (1999), 87-98.

[19] D. V. Lindley, Bayesian Statistics – A review. SIAM, Philadelphia (1972).

[20] X. Liu, and L.-C. Tang, A Bayesian Optimal Design for Accelerated degradation tests, John Wiley & Sons, Ltd., (2010).

[21] M. D. McKay, W. J. Conover, and R. J. Beckman. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. Technometrics, **21** (1979) 239-245.

[22] W. Q. Meeker, and G. J. Hahn, Asymptotically optimum over-stress tests to estimate the survival probability at a condition with a low expected failure probability, Technometrics **19** (1977), 381–399.

[23] W. Q. Meeker, and L. A. Escobar, Statistical Methods for Reliability Data, New York: Wiley (1998).

[24] W. Q. Meeker, and W. Nelson, Optimum Accelerated Life Tests for Weibull and Extreme Value Distributions, IEEE Transactions on Reliability, R-**24** (1975), 321–332.

[25] E. Monroe, R. Pan, C. Anderson-Cook, D. Montgomery, and C. Borror, A Generalized Linear Model Approach to Designing Accelerated Life Test Experiments, Quality Reliability Engineering International **27** (2010)**,** 595-607.

[26] E. Monroe, R. Pan, C. Anderson-Cook, D. Montgomery, and C. Borror, Sensitivity analysis of optimal designs for accelerated life testing, Journal of Quality Technology **42** (2010)**,** 121-135.

[27] M. D. Morris and T. J. Mitchell. Exploratory designs for computational experiments. J. of Statistical Planning and Inference **43** (1995) 381-402.

[28] P. Müller, Simulation-Based Optimal Design, in Bayesian Statistics 6, eds. J. M. Bernardo, J. O. Berger, A. P. David, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, (2000) pp. 459–474.

[29] P. Müller, and G. Parmigiani, Optimal Design via Curve Fitting of Monte Carlo Experiments, Journal of the American Statistical Association **90** (1995), 1322–1330.

[30] W. Nelson, Accelerated Testing, John Wiley & Sons: New York, (1990).

[31] W. Nelson, and T. Kielpinski, Theory for Optimum Censored Accelerated Life Tests for Normal and Lognormal Distributions, Technometrics **18** (1976), 105–114.

[32] W. Nelson, and W. Q. Meeker, Theory for optimum accelerated Censored Life Tests for Weibull and Extreme Value Distributions, Technometrics **20** (1978), 171-177.

[33] A. B. Owen. Orthogonal Arrays for Computer Experiments, Integration and Visualization. Statistica Sinica **2** (1992) 439-452.

[34] K. Palmer, and K. Tsui. A minimum bias Latin hypercube design. IIE Transactions **33** (2001), 793-808.

[35] J. W. Park, and B. J. Yum, Optimal Design for Accelerated Life Tests with Two Stresses, Naval Research Logistics **43** (1996)**,** 863-884.

[36] J. Park. Optimal Latin-hypercube designs for computer experiments. J. of Statistical Planning and Inference **39** (1994) 95-111.

[37] N. G. Polson, A Bayesian Perspective on the Design of Accelerated Life Tests, in Advances in Reliability, ed. A. P. Basu, New York: Elsevier (1993) pp. 321–330.

[38] H. Raiffa, R. Schlaiffer, Applied Statistical Decision Theory. Division of Research, Harvard Business School, Boston (1961).

[39] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, Design and Analysis of Computer Experiments, Statistical Science Vol. **4** (4) (1989), 409-435.

[40] R. R. Sitter, and B. Torsney, Optimal designs for binary response experiments with two design variables, Statistica Sinica **5** (1995)**,** 405-419.

[41] D. M. Steinberg, and D. k. j. Lin. A construction method for orthogonal Latin hypercube designs. Biometrika **93** (2006) 279-288.

[42] D. Sun, R. Tsutakawa, and W. Lu, Bayesian design of experiment for quantal responses: What's promised versus what's delivered, J. Statist. Planning and Inference **52** (1996), 289-306.

[43] L. C. Tag, Y. S. Sun, T. N. Goh, and H. L. Ong, Planning multiple-level constant stress accelerated life tests, Proceedings of the Annual Reliability and Maintainability Symposium (2002) 338- 342.

[44] B. Tang. OA-based Latin hypercubes. IIQO Research Report RR-91-10 (1991). The Institute for Improvement in Quality and Productivity, University of Waterloo. Submitted to JASA.

[45] B. Tang. Orthogonal Array-Based Latin hypercubes. J. of the American Statistical Association Vol. **88** (1993), No. 424, pp. 1392-1397.

[46] P. A. Tobias, and D. C. Trindade, Applied Reliability, 2nd ed., New York: Van Nostrand Reinhold (1995).

[47] I. Verdinelli, N. G. Polson, and N. D. Singpurwalla, Shannon Information and Bayesian Design for Prediction in Accelerated Life Testing, in Reliability and Decision Making, eds. R. E. Barlow, C. A. Clariotti, and F. Spizzichino, London: Chapman & Hall, (1993) pp. 247–256.

[48] P. Whittle, Some General Points in the Theory of Optimal Experimental Design, Journal of the Royal Statistical Society, Ser. B, **35** (1973), 123–130.

[49] K. Q. Ye. Column orthogonal Latin hypercubes and their application in computer experiments. J. Amer. Statist. Assoc. **93** (1998) 1430-1439.

[50] K. Q. Ye, W. Li, and A. Sudjianto. Algorithmic construction of optimal symmetric Latin hypercube designs. J. of Statistical Planning and Inference **90** (2000) 145-159.

[51] Y. Zhang, and W. Q. Meeker, Bayesian Methods for Planning Accelerated Life Tests, Technometrics, Vol. **48** (1), (2006).

CHAPTER 3

SIMULATION-BASED BAYESIAN OPTIMAL ACCELERATED LIFE TEST DESIGN FOR MODEL DISCRIMINATION

Ehab Nasir[1], Rong Pan[2]

[1] Industrial Engineer, Intel Corporation, Chandler, Arizona.

[2] Associate Professor, School of Computing, Informatics, and Decision Systems Engineering, Arizona State University.

Abstract

Accelerated life test (ALT) planning in Bayesian framework is studied in this paper with a focus of differentiating competing acceleration models, when there is uncertainty as to whether the relationship between log mean (life) and the stress (with possibly transformation) is linear or exhibits some curvature. The proposed criterion is based on the Hellinger distance measure between predictive distributions. The optimal stress-factor setup and unit allocation are determined at three stress levels subject to test-lab equipment and test-duration constraints. Optimal designs are validated by their recovery rates, where the true, data-generating, model is selected under the DIC (Deviance Information Criterion) model selection rule, and by comparing their performance with other test plans including the typically used three stress-levels good compromise plan, best traditional plan and well known 4:2:1 compromise ALT test plans. Results show that the proposed optimal design method has the advantage of substantially increasing a test plan's ability to distinguish among competing ALT models, thus providing better guidance as to which model is appropriate for the follow-on testing phase in the experiment.

Key Words - Reliability test plans, Hellinger distance, Model selection, Deviance information criterion (DIC), Non-parametric curve fitting.

3.1. Motivation for Work

Most work of the optimal Accelerated Life Testing (ALT) designs in literature has focused on finding test plans that allow more precise estimate of a reliability quantity, such as life percentile, at a lower stress level (it is usually the use stress level). See for example Nelson and Kielpinski [18, 28], Meeker [19], Nelson and Meeker [23, 30], Bai and Chung [6], Bai, Chung and Chun [7]. The associated confidence intervals of an estimate reflect the uncertainty arising from limited sample size and censoring at test, but do not account for model form inadequacy. Through the model-based extrapolation that characterizes ALTs, model errors can be quickly amplified and potentially dominate other sources of errors in reliability prediction. Implicit in the design criteria used in current ALTs is the assumption that the form of the acceleration model is correct. In many real-world problems this assumption could be unrealistic. A more realistic goal of an initial stage of ALT experimentation is to find an optimal design that helps in selecting a model among rival or competing model forms; i.e., a design that may assist in model discrimination. The ability to choose between competing model forms in an early experimentation stage has an important impact on the effective design of subsequent experimentation phases. For example, the use-condition extrapolation of ALT using model form (1) can be quite different than that of model forms (2) or (3).

$$\ln(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad (1)$$

$$\ln(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 \quad (2)$$

$$\ln(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 \quad (3)$$

Thus, the ALT designs that are good for model form discrimination could be quite different from those that are more appropriate for life percentile prediction under a specific model.

Extrapolation in both stress and time is a typical characteristic of ALT inference. The most common accelerated failure time regression models (based, for example, on Lognormal or Weibull fit to the failure time distribution at a given stress level) are only adequate for modeling some simple chemical processes that lead to failure, see Meeker and Escobar [23]. However, for modern electronic devices, more sophisticated models with basis in the physics of failure

mechanisms are required. These complicated models are expected to have more parameters with possible interactions among stress factors. Therefore, investigating ALT designs with model selection capability is needed more than ever before.

Meeker [22] in his discussion of figures of merit when developing an ALT plan emphasizes the usefulness of a test plan's robustness to the departure from the assumed model by evaluating the test plan properties under alternative, typically more general, models. For example, when planning a single-factor experiment under a linear model, it is useful to evaluate the test plan properties under a quadratic model. Also, when planning a two-factor experiment under the assumption of a linear model with no interaction, it is useful to evaluate the test plan properties under a linear model with an interaction term. We strongly believe that it is worthwhile to consider these recommended practices ahead of time when the test plan is being devised in the first place by allowing a design criterion that is capable of model form discrimination.

3.2. Previous Work

A considerable work has been done in the development of experimental designs for discrimination among regression models. See, for example, Hunter and Reiner [19], Box and Hill [9], Hill et al. [18], Atkinson and Cox [2]. A comprehensive review of early contributions is given by Hill [17]. Atkinson and Fedorov [4, 5] described the T-optimality criterion (non-Bayesian) where it is assumed that the true model and its parameters are known. Early work of discrimination among non-linear models resulted in sequential experimentation procedures. See for example, A. C. Atkinson, and D. R. Cox [2], A. C. Atkinson, and A. N. Donev [3], A. C. Atkinson, and V. V. Fedorov [4], A. C. Atkinson, and V. V. Fedorov [5]

More recently, many authors focused on the development of T-optimum criterion for model discrimination; see, for example, Atkinson and Donev [3], Ponce de Leon and Atkinson [21]. In addition, Ucinski and Bogacka [34] obtained optimal designs for discrimination of multi-response dynamic models; Dette and Titoff [14] derived new properties of T-optimal designs and showed that in nested linear models, the number of support points in a T-optimal design is usually

too small to enable the estimate of all parameters in the full model; Agboto, Li and Nachtsheim [1] reviewed T-optimality among other new optimality criteria for constructing two-level optimal discrimination designs for screening experiments.

Bayesian criteria were also considered in model discrimination; see, for example, Meyer et al. [26] where they considered a Bayesian criterion that is based on the Kullback-Leibler information to choose follow-up run after a factorial design to de-alias rival models. Bingham and Chipman [8] proposed a Bayesian criterion that is based on the Hellinger distance between predictive densities for choosing optimal designs for model selection with prior distributions specified for model coefficients and errors. For a comprehensive review on Bayesian experimental design reader is referred to Chaloner and Verdinelli [13].

All of the above attempts at model discrimination work have been in the context of traditional experimental design, i.e., standard experimental designs such as factorial, fractional factorial, Box central composite, etc. None to our knowledge has been explicitly targeting model discrimination in planning of accelerated life test (ALT) experiments, where failure time censoring is commonly expected. Nelson [30] (p. 350) has cautioned that the statistical theory for traditional experimental design is correct only for complete data, one should not assume that properties of standard experimental designs hold for *censored* and interval-censored data as they usually do not hold. For example, aliasing of effects may depend on the censoring structure. In addition, the variance of an estimate of a model coefficient depends on the amount of censoring at all test conditions and on the true value of (possibly all) model coefficients. Thus, the censoring times at each test condition are part of the experimental design and affect its statistical properties. As such, our current work draws its importance from its attempt at contributing to model discrimination literature for accelerated life test planning when censoring is inevitable.

3.3. Proposed Methodology

3.3.1. Rational for Model Discrimination Methodology

Suppose that our objective is to arrive at an ALT test plan that is capable of discriminating among competing acceleration models. Assume that there are two rival models and it is better that the experimental data can help in choosing one model. Intuitively, a good design should be expected to generate far apart results based on the two competing models, and then the experimenter can select the model based on the actual observations from the experiment. In ALT, the lifetime percentile is typically of interest; therefore the larger the distance (disagreement) in prediction the better our ability to discriminate (distinguish) among these competing models. That is, a good design will amplify the difference among models. Therefore, we propose to use the relative prediction performance of each model over the range of its parameters to identify the optimal design. Figure 3.1 shows how important it is for the experimenter to arrive at the best representative model to reduce prediction errors at use conditions (UCs). For example if $M_1$is the true model, and one selects M_2 then under ALT extrapolation the error in prediction of a quantile of interest at use conditions, $\Delta\hat{\tau}_p(UC)$ is even worse and vice versa if M_2 is the true model and one picks M_1 to proceed with.

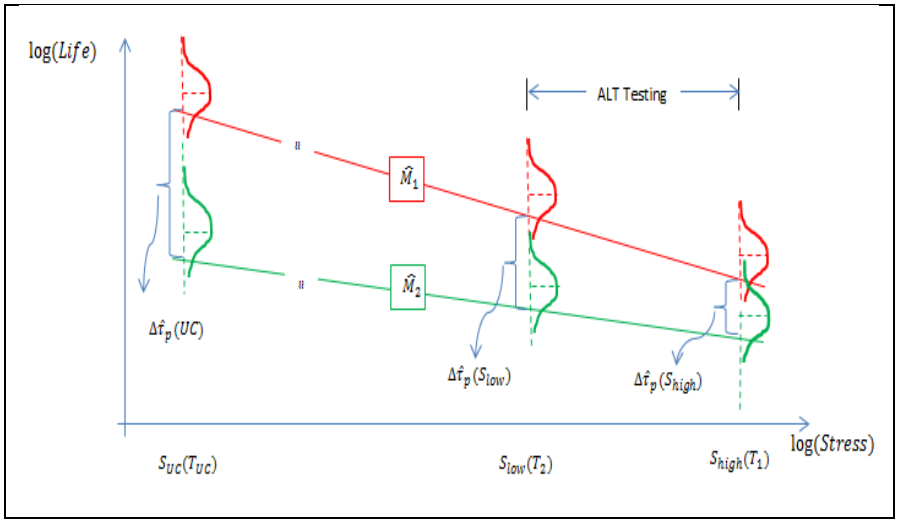


Figure 3.1: $\widehat{M}_1$ Versus $\widehat{M}_2$ at UCs - Importance of Identifying Correct Model

Before discussing our proposed design criterion, we lay the ground for “distance” as measure between probability densities. In the following Section 3.3.2 we introduce the “distance” measure and show our selection of Hellinger distance as a measure of disagreement between predictive densities.

3.3.2. Distance (Divergence) Measure of Probability Distributions

There is a substantial number of distance measures applied in many different fields such as physics, biology, psychology, information theory, etc. See Sung-Hyuk Cha [11] for a comprehensive survey on distance/similarity measures between probability density functions. From the mathematical point of view, distance is defined as a quantitative measure of how far apart two objects are. In statistics and probability theory, a statistical distance quantifies the dissimilarity between two statistical objects, which can be two random variables or two probability distributions. A distance between two populations can be interpreted as measuring the distance between two probability distributions, hence it is essentially the difference of probability measures.

3.3.2.1. Distance Measure

A measure $D(x, y)$ between two points x, y is said to be a distance measure or simply distance if

I. $D(x, y) > 0$ when $x \neq y$ and $D(x, y) = 0$ if and only if $x = y$,

II. $D(x, y) = D(y, x)$,

III. $D(x, y) + D(y, z) \geq D(x, z)$.

(4)

Conditions (I) through (III) imply, respectively, that the distance must be non-negative (positive definite), symmetric and sub-additive (triangle inequality: the distance from point x to z directly must be less than or equal to the distance in reaching point z indirectly through point y. Note that distance $D(x, y)$ is also called $metric$.

The definition of a metric space follows from the definition of distance or metric; i.e., a space X is said to be a metric space if for every pair of points (x, y) in X there is defined a distance $D(x, y)$ satisfying conditions (I) through (III) in (4). Many statistical distances are not metrics, because they lack one or more of the properties of proper metric. For example, pseudo-metrics can violate the "positive definiteness" property; quasi-metrics can violate the "symmetry" property; and semi-metrics can violate the "triangle inequality" property. Some statistical distances are referred to as divergences.

3.3.2.2. Hellinger Distance

The choice of a distance measure depends on the measurement type or representation of quantities under study. In this study, the Hellinger distance $(\mathrm{D_H})$, Deza and Deza [15], is chosen to measure the distance between the two probability distributions that represent the distributions of $\hat{t}_p$ at lower and higher ALT stress test conditions. Computing the distance between two probability distributions can be regarded as the same as computing the Bayes (or minimum misclassification) probability, see Duda, Hart and Stork [16], and Cha and Srihari [12]. This is equivalent to measuring the overlap between two probability distributions as distance, Cha and Srihari [12].

For the discrete probability distributions $P = (p_1 \cdots p_k)$ and $Q = (q_1 \cdots q_k)$, the Hellinger distance (D_H) is defined as:

$$D_H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2} \tag{5}$$

This is directly related to the Euclidean norm of the difference of the square root vectors,

$$D_H(P, Q) = \frac{1}{\sqrt{2}} \left\| \sqrt{P} - \sqrt{Q} \right\|_2 \tag{6}$$

For the continuous probability distributions, the squared Hellinger distance is defined as:

$$D_H^2(P, Q) = \frac{1}{2} \int \left(p_x^{\frac{1}{2}} - q_x^{\frac{1}{2}} \right)^2 dx$$

$$= 1 - \int \sqrt{p_x q_x}\ dx \tag{7}$$

Hellinger distance follows the triangle inequality, that is $0 \leq D_H(P,Q) \leq 1$. The maximum distance of 1 is attained when P assigns probability zero to every set to which Q assigns a positive probability, and vice versa. The Hellinger distance is related to Bhattacharyya coefficient $BC(P,Q)$ as it can be defined as

$$D_H(P,Q) = \sqrt{1 - BC(P,Q)} \tag{8}$$

See Aman Ullah [35].

Some examples of Hellinger distance are given below:

- The squared Hellinger distance between two univariate normal densities $P \sim N(\mu_1, \sigma_1^2)$ and $Q \sim N(\mu_2, \sigma_2^2)$ is given by $D_H^2(P,Q) = 1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}} \exp(-\frac{1}{4}\frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2})$
- The squared Hellinger distance between two exponential densities $P \sim exp(a)$ and $Q \sim exp(b)$ is given by $D_H^2(P,Q) = 1 - \frac{\sqrt[2]{ab}}{a+b}$
- The squared Hellinger distance between two Weibull densities $P \sim Weib(\alpha, a)$ and $Q \sim Weib(\alpha, b)$, where α is a common shape parameter and a, b are the scale parameters respectively, is given by

  $D_H^2(P,Q) = 1 - \frac{2(ab)^{\alpha/2}}{a^\alpha + b^\alpha}$
- The squared Hellinger distance between two Poisson densities with rate parameters a and b, so that $P \sim Pois(a)$ and $Q \sim Pois(b)$ is given by $D_H^2(P,Q) = 1 - \exp(\frac{-1}{2}(\sqrt{a} - \sqrt{b})^2)$

3.3.3. Criterion for Model Discrimination

As stated in Section 3.3.1, we proposed the use of the relative prediction performance of each model over the range of its parameters to identify the optimal design. Therefore, fail/censored data generated by each model is utilized to derive posterior predictions of a quantity

of interest (life percentile) at stress condition $\tau_p(S)$ by all competing models. Then it is followed by pairwise computation of the posterior prediction distance (disagreement) for all models. Maximization of those distances serves as a measure of model distinguishability in that sense. Under the key assumption of linear acceleration in ALT, it is expected that, the distance (disagreement) in prediction at use condition (UC) to be no less than those obtained at stress condition (due to extrapolation errors). Figure 3.3 illustrates proposed evaluation flow. In what follows we formalize the problem in a Bayesian framework of experimental design. In Bayesian framework of experimental design, the problem of optimal design can be thought of as finding a design d^* that maximizes a utility function $U(d)$ that quantifies the objective of the experiment (i.e., model form distinguishability in our case).

3.3.3.1. Criterion Formulation

Suppose that under design d, the experimental outcome may be generated by one of the following two models:

- Model 1, M_1, with its parameter vector θ_1, its outcome denoted by $Y_1 = (y_{11}, \dots, y_{N1})$
- Model 2, M_2, with its parameter vector θ_2, its outcome denoted by $Y_2 = (y_{12}, \dots, y_{N2})$

Consider as an initial utility function to be optimized (maximized), the difference in prediction of life percentile of interest τ_p at the low stress $\tau_p(S_1)$ of the ALT test setup across all pairs of competing models. Ultimately, interest lies in the prediction of the 1st percentile of life distribution at use condition, $\tau_{0.01}$. Since the low stress level is the closest to the use stress level, a large difference in prediction at the low level will give rise to an even larger difference in prediction at the use level (due to extrapolation errors). Therefore, a design is preferable in discrimination sense due to the fact that it causes competing models to predict same quantity of interest differently under the same data set. However, selection of the low stress level to optimize the local utility function may run the risk of not enough fails obtained to sufficiently estimate life distribution percentiles. Therefore, we modify the initially proposed utility function to be optimized,

by considering the simultaneous difference in prediction of life percentile of interest τ_p at the low stress $\tau_p(S_1)$ and high stress $\tau_p(S_2)$ test setup across all pairs of competing models, that is $D_{S_1}(\cdot)$ and $D_{S_2}(\cdot)$ respectively. At the high stress level it is expected to have sufficient fails to properly estimate a life distribution (due to less censoring). In this study, we consider constant-stress ALT plans where it is assumed that there is no interaction between stress variables. It is also assumed that spread in log (life) is constant, that is does not depend on stress, hence parallel lines through τ_p at different stress levels for each model, see Figure 3.2.

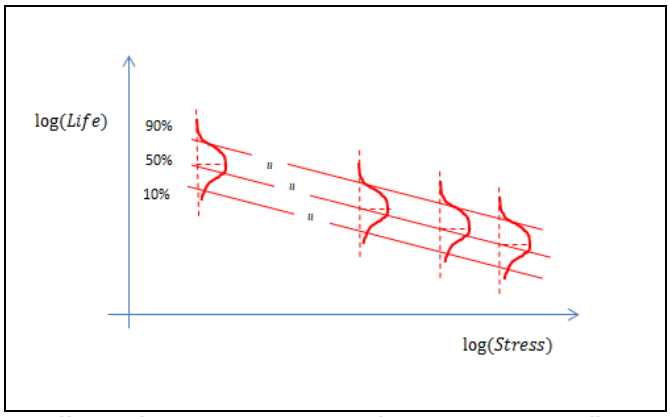


Figure 3.2: Failure Time versus Stress (Constant Spread)

For the demonstration example with two models, M_1 and M_2, the pairwise local utilities are as follows:

$$u_{2|1}\big(d, M_1(\theta_1, Y_1), M_2(\theta_2, Y_1)\big) = D_{S_1}\left(\hat{\tau}_{p,(M2|Y1)}, \hat{\tau}_{p,(M1|Y1)}\right) + D_{S_2}\left(\hat{\tau}_{p,(M2|Y1)}, \hat{\tau}_{p,(M1|Y1)}\right) =$$

$$u_{2|1}$$

(9)

$$u_{1|2}(d, M_1(\theta_1, Y_2), M_2(\theta_2, Y_2)) = D_{S_1}\left(\hat{\tau}_{p,(M1|Y2)}, \hat{\tau}_{p,(M2|Y2)}\right) + D_{S_2}\left(\hat{\tau}_{p,(M1|Y2)}, \hat{\tau}_{p,(M2|Y2)}\right) =$$

$$u_{1|2}$$

(10)

Equation (9) represents the difference in τ_p prediction of model (M_2) conditional on data from model (M_1) relative to model (M_1) prediction of the same quantity, while equation (10) represents the difference in τ_p prediction of model (M_1) conditional on data from model (M_2) relative to

model (M_2) prediction of the same quantity. That is the relative prediction performance of each model over the range of its parameter vector.

At the time of designing an experiment, the experimental outcome is yet to observe, so we do not know which model form and its parameter vector are correct. Therefore,

a) We assess the utility $u_{i|j}(\cdot)$ of a design by its expectation with respect to the sampling distribution of the data $p(y_1|\theta_1, d)$, and $p(y_2|\theta_2, d)$, and the prior distribution of the parameter vectors $\pi(\theta_1)$ and $\pi(\theta_2)$. That is calculating the pre-posterior expectation.

$$E(u_{2|1}) = \iint u_{2|1}\, p(y_1|\theta_1, d)\, \pi(\theta_1|d) dy_1 d\theta_1 \tag{11}$$

$$E(u_{1|2}) = \iint u_{1|2}\, p(y_2|\theta_2, d)\, \pi(\theta_2|d) dy_2 d\theta_2 \tag{12}$$

Equation (11) gives an expression of the expected pre-posterior prediction difference in τ_p of model (M_2) conditional on data from model (M_1) relative to model (M_1) prediction of the same quantity. Similarly, equation (12) gives an expression of the expected pre-posterior prediction difference in τ_p of model (M_1) conditional on data from model (M_2) relative to model (M_2) prediction of the same quantity.

b) Since it is not known which of the two models (M_1) or (M_2) is the true model, we combine a weighted expected utilities $E(u_{i|j})$ to obtain the desired global utility function $U(d)$ to be maximized. The weighing is achieved by priors assigned to the models, $\pi(M_1)$ and $\pi(M_2)$ respectively.

$$\begin{aligned} U(d) &= \sum_{\substack{i,j=1,2 \\ i \neq j}} \pi(M_i).E(u_{i|j}) \\ &= \pi(M_1).E(u_{2|1}) + \pi(M_2).E(u_{1|2}) \end{aligned} \tag{13}$$

Equation (13) can be interpreted as a measure of model distinguishability between two models. The larger the value of $U(d)$, the dissimilar the two models are to each other. Extending (13) to account for situations where more than two models are to be distinguished among is straightforward.

As can be seen from equations (11)-(13), arriving at an optimal design d^* that maximizes (13) is a nontrivial task due to the high dimensional integration and optimization required. There is no closed form solution to (13). Numerical evaluation of the multiple integral for a given choice of design (d) will be needed, which in itself a formidable task given the fact that the integration is defined over the data space and parameter space. The obtained estimate of $U(d)$ must then be maximized over the design variable d, which is in often cases a multidimensional vector. We use a Monte Carlo simulation-based approach to find the optimal design d^*.

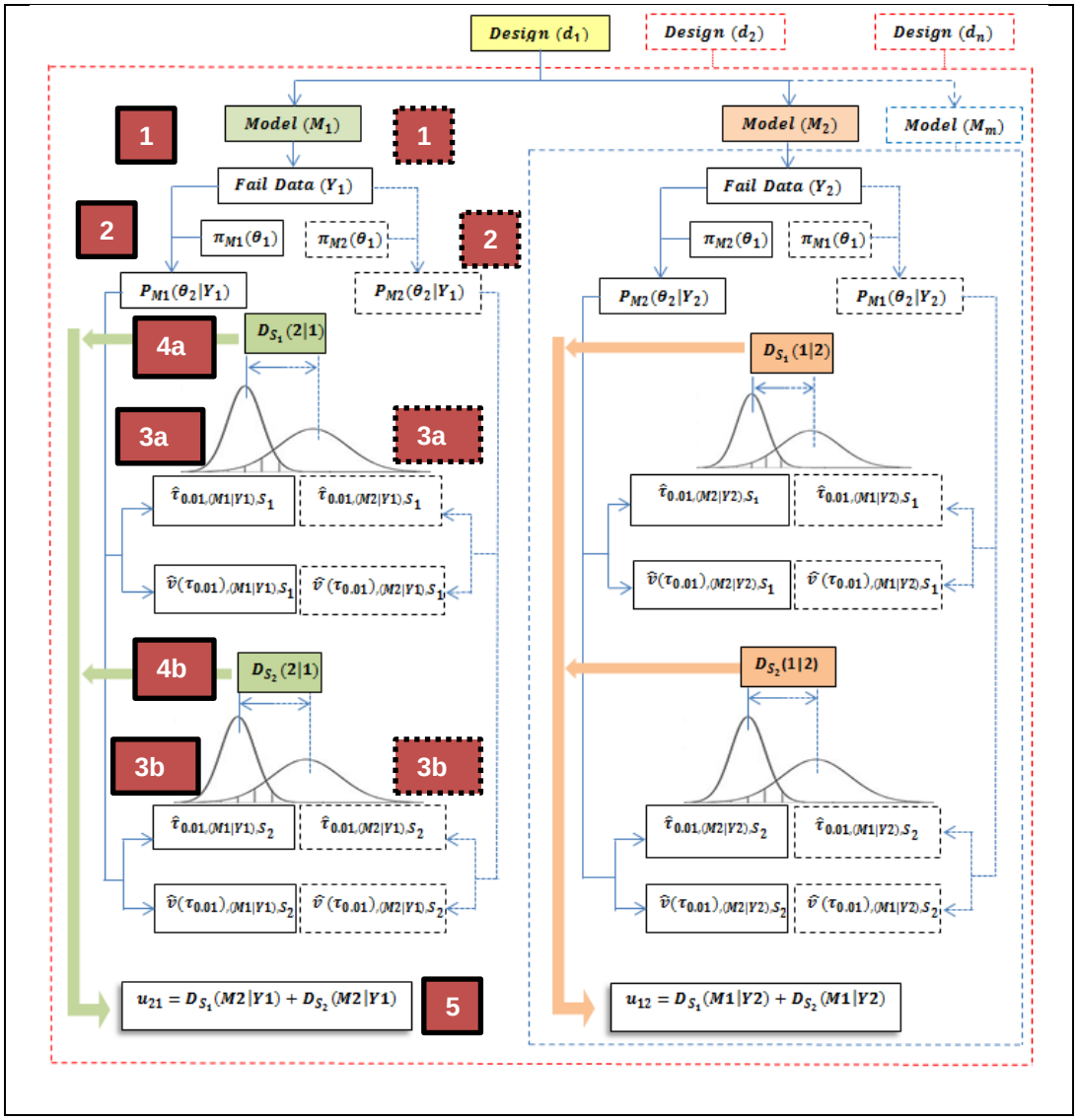


Figure 3.3: High Level Methodology Flow

Figure 3.3 represents a high level flow of the proposed methodology. Fail data Y_1 resulting from acceleration model M_1; step 1, are combined through Bayes' theorem with prior info available on parameters $\pi_{M1}(\theta_1)$ to generate posterior estimates of the parameters given observed data $P_{M1}(\theta_2|Y_1)$; step 2. Then posterior distribution of predictions of life percentile of interest τ_p is obtained using Gibbs sampler at both high and low stress conditions; step 3 (a, b). Same steps are repeated on same data set Y_1 using rival model M_2 (all dashed boxes in Figure 3.3). That gives the relative prediction performance of each model over the range of its parameters conditioned on same data set. The sum of Hellinger distances between prediction distributions are obtained as local utility; step 4 (a, b). This process is repeated for models M_2 through M_m. Local utilities are then weighted by model priors into a global utility to be maximized. More detailed steps of Figure 3.3 are illustrated in section 5.5.

3.4. Model Selection Under Optimal Discriminant Design

In section 3, we outlined the proposed methodology for obtaining an ALT model-discriminant optimal design. In this section, the tools that are used to validate that obtained optimal designs are introduced. It is shown that these designs are indeed optimal under desired optimality criterion as they maximize the proportion of times (recovery rate) in which the true, data-generating, model is selected under an appropriate model selection rule.

3.4.1. Deviance Information Criterion (DIC)

The Deviance information criterion (DIC) was introduced by Spiegelhalter et al. [33] as an easily computed and rather universally applicable Bayesian criterion for posterior predictive model comparison. It compromises between data fit and model complexity, like many other non-Bayesian criteria. It generalizes Akaike's information criterion (AIC) that appears as a special case under a vague prior (negligible prior information), and Bayesian information criterion (BIC), also known as Schwarz criterion. DIC is particularly useful in Bayesian model selection problems

where the posterior distributions of the models have been obtained by Markov chain Monte Carlo (MCMC) simulation. It is an asymptotic approximation as the sample size gets large in a similar behavior to (AIC) and (BIC). It also requires that the posterior distribution be approximately multivariate normal. Claeskens and Hjort (Ch. 3.5) [10] show that the (DIC) is large-sample equivalent to the natural model-robust version of the (AIC).

Define the following

- Deviance as $Dev(\theta) = -2\log[p(y|\theta)] + C$, where y are the data, θ are vector of model unknown parameters, $p(y|\theta)$ is the likelihood function and C is a constant term that cancels out when comparing models.
- Expectation as $\overline{Dev} = E_\theta[Dev(\theta)]$. This measures how well a model fits the data, the larger its value, the worse the fit.
- Effective number of model parameters as $p_D = \overline{Dev} - Dev(\bar{\theta})$, where $\bar{\theta}$ is the expectation of θ. The larger p_D, the easier for the model to fit the data.

Then, the Deviance information criterion (DIC), is defined as a classical estimate of fit, plus twice the effective number of parameters, to give

$$\begin{aligned} DIC &= Dev(\bar{\theta}) + 2p_D \\ &= \overline{Dev} + p_D \end{aligned} \tag{14}$$

3.4.2. Interpretation of Values of (DIC)

When comparing models; models with smaller (DIC) are preferred to models with larger (DIC). Models are penalized both by the value of $\overline{Dev}$, which favors a good fit, but also (in common with AIC and BIC) by the effective number of parameters p_D. Since $\overline{Dev}$ decreases as the number of parameters in a model increases, the p_D term compensates for this effect by favoring models with a smaller number of parameters.

3.4.3. Calculation of (DIC)

The (DIC) is preferable over other criteria in Bayesian model selection due to the fact that the (DIC) can be easily calculated from the samples generated by a Markov chain Monte Carlo simulation. (AIC) and (BIC) would require calculating the likelihood at its maximum over θ, which is not readily available from the MCMC simulation. To calculate (DIC), one computes $\overline{Dev}$ as the average of $Dev(\theta)$ over the samples of θ, and $Dev(\bar{\theta})$ as the value of Dev evaluated at the average of the samples of θ. Then the (DIC) follows directly from these approximations. OpenBUGS or WinBUGS can also be used to calculate the (DIC), from the "Inference" menu simply add a (DIC) monitor (similar to adding monitors for all the other quantities of interest) after burn-in.

3.5. METHODOLOGY ILLUSTRATION

3.5.1. Description of Design Problem

Reliability engineer is interested in studying the intermetallic growth of Au-Al interface in a semi-conductor assembly. It is known that fail mechanism of interest is activated by temperature stress so an accelerated life test is desired in order to estimate the device lifetime. However, there is uncertainty as to whether the relationship between log (life) and the stress (possibly transformed) is linear or exhibit some curvature as indicated by an early look-ahead data set. As a result, current interest lies in an accelerated life test plan that is capable of discriminating between linear and quadratic acceleration models in temperature stress. There are also constraints imposed by available budget for testing (test units), and stress-lab equipment availability and capability as shown below.

- Bake stress chambers are available for 42 days (1,008 hours maximum test time).
- Two types of bake ovens are available with different temperature range capabilities.
  - Lower stress bake oven can be set to run temperature range from 60°C to 115°C.

- Higher stress bake oven can be set to run temperature range from 100°C to 250°C.
- Equipment's tolerance estimated at +/-5°C.

- Experimental budget allows for no more than 20 runs.

The engineer's objective is to determine optimal unit allocation and stress level settings so as to discriminate between the two competing acceleration models.

3.5.2. Competing Acceleration Models

Based on past experience with similar fail mechanism, the reliability engineer believes that the Weibull distribution would adequately describe Au-Al intermetallic growth life in a semiconductor package, which implies a smallest extreme value (SEV) distribution for the log-life. That is, if T is assumed to have a Weibull distribution, $\mathrm{T}\sim\mathrm{WEIB}(\alpha,\beta)$, then $\log(t)\sim SEV(\mu,\sigma)$, where $\sigma=\frac{1}{\beta}$ is the scale parameter and $\mu=\log(\alpha)$ is the location parameter. The Weibull CDF and PDF can be written as

$$F(t|\alpha,\beta)=\Phi_{SEV}\left(\frac{\log(t)-\mu}{\sigma}\right)=1-exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right] \quad (15)$$

$$f(t|\alpha,\beta)=\frac{1}{\sigma t}\Phi_{SEV}\left(\frac{\log(t)-\mu}{\sigma}\right)=\frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1}exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right],\ t>0 \quad (16)$$

In above parameterization, $\beta>0$ is the shape parameter and $\alpha>0$ is the scale parameter as well as the 0.632 quantile.

The Arrhenius Life-Temperature relationship was expected to describe the acceleration behavior.

$$t(Temp)=A\,.\exp\left(\frac{E_a}{K\times Temp}\right), \quad (17)$$

where,

- $t(Temp)$ is the life characteristic related to temperature.
- A is a constant, and (E_a) is the activation energy of the chemical reaction in electron volts.

- $Temp$ is temperature in Kelvin (^{o}C+273.15).
- K is Boltzmann's constant $(8.617385\ E^{-5}\ eV/K)$

Reliability engineer would like to consider two life-stress relationships to discriminate between experimentally; namely the linear relationship M_1 and the quadratic relationship M_2.

M_1 model can be expressed in linearized form by taking the logarithmic of both sides as

$$\mu_1 = \beta_0 + \beta_1 x, \tag{18}$$

By standardizing the accelerating variable, the above model (18) can be expressed as

$$\mu_1 = \gamma_0 + \gamma_1 \xi\ , \tag{19}$$

where the standardized variables are expressed as

$$\xi = \frac{(x - x_{low})}{(x_{high} - x_{low})}, \qquad \xi \in [0, 1] \tag{20}$$

New coefficients are related to previous ones through

$$\gamma_0 = \beta_0 + \beta_1 x_{low},$$

$$\gamma_1 = \beta_1 (x_{high} - x_{low}) \tag{21}$$

At $x = x_{low}, \xi = 0$, and at $x = x_{high}, \xi = 1$. Thus,

$$\mu_{1\,low} = \gamma_0$$

$$\mu_{1\,high} = \gamma_0 + \gamma_1 \tag{22}$$

M_2 model can be expressed in linearized form by taking the logarithmic of both sides as

$$\mu_2 = \beta_0 + \beta_1 x + \beta_2 x^2, \tag{23}$$

By standardizing the accelerating variable, the above model (23) can be expressed as

$$\mu_2 = \gamma_0 + \gamma_1 \xi + \gamma_2 \xi^2 , \tag{24}$$

where

$$\gamma_0 = \beta_0 + \beta_1 x_{low} + \beta_2 x^2{}_{low},$$

$$\gamma_1 = \beta_1 (x_{high} - x_{low})$$

$$\gamma_2 = \beta_2 (x^2{}_{high} - x^2{}_{low}) \tag{25}$$

At $x = x_{low}, \xi = 0$, and at $x = x_{high}, \xi = 1$. Thus,

$$\mu_{2\,low} = \gamma_0$$

$$\mu_{2\,high} = \gamma_0 + \gamma_1 + \gamma_2 \tag{26}$$

For both models, for Type-I censored data (time censoring), the probability of obtaining a censored observation at time t_c is given by

$$\Pr(\mathrm{t} > \mathrm{t}_c) = exp\left[-\left(\frac{t_c}{\alpha}\right)^{\beta}\right], \qquad t_c > 0 \tag{27}$$

3.5.3. Design Criterion

The optimization criterion and its formulation were discussed in detail in section 3.3 and its subsections. Criterion summarized in equation (13), reproduced below for the two models under consideration (M_1) or (M_2) with model priors; $\pi(M_1)$ and $\pi(M_2)$ respectively.

$$U(d) = \sum_{\substack{i,j=1,2 \\ i \neq j}} \pi(M_i).E(u_{i|j})$$

$$= \pi(M_1).E(u_{2|1}) + \pi(M_2).E(u_{1|2})$$

3.5.4. Prior Distributions Elicitation

Engineer assumed an equal weight for both models to begin with. That is, $\pi(M_1) = \pi(M_2) = 0.5$ (or 50%). For model M_1, equation (19) shows parameter vector θ_1 as$(\gamma_0, \gamma_1, \sigma)^T$, and for model M_2, equation (24) shows parameter vector θ_2 as$(\gamma_0, \gamma_1, \gamma_2, \sigma)^T$. One would need to

specify a prior distribution for each of the parameters or $p_{M1}(\theta)$ and $p_{M2}(\theta)$ .We would initially use the parameters in their original units (before transformation) to relate to the engineer's prior knowledge. Standardization is applied once prior distributions in original units have been effectively solicited from engineers.

Given historical learning and previous experience with similar fail mechanism, the reliability engineer believes that appropriate independent prior distributions on the parameters can be specified as follows: for the activation energy, a uniform distribution that gives an equal likelihood for values that range from 1.0 to 1.05 eV would be appropriate to use. Note that in the case of the quadratic model M_2 this parameter may no longer directly correspond to the activation energy of the chemical reaction. Not much was known about the intercept, and the quadratic coefficient in M_2 so both were given a vague (diffuse) normal distribution with mean of 0.0 and low precision of $1.0E^{-6}$ ($\sigma = 1000\ or\ \sigma^2 = 1E^{+6}$). A positive density support was assumed for the Weibull shape parameter as gamma distribution with shape of 2 and scale of 1.

3.5.5. Construction of Optimal Design

Our optimization algorithm is Monte Carlo simulation-based in which the optimal design d^* is arrived at by evaluating the design criterion in (13) for each of the candidate designs, and selecting the design that maximizes the design criterion (utility function of interest). We summarize the algorithm steps as follows:

1. For a given experimental run budget (N), and number of stress-factors to study (k), construct a Latin hypercube $LHD\ (N, k)$ design at each stress level. Then generate a modified $mLHD$ per Nasir and Pan [29] to create the finalized design grid.
2. Over the finalized design grid, for each design $d_i\ (i = 1, \dots, d)$ randomly simulate fail data from the joint density $(\theta, y)_{d_i, M_i}$ of each of the rival models $M_i\ (i = 1, \dots, m)$.

$$(\theta, y)_{d_i, M_i} \sim p_{d_i, M_i}(\theta, Y) \sim p(\theta)_{M_i}.p_{d_i, M_i}(y|\theta) \tag{28}$$

That is, independently generate random fail data using the competing acceleration models (using equation (19) for model M_1 and equation (24) for model M_2). Consider all possible combinations of sample sizes (unit allocation) at each stress factor-level combinations.

Computational time can be reduced if units are allocated at increments >1 to each of the stress levels.

3. Simulated experiments (failure times) are compared against a predetermined test duration t_c to determine if a test unit failure time is censored.
4. Calculate the relative prediction performance of each model over the range of its parameters. This is done by using Gibbs sampler (WinBUGS) to compute posterior predictions of, $\tau_p(x_S)$, the 100 p^{th} quantile of the lifetime distribution at both high and low stress conditions ($S = S_{High}, and\ S = S_{Low}$). That is within experimental region where data are observed. A typical reliability interest is when $p = 0.01$, so in the case of models M_1 and M_2, the outcome of this step is the posterior distribution of predicted values for each model given same data set.

    4.1. For Y_1 (fail data generated under model M_1) at $S = S_{High}$

    $$\hat{\tau}_{0.01,(M1|Y1),S_{High}} \text{ , and } Var\ (\hat{\tau}_{0.01})_{(M1|Y1),S_{High}} \quad (29)$$

    $$\hat{\tau}_{0.01,(M2|Y1),S_{High}} \text{ , and } Var\ (\hat{\tau}_{0.01})_{(M2|Y1),S_{High}} \quad (30)$$

    4.2. For Y_1 (fail data generated under model M_1) at $S = S_{Low}$

    $$\hat{\tau}_{0.01,(M1|Y1),S_{Low}} \text{ , and } Var\ (\hat{\tau}_{0.01})_{(M1|Y1),S_{Low}} \quad (31)$$

    $$\hat{\tau}_{0.01,(M2|Y1),S_{Low}} \text{ , and } Var\ (\hat{\tau}_{0.01})_{(M2|Y1),S_{Low}} \quad (32)$$

    4.3. For Y_2 (fail data generated under model M_2) at $S = S_{High}$

    $$\hat{\tau}_{0.01,(M2|Y2),S_{High}} \text{ , and } Var\ (\hat{\tau}_{0.01})_{(M2|Y2),S_{High}} \quad (33)$$

    $$\hat{\tau}_{0.01,(M1|Y2),S_{High}} \text{ , and } Var\ (\hat{\tau}_{0.01})_{(M1|Y2),S_{High}} \quad (34)$$

    4.4. For Y_2 (fail data generated under model M_2) at $S = S_{Low}$

    $$\hat{\tau}_{0.01,(M2|Y2),S_{Low}} \text{ , and } Var\ (\hat{\tau}_{0.01})_{(M2|Y2),S_{Low}} \quad (35)$$

    $$\hat{\tau}_{0.01,(M1|Y2),S_{Low}} \text{ , and } Var\ (\hat{\tau}_{0.01})_{(M1|Y2),S_{Low}} \quad (36)$$

5. Use Hellinger distance measure (D_H), or any appropriate distance measure between probability distributions for that matter, to calculate pairwise local utilities $(u_{2|1})$ and $(u_{1|2})$ as in (9) and (10), reproduced below for convenience

5.1. For model M_2 conditional on data from model M_1

$$u_{2|1} = D_{H_{S_{High}}}\left(\hat{\tau}_{0.01,(M2|Y1)}, \hat{\tau}_{0.01,(M1|Y1)}\right) + D_{H_{S_{Low}}}\left(\hat{\tau}_{0.01,(M2|Y1)}, \hat{\tau}_{0.01,(M1|Y1)}\right)$$

5.2. For model M_1 conditional on data from model M_2

$$u_{1|2} = D_{H_{S_{High}}}\left(\hat{\tau}_{0.01,(M1|Y2)}, \hat{\tau}_{0.01,(M2|Y2)}\right) + D_{H_{S_{Low}}}\left(\hat{\tau}_{0.01,(M1|Y2)}, \hat{\tau}_{0.01,(M2|Y2)}\right)$$

6. Since it is unknown which of the two models is the true data generating model, we combine the Monte Carlo samples in local utilities $u_{2|1}$ and $u_{1|2}$ to obtain the desired total observed utility function $u(d) = u_{2|1} + u_{1|2}$ to be maximized for an optimal design.
7. Approximate pre-posterior global utility $U(d) = E[u(d)]$ by fitting a smooth surface to the combined Monte Carlo sample generated in step (6) as a function of selected design.
8. The optimal design d^* is found by maximizing the fitted surface (maximum pre-posterior Hellinger distance between predictive densities).

Since the expected utility $U(d)$ surface is generally of continuous nature, the direct application of the Monte Carlo simulation, in step (7) of the algorithm will require large scale simulations to be applied, and will only be computationally inefficient due to the large number of iterations needed and duplication of effort in neglecting valuable information already generated at a nearby design points. That is, repeated simulations at close by points on the design grid. Therefore, to reduce computational cost, step (8) of the algorithm utilizes the non-parametric surface fitting approach originally proposed by Muller and Parmigiani [28] and Muller [27] for finding optimal designs.

3.5.6. Results for Discriminating Linear vs. Quadratic ALT Models

Table 3.1 lists the temperature stress ranges that were used in the planning of the ALT experiment. The surface fitting smoothing approach for finding optimal design requires simulation of experiments (d_i, θ_i, y_i) on a design grid. Full grid of the three temperature ranges can be used in the simulation. However, we instead use a modified Latin Hypercube design to replace the full grid and reduce computational cost at no loss of coverage and to allow available experimental

budget. Table 3.2 shows the design grid created using a modified Latin Hypercube $mLHD$ for the available budget of 20 experimental runs.

Bake Stress	Temperature Range in °C (Oven tolerance $\pm$ 5 °C)	
	Lower	Upper
T_{High}	150	180
T_{Middle}	115	145
T_{Low}	80	110

Table 3.1: Temperature Stress Range Used in Experiment

Run #	Low Temp Oven Setup	High Temp Oven Setup		Run Source
	Temp°C (low)	Temp°C (Mid)	Temp°C (High)	
1	85	140	155	mLHD
2	80	125	150	mLHD
3	90	115	170	mLHD
4	100	120	160	mLHD
5	95	145	165	mLHD
6	100	130	180	mLHD
7	110	135	160	mLHD
8	95	130	165	mLHD
9	110	120	175	mLHD
10	90	125	150	mLHD
11	105	135	175	mLHD
12	80	140	170	mLHD
13	80	115	150	AUG-C1
14	110	145	180	AUG-C2
15	110	115	150	AUG-C3
16	110	115	180	AUG-C4
17	80	115	180	AUG-C5
18	80	145	150	AUG-C6
19	80	145	180	AUG-C7
20	110	145	150	AUG-C8

Table 3.2: mLHD Grid with 12 Runs and 8 Corner Augmentations

Following simulation algorithm steps outlined in Section 3.5.5, the optimal design under criterion (13) for discriminating between linear and quadratic acceleration models in single accelerating variable (temperature), and under practical constraints outlined in Section 3.5.1 is summarized in Figures 3.4 through 3.6.

Figure 3.4 displays the pre-posterior expected value of the utility function ($Log[U(d)]$) as a function of the stress magnitude and percent unit allocation to each of the three stress levels used in planning of the experiment. The utility function is maximized when

1. Higher temperature level is set at the highest value $(180°C)$ of its range$(150°C - 180°C)$, with approximate unit allocation of 12%.
2. Middle temperature level is set at the intermediate value $(130°C)$ of its range $(115°C - 145°C)$, with approximate unit allocation of 55%.
3. Lower temperature level is set at value $(100°C)$, slightly above the intermediate value of its range $(80°C - 110°C)$, with approximate unit allocation of 33%.

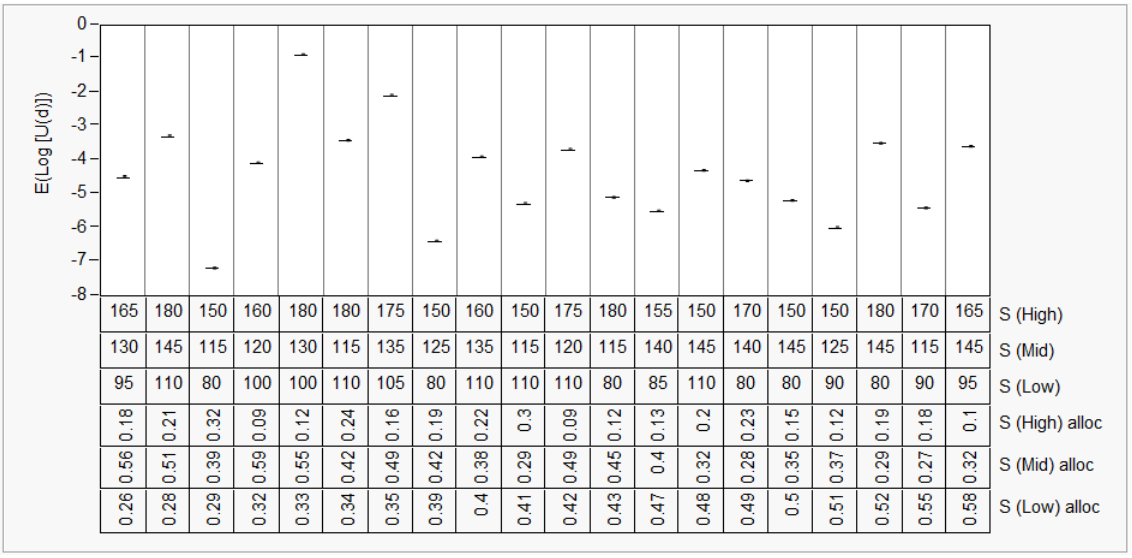
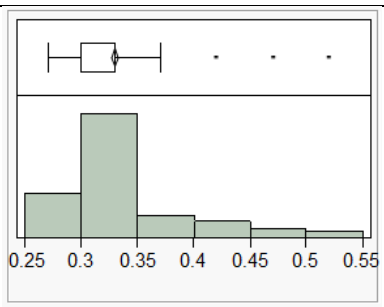


Figure 3.4: Pre-posterior Expected Log [U(d)] as a Function of Stress and Unit Allocation

Figures 3.5 and 3.6 display, respectively, the distributions for unit allocation and stress levels at simulated optimal designs from 1500 simulation runs. Approximate mean values are

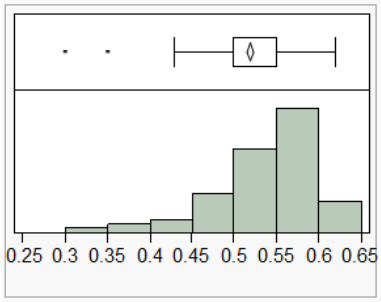
ones considered in Figure 3.4. Some designs were close to optimal with slightly different unit allocation and stress levels setup as shown in Figure 3.4.



Mean ≈ 0.330 (33%)

Moments	
Mean	0.3301667
Std Dev	0.0568841
Std Err Mean	0.0014687
Upper 95% Mean	0.3330477
Lower 95% Mean	0.3272857
N	1500

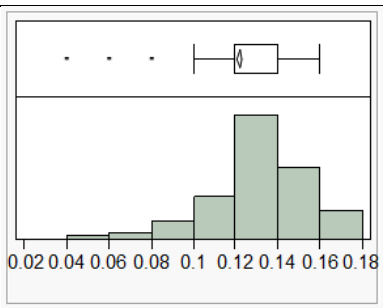
Unit allocation at Lower Stress



Mean ≈ 0.519 (52%)

Moments	
Mean	0.5186
Std Dev	0.0601979
Std Err Mean	0.0015543
Upper 95% Mean	0.5216488
Lower 95% Mean	0.5155512
N	1500

Unit allocation at Middle Stress

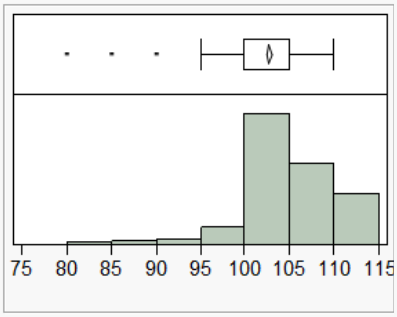


Mean ≈ 0.122 (12%)

Moments	
Mean	0.1216667
Std Dev	0.022662
Std Err Mean	0.0005851
Upper 95% Mean	0.1228144
Lower 95% Mean	0.1205189
N	1500

Unit allocation at Higher Stress

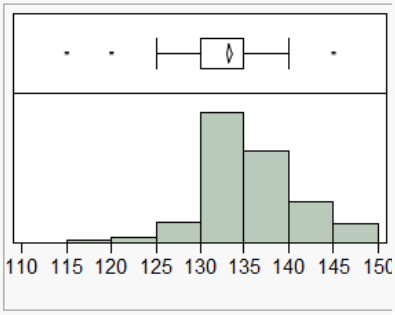
Figure 3.5: Distributions for Unit Allocation at Optimal Design



Mean ≈ 102.7 (°C)

Moments	
Mean	102.70667
Std Dev	4.6319462
Std Err Mean	0.1195963
Upper 95% Mean	102.94126
Lower 95% Mean	102.47207
N	1500

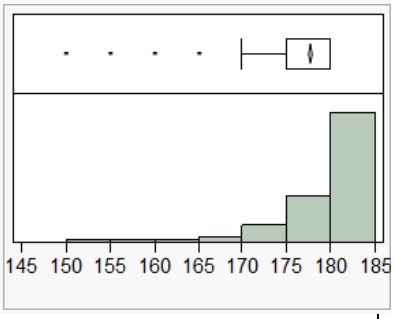
Lower Stress Setup



Mean ≈ 133.2 (°C)

Moments	
Mean	133.24667
Std Dev	5.0803243
Std Err Mean	0.1311734
Upper 95% Mean	133.50397
Lower 95% Mean	132.98936
N	1500

Middle Stress Setup



Mean ≈ 177.6 (°C)

Moments	
Mean	177.60667
Std Dev	3.9220161
Std Err Mean	0.101266
Upper 95% Mean	177.8053
Lower 95% Mean	177.40803
N	1500

High Stress Setup

Figure 3.6: Distributions for Stress Levels at Optimal Design

3.5.6.1. Some Remarks on the Obtained Optimal Discriminant Test Plan

The following remarks are drawn in comparison to the well-known “Optimum ALT Plans” (use two test stress levels, no intermediate level, with unequal numbers of test units; more allocation towards lower level), and the more practical “Good Compromise ALT Plans” (use three or four stress levels with unequal allocation of test units; more allocation towards lower level). Although the primary objective of these plans (estimation accuracy in most), is quite different than ours (model discrimination), pointing out similarities and dissimilarities between the two groups of plans is of an added value in our judgment; Nelson [30] has once said “a good plan should be multi-purpose and robust and provide accurate estimates.”

1. The test plan allocates the larger proportion of units to the intermediate stress level (~55%). This is favorable for robustness and early failures and will be most sensitive for detecting nonlinearity of the relationship (minimize variance of the estimate of the quadratic coefficient).
2. The test plan allocates more test units to the lower stress level (~33%) than to the higher stress level (~12%). This is favorable for more accurate extrapolation with respect to stress, as suggested by optimum plans.
3. The test plan sets the high temperature value to the highest possible in its allowable range, this is known to be a good practice when interest lies in minimizing the standard error of the estimate of any percentile at the design stress (a very common objective of ALTs).
4. The test plan does not set the lower temperature value to the lowest possible in its allowable range as suggested by the optimum test plan (effective if the design stress is close to the test range), but rather choses an intermediate value. One drawback to having to test at the lowest extreme of the test range is the longer test time needed for units to fail.
5. In comparison to Meeker-Hahn [33] (4:2:1) plans that use unequal allocation (close to optimum for short extrapolation), the test plan that is as close to the (4:2:1) allocation as possible (3.85:1.89:1.26) ranks low with respected to our discrimination criterion. From Figure 3.4, it sets [stress, allocation] pairs at [90°C, 55%] for high stress, [115°C, 27%] for intermediate stress, and [170°C, 18%] for lower stress.

3.5.6.2. Recovery Rate for the Obtained Optimal Discriminant Test Plan

Optimal model discriminant designs are expected to maximize the proportion of times in which the true, data-generating, model is selected under an appropriate model selection criterion. We have chosen to use the (DIC) model selection rule as explained in details in section 3.4. Other methods of model selection such as BF (Bayes Factor) or BIC (Bayesian Information Criterion) could have also used. The following definitions were used in creating plan comparison in Figure 3.7 as function of total sample size across all three-stress levels.

- True Model = acceleration model from which data was simulated. That is, equation (19) for the linear model, and equation (24) for the quadratic model.
- Assumed Model = actual acceleration model fitted to the simulated data.
- % Recovery Rate = fraction of times the true model recovered (correctly identified) under DIC-based model selection under optimal stress setup and unit allocation per each plan.

Assumption used in the optimal stress setup and unit allocation for each plan:

- Same prior distributions given to same parameters across all models.
- All plans used three levels of stress (temperature) in the range of $(150°\mathrm{C} - 180°\mathrm{C})$ for high temp, $(115°\mathrm{C} - 145°\mathrm{C})$ for middle temp, and $(80°\mathrm{C} - 110°\mathrm{C})$ for low temperature stress. All plans share the same fixed experimental budget (sample size).
- Stress setup and unit allocation were determined as follows
    - **Model discriminant plan**: *unequally* spaced test levels with *unequal* allocation that puts more units at the middle of the test range. Optimal design setup used: highest temp of $(180°\mathrm{C})$ with 12% allocation, intermediate temp of $(130°\mathrm{C})$ with 55% allocation, and lower temperature of $(100°\mathrm{C})$, slightly above the intermediate value in the low temp range, with 33% allocation.
    - **Good compromise plan**: *equally* spaced test levels with *unequal* allocation that puts more units at the extremes of the test range and fewer in the middle. We've used 50% at lower level, 30% at higher level, and remaining 20% at the middle

level. For the equal spacing of stress levels $180°C$ was selected as highest possible, $110°C$ as lowest, and $145°C$ as the intermediate stress.

- **Best traditional plan**: *equally* spaced test levels with *equal* allocation. Typically, highest stress needs to be specified. We selected, highest possible of $180°C$, while lowest stress is selected to minimize std. error of ML estimate of log mean life at design stress. We arbitrary select lowest possible of $110°C$ without optimization, thus setting the intermediate stress at an equal space of $145°C$. Equal allocation puts approximately 33.33% of units at each stress level.
- **Meeker-Hahn [24] (4:2:1) compromise plan**: High stress typically specified from practical constraint, we assumed it to be at $180°C$. Low stress is chosen to minimize asymptotic variance of MLE of a life percentile of log life at design stress, we arbitrary select lowest possible of $110°C$ without optimization. Middle stress is set at midway between the others, that is $145°C$. Allocation of samples follows $\frac{4}{7}$ or (57%) to low stress, $\frac{2}{7}$ or (29%) to middle stress, and $\frac{1}{7}$ or (14%) to high stress.

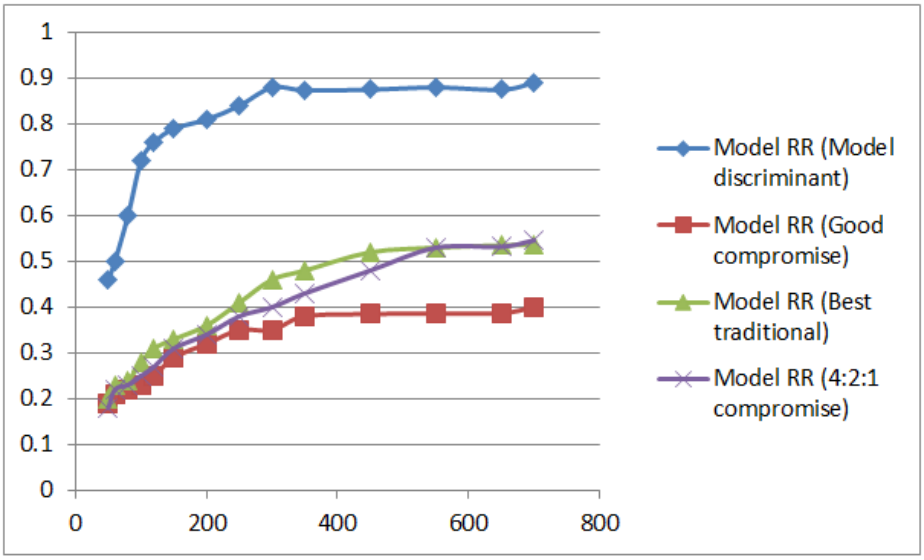


Figure 3.7: Recovery Rate versus Sample Size Comparison across Test Plans

As noted in Section 3.5.6.1 the primary objective/s of the test plans we're comparing our plan to are parameter estimation and prediction accuracy of a quantity of interest at design stress, which is different than our test plan's objective (model-form discrimination). Therefore, the apparent superiority w.r.t. to the recovery rate under DIC model selection of our test plan as shown in figure 3.7 should come as no surprise, and should not be interpreted as overall superiority as objectives are different, by comparison across the different plans we intended to compare the effect of recommended stress setup and unit allocation for these plans on model discrimination. It appears that optimizing for one objective will take its toll on another. Model-discriminant plan tends to allocate higher percentage of units to the middle stress. That is intuitively appealing for robustness and early failures.

3.6. Conclusion and Areas for Future Research

In this study, we have presented a simulation-based Latin Hypercube Bayesian accelerated life test planning (ALT) method with the objective of discrimination between competing acceleration model forms. Our proposed criterion is based on the Hellinger distance measure between predictive distributions. We applied the criterion to accelerated life test planning in which interest lies in a test plan that is capable of differentiating between linear and quadratic acceleration models in one-accelerating variable (temperature) when there is uncertainty as to whether the relationship between log mean (life) and the stress (possibly transformed) is linear or exhibit some curvature. We used the criterion to determine optimal stress-factor setup and unit allocation at three stress levels subject to test-lab equipment and time constraints. Optimal designs were defined as those that maximize the proportion of times (recovery rate) in which the true, data-generating, model is selected under the DIC (Deviance Information Criterion) model selection rule. We compared our optimal test plan with that of the typically used three stress-levels good compromise, best traditional and well known 4:2:1 compromise ALT test plans. Results showed that our obtained optimal design has the advantage of substantially increasing the test plan ability to distinguish among competing model forms. Thus

providing better guidance as to which model is appropriate for the follow-on testing phase in the experiment.

As with many of the complex MCMC problems, the main limitation is typically the computationally intensive calculations and the need for point-wise evaluation of utility function. This has been eased by the use a modified Latin Hypercube sampling scheme reinforced with design point augmentation, and followed by the application of curve-fitting optimization approach. The simulation-based Bayesian approach described in this paper could be extended to model-discriminant ALT planning problems with more than one accelerating variable and more complicated acceleration models.

References

[1] V. Agboto, W. Li, and C. Nachtsheim, Screening designs for model discrimination, Journal of Statistical Planning and Inference, vol. **140**, no. 3 (2010), pp. 766-780.

[2] A. C. Atkinson, and D. R. Cox, Planning experiments for discriminating between models, J. Roy. Statist. Soc. Ser. B **36** (1974), 321-348.

[3] A. C. Atkinson, and A. N. Donev, Optimum experimental designs, Oxford, England: Oxford University Press.

[4] A. C. Atkinson, and V. V. Fedorov, The design of experiments for discriminating between two rival models, Biometrika **62** (1975), 1, pp. 57-70.

[5] A. C. Atkinson, and V. V. Fedorov, Optimal design: Experiments for discriminating between several models, Biometrika **62** (1975), 2, pp. 289-303.

[6] D. S. Bai, and S. W. Chung, An optimal design of accelerated life test for Exponential distribution, Reliability Engineering and System Safety **31** (1991), 57-64.

[7] D. S. Bai, S. W. Chung, and Y. R. Chun, Optimal design of partially accelerated life tests for the Lognormal distribution under type I censoring, Reliability Engineering and System Safety **40** (1993), 85-92.

[8] D. R. Bingham, and H. A. Chipman, Optimal designs for model selection, Technical Report **388**, University of Michigan and University of Waterloo (2002).

[9] G. E. P. Box, W. J. Hill, Discrimination among mechanistic models, Technometrics **9** (1967), 57-71.

[10] G. Claeskens, and N. L. Hjort, Model Selection and Model Averaging, Cambridge. (2008) Section 3.5.

[11] Sung-Hyuk Cha, Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions, International Journal of Mathematical Models and Methods in Applied Sciences **1** (4) (2007), 300-307.

[12] Sung-Hyuk Cha, and S. N. Srihari , On measuring the distance between two histograms, The Journal of the Pattern Recognition Society, **35** (2002), 1355-1370.

[13] K. Chaloner and I. Verdinelli. Bayesian Experimental Design: A Review. Statistical Science, **10** (1995), 273-304.

[14] H. Dette, and S. Titoff, Optimal discrimination designs, The Annals of Statistics, vol. **37** (2009), no. 4, pp. 2056-2082.

[15] E. Deza, and M. M. Deza, Dictionary of Distances, Elsevier, (2006).

[16] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2nd ed. Wiley, (2001).

[17] P. D. H. Hill, A review of experimental design procedures for regression model discrimination, Technometrics **20** (1978), 1, pp 15-21.

[18] W. J. Hill, W. G. Hunter, D. W. Wichern, A joint design criterion for the dual problem of model discrimination and parameter estimation, Technometrics **10** (1968), 145-160.

[19] W. G. Hunter, A. M. Reiner, Designs for discriminating between two rival models, Technometrics **7** (1965), 307-323.

[20] T. J. Kielpinski and W. Nelson, Optimum censored accelerated life tests for normal and lognormal life distributions, IEEE Trans. Reliability **R-24**, 5 (1975), 310-320.

[21] P. D. Leon and A. C. Atkinson, Optimum experimental design for discriminating between two rival models in the presence of prior information, Biometrika **78** (1991), 601-608.

[22] W. Q. Meeker, A comparison of accelerated life test plans for Weibull and Lognormal distributions and type I censored data, Technometrics **26** (1984), 157–172.

[23] W. Q. Meeker, and L. A. Escobar, A Review of Recent Research and Current Issues in Accelerated Testing, International Statistical Review **61,** 1, (1993), 147-168.

[24] W. Q. Meeker, and G. J. Hahn, How to Plan an Accelerated Life Test: Some Practical Guidelines, American Society for Quality Control, Milwaukee, WI, (1985).

[25] W. Q. Meeker, and W. Nelson, Optimum accelerated life tests for the Weibull and extreme value distributions. IEEE Trans. Reliability **R-24**, 5 (1975), 321-332.

[26] R. D. Meyer, D. S. Steinberg, G. E. P. Box, Follow-up designs to resolve the confounding in multifactor experiments, Technometrics **38** (1996), 303-313.

[27] P. Müller, Simulation-Based Optimal Design, in Bayesian Statistics 6, eds. J. M. Bernardo, J. O. Berger, A. P. David, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, (2000) pp. 459–474.

[28] P. Müller, and G. Parmigiani, Optimal Design via Curve Fitting of Monte Carlo Experiments, Journal of the American Statistical Association **90** (1995), 1322–1330.

[29] E. Nasir, R. Pan, Simulation-Based Bayesian Optimal Design for Multi-factor Accelerated Life Tests, Communication in Statistics – Simulation and Computation (Submitted) (2013).

[30] W. Nelson, Accelerated Testing, John Wiley & Sons: New York, (1990).

[31] W. Nelson, and T. J. Kielpinski, Theory for optimum censored accelerated life tests for normal and lognormal life distributions, Technometrics **18** 1 (1976) 105-114.

[32] W. Nelson, and W. Q. Meeker, Theory for optimum accelerated censored life tests for Weibull and extreme value distributions, Technometrics **20**, 2 (1978) 171-177.

[33] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde, Bayesian measures of model complexity and fit, J. R. Statist. Soc. B **64** (2002) P4, pp. 583-639.

[34] D. Ucinski and B. Bogacka, T-optimum design for discrimination between two multi-response dynamic models, Journal of the Royal Statistical Society: Statistical Methodology **67**(B) (2005), 3-18.

[35] A. Ullah, Entropy, divergence and distance measures with econometric applications, Journal of Statistical Planning and Inference **49** (1996), 137-162.

CHAPTER 4

A SEQUENTIAL BAYESIAN APPROACH TO MODEL DISCRIMINATION IN ACCELERATED LIFE TEST PLANNING

Ehab Nasir[1], Rong Pan[2]

[1] Industrial Engineer, Intel Corporation, Chandler, Arizona.

[2] Associate Professor, School of Computing, Informatics, and Decision Systems Engineering, Arizona State University.

Abstract

In planning accelerated life tests (ALTs), very often there exist uncertainty regarding the form of the model that describes the relationship between the parameters of the life distribution of units under test and the, possibly transformed, stress. In this paper, we propose a sequential Bayesian model-discriminant scheme. The idea is based on an initial experiment planned and conducted at an optimal setup identified under a model-discriminant design criterion; the Hellinger distance measure between predictive densities, to provide the maximum possible information with respect to model discrimination while using less experimental budget. If a "winning" model could not be identified at the required statistical significance level, then a subsequent model-discriminant experimentation is planned and conducted while budget allows. The subsequent test will leverage the most current information to allow for Bayesian model comparison through posterior model probabilities and their corresponding ratios in what is known as Bayes factor. Sequential testing is terminated upon exhaustion of available experimental budget and/or when strong evidence in favor of one model over another is demonstrated through Bayes factor. In our demonstration example we consider the case for single variable constant-stress accelerated life test at three levels where a sequential Bayesian framework is proposed to optimally discriminate between linear and quadratic model forms. Advantage of proposed methodology as compared to non-sequential testing under similar model-discriminant design criterion is evaluated through model recovery rate under the Bayes factor model selection rule. Results showed that

performance of sequential model discriminant in ALT is adversely impacted by the amount of censoring in the data, in the case of linear vs. quadratic model form with testing at three levels of constant stress, sequential testing can improve model recovery rate by approximately 8% in the case of complete data. In the case of censoring, two buckets were considered for right censored data; namely (30-40%) and (50-60%), and results showed that both testing schemes suffered in their ability to discriminate between models and there was no apparent advantage in adopting sequential testing. This finding is troublesome as majority of ALT data are characterized by censoring and complete data are rare.

Key Words - Sequential Accelerated Life Test Design, Bayesian Design, Model Discrimination, Hellinger distance, Bayes Factor.

4.1. Introduction

Accelerated life tests (ALTs) are widely used throughout industry to gather valuable information on the life distribution of material of construction and product performance at use conditions (design stress) in relatively short duration of testing time. In an ALT, experimental units or prototypes are subjected to higher than normal operating conditions (stresses) to induce early failures that can be used to make inference about life distributional quantities of interest at normal operation conditions. In planning accelerated life tests (ALTs), very often there exist uncertainty in the form of the model used to describe the relationship between the parameters of the life distribution and applied stress. Moreover, in a typical ALT planning, initial values of the unknown model parameters are specified as “best-guessed at” values so as to obtain a locally optimum test plan under an unverified assumption of correctness of model being used. Very often there would be a miss-specification error of model parameters, and when combined with miss-specification error of model being used, a high enough margin of combined error would result in preventing the test plan from achieving its desired statistical efficiency. Therefore, an experiment designed with the objective of model discrimination between competing models will be of great value in providing information about appropriateness of model being used. Better yet, if this

information can be obtained in the most efficient way at the lowest possible cost. With that in mind, we investigate the benefit of sequential testing in ALT model discrimination. In sequential testing the experimental budget is not fixed upfront. Instead information are evaluated as they become available and as soon as significant evidence in favor of one model over the rest become evident a “winning” model may be selected at a much earlier stage at lower cost than would be possible with a onetime large experiment at consequently higher financial and/or experimentation cost.

The Bayesian theory allows us to implement a sequential scheme and make use of the most recent information obtained from data. In addition to that, any available prior information about the models under investigation can be considered in the analysis. The MCMC-based methods have the advantage of its capacity to handle both linear and nonlinear models. Hence, our proposed methodology is utilizing Bayesian MCMC methods.

4.2. Related Work

Experimental design theory for precise estimation of model parameters has been the focus of majority of the available research and development efforts. For example, Atkinson and Donev [4], Fedorov and Hackl [16], Box and Hill [9], Kiefer [20], Chaloner and Verdinelli [11]. Various design criteria were considered, among which the one that has received the most attention is that of identifying an experimental design that makes the variances of a model’s parameter estimates as small as possible. It was termed as the D-optimal criterion, under which the optimal design will maximize the determinant of the variance-covariance matrix. Implicit in the D-optimal design is the assumption that the model is true in that it is the correct model that generated the data. Obviously, this assumption is far from true in most of real-world problems, as the true model may never be known for certainty but close approximation to it is expected to be of value. On the other hand, the design problem for discrimination between rival models has received less attention and has been developed for simple models only. Previous work in literature has discussed the model discrimination experimental designs for linear models, and various criteria were considered as well. For example, Atkinson and Fedorov [5, 6], Burke et al.

[10], Stewart et al. [31], Box and Hill [9], Ponce de Leon and Atkinson [28], Muller and Ponce de Leon [25], Felsenstein [18], Fedorov and Khabarov [17], Ucinski and Bogacka [33], Dette and Titoff [15], Agboto, Li and Nachtsheim [1]. The focus of these criteria shifted to designs that maximally discriminate between two or more models. The most popular such criterion, termed as T-optimal, was first introduced by Atkinson and Fedorov [5, 6] for the single response case. It has the statistical interpretation as the power of a test for the fit of a second model when the first model is true. Bayesian criteria were also considered in model discrimination. See, for example, Ponce de Leon and Atkinson [28] where T-optimality was extended to Bayesian T-optimality to include prior information in the design process. Meyer et al. [23] considered a Bayesian criterion that is based on the Kullback-Leibler information to choose follow-up run after a factorial design to de-alias rival models; Bingham and Chipman [7] proposed a Bayesian criterion that is based on the Hellinger distance between predictive densities for choosing optimal designs for model selection with prior distributions specified for model coefficients and errors. Chaloner and Verdinelli [11] provided a comprehensive review on Bayesian experimental design. However, applications of T-optimal designs have been limited possibly due to the poor estimation properties of these designs, the computational burden of implementation and the requirement to assume a true model. Modifications to the D-optimal criterion have also been suggested for model discrimination purposes. See Atkinson [2], Lim and Studden [21] and Atkinson and Cox [3]. Other approaches of interest have been proposed by Dette [14] in the context of polynomial regression models.

Sequential testing and design of experiments have been also studied previously, one may refer to Chernoff [12], Pilz [27], Wetherill and Glazebrook [34] and Michlin et at. [24]. Sequential testing in the context of ALT planning and inference has been studied by Liu and Tang [22], and Tang and Liu [32] for the single-variable constant-stress accelerated test. However, as far as we know, no previous work has aimed at sequential testing in the context of ALT model discrimination. That is an additional motivation to the work presented in this study.

In the remainder of the paper, we present the proposed framework of the sequential

model-discriminant testing scheme in Section 3, a case study in Section 4, and discussion and concluding remarks in Section 5.

4.3. Proposed Methodology

4.3.1. Sequential Testing Scheme

We consider the problem of discriminating between rival models when planning accelerated life test (ALTs). Very often there exist uncertainty regarding the form of the model that describes the relationship between the parameters of the life distribution and the applied stress. Assume that uncertainty in model form can be summarized by a finite number of rival models K that are described by the random variable $M \in \{m = 1, \dots, K\}$ with associated prior probability $\pi(M = m)$ of a particular model m being true. Each model has its own set of parameters θ_m with a likelihood function $f(y|M = m, \theta_m)$ given data y. Relevant prior distributions are placed on the parameters of each model, θ_m and are denoted as $\pi(\theta_m|M = m)$. In this paper we propose a sequential Bayesian model-discriminant scheme. Bayesian methods allow us to implement a sequential scheme and make use of the most recent information obtained from data. Figure 4.1 shows a schematic illustration of the proposed sequential model-discriminant ALT planning.

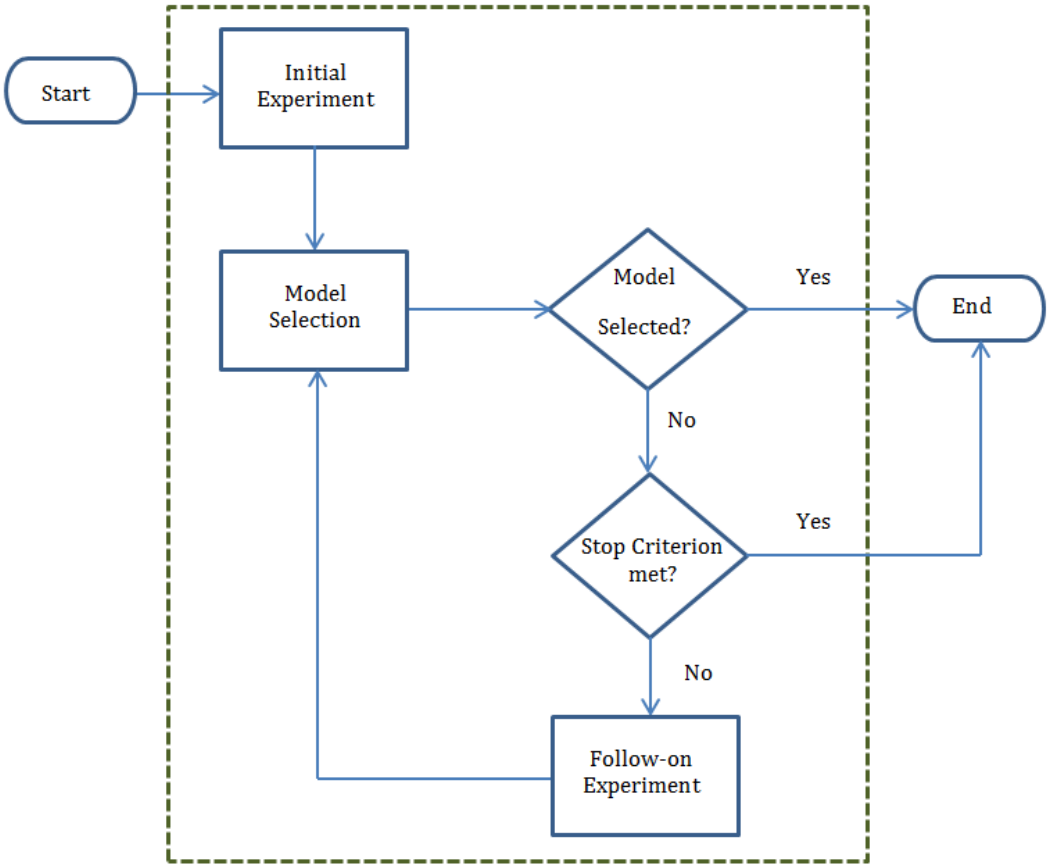


Figure 4.1: Sequential Model-discriminant ALT Scheme

The proposed sequential scheme consists of the following main steps:

1. Initial experiment design
2. Model selection analysis
3. Follow-on experiment design
4. Exit (stop) criterion

The scheme starts with an initial experiment that is planned and run at an optimal setup identified under a model-discriminant design criterion. This will provide the maximum possible information with respect to discrimination between models while using less experimental budget. We use a criterion that is based on Hellinger distance between predictive distributions; details of design optimality criterion are covered in section 4.3.3. If the output of the initial experiment enabled

clear differentiation between competing models, then no further experimentation is needed and a "winning" model is selected. Details of model selection procedure are covered in section 4.3.2. If no winning model can be selected, then further experimentation, while budget is available, is recommended and a follow-on experiment is conducted. The test setup for the follow-on experiment at which new data are to be collected is identified under similar model-discriminant optimality criterion as one utilized in the initial experiment with the advantage of inputs that are now updated with the outcome from previous experiment in a natural way following Bayes rule. The sequential testing is terminated upon exhaustion of available experimental budget and/or when a pre-determined stopping criterion is met, for example when one model displays strong evidence against the rest of rival models as indicated by an appropriate value of Bayes factor BF.

4.3.2. Model Selection

Thinking in terms of models $M = m$ rather than parameters θ_m, the posterior probability $f(M = m|y)$ of a model $M \in \{m = 1, \dots, K\}$ given data y is given by Bayes' theorem as:

$$f(M = m|y) = \frac{f(y|M=m).\pi(M=m)}{f(y)} \quad (1)$$

Data-dependent term $f(y|M = m)$ is a likelihood, and represents the probability that data is generated under model $M = m$. Bayesian model comparison is based on evaluating this data-dependent term. In the remainder of this article $M = m$ will be abbreviated with m for simplicity. For example, in the case of two rival models m_1 and m_2 (or two hypotheses H_1 and H_2), Bayesian comparison is performed via the posterior model probabilities and their corresponding ratio

$$\frac{f(m_1|y)}{f(m_2|y)} = \frac{f(y|m_1)}{f(y|m_2)} \; X \; \frac{\pi(m_1)}{\pi(m_2)} = BF_{12} \; X \; \frac{\pi(m_1)}{\pi(m_2)} \quad (2)$$

That is

$$Posterior\ model\ odds = Bayes\ factor\ (BF)\ X\ prior\ model\ odds$$

As can be seen from above expression, Bayes factor (BF_{12}) of model m_1 versus model m_1 is defined by the ratio of the integrated (or marginal likelihoods) of the two models being compared

$f(y|m_1)$ and $f(y|m_2)$, whereas the prior model odds is the ratio of the prior model probabilities $\pi(m_1)$ and $\pi(m_2)$. In light of our ignorance (lack of knowledge) as to which model is true, the $prior\ model\ odds = 1$ since $\pi(m_1) = \pi(m_2) = 0.5$. This results in $Posterior\ model\ odds = Bayes\ factor\ (BF)$. Hence, Bayes factor (BF) calculation depends on how to compute integrated likelihoods,

$$BF_{12} = \frac{\int f(y|\theta_1,m_1)\,\pi(\theta_1|m_1)\,d\theta_1}{\int f(Y|\theta_2,m_2)\pi(\theta_2|m_2)\,d\theta_2} \quad (3)$$

In simply suitable models, one can calculate the integrals in (3) analytically. However, most of the times (3) is hard to evaluate, especially in high dimensions in problems lacking neat, closed-form solution. See Kass and Raftery [19] for a survey.

4.3.2.1. Interpretation of Values of Bayes Factor (BF)

When no prior information is available on model structure, then equal prior model probabilities are considered resulting in model comparison being based solely on Bayes factors. If one considers model comparison as hypothesis testing in which interest lies in evaluating the null hypothesis H_0 (corresponding to model m_0) against the alternative H_1 (corresponding to model m_1), then both the posterior model odds and the corresponding Bayes factor BF_{10} evaluate the evidence against the null hypothesis, similar to classical significance tests. Suggested interpretation of Bayes factor is provided by Kass and Raftery [19]; see also Table 4.1.

$\log(BF_{10})$	BF_{10}	$Evidence\ against\ H_0$
0 – 1	1 – 3	Negligible
1 – 3	3 – 20	Positive
3 – 5	20 – 150	Strong
> 5	> 150	Very strong

Table 4.1: Bayes Factor Interpretation as Given by Kass and Raftery

4.3.2.2. Computation of the Marginal Likelihood

There are a number of methods that seek to estimate the marginal density $f(y|M = m)$ for each model, and subsequently calculate Bayes factor (BF) using equation (3). Majority of them operate on a posterior sample that has already been produced by some non-iterative or Markov chain Monte Carlo (MCMC) sampling methods. A review on the different computational methods for the estimation of the marginal likelihood and their comparison can be found in Bos [8]. Chib [13] proposed estimating the marginal likelihood by sampling from the posterior distribution using the Gibbs sampler. We use the most popular approximation of the marginal likelihood known as Laplace approximation which is given by

$$f(y|m) \approx (2\pi)^{\frac{d_m}{2}} \left|\tilde{\Sigma}_m\right|^{\frac{1}{2}} f\left(y\middle|\tilde{\theta}_m, m\right) f\left(\tilde{\theta}_m\middle|m\right), \tag{4}$$

where $\tilde{\theta}_m$ is the posterior mode of the parameters of model m and $\tilde{\Sigma}_m = \left(H_m(\tilde{\theta}_m)\right)^{-1}$, with $H_m(\tilde{\theta}_m)$ being equal to the minus of the second derivative matrix of the log-posterior density $\log f(\theta|y, m)$ evaluated at the posterior mode $\tilde{\theta}_m$.

To avoid analytic calculation of $\tilde{\theta}_m$ and $\tilde{\Sigma}_m$, we use the Laplace-Metropolis estimator proposed by Raftery [29] and Lewis and Raftery [30]. Using this approach, we estimate $\tilde{\theta}_m$ and $\tilde{\Sigma}_m$ from the output of MCMC algorithm by the posterior mean $\bar{\theta}_m$ and variance-covariance matrix S_m of the simulated values, respectively. Hence the Laplace-Metropolis estimator is given by

$$f(y|m) \approx (2\pi)^{\frac{d_m}{2}} |S_m|^{\frac{1}{2}} f\left(y\middle|\bar{\theta}_m, m\right) f\left(\bar{\theta}_m\middle|m\right), \tag{5}$$

where

$$\bar{\theta}_m = \frac{1}{T} \sum_{t=1}^{T} \theta_m^{(t)} \text{ and } S_m = \frac{1}{T-1} \sum_{t=1}^{T} (\theta_m^{(t)} - \bar{\theta}_m)(\theta_m^{(t)} - \bar{\theta}_m)^T.$$

To estimate the Laplace-Metropolis estimator using WinBUGS, the following steps are used:

1. Generate an MCMC sample output in WinBUGS.
2. Estimate the following from the MCMC sample output:

a. The posterior mean of the parameters of interest from each model denoted by $\bar{\theta}_m$.

b. The posterior standard deviation of the parameters of interest from each model denoted by: $S_{\theta_m} = (s_1, \ldots, s_d)$.

c. The posterior correlation between the parameters of interest from each model by denoted by R_{θ_m}. This is done using the command *Correlation* from the *Inference* menu in WinBUGS.

3. Calculate the expression

$$\log \hat{f}(y|m) = \frac{1}{2} d_m \log(2\pi) + \frac{1}{2} \log\left|R_{\theta_m}\right| + \sum_{j=1}^{d_m} \log s_j + \sum_{i=1}^{n} \log f(y_i|\bar{\theta}_m, m) + \log f(\bar{\theta}_m|m),$$

where s_j are the posterior standard deviations of θ_j parameter estimated from the MCMC output.

4.3.2.3. Model Recovery Rate

Model discriminant optimal designs are expected to maximize the proportion of times in which the true, data-generating, model is selected under an appropriate model selection criterion. In this study we have used BF (Bayes Factor) for model selection. However, other criteria such as DIC or BIC could have also been used. In determining model recovery rate under the optimal design, the following definitions were used:

- True model = acceleration model from which data was simulated. In the demonstration example it is Equation (15) for the linear model, and Equation (20) for the quadratic model.
- Assumed model = actual acceleration model fitted to the simulated data.

- Recovery rate (RR) = fraction of times the true model recovered (correctly identified) given BF model selection criterion under optimal test plan (i.e. stress setup and unit allocation).

4.3.3. Description of Design Criterion

Our optimality criterion is based on a simple, yet intuitive idea, that a good design is expected to generate data set that will cause competing models to predict same quantity of interest differently. The larger the distance (disagreement) in prediction the better our ability to discriminate (distinguish) between rival models. That is, a good design should amplify the difference among models and ease the selection of a winning one. We have proposed the use of this criterion in Nasir and Pan [26] in the context of non-sequential ALT model selection. The criterion uses the relative prediction performance of each model over the range of its parameters to identify the optimal design (different models may have different set of parameters). Criterion formulation is explained in details in Chapter 3 and its subsections.

Herein, we briefly describe the criterion, for the case of two competing models, M_1 and M_2 with model priors; $\pi(M_1)$ and $\pi(M_2)$ respectively. Under design d, model M_1 has parameter vector θ_1 and experimental outcome $Y_1 = (y_{11}, \dots, y_{N1})$, while model M_2 has parameter vector θ_2 and experimental outcome $Y_2 = (y_{12}, \dots, y_{N2})$. The utility function to be optimized (maximized), is the difference in prediction of life percentile of interest τ_p at the low stress $\tau_p(S_1)$ and high stress $\tau_p(S_2)$ test setup across all pairs of competing models. That is $D_{S_1}(\cdot)$ and $D_{S_2}(\cdot)$ respectively. The pairwise local utilities, $u_{2|1}$ and $u_{1|2}$, were defined as

$$u_{2|1}\big(d, M_1(\theta_1, Y_1), M_2(\theta_2, Y_1)\big) = D_{S_1}\left(\hat{\tau}_{p,(M2|Y1)}, \hat{\tau}_{p,(M1|Y1)}\right) + D_{S_2}\left(\hat{\tau}_{p,(M2|Y1)}, \hat{\tau}_{p,(M1|Y1)}\right) = u_{2|1} \quad (6)$$

$$u_{1|2}(d, M_1(\theta_1, Y_2), M_2(\theta_2, Y_2)) = D_{S_1}\left(\hat{\tau}_{p,(M1|Y2)}, \hat{\tau}_{p,(M2|Y2)}\right) + D_{S_2}\left(\hat{\tau}_{p,(M1|Y2)}, \hat{\tau}_{p,(M2|Y2)}\right) = u_{1|2} \quad (7)$$

Equation (6) represents the difference in τ_p prediction of model (M_2) conditional on data from model (M_1) relative to model (M_1) prediction of the same quantity, while equation (7) represents

the difference in τ_p prediction of model (M_1) conditional on data from model (M_2) relative to model (M_2) prediction of the same quantity. That is the relative prediction performance of each model over the range of its parameter vector.

The utility $u_{i|j}(\cdot)$ of the design is assessed by its expectation with respect to the sampling distribution of the data $p(y_1|\theta_1, d)$, and $p(y_2|\theta_2, d)$, and the prior distribution of the parameter vectors $\pi(\theta_1)$ and $\pi(\theta_2)$. That is by calculating the pre-posterior expectation.

$$E(u_{2|1}) = \iint u_{2|1}\, p(y_1|\theta_1, d)\, \pi(\theta_1|d) dy_1 d\theta_1 \tag{8}$$

$$E(u_{1|2}) = \iint u_{1|2}\, p(y_2|\theta_2, d)\, \pi(\theta_2|d) dy_2 d\theta_2 \tag{9}$$

Equation (8) gives an expression of the expected pre-posterior prediction difference in τ_p of model (M_2) conditional on data from model (M_1) relative to model (M_1) prediction of the same quantity. Similarly, equation (9) gives an expression of the expected pre-posterior prediction difference in τ_p of model (M_1) conditional on data from model (M_2) relative to model (M_2) prediction of the same quantity.

A critical distinction from the non-sequential (one-time large DOE) design, as compared to sequential design, is the sequential updating of the prior distribution of the model parameters from one stage to another. At the onset of each sequential experiment, the current posterior distribution becomes the new prior distribution going into the new experiment.

Since it is not known which of the two models (M_1) or (M_2) is the true model that will generate the experimental outcome, a combined weighted expected utilities $E(u_{i|j})$ were used to obtain the desired global utility function $U(d)$ to be maximized. The weighing is achieved by priors assigned to the models, $\pi(M_1)$ and $\pi(M_2)$ respectively.

$$\begin{aligned} U(d) &= \sum_{\substack{i,j=1,2 \\ i \neq j}} \pi(M_i).E(u_{i|j}) \\ &= \pi(M_1).E(u_{2|1}) + \pi(M_2).E(u_{1|2}) \end{aligned} \tag{10}$$

Equation (10) is interpreted as a measure of model distinguishability between two models. The larger the value of $U(d)$, the dissimilar the two models are to each other. Extending (10) to

account for situations where more than two models are to be distinguished among is straightforward.

4.4. Case Study

The demonstration example presented here has been previously used in Chapter 3; Nasir and Pan [26], in the context of non-sequential model discriminant ALT planning. The intent is to provide a comparison of sequential to non-sequential testing with respect to model recovery rate for identifying correct model. In this study the total sample size and testing duration are pre-determined given budget and testing equipment availability constraints.

Description of the design problem and the linear versus quadratic competing models can be found in Chapter 3 of this dissertation. Model recovery rate for both testing schemes is evaluated in terms of Bayes factor.

4.4.1. Model Discrimination Results

We make use of computer simulation to generate experimental data and explore the performance of proposed method. As illustrated in Figure 4.1, the sequential scheme starts with an initial experiment that is planned and run at an optimal setup identified under a model-discriminant design criterion providing the maximum possible information with respect to discrimination between models while using less experimental budget. We use a criterion that is based on Hellinger distance between predictive distributions. Under this criterion, the optimal experimental setup for the discrimination between the linear and quadratic model forms given the constraints described in “Description of Design Problem” section was identified in Chapter 3 as

1. Higher temperature level is to be set at the highest value $(180°\mathrm{C})$ of its range$(150°\mathrm{C} - 180°\mathrm{C})$, with approximate unit allocation of 12%.
2. Middle temperature level is to be set at the intermediate value $(130°\mathrm{C})$ of its range $(115°\mathrm{C} - 145°\mathrm{C})$, with approximate unit allocation of 55%.
3. Lower temperature level is to be set at value $(100°\mathrm{C})$, slightly above the intermediate value of its range $(80°\mathrm{C} - 110°\mathrm{C})$, with approximate unit allocation of 33%.

Above set up was shown to be the condition at which the pre-posterior expected value of the utility function ($Log[U(d)]$) is maximized as a function of the stress magnitude and percent unit allocation to each of the three stress levels used in the non-sequential planning of the experiment. See Nasir and Pan [26] for details.

When adopting the same design problem in the context of sequential testing for model discrimination, we are faced with the initial question of how to conduct a meaningful comparison between the two test planning schemes. Questions such as test unit allocation, stress-level set up and testing durations come to mind as key considerations. Therefore, we assume the following:

- Sample size: both testing schemes have the same budget of a total of 300 units available for testing. Non-sequential testing will consume all of the 300 units in one large experiment, while sequential testing will be carried out in two phases for the total sample size of 300 units. Sequential testing will split total sample size equally by the number of phases planned for testing. For example, in a two-phase sequential test, phase one uses 150 units as same as phase-two.
- Unit allocation: since both testing schemes utilize the same utility function for optimization, a good starting point for both is optimal allocation identified in Chapter 3. That is 12%, 55%, and 33% at high, middle, and low stress-level respectively.
- Test duration: total length of testing time is same for both testing schemes. However, sequential testing will split total duration equally by the number of phases planned for testing. For example, in a two-phase sequential test, given demonstration case at hand, phase one has 21 days available for testing as same as phase-two. Non-sequential testing is conducted in 42 days.
- Stress-level setup: non-sequential test will adopt optimal setup identified in Chapter 3. The sequential test optimizes stress level setup subject to fixed unit allocation at stress levels (12% at high, 55% at middle, and 33% at low), and test duration constraints. Algorithm for constructing optimal design, as in Chapter 3 is used.

With above assumptions in place, we consider two cases for fail data, namely complete data (100% failure) and right censored data.

4.4.1.1. Complete Data

Summary of performance comparison in model recovery rates between the non-sequential and sequential model discriminant testing schemes for the case of compete data (100% failure) is shown in Table 4.2.

Total budget	Testing Scheme					
	Non-Sequential			Sequential		
	Exp #	Sample Size/Exp	Recovery Rate	Exp #	Sample Size/Exp	Recovery Rate
300	1	300	0.85	1	150	0.74
	2	n/a	n/a	2	150	0.92

Table 4.2: Recovery Rate Comparison for Sequential vs. Non-sequential Model Discriminant Testing for Complete Data

From summary results in Table 4.2, it can be seen that in the case of total budget of 300 units, sequential testing in two phases, each having 50% of available budget, results in a lower recovery rate as compared to non-sequential testing in phase-one (74% vs. 85%). However, it compensates for that drop in second phase with recovery rate of 92% as compared to 85%. This is influenced by the fact that the starting point for phase-two experimentation has already been updated with information obtained from phase-one testing. For the total budget of 300 units, there was a roughly 8% gain in recovery rate by conducting sequential testing. Table 4.3 shows the optimal stress setup and %unit allocation for both testing schemes. In phase-one of sequential testing with a reduced budget of 150 units (as compared to 300 units) and fixed unit allocation at the three stress levels (12%, 55%, 33%), the plan optimizes the stress levels to test at higher levels for all three stress levels $S_H, S_M \; and \; S_L$ as compared to non-sequential test. This can be

explained as a compensation for testing for shorter time duration with less total sample size (1/2 of that for non-sequential for both time and sample size). However, phase-two of the sequential testing is constraint only by total sample size available (150 units), and optimizes both of the stress levels and unit allocation. It can be noticed that given updated model parameters from phase-one, the plan tests at reduced stress levels with more unit allocation towards lower level.

Total budget	Testing Scheme					
	Non-Sequential			Sequential		
	Exp #	Stress level	Unit allocation	Exp #	Stress level	Unit allocation
300	1	$S_H(180℃)$	12%	1	$S_H(180℃)$	12%
		$S_M(130℃)$	55%		$S_M(145℃)$	55%
		$S_L(100℃)$	33%		$S_L(105℃)$	33%
	2	n/a	n/a	2	$S_H(155℃)$	10%
					$S_M(115℃)$	31%
					$S_L(85℃)$	59%

Table 4.3: Stress Setup and Unit Allocation for Sequential vs. Non-sequential Model Discriminant Testing for Complete Data

4.4.1.2. Right Censored Data

Summary of performance comparison in model recovery rates between the non-sequential and sequential model discriminant testing schemes for the case of right censored data is shown in Tables 4.4 - 5. Table 4.4 shows comparison results when censoring is in the range of 30-40%, while Table 4.5 shows comparison results when censoring is in the range of 50-60%.

Total budget	Testing Scheme					
	Non-Sequential			Sequential		
	Exp #	Sample Size/Exp	Recovery Rate	Exp #	Sample Size/Exp	Recovery Rate
300	1	300	0.67	1	150	0.61
	2	n/a	n/a	2	150	0.68

Table 4.4: Recovery Rate Comparison for Sequential vs. Non-sequential Model Discriminant Testing with 30-40% Censoring in Fail Data

Total budget	Testing Scheme					
	Non-Sequential			Sequential		
	Exp #	Sample Size/Exp	Recovery Rate	Exp #	Sample Size/Exp	Recovery Rate
300	1	300	0.63	1	150	0.57
	2	n/a	n/a	2	150	0.64

Table 4.5: Recovery Rate Comparison for Sequential vs. Non-sequential Model Discriminant Testing with 50-60% Censoring in Fail Data

From summary results in Tables 4.4-4.5, it can be seen that the gain in model recovery rate obtained by testing sequentially diminishes as the amount of censoring in the data increases. As a matter of fact, the ability of the test plan to discriminate model form adversely impacted even in the case of the non-sequential testing. This is troublesome finding as majority of ALT data are characterized by censoring. Complete data are rare, especially for high reliable components.

4.5. Conclusion and Areas for Future Research

In this paper, we have considered the problem of discriminating between rival models when planning accelerated life test (ALTs) as often there exist uncertainty regarding the form of the model that describes the relationship between the parameters of the life distribution and the applied stress. We assumed that uncertainty in model form can be summarized by a finite number of rival models K that are described by the random variable $M \in \{m = 1, \dots, K\}$ with associated prior probability $\pi(M = m)$ of a particular model m being true. We proposed a sequential Bayesian model-discriminant scheme to address the design problem as Bayesian methods allow us to implement a sequential scheme and make use of the most recent prior information about the models under investigation.

The idea was based on an initial experiment planned and conducted at an optimal setup identified under a model-discriminant design criterion; the Hellinger distance measure between predictive densities was used to provide the maximum possible information with respect to model discrimination while using less experimental budget. If no model can be identified, then ta subsequent model-discriminant experimentation is planned and conducted while budget allows by leveraging the most current information to allow for Bayesian model comparison through posterior model probabilities and their corresponding ratios. Sequential testing is terminated upon exhaustion of available experimental budget and/or when strong evidence in favor of one model over another is demonstrated through Bayes factor. In our demonstration example we considered the case for single variable constant-stress accelerated life test at three levels where proposed methodology was applied to optimally discriminate between linear and quadratic model forms. Advantage of proposed methodology as compared to non-sequential testing under similar model-discriminant design criterion was evaluated through model recovery rate under Bayes factor model selection rule. Results showed that performance of sequential model discriminant in ALT is adversely impacted by the amount of censoring in the data, in the case of linear vs. quadratic model form with testing at three levels of constant stress, sequential testing can improve model recovery rate by approximately 8% in the case of complete data. In the case of censoring two

buckets were considered for right censoring, namely (30-40%) and (50-60%). Results showed that both testing schemes suffered in their ability to discriminate between rival models and there was no apparent advantage in adopting sequential testing. This finding is troublesome as majority of ALT data are characterized by censoring and complete data are rare.

Much more interesting work remains to be investigated in this area, to name a few: sample size optimization across the different phases of sequential testing, optimal number of phases for sequential testing, effectiveness of proposed scheme in discrimination of other forms of models, and impact of different model selection criteria on recovery rate.

References

[1] V. Agboto, W. Li, and C. Nachtsheim, Screening designs for model discrimination, Journal of Statistical Planning and Inference, vol. 140, no. 3 (2010), pp. 766-780.

[2] A. C. Atkinson, Planning experiments to detect inadequate regression models, Biometrika 59 (1972), 275-293.

[3] A. C. Atkinson, and D. R. Cox, Planning experiments for discriminating between models, J. Roy. Statist. Soc. Ser. B 36 (1974), 321-348.

[4] A. C. Atkinson, and A. N. Donev, Optimum experimental designs, Oxford, England: Oxford University Press (1992).

[5] A. C. Atkinson, and V. V. Fedorov, The design of experiments for discriminating between two rival models, Biometrika 62 (1975a), 1, pp. 57-70.

[6] A. C. Atkinson, and V. V. Fedorov, Optimal design: Experiments for discriminating between several models, Biometrika 62 (1975b), 2, pp. 289-303.

[7] D. R. Bingham, and H. A. Chipman, Optimal designs for model selection, Technical Report 388, University of Michigan and University of Waterloo (2002).

[8] C. S. Bos, A comparison of Marginal Likelihood Computation Methods, Technical Report, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands (2002).

[9] G. E. P. Box, W. J. Hill, Discrimination among mechanistic models, Technometrics 9 (1967), 57-71.

[10] A. L. Burke, T. A. Duever, and A. Pendilis, Model discrimination via designed experiments: Discriminating between the terminal and penultimate models on the basis of composition data, Macromolecules 27 (1994), 386-399.

[11] K. Chaloner and I. Verdinelli. Bayesian Experimental Design: A Review. Statistical Science, 10 (1995), 273-304.

[12] H. Chernoff, Sequential Analysis and Optimum Design, Society for Industrial and Applied Mathematics (SIAM), Philadelphia (1972).

[13] S. Chib, Marginal likelihood from the Gibbs output, Journal of the American Statistical Association 90 (1995), 1313-1321.

[14] H. Dette, Optimal designs for identifying the degree of a polynomial regression, Annals of Statistics 23 (1995), 1248-1267.

[15] H. Dette, and S. Titoff, Optimal discrimination designs, The Annals of Statistics, vol. 37 (2009), no. 4, pp. 2056-2082.

[16] V. V. Fedorov and P. Hackl, Model-Oriented Design of Experiments, Lecture Notes in Statistics (1997), New York: Springer-Verlag.

[17] V. V. Fedorov and V. Khabarov, Duality of optimal design for model discrimination and parameter estimation, Biometrika 73, 1, (1986), 183-190.

[18] K. Felsenstein, Optimal Bayesian design for discriminating among rival models, Computational Statistics & Data Analysis 14 (1992), 427-436.

[19] R. Kass, and A. Raftery, Bayes factors, Journal of the American Statistical Association 90 (1995), 773-795.

[20] J. Kiefer and J. Wolfowitz, Optimum design in regression problems, The Annals of Mathematical Statistics 30 (1959), 271-294.

[21] Y. B. Lim and W. J. Studden, Efficient Ds-optimal designs for multivariate polynomial regression on the q-cube, Annals of Statistics 16 (1988), 1225-1240.

[22] X. Liu and L. C. Tang, A Sequential Constant-Stress Accelerated Life Test Scheme and its Bayesian Inference, Quality and Reliability Engineering International 25 (2009), 91-109.

[23] R. D. Meyer, D. S. Steinberg, G. E. P. Box, Follow-up designs to resolve the confounding in multifactor experiments, Technometrics 38 (1996), 303-313.

[24] Y. H. Michlin, L. Meshkov and I. Grunin, Improvement on Sequential Testing in MIL-HDBK-781A and IEC 61124, IEEE Transactions on Reliability 57 (2008), 379-387.

[25] W. G. Muller and A. C. Ponce de Leon, Discriminating between two binary data models: sequentially designed experiments, Journal of Statistical Computation and Simulation 55 (1996), 87-100.

[26] E. Nasir, R. Pan, Simulation-based Bayesian Optimal ALT Designs for Model Discrimination, Reliability Engineering & System Safety (Submitted) (2014).

[27] J. Pilz, Bayesian Estimation and Experimental Design in Linear Regression Models, John Wiley & Sons, Germany (1991).

[28] A. C. Ponce de Leon and A. C. Atkinson, Optimum experimental design for discriminating between two rival models in the presence of prior information, Biometrika 78 (1991), 601-608.

[29] A. Raftery, Hypothesis testing and model selection, in W. Gilks, S. Richardson, and D. Spiegelhalter, eds., Markov Chain Monte Carlo in Practice, Chapman & Hall, Suffolk, UK, (1996b) pp. 163-188.

[30] A. Raftery, D. Madigan, J. Hoeting (1997), Bayesian model averaging for linear regression models, Journal of the American Statistical Association 92 (1997), 179-191.

[31] W. E. Stewart, Y. Shon, and G. E. P. Box, Discrimination and goodness of fit of multi-response mechanistic models, AIChE Journal 44, 6, (1998), 1404-1412.

[32] L. C. Tang and X. Liu, Planning and Inference for a Sequential Accelerated Life Test, Journal of Quality Technology 42, No.1 (2010).

[33] D. Ucinski and B. Bogacka, T-optimum design for discrimination between two multi-response dynamic models, Journal of the Royal Statistical Society: Statistical Methodology 67(B) (2005), 3-18.

[34] G. B. Wetherill and K. D. Glazebrook, Sequential Methods in Statistics, 3rd edition. London, UK: Chapman and Hall (1986).

CHAPTER 5

CONCLUSION

Accelerated life tests (ALTs) are widely used throughout industry. Engineers rely on data from ALTs to drive critical business decisions. Continual research in new methods for ALT planning and ALT data analysis is crucial for many industries. The challenge with ALTs is a combined effect of factors that can be classified into two main groups, namely:

1. The typical characteristics of ALTs
    - Statistical theory for traditional experimental design and properties of standard experimental designs do not hold for censored and interval-censored data (majority of data). It is correct only for complete data (100% failure) which is rare in industry nowadays.
    - Extrapolation in both stress and time.
    - Model dependence on parameters and the implicit assumption in design criteria used in current ALTs that the form of the acceleration model is correct.
2. The business environment need
    - Informative data (precise estimate/prediction) for on-time business decision
    - At lowest possible cost
    - In least possible amount of time

In this dissertation, we contribute to the ALT planning literature by proposing three ALT planning methods in a Bayesian framework with the third method being an extension to the second proposed methodology. We strongly believe that reliability engineers have accumulated a tremendous wealth of knowledge regarding failure modes, mechanisms, reliability models and associated model parameters from experimentation and documentation over the years. Unless it is a breaking through technology, most of new product introductions are based on incremental improvements in technology. As such available knowledge from experimentation should be readily available for use to reduce the amount and duration of new testing required to meet a business need. The vehicle for incorporating available information into test planning is Bayesian methods, hence our favorable choice to implement a Bayesian framework in our study.

In Chapter 2, we proposed a simulation-based approach for the design of ALT plans with multiple stresses utilizing Latin hypercube sampling scheme to overcome the practical difficulty arising from the increased number of experiments required due to the increased number of stress factor-level combinations to be studied. When applied to an industrial case study it was able to incorporate available prior information on model parameters along with new data to update information going into the planning problem. A comparison to its counterpart of full grid simulation quantified the computational cost gain at no loss of statistical efficiency, and a comparison to the large-sample approximation method revealed the flexibility of the proposed approach in determining optimal stress settings with less assumptions being made and more intuitive unit allocations.

In Chapter 3, we provided an approach for optimum ALT design with the objective of model discrimination among rival model forms. Our proposed criterion was based on the Hellinger distance measure between predictive distributions. The optimal stress-factor setup and unit allocation were determined at three stress levels subject to stress lab equipment and test-duration constraints. We compared the performance of obtained test plans with other test plans including the typically used three stress-levels good compromise plan, best traditional plan and well known 4:2:1 compromise ALT test plans. Results showed that when approach is applied to the case of linear versus quadratic ALT models, our optimal design method has the advantage of substantially increasing a test plan's ability to distinguish among competing ALT models and provide better guidance as to which model is appropriate for the experiment.

In Chapter 4 we extended the approach of model discrimination discussed in Chapter 3 to sequential model discrimination in accelerated life test planning. The idea was based on an initial experiment planned and conducted at an optimal setup identified under a model-discriminant design criterion; the Hellinger distance measure between predictive densities, to provide the maximum possible information with respect to model discrimination while using less experimental budget. If an appropriate model could not be identified, then a follow on model-discriminant experimentation is planned and conducted while budget allows by leveraging the most current information to allow for Bayesian model comparison through posterior model

probabilities and their corresponding ratios. Results showed that performance of sequential model discriminant in ALT is adversely impacted by the amount of censoring in the data, in the case of linear vs. quadratic model form with testing at three levels of constant stress, sequential testing can improve model recovery rate by approximately 8% in the case of complete data. In the case of censoring, two buckets were considered for right censored data; namely (30-40%) and (50-60%), and results showed that both testing schemes suffered in their ability to discriminate between models and there was no apparent advantage in adopting sequential testing. This finding is troublesome as majority of ALT data are characterized by censoring and complete data are rare.

Much more work remains to be investigated in this area, to name a few: sample size optimization across the different phases of sequential testing, optimal number of phases for sequential testing, effectiveness of proposed scheme in discrimination of other forms of models, impact of different model selection criteria on recovery rate, and formal introduction of cost in utility function when finding optimal designs.

The problem of ALT planning and execution to meet business needs is of great interest and poses tremendous challenges at the same time. Engineers would like to know how to plan effective ALTs in ever increasing complex industrial environments where a decision based on the trio of good data at low cost in minimal time is the measure of success. We will continue our investigation into the subject in future research with focus on improvement areas identified.

REFERENCES

V. Agboto, W. Li, and C. Nachtsheim, Screening designs for model discrimination, Journal of Statistical Planning and Inference, vol. **140**, no. 3 (2010), pp. 766-780.

A. C. Atkinson, Planning experiments to detect inadequate regression models, Biometrika 59 (1972), 275-293.

A. C. Atkinson, and D. R. Cox, Planning experiments for discriminating between models, J. Roy. Statist. Soc. Ser. B **36** (1974), 321-348.

A. C. Atkinson, and A. N. Donev, Optimum experimental designs, Oxford, England: Oxford University Press (1992).

A. C. Atkinson, and V. V. Fedorov, The design of experiments for discriminating between two rival models, Biometrika **62** (1975a), 1, pp. 57-70.

A. C. Atkinson, and V. V. Fedorov, Optimal design: Experiments for discriminating between several models, Biometrika **62** (1975b), 2, pp. 289-303.

D. S. Bai, and S. W. Chung, An optimal design of accelerated life test for exponential distribution, Reliability Engineering and System Safety **31** (1991), 57-64.

D. S. Bai, and S. W. Chung, Optimal design of partially accelerated life test for exponential distribution under Type-I censoring, IEEE Transactions on Reliability **41** (1992), 400-406.

D. S. Bai, and M. S. Kim, Optimum simple step-stress accelerated life tests for the Weibull distribution and type-I censoring, Naval Research Logistics **40** (1993), 193-210.

D. S. Bai, S. W. Chung, and Y. R. Chun, Optimal design of partially accelerated life tests for the Lognormal distribution under type I censoring, Reliability Engineering and System Safety **40** (1993), 85-92.

D. W. Behnken, and D. G. Watts, Bayesian estimation and design of experiments for growth rates when sampling from the Poisson distribution, Biometrics **28** (1972)**,** 999-1009.

J. O. Berger, Statistical Decision Theory and Bayesian Analysis, 2nd ed. N.Y.: Springer- Verlag, (1985).

D. R. Bingham, and H. A. Chipman, Optimal designs for model selection, Technical Report **388**, University of Michigan and University of Waterloo (2002).

C. S. Bos, A comparison of Marginal Likelihood Computation Methods, Technical Report, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands (2002).

G. E. P. Box, W. J. Hill, Discrimination among mechanistic models, Technometrics **9** (1967), 57-71.

A. L. Burke, T. A. Duever, and A. Pendilis, Model discrimination via designed experiments: Discriminating between the terminal and penultimate models on the basis of composition data, Macromolecules **27** (1994), 386-399.

Sung-Hyuk Cha, Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions, International Journal of Mathematical Models and Methods in Applied Sciences **1** (4) (2007), 300-307.

Sung-Hyuk Cha, and S. N. Srihari , On measuring the distance between two histograms, The Journal of the Pattern Recognition Society, **35** (2002), 1355-1370.

K. Chaloner, and K. Larntz, Bayesian Design for Accelerated Life Testing, Journal of Statistical Planning and Inference **33** (1992), 245–259.

K. Chaloner, and K. Larntz, Optimal Bayesian Design Applied to Logistic Regression Experiments, Journal of Statistical Planning and Inference **21** (1989), 191–208.

K. Chaloner and I. Verdinelli. Bayesian Experimental Design: A Review. Statistical Science, **10** (1995), 273-304.

P. S. Chan, H. Y. So, H. K. T. Ng, and W. Gao, Optimal Sample Size Allocation for Multi-level Stress Testing with Extreme Value Regression Under Type-I Censoring, Naval Research Logistics Manuscript, (2012).

H. Chernoff, Sequential Analysis and Optimum Design, Society for Industrial and Applied Mathematics (SIAM), Philadelphia (1972).

S. Chib, Marginal likelihood from the Gibbs output, Journal of the American Statistical Association **90** (1995), 1313-1321.

T. Cioppa, and T. Lucas. Efficient Nearly Orthogonal and Space-Filling Latin Hypercubes. Technometrics **49** (2007), 1, pg. 45.

G. Claeskens, and N. L. Hjort, Model Selection and Model Averaging, Cambridge. (2008) Section 3.5.

M. Clyde, P. Müller, and G. Parmigiani, Optimal Design for Heart Defibrillators, in Case Studies in Bayesian Statistics 2, eds. C. Gatsonis, J. Hodges, R. E. Kass, and N. Singpurwalla, New York: Springer-Verlag, (1995) pp. 278–292.

H. Dette, Optimal designs for identifying the degree of a polynomial regression, Annals of Statistics **23** (1995), 1248-1267.

H. Dette, and S. Titoff, Optimal discrimination designs, The Annals of Statistics, vol. **37** (2009), no. 4, pp. 2056-2082.

E. Deza, and M. M. Deza, Dictionary of Distances, Elsevier, (2006).

R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2nd ed. Wiley, (2001).

A. Erkanli, and R. Soyer, Simulation-Based Designs for Accelerated Life Tests, Journal of Statistical Planning and Inference **90** (2000), 335–348.

L. A. Escobar, and W. Q. Meeker, Planning accelerated life tests with two or more experimental factors, Technometrics Vol. **37** (4) (1995), 411-427.

K. Fang, C. Ma, and P. Winker. Centered L_2-Discrepency of Random Sampling and Latin Hypercube Design, and Construction of Uniform Designs. Mathematics of Computation Vol. **71**(2000), 237, 275-296.

V. V. Fedorov and P. Hackl, Model-Oriented Design of Experiments, Lecture Notes in Statistics (1997), New York: Springer-Verlag.

V. V. Fedorov and V. Khabarov, Duality of optimal design for model discrimination and parameter estimation, Biometrika **73**, 1, (1986), 183-190.

K. Felsenstein, Optimal Bayesian design for discriminating among rival models, Computational Statistics & Data Analysis **14** (1992), 427-436.

A. A. Giunta, S. F. Wojtkiewicz Jr., and M. S. Eldred, Overview of Modern Design of Experiments Methods for Computational Simulations, AIAA (2003)-0649.

M. Hamada, H. F. Martz, C. S. Reese, and A. G. Wilson, Finding Near-Optimal Bayesian Experimental Designs via Genetic Algorithms, The American Statistician **55** (2001), 175–181.

P. D. H. Hill, A review of experimental design procedures for regression model discrimination, Technometrics **20** (1978), 1, pp 15-21.

W. J. Hill, W. G. Hunter, D. W. Wichern, A joint design criterion for the dual problem of model discrimination and parameter estimation, Technometrics **10** (1968), 145-160.

W. G. Hunter, A. M. Reiner, Designs for discriminating between two rival models, Technometrics **7** (1965), 307-323.

V. Joseph, and Y. Hung. Orthogonal-Maximin Latin Hypercube Designs. Statistica Sinica **18**, (2008), 171-186.

R. Kass, and A. Raftery, Bayes factors, Journal of the American Statistical Association **90** (1995), 773-795.

M. Keramat, and R. Kielbasa. Modified Latin Hypercube Sampling Monte Carlo (MLHSMC) Estimation for Average Quality Index. Analog Integrated Circuits and Signal Processing, **19** (1999), 87-98.

J. Kiefer and J. Wolfowitz, Optimum design in regression problems, The Annals of Mathematical Statistics **30** (1959), 271-294.

T. J. Kielpinski and W. Nelson, Optimum censored accelerated life tests for normal and lognormal life distributions, IEEE Trans. Reliability **R-24**, 5 (1975), 310-320.

P. D. Leon and A. C. Atkinson, Optimum experimental design for discriminating between two rival models in the presence of prior information, Biometrika **78** (1991), 601-608.

Y. B. Lim and W. J. Studden, Efficient Ds-optimal designs for multivariate polynomial regression on the q-cube, Annals of Statistics **16** (1988), 1225-1240.

D. V. Lindley, Bayesian Statistics – A review. SIAM, Philadelphia (1972).

X. Liu, and L.-C. Tang, A Bayesian Optimal Design for Accelerated degradation tests, John Wiley & Sons, Ltd., (2010).

X. Liu and L. C. Tang, A Sequential Constant-Stress Accelerated Life Test Scheme and its Bayesian Inference, Quality and Reliability Engineering International **25** (2009), 91-109.

M. D. McKay, W. J. Conover, and R. J. Beckman. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. Technometrics, **21** (1979) 239-245.

W. Q. Meeker, A comparison of accelerated life test plans for Weibull and Lognormal distributions and type I censored data, Technometrics **26** (1984), 157–172.

W. Q. Meeker, and L. A. Escobar, A Review of Recent Research and Current Issues in Accelerated Testing, International Statistical Review **61,** 1, (1993), 147-168.

W. Q. Meeker, and G. J. Hahn, How to Plan an Accelerated Life Test: Some Practical Guidelines, American Society for Quality Control, Milwaukee, WI, (1985).

W. Q. Meeker, and G. J. Hahn, Asymptotically optimum over-stress tests to estimate the survival probability at a condition with a low expected failure probability, Technometrics **19** (1977), 381–399.

W. Q. Meeker, and L. A. Escobar, Statistical Methods for Reliability Data, New York: Wiley (1998).

W. Q. Meeker, and W. Nelson, Optimum Accelerated Life Tests for Weibull and Extreme Value Distributions, IEEE Transactions on Reliability, R-**24** (1975), 321–332.

R. D. Meyer, D. S. Steinberg, G. E. P. Box, Follow-up designs to resolve the confounding in multifactor experiments, Technometrics **38** (1996), 303-313.

Y. H. Michlin, L. Meshkov and I. Grunin, Improvement on Sequential Testing in MIL-HDBK-781A and IEC 61124, IEEE Transactions on Reliability **57** (2008), 379-387.

E. Monroe, R. Pan, C. Anderson-Cook, D. Montgomery, and C. Borror, A Generalized Linear Model Approach to Designing Accelerated Life Test Experiments, Quality Reliability Engineering International **27** (2010)**,** 595-607.

E. Monroe, R. Pan, C. Anderson-Cook, D. Montgomery, and C. Borror, Sensitivity analysis of optimal designs for accelerated life testing, Journal of Quality Technology **42** (2010)**,** 121-135.

M. D. Morris and T. J. Mitchell. Exploratory designs for computational experiments. J. of Statistical Planning and Inference **43** (1995) 381-402.

P. Müller, Simulation-Based Optimal Design, in Bayesian Statistics 6, eds. J. M. Bernardo, J. O. Berger, A. P. David, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, (2000) pp. 459–474.

P. Müller, and G. Parmigiani, Optimal Design via Curve Fitting of Monte Carlo Experiments, Journal of the American Statistical Association **90** (1995), 1322–1330.

W. G. Muller and A. C. Ponce de Leon, Discriminating between two binary data models: sequentially designed experiments, Journal of Statistical Computation and Simulation **55** (1996), 87-100.

E. Nasir, R. Pan, Simulation-Based Bayesian Optimal Design for Multi-factor Accelerated Life Tests, Communication in Statistics – Simulation and Computation (Submitted) (2013).

E. Nasir, R. Pan, Simulation-based Bayesian Optimal ALT Designs for Model Discrimination, Reliability Engineering & System Safety (Submitted) (2014).

W. Nelson, Accelerated Testing, John Wiley & Sons: New York, (1990).

W. Nelson, and T. Kielpinski, Theory for Optimum Censored Accelerated Life Tests for Normal and Lognormal Distributions, Technometrics **18** (1976), 105–114.

W. Nelson, and W. Q. Meeker, Theory for optimum accelerated Censored Life Tests for Weibull and Extreme Value Distributions, Technometrics **20** (1978), 171-177.

A. B. Owen. Orthogonal Arrays for Computer Experiments, Integration and Visualization. Statistica Sinica **2** (1992) 439-452.

K. Palmer, and K. Tsui. A minimum bias Latin hypercube design. IIE Transactions **33** (2001), 793-808.

J. W. Park, and B. J. Yum, Optimal Design for Accelerated Life Tests with Two Stresses, Naval Research Logistics **43** (1996)**,** 863-884.

J. Park. Optimal Latin-hypercube designs for computer experiments. J. of Statistical Planning and Inference **39** (1994) 95-111.

J. Pilz, Bayesian Estimation and Experimental Design in Linear Regression Models, John Wiley & Sons, Germany (1991).

N. G. Polson, A Bayesian Perspective on the Design of Accelerated Life Tests, in Advances in Reliability, ed. A. P. Basu, New York: Elsevier (1993) pp. 321–330.

A. C. Ponce de Leon and A. C. Atkinson, Optimum experimental design for discriminating between two rival models in the presence of prior information, Biometrika **78** (1991), 601-608.

A. Raftery, Hypothesis testing and model selection, in W. Gilks, S. Richardson, and D. Spiegelhalter, eds., Markov Chain Monte Carlo in Practice, Chapman & Hall, Suffolk, UK, (1996b) pp. 163-188.

A. Raftery, D. Madigan, J. Hoeting (1997), Bayesian model averaging for linear regression models, Journal of the American Statistical Association **92** (1997), 179-191.

H. Raiffa, R. Schlaiffer, Applied Statistical Decision Theory. Division of Research, Harvard Business School, Boston (1961).

J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, Design and Analysis of Computer Experiments, Statistical Science Vol. **4** (4) (1989), 409-435.

R. R. Sitter, and B. Torsney, Optimal designs for binary response experiments with two design variables, Statistica Sinica **5** (1995)**,** 405-419.

D. M. Steinberg, and D. k. j. Lin. A construction method for orthogonal Latin hypercube designs. Biometrika **93** (2006) 279-288.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde, Bayesian measures of model complexity and fit, J. R. Statist. Soc. B **64** (2002) P4, pp. 583-639.

W. E. Stewart, Y. Shon, and G. E. P. Box, Discrimination and goodness of fit of multi-response mechanistic models, AIChE Journal **44**, 6, (1998), 1404-1412.

D. Sun, R. Tsutakawa, and W. Lu, Bayesian design of experiment for quantal responses: What's promised versus what's delivered, J. Statist. Planning and Inference **52** (1996), 289-306.

L. C. Tang and X. Liu, Planning and Inference for a Sequential Accelerated Life Test, Journal of Quality Technology **42**, No.1 (2010).

L. C. Tag, Y. S. Sun, T. N. Goh, and H. L. Ong, Planning multiple-level constant stress accelerated life tests, Proceedings of the Annual Reliability and Maintainability Symposium (2002) 338- 342.

B. Tang. OA-based Latin hypercubes. IIQO Research Report RR-91-10 (1991). The Institute for Improvement in Quality and Productivity, University of Waterloo. Submitted to JASA.

B. Tang. Orthogonal Array-Based Latin hypercubes. J. of the American Statistical Association Vol. **88** (1993), No. 424, pp. 1392-1397.

P. A. Tobias, and D. C. Trindade, Applied Reliability, 2nd ed., New York: Van Nostrand Reinhold (1995).

D. Ucinski and B. Bogacka, T-optimum design for discrimination between two multi-response dynamic models, Journal of the Royal Statistical Society: Statistical Methodology **67**(B) (2005), 3-18.

A. Ullah, Entropy, divergence and distance measures with econometric applications, Journal of Statistical Planning and Inference **49** (1996), 137-162.

I. Verdinelli, N. G. Polson, and N. D. Singpurwalla, Shannon Information and Bayesian Design for Prediction in Accelerated Life Testing, in Reliability and Decision Making, eds. R. E. Barlow, C. A. Clariotti, and F. Spizzichino, London: Chapman & Hall, (1993) pp. 247–256.

G. B. Wetherill and K. D. Glazebrook, Sequential Methods in Statistics, 3rd edition. London, UK: Chapman and Hall (1986).

P. Whittle, Some General Points in the Theory of Optimal Experimental Design, Journal of the Royal Statistical Society, Ser. B, **35** (1973), 123–130.

K. Q. Ye. Column orthogonal Latin hypercubes and their application in computer experiments. J. Amer. Statist. Assoc. **93** (1998) 1430-1439.

K. Q. Ye, W. Li, and A. Sudjianto. Algorithmic construction of optimal symmetric Latin hypercube designs. J. of Statistical Planning and Inference **90** (2000) 145-159.

Y. Zhang, and W. Q. Meeker, Bayesian Methods for Planning Accelerated Life Tests, Technometrics, Vol. **48** (1), (2006).