ROBUST AND OPTIMAL DESIGN STRATEGIES FOR NONLINEAR MODELS USING GENETIC ALGORITHMS

by

Sydney Kwasi Akapame

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Statistics

MONTANA STATE UNIVERSITY
Bozeman, Montana

April, 2014

DEDICATION

I dedicate this dissertation to my parents, Clement Akapame and Nina Crabbe, and all my loved ones.

ACKNOWLEDGEMENTS

I would like to give the utmost thanks to God who has faithfully guided and helped me throughout my studies at Montana State University. It was a long journey but he kept me steadfast. For this, I am most grateful.

I cannot thank my advisor, Dr. John Borkowski, enough for his patient guidance and overall excellent personality which made my research experience enjoyable. John, it was such a pleasure working with you. You introduced me to so many new things - optimal designs and genetic algorithms - and you were never economical with a kind word! I am very thankful that you believed in me enough to work with me.

Megan, your attention to detail and willingness to help me with my questions whenever I walked into your office cannot be overlooked! You graciously read the first draft of my dissertation and brought important issues to my attention. You played no mean part in improving the quality of this dissertation. Thanks for serving on my committee.

Mark, it is in your generalized linear models (GLMs) class that I actually started working on my research! You got John to get me to do a paper on optimal designs for GLMs as my project. I cannot over-emphasize how useful that was! You have also contributed considerably to my writing as a statistician. Thanks for serving on my committee.

My gratitude goes to all my other committee members: Jim Robison-Cox, Steve Cherry and Prasanta Bandyopadhyay. Working with you has been such a pleasure. Your invaluable comments made this dissertation possible.

Finally, I am grateful to Rejoice, Josie and Becky and my loving church family.

TABLE OF CONTENTS

TABLE OF CONTENTS – CONTINUED

TABLE OF CONTENTS – CONTINUED

LIST OF TABLES

Table Page

LIST OF TABLES – CONTINUED

Table Page

LIST OF FIGURES

Figure Page

LIST OF FIGURES – CONTINUED

Figure Page

LIST OF FIGURES – CONTINUED

Figure Page

LIST OF FIGURES – CONTINUED

Figure Page

LIST OF FIGURES – CONTINUED

Figure Page

ABSTRACT

Experimental design pervades all areas of scientific inquiry. The central idea behind many designed experiments is to improve or optimize inference about the quantities of interest in a statistical model. Thus, the strengths of any inferences made will be dependent on the choice of the experimental design and the statistical model. Any design that optimizes some statistical property will be referred to as an optimal design. In the main, most of the literature has focused on optimal designs for linear models such as low-order polynomials. While such models are widely applicable in some areas, they are unsuitable as approximations for data generated by systems or mechanisms that are nonlinear. Unlike linear models, nonlinear models have the unique property that the optimal designs for estimating their model parameters depend on the unknown model parameters. This dissertation addresses several strategies to choose experimental designs in nonlinear model situations.

Attempts at solving the nonlinear design problem have included locally optimal designs, sequential designs and Bayesian optimal designs. Locally optimal designs are optimal designs conditional on a particular guess of the parameter vector. Although these designs are useful in certain situations, they tend to be sub-optimal if the guess is far from the truth. Sequential designs are based on repeated experimentation and tend to be expensive. Bayesian optimal designs generalize locally optimal designs by averaging a design optimality criterion over a prior distribution, but tend to be sensitive to the choice of prior distribution. More importantly, in cases where multiple priors are elicited from a group of experts, designs are required that are robust to the class (or range) of prior distributions. New robust design criteria to address the issue of robustness are proposed in this dissertation. In addition, designs based on axiomatic methods for pooling prior distributions are obtained.

Efficient algorithms for generating designs are also required. In this research, genetic algorithms (GAs) are used for design generation in the MATLAB® computing environment. A new genetic operator suited to the design problem is developed and used. Existing designs in the published literature are improved using GAs.

CHAPTER 1

INTRODUCTION

The underlying mechanisms that generate data for most physical or chemical processes are often inherently nonlinear. To ease data analysis, researchers have often resorted to linearizing the usually complex nonlinear models and then using ordinary least squares techniques to obtain parameter estimates for inference. However, the widespread use of novel optimization methods and increasing access to high-end computing resources make it possible for researchers to directly use nonlinear models in data analysis when relevant. Nonlinear models arise frequently in the physical and biological sciences and as a result, the design of efficient experiments when planning to fit nonlinear statistical models is of great interest to researchers in these areas.

Designs optimal for specific experimental objectives have been in the literature for decades. For example, optimal designs for prediction purposes entered the literature in 1918 (Smith, 1918). Designs for other experimental objectives including, but not limited to, parameter estimation, model discrimination and lack-of-fit have since been widely studied. The majority of the work has been in the context of linear models. Work on nonlinear models is relatively new, with the earliest work credited to Box and Lucas (1959). The scarcity of work on experimental design for nonlinear models in light of their usefulness makes research in the area imperative.

The optimal design problem in terms of design of experiments for linear models differs substantially from that of nonlinear models. This has perhaps led to the disproportionate amount of work in favor of linear models. To illustrate the problem, consider the following model

$$\boldsymbol{y} = \boldsymbol{\eta}(\boldsymbol{x}, \boldsymbol{\theta}) + \boldsymbol{\epsilon} \tag{1.1}$$

where $\boldsymbol{\eta}(\boldsymbol{x}, \boldsymbol{\theta})$ is the expectation function and $\boldsymbol{\epsilon}$ is a zero-mean, constant variance error vector. If $\boldsymbol{\eta}(\boldsymbol{x}, \boldsymbol{\theta})$ is linear in parameter vector $\boldsymbol{\theta}$, then $\boldsymbol{\eta}(\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{X}\boldsymbol{\theta}$ for model matrix $\boldsymbol{X}$. The result of this is that the Fisher information matrix for $\boldsymbol{\theta}$ will not depend on the unknown $\boldsymbol{\theta}$, hence the optimal experimental design is not a function of $\boldsymbol{\theta}$. If $\boldsymbol{\eta}(\boldsymbol{x}, \boldsymbol{\theta})$ is nonlinear in $\boldsymbol{\theta}$, then the Fisher information matrix and, hence, the optimal design is a function of $\boldsymbol{\theta}$. This dependence of the design on $\boldsymbol{\theta}$ in nonlinear situations poses a problem because $\boldsymbol{\theta}$ is unknown. In other words, prior knowledge of $\boldsymbol{\theta}$ is required in order to design optimal experiments to estimate $\boldsymbol{\theta}$. This is sometimes called the parameter-dependency problem.

Some approaches to addressing the parameter-dependency problem have been proposed in the literature. Locally optimal designs were introduced by Box and Lucas (1959) who argued that prior knowledge of $\boldsymbol{\theta}$ is always available in practical situations. Locally optimal designs are optimal with respect to a particular guess of $\boldsymbol{\theta}$ and therefore tend to be sub-optimal if the prior knowledge or guess is further from the truth. Making an initial guess of $\boldsymbol{\theta}$ and then alternating the processes of design, experimentation and analysis until some pre-specified termination point is reached is the objective of another approach called sequential experimental design. This may be ideal but infeasible if data collection is expensive, for example.

A natural approach to the problem is the Bayesian approach which involves specifying a prior distribution for $\boldsymbol{\theta}$ and averaging the optimality criterion over it. The problem with the Bayesian paradigm is the sensitivity of the resulting optimal design to the choice of prior distribution used in design construction. This problem is similar to that of locally optimal designs when $\boldsymbol{\theta}$ is misspecified. In addition, if more than one prior is plausible for $\boldsymbol{\theta}$, then it is desirable to have a design that is robust to the specified prior distributions. Thus, the need for robustness of design is more important than optimality of designs in the context of nonlinear models. The problem of

multiple priors has been investigated to some extent for linear models (e.g., Toman (1992), Toman and Gastwirth (1993) and DasGupta and Studden (1991)), but not for nonlinear models.

In addition to the issue of robustness in nonlinear design is the issue of availability of efficient algorithms for generating designs. A survey of the literature shows that the reluctance of most researchers to use optimal designs can be traced to the unavailability of easily implementable algorithms. Thus, efficient algorithms are required to allow practical implementation of the methods.

The objective of this dissertation is two-fold: (1) to address issues of robustness related to the design of experiments for nonlinear models and, (2) to provide efficient genetic algorithms for design generation. As a result, Chapter 2 reviews response surface methodology designs, optimal experimental designs and some algorithmic methods for generating designs. It also presents practical examples of design optimality criteria for linear models. Notation used throughout the dissertation is also introduced.

Chapter 3 introduces new robust design criteria for nonlinear models. It also discusses aggregation methods for prior distributions in light of their applicability to robustness of design. A new reproduction operator to speed the search of a genetic algorithm for design generation is also introduced. Implementation of the new robust criteria is also discussed.

Chapter 4 presents several examples of improvements to existing designs in terms of commonly used design optimality criteria which are obtained using genetic algorithms that implement the new reproduction operator. Chapter 5 presents applications of the new robust design criteria in Chapter 3 to the one-compartment and Michaelis-Menten models used in pharmacokinetics and enzyme kinetics respectively. It is assumed in the applications that parameter estimation is of interest.

Concluding remarks and a discussion of future research is the subject of Chapter 6. The MATLAB® code used for generating the designs in this dissertation is found in Appendix C.

CHAPTER 2

LITERATURE REVIEW

1. Introduction

Response surface methodology (RSM) deals with the exploration and optimization of response surfaces. Consider the case where the response is y and there is a set of predictor variables $x_1, x_2, ..., x_k$. In some instances, the relationship between y and $X = \{x_1, x_2, .., x_k\}$ may be known exactly based on the underlying engineering, chemical or physical principles. As a result, the model of interest can be written in the form $y = g(x_1, x_2, .., x_k) + \epsilon$, where ϵ represents the error in the system. This type of relationship is often called a mechanistic model. In most situations, however, the exact relationship between y and x is unknown and so an empirical model $y = f(x_1, x_2, .., x_k) + \epsilon$ is estimated yielding $\widehat{y} = \widehat{f}(x_1, x_2, .., x_k)$. The empirical model is called a **response surface model**. For example, suppose the following true mechanistic model is unknown, assuming $E(\epsilon) = 0$,

$$E(y) = \exp(0.5x_1 - 1.5x_2) + 5.$$

A designed experiment produced data leading to fitting the following approximating second-order model

$$\widehat{y} = 5.89 + 0.98x_1 - 2.38x_2 - 1.09x_1x_2 + 0.28x_1^2 + 1.41x_2^2.$$

The response surfaces for $E(y)$ and $\widehat{y}$ are in Figure 2.1. The two response surfaces are almost indistinguishable. A closer look suggests that the true model has a maximum that is slightly higher than the approximating model. The maxima for both models

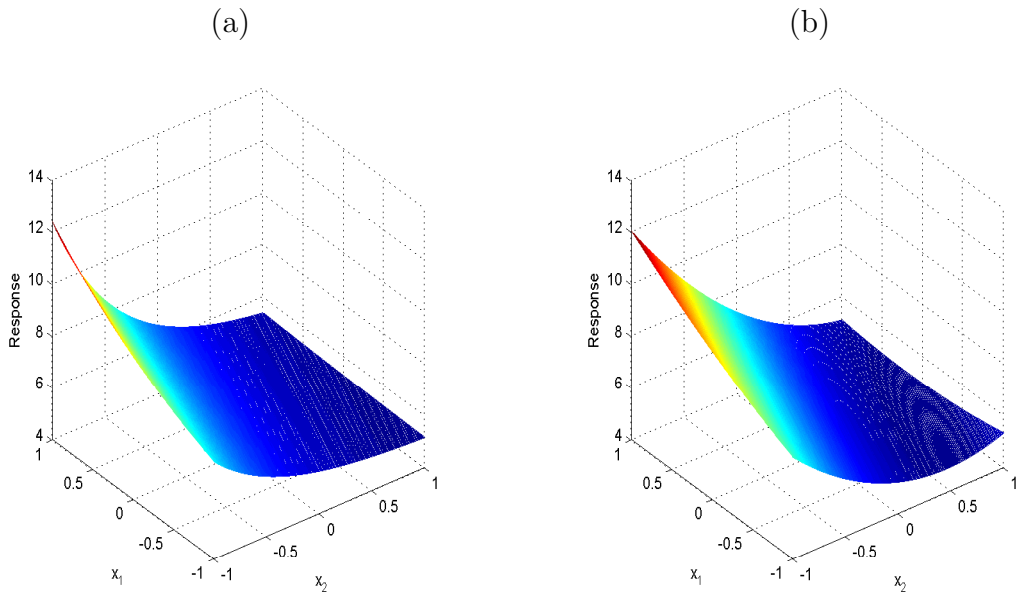

Figure 2.1: Plots of the true response surface (a) and the response surface for the approximating function (b) respectively.

occurs at $(x_1, x_2) = (-1, 1)$. Collecting data efficiently to fit the approximating function is central to the practice of RSM.

1.1. Goals of RSM

The primary goals of RSM (Myers, Montgomery, and Anderson-Cook, 2009) can be summarized as follows:

- Developing an experimental strategy for exploring the space of the process or independent variables with respect to a response of interest,
- Empirical statistical modeling to develop an appropriate approximating relationship between response and process variables and,
- Finding the levels or values of the process variables to optimize desirable values of the responses, such as maxima, minima, or specific target values.

1.2. Major Applications of RSM

RSM has been widely used for solving problems in many fields including industrial engineering, and the biological and social sciences (Myers, Khuri, and Carter, 1989). These are discussed briefly below:

- Industrial Engineering Applications: The use of RSM in industry is motivated by the quest for quality. Response surface designs such as central composite designs, Box-Behnken designs, and fractional factorial designs are widely used in industry. Applications vary from polymer optimization to the exploration of a detergent system. RSM design and data analysis are used to obtain the general vicinity of best operating conditions within a region of interest (Box, 1957). Various industrial-pollution studies have employed response surface methodology. For example, Huck, Murphy, Reed, and LeClair (1977) determined the polymer properties and mixing conditions required to produce optimal flocculation for mine waters of specified strengths containing iron, zinc, and copper either singly or in combination. Wallis (1978) reports the use of RSM in studies related directly to power station cooling systems.

- Biological Applications: RSM techniques have been found useful for studying the relationship between the chemical structure of a compound and its biological activity. Mager (1982a) and Mager (1982b) studied the structure-neurotoxicity relationship of organophosphorus pesticides and used a canonical analysis of the fitted equation to elucidate properties of the response surface. Dincer and Ozdurmus (1977) used the method of steepest descent to determine the most suitable combination of four independent formulation and process variables for the disintegration time of coated tablets in simulated intestinal fluid. Carter, Wampler, and Stablein (1983) have used RSMs to elucidate the actions and

interactions of cytotoxic drugs in combination and to estimate the optimal levels of each drug for the treatment of cancer with and without side-effect constraints. Belloto, Dean, Moustafa, Molokhia, Gouda, and Sokoloski (1985) used RSM to study the solubility of pharmaceutical formulations. Maddox and Richert (1977) and Shek, Ghani, and Jones (1980) demonstrate other uses.

- Social Science Applications: Economics, operations research and system simulation are just a few of the areas that have benefited immensely from RSM. Shechter and Heady (1970) used response surface techniques to design and analyze experiments from a simulation model dealing with the feed-grain program. Biles (1975) illustrates the use of RSM techniques in inventory management. Montgomery and Bettencourt (1977) provide an example in which a simulation of a military tank duel is analyzed to ascertain the values of two design variables that will optimize four dependent variables simultaneously. They used a nonlinear programming technique to analyze data taken from a rotatable central composite design (CCD).

2. Brief Overview of Classical RSM Designs

2.1. Orthogonal Designs (2^k Designs)

Factorial designs are widely used in experiments involving several factors where it is necessary to investigate the joint effects of the factors on a response variable. An important and common case occurs when each of the k factors has exactly 2 levels. A factorial design involving these two-level factors is referred to as a 2^k factorial design because each replicate of the design has exactly 2^k experimental runs or trials. 2^k factorial designs are very important in response surface work for the following reasons (Myers et al., 2009):

- A 2^k design is useful at the start of a response surface study as a screening experiment to identify the critical or important process or system variables.

- In a response surface study where the maxima or minima or a process is desired, 2^k designs can often be used to fit first-order response surface models and to generate the factor effect estimates required to perform the evolutionary operation methods of steepest ascent (for a maximum) or descent (for a minimum), as well as models including interaction effects.

- The 2^k design forms the basic building block to create other response surface designs. For example, augmenting a 2^k design with axial runs and center points results in a central composite design (CCD) which is one of the most important designs for fitting second-order response surface models.

An example of a 2^2 design for a two-predictor model with the levels of the explanatory variables coded as ± 1 is given below. $\boldsymbol{D}$ is the design matrix for a first-order model with two predictors.

$$\boldsymbol{D} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \\ -1 & -1 \\ -1 & 1 \end{bmatrix} \tag{2.1}$$

2.2. 2^{k-p} Fractional Factorial Designs

The number of runs required for a 2^k factorial design exponentially outgrows the resources of the experimenter as the number of factors increases. If the experimenter can reasonably assume that certain higher-order interactions are negligible, then information on the main effects and low-order interactions can be obtained by running only a fraction of the complete factorial experiment. A design containing a

subset of the factor level combinations of a full factorial is called a fractional factorial design (Finney, 1943). Fractional factorial designs are especially useful for screening experiments where the goal is to identify the most important factors among a large set of factors. The successful use of two-level fractional factorial designs is based on three main ideas:

- The sparsity of effects principle: when there are many variables under consideration, it is typical for the system or process to be dominated by main effects and low-order interactions.
- The projective property: a fractional factorial design can be projected to stronger designs in a subset of the significant factors.
- Sequential experimentation: It is possible to combine the runs from two or more fractional factorial designs to sequentially form a larger design to estimate the factor effects and interactions of interest.

A 2^{k-p} fractional factorial design is a 2^pth fraction of a 2^k factorial design where p is a positive integer less than k. For example, a 2^{k-1} fractional factorial design is $\frac{1}{2}$ of a 2^k factorial design. The design can be generated by aliasing the highest order interaction with the intercept. An example of a 2^{4-1} fractional factorial design, that is a $\frac{1}{2}$ fraction of a 2^4 design is given below. The design is generated by aliasing the highest order interaction, $x_1x_2x_3x_4$, with the intercept. That is, for any row, the product of the $x_1, x_2, x_3,$ and x_4 columns is 1.

$$\boldsymbol{D} = \begin{bmatrix} -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}. \tag{2.2}$$

For the design above, a model containing a subset of two-factor interactions can be fit. Two-factor interactions are the highest order interactions that can be fit due to the fact that the three-factor interactions are *aliased* with the main effects. That is, the product $x_i x_j x_k$ of any three columns equals the remaining fourth column. A design is of resolution R if no m-factor effect is aliased with another effect containing less than $R - m$ factors. Thus, this design is a resolution IV design. Refer to Myers et al. (2009) for a detailed discussion of aliasing and design resolution.

2.3. Plackett-Burman Designs

Plackett-Burman designs (Plackett and Burman, 1946) are a special class of 2-level fractional factorial designs for studying a maximum of $k = N - 1$ factors in N experimental runs, where N is a multiple of 4. If N is a power of 2, these designs are resolution III fractional factorial designs. Most Plackett-Burmann designs have complex aliasing structures and are recommended only for screening experiments.

2.4. Designs for the Second-Order Model

Variable screening, an essential phase of RSM, makes extensive use of two-level factorials and their fractions. The experimenter, however, may also be interested in fitting a second-order response surface model in the design variables $x_1, x_2, .., x_k$ as an approximation to the unknown mechanistic model. This response surface analysis may involve optimization through the use of a ridge anlysis or a canonical analysis. Regardless of the form of the analysis, the experimental design should allow the experimenter to fit the second-order model

$$E(y) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i<j} \sum \beta_{ij} x_i x_j. \tag{2.3}$$

The model in (2.3) has $P = 1 + 2k + k(k-1)/2 = (k+1)(k+2)/2$ parameters because of an intercept, k first-order terms, k quadratic terms, and $k(k-1)/2$ two-factor interactions. Thus, there must also be at least $N = P$ points and at least 3 levels of each design variable because a design variable with only two-levels will inevitably result in a rank-deficient model matrix. In the case of first-order designs, the dominant desirable design property is orthogonality. However, orthogonality ceases to be an issue for second-order designs, and while estimation of individual coefficients is still important, it becomes secondary to the properties of the scaled or standardized prediction variance. This stems from the fact there is often less concern with what variables belong in the model than with the quality of $\widehat{y}(x)$ as a predictor or, equivalently as, an estimator for $E(y)$.

2.4.1. The Class of Central Composite Designs (CCDs):

CCDs are the most popular class of second-order designs and were introduced by Box and Wilson (1951).

Much of the motivation for the CCD evolves from its use in sequential experimentation (Myers et al., 2009). Assuming $k \geq 2$ design variables, the CCD consists of:

i. A 2^k full factorial or a 2^{k-p} fractional factorial design of at least resolution V. Each point has the form $(x_1, x_2, ..., x_k) = (\pm 1, \pm 1, .., \pm 1)$,

ii. $2k$ axial points of the form $(x_1, .., x_i, ., x_k) = (0, .., \pm\alpha .., 0)$ for $1 \leq i \leq k$, and

iii. n_c center points $(x_1, x_2, ..., x_k) = (0, 0, .., 0)$.

If $\alpha = 1$ for the axial points, then the design is referred to as a *face-centered cube* design. Each of the three types of points in a CCD play different roles. The factorial points allow estimation of the first-order and interaction terms. Axial points allow the estimation of the squared terms and, the center points provide an internal estimate of pure error used to test for lack of fit (when replicated) and also contribute toward estimation of the squared terms.

The structure of a CCD is given below with $\alpha = \sqrt{2}$ for 2 design variables and $n_c = 1$ center point.

$$\boldsymbol{D} = \begin{bmatrix} \pm\mathbf{1} & \pm\mathbf{1} \\ \pm\sqrt{\mathbf{2}} & \mathbf{0} \\ \mathbf{0} & \pm\sqrt{\mathbf{2}} \\ 0 & 0 \end{bmatrix}. \tag{2.4}$$

2.4.2. Box-Behnken Designs (BBDs): Box and Behnken (1960) introduced this class of experimental designs for second-order models. Given $k \geq 3$ design variables, most BBDs are constructed by combining two-level factorial or fractional-factorial designs with *balanced incomplete block designs* or BIBDs. Every balanced incomplete block design (and hence the BBD considered) is associated with the following design parameters:

k = number of design variables,

b = number of blocks in the BIBD,

t = number of design variables per block,

r = number of blocks in which a design variable appears, and

$\lambda = \frac{r(t-1)}{b-1}$ is the number of blocks each pair of design variables appears in the design.

The algorithm for constructing a BBD is the following:

1. The t columns defining a 2^t factorial design with levels ± 1 replace the t design variables appearing in each block of the BIBD,

2. The remaining $k - t$ columns are set to 0, and

3. The design is augmented with n_c mid-level center points $(0, ..., 0)$.

An example of a design matrix from a BBD with $k = 3$ design variables, $n_c = 1$ center point, and generated from a BIBD with $b = 3$ blocks and $t = 2$ treatments per block is given below:

$$\boldsymbol{D} = \begin{bmatrix} \pm\mathbf{1} & \pm\mathbf{1} & \mathbf{0} \\ \pm\mathbf{1} & \mathbf{0} & \pm\mathbf{1} \\ \mathbf{0} & \pm\mathbf{1} & \pm\mathbf{1} \\ 0 & 0 & 0 \end{bmatrix}. \tag{2.5}$$

The total size of a BBD is $N = fkr/t + n_c = fb + n_c$, where $f = 2^t$. For the BBD shown above, $N = 13$. Using the same strategy, BBDs can be constructed for $k = 4, 5$. For $k > 5$ design variables, the construction of the design may be based on combining fractional factorial designs with *partially balanced incomplete block designs* and using fractional factorial designs. In this case, each treatment does not have to occur the same number of times with every other treatment. Myers et al. (2009)

provide a discussion of the cases where $k = 6$ and $k = 7$. A BBD also has two interesting characteristics:

1. It is nearly rotatable and, for $k = 4$ and $k = 7$, it is exactly rotatable. A design is rotatable if the scaled prediction variance has the same value at any two locations that are the same distance from the design center, and

2. It is a spherical design. That is, the experimental region is assumed to be spherical. In the case where $k = 3$, all the points are the midpoints of the edges of the cube and there are no factorial points or cube face points. This contrasts sharply with the face-centered cube CCD (Myers et al., 2009) which gives a good coverage of the cube. This suggests that the use of the BBD should be confined to situations in which the experimenter is not interested in predicting response at the extremes (that is, the corners of the design region).

3. Prediction Variance Properties of Response Surface Designs

Variance-optimal designs are designs that produce estimates of the model parameters with minimum variance. More often than not, the prediction variance property of a design is of critical importance. Consider the linear model

$$y = X\beta + \epsilon \tag{2.6}$$

where y is the $n \times 1$ vector of responses, X is the $n \times p$ model matrix, β is a $p \times 1$ vector of model parameters, and ϵ is an $n \times 1$ error vector. The ordinary least squares (OLS) estimator of the parameter vector β is

$$\widehat{\beta} = (X^T X)^{-1} X^T y \tag{2.7}$$

and, assuming homogeneity of error variance,

$$Var(\widehat{\beta}) = \sigma^2(X^TX)^{-1}. \tag{2.8}$$

Suppose that the prediction of the response is desired at a particular point, $x = (x_1, ..., x_k)$. That is, x is the vector of the design variables at which prediction is desired. Let $f(x)$ be the model vector formed by expanding x to contain the P terms associated with the model parameters in β. The *prediction variance* (PV) at point x is

$$Var(\hat{y}(x)) = \sigma^2 f^T(x)(X^TX)^{-1}f(x) = PV(x). \tag{2.9}$$

Three of the most important implications of the definition are

1. $Var(\hat{y}(x))$ varies from location to location in the design space,

2. $Var(\hat{y}(x))$ depends on the choice of model, and

3. $Var(\hat{y}(x))$ depends on the choice of the experimental design.

In design comparison studies, a scaled prediction variance, denoted $SPV(x)$, which takes the sample size into account is often used. It is defined as:

$$SPV(x) = \frac{NVar(\hat{y}(x))}{\sigma^2} = Nf^T(x)(X^TX)^{-1}f(x)$$

Division by σ^2 makes $SPV(x)$ scale-free and multiplication by N allows it to reflect variance on a per observation basis. That is, if two designs are being compared, scaling by N penalizes the design with the larger design size.

3.1. Prediction Variance Examples

Example 1. Consider the $N = 9$ point CCD with $\alpha = \sqrt{2}$, and 1 center point. The model matrix X and information matrix X^TX are

$$X = \begin{bmatrix} 1 & -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & \sqrt{2} & 0 & 0 & 2 & 0 \\ 1 & -\sqrt{2} & 0 & 0 & 2 & 0 \\ 1 & 0 & \sqrt{2} & 0 & 0 & 2 \\ 1 & 0 & -\sqrt{2} & 0 & 0 & 2 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \Rightarrow X^TX = \begin{bmatrix} 9 & 0 & 0 & 0 & 8 & 8 \\ 0 & 8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 \\ 8 & 0 & 0 & 0 & 12 & 4 \\ 8 & 0 & 0 & 0 & 4 & 12 \end{bmatrix}$$

If $x = (x_1, x_2)$, then $f^T(x) = [1 \quad x_1 \quad x_2 \quad x_1x_2 \quad x_1^2 \quad x_2^2]$, and the prediction variance

$$PV(x) = \sigma^2 f^T(x)(X^TX)^{-1}f(x) = \sigma^2\left(1 - \frac{7}{8}\rho^2 + \frac{11}{32}\rho^4\right), \text{ where } \rho = \sqrt{x_1^2 + x_2^2}. \tag{2.10}$$

This design is rotatable because it is a function solely of the distance ρ.

Example 2. Consider the BBD with $k = 3, \lambda = 2, n_c = 3$ and $N = 15$ points. The model matrix X is

$$\begin{bmatrix} \mathbf{1} & \mathbf{\pm 1} & \mathbf{\pm 1} & \mathbf{0} & \mathbf{\pm 1} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{\pm 1} & \mathbf{0} & \mathbf{\pm 1} & \mathbf{0} & \mathbf{\pm 1} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{1} \\ \mathbf{1} & \mathbf{0} & \mathbf{\pm 1} & \mathbf{\pm 1} & \mathbf{0} & \mathbf{0} & \mathbf{\pm 1} & \mathbf{0} & \mathbf{1} & \mathbf{1} \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Following the same steps as performed for the CCD, but with $x = (x_1, x_2, x_3)$,

$$f^T(x) = [1 \quad x_1 \quad x_2 \quad x_3 \quad x_1x_2 \quad x_1x_3 \quad x_2x_3 \quad x_1^2 \quad x_2^2 \quad x_3^2]$$

and a 10×10 $(X^TX)^{-1}$ matrix, the prediction variance function is given by

$$Var(\hat{y}(\boldsymbol{x})) = \sigma^2 \left(\frac{1}{3} - \frac{5}{24}\rho_1 + \frac{13}{48}\rho_2 + \frac{7}{24}\rho_3 \right)$$

where $\rho_1 = x_1^2 + x_2^2 + x_3^2$, $\rho_2 = x_1^4 + x_2^4 + x_3^4$, and $\rho_3 = x_1^2x_2^2 + x_1^2x_3^2 + x_2^2x_3^2$. This design is not rotatable because it is not solely a function of the distance $\sqrt{\rho_1}$.

4. Design Optimality Criteria for Linear Models

The decision of which experimental design to run, given a set of candidate designs, is critical to realizing research goals. Orthogonality of the design matrix, which ensures that the parameter estimates for a particular model are uncorrelated, was of great importance upon the development of the first full and fractional factorial experimental designs. Orthogonality, as well as balance and estimability, continued to be the design criteria of choice until the development of response surface methodology. Orthogonality, although desirable, may be impracticable, unfortunately. For a second-

order model to be fit, a response surface design requires at least three levels of variable settings. To require orthogonality for first-order (x_i) and interaction ($x_i x_j$) terms may also require a large number of experimental runs, specifically, a total of at least 3^k design points for k design variables, while orthogonality will be lost for the second-order (x_i^2) terms.

The impracticality of the large design sizes and loss of orthogonality led to the introduction of alternative criteria for comparing and evaluating response surface designs in the **optimal design theory** work by Kiefer (1959, 1961) and Kiefer and Wolfowitz (1959). Design optimality criteria are primarily concerned with *optimal properties* of the information matrix $X^T X$. By studying the optimality criteria, the experimenter can determine the adequacy of a proposed experimental design prior to running it. Note that optimality criteria based on $X^T X$ are model dependent. Several of the most popular and commonly-used design optimality criteria in the literature are discussed below. A design will be represented by the probability measure ξ on a finite support $\mathcal{X}$. A probability measure is a real-valued function whose total mass is 1, that is, $\int_{\mathcal{X}} \xi \, \mathrm{d}x = 1$. Thus, the mass assigned to any support point need not be rational, implying the design is not implementable in practice. In the discussion of optimality criteria given below, only implementable N-point (i.e, exact) designs are considered. A more thorough discussion of exact designs is presented in Section 2.

4.1. D-optimality Criterion

The D-criterion of a design, ξ, is

$$D(\xi) = |X^T X|. \tag{2.11}$$

The D-optimality criterion is the most widely used in the literature. The D-optimum design ξ maximizes the determinant $|X^TX|$ or equivalently, minimizes $|(X^TX)^{-1}|$. D-optimal designs focus on efficient parameter estimation. As a result, maximizing $|X^TX|$ leads directly to minimizing the diagonal and off-diagonal elements of $(X^TX)^{-1}$ which are, respectively, directly proportional to the variances and covariances of the parameter estimates. D-optimum designs minimize the generalized variance $(X^TX)^{-1}$ of the parameter estimates. Formally, the D-optimum design ξ_D is defined as

$$\xi_D = \arg\max_{\xi} |X^TX| = \arg\max_{\xi} D(\xi). \tag{2.12}$$

4.2. A-optimality Criterion

An A-optimum design minimizes the trace of $(X^TX)^{-1}$. That is, the A-criterion for a design ξ is

$$A(\xi) = tr(X^TX)^{-1}. \tag{2.13}$$

Thus, A-optimal designs focus on minimizing the sum or average of the variances of the parameter estimates. The A-optimality criterion differs from D-optimality in the sense that A-optimal designs focus only on the variances of estimates and not their covariances. Thus, the A-optimal design, ξ_A, is defined as

$$\xi_A = \arg\min_{\xi} \left\{tr(X^TX)^{-1}\right\} = \arg\min_{\xi} A(\xi). \tag{2.14}$$

4.3. G-optimality Criterion

The primary goal of many designed experiments is to allow for efficient prediction throughout the design space R. G-optimal designs minimize the maximum prediction variance or scaled prediction variance over the design region. They seek to protect

against the worst case prediction variance. Formally, the G-criterion of a design ξ is given by

$$G(\xi) = \max_{x \in R} N f^T(x)(X^T X)^{-1} f(x). \tag{2.15}$$

The G-optimal design, ξ_G, is then

$$\xi_G = \arg\min_{\xi} \max_{x \in R} \left\{ N f^T(x)(X^T X)^{-1} f(x) \right\} = \arg\min_{\xi} G(\xi). \tag{2.16}$$

4.4. IV-optimality Criterion

The IV-optimality criterion also addresses properties of the prediction variance. The Integrated Variance (IV) optimal designs minimize the average scaled prediction variance over the design space R. Averaging is accomplished via integration over R. The IV-criterion is given formally by

$$IV(\xi) = \frac{1}{A} \int_R N f^T(x)(X^T X)^{-1} f(x) \, \mathrm{d}x \tag{2.17}$$

where A is the volume of the design region R. Thus the IV-optimal design is

$$\xi_{IV} = \arg\min_{\xi} \frac{1}{A} \int_R N f^T(x)(X^T X)^{-1} f(x) \, \mathrm{d}x = \arg\min_{\xi} IV(\xi). \tag{2.18}$$

4.5. E-optimality Criterion

Often, the objective of the experimenter is to minimize the volume of the confidence ellipsoid which is achieved by a D-optimal design. A long, thin ellipsoid, besides indicating that some parameters are imprecisely estimated, also indicates that some linear combinations of the parameters will be estimated poorly. To address this problem, E-optimal designs attempt to minimize the imprecision associated with

these linear combinations. If $\lambda_i, \;\; i = 1, .., p$, are the eigenvalues of $(X^T X)^{-1}$, then the E-criterion of a design ξ is

$$E(\xi) = \max_i \lambda_i. \tag{2.19}$$

The E-optimal design, ξ_E, is then

$$\xi_E = \arg \min_{\xi} \max_i \lambda_i \tag{2.20}$$

Interest is in minimizing the maximum eigenvalue because it is directly related to the longest axis of the confidence ellipsoid.

4.6. Subset D-optimality Criterion

A D_S-optimal design (Hill and Hunter, 1974) is appropriate when primary interest is not in the complete set of p model parameters, but only a subset of s ($s < p$) parameters. The terms of the model can be divided into two groups (Atkinson, Donev, and Tobias, 2007):

$$E(Y) = f_1^T(x)\beta_1 + f_2^T(x)\beta_2 \tag{2.21}$$

where β_1 is the $s \times 1$ parameter vector of interest. The elements of the $(p - s) \times 1$ parameter vector β_2 are treated as nuisance parameters. A typical application of this optimality criterion occurs when an experiment is designed to check the goodness of fit of a model. The tentative model with terms $f_2(x)$ is embedded in the more general model which also includes $f_1(x)$ terms. In order to test whether the simpler model is adequate, β_1 must be estimated with minimum variance, providing the most powerful test of $\beta_1 = 0$. Atkinson et al. (2007) provide an expression for the variance of β_1

for a design, ξ. First, we will call the information matrix of the more general (full) model $M(\xi)$. This is partitioned as

$$M(\xi) = \begin{bmatrix} M_{11}(\xi) & M_{12}(\xi) \\ M_{12}^T(\xi) & M_{22}(\xi) \end{bmatrix}. \tag{2.22}$$

Here, $M_{11}(\xi)$ is the portion of the information matrix that corresponds to β_1. The criterion is given by

$$D_S(\xi) = \frac{|M(\xi)|}{|M_{22}(\xi)|}$$

The expression for the variance function is

$$d_s(x, \xi) = f^T(x)M^{-1}(\xi)f(x) - f_2^T(x)M_{22}^{-1}(\xi)f_2(x). \tag{2.23}$$

The D_S-optimal design ξ_{Ds} is defined as

$$\xi_{Ds} = \arg\max_{\xi} D_S(\xi) \text{ and } d_s(x, \xi_{Ds}) \leq s. \tag{2.24}$$

4.7. T-optimality Criterion

T-optimal designs (Atkinson and Fedorov, 1975) are used to discriminate between models. Consider two models, $\eta_1(x, \theta_1)$ and $\eta_2(x, \theta_2)$ where the former model is assumed to be the true or data-generating model. Both models, for instance, could be suitable for modeling the decay of a chemical substance. For the η_1 model, θ_1 is assumed to be known, so that $\eta_1(x, \theta_1) = \eta_1(x)$. The T-criterion for an N-point design ξ is given by

$$T(\xi) = \sum_{i=1}^{N} \left(\eta_1(x_i) - \eta_2(x_i, \widehat{\theta}_2)\right)^2, \tag{2.25}$$

and the T-optimal design, ξ_T, is defined as

$$\xi_T = \arg\max_{\xi} \sum_{i=1}^{N} \left\{ \left(\eta_1(x_i) - \eta_2(x_i, \widehat{\theta}_2) \right)^2 \right\} = \arg\max_{\xi} \left\{ T(\xi) \right\}. \tag{2.26}$$

4.8. Numerical Examples

In this section, practical applications of several optimality criteria are presented. For example, consider a 3^2 design and the interaction model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon.$$

The model matrix X, $X^T X$, and $(X^T X)^{-1}$ for the 3^2 design are

$$X = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad X^T X = \begin{bmatrix} 9 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix} \quad (X^T X)^{-1} = \begin{bmatrix} \frac{1}{9} & 0 & 0 & 0 \\ 0 & \frac{1}{6} & 0 & 0 \\ 0 & 0 & \frac{1}{6} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \end{bmatrix}.$$

Note that a 3^2 design is a face-centered central composite design with 1 centerpoint.

- D-criterion: $|X^T X| = 9 \cdot 6 \cdot 6 \cdot 4 = 1296$

- A-criterion: $tr\{(X^TX)^{-1}\} = \frac{1}{9} + \frac{1}{6} + \frac{1}{6} + \frac{1}{4} = \frac{25}{36}$

- G-criterion: $\max_{x \in R} N f^T(x)(X^TX)^{-1} f(x) = G(\xi)$ where

$$\begin{aligned} G(\xi) &= \max_{x \in R} \left\{ 9 \cdot [1 \;\; x_1 \;\; x_2 \;\; x_1x_2](X^TX)^{-1}[1 \;\; x_1 \;\; x_2 \;\; x_1x_2]^T \right\} \\ &= \max_{x \in R} \left\{ 1 + \frac{3}{2}x_1^2 + \frac{3}{2}x_2^2 + \frac{9}{4}x_1^2x_2^2 \right\} = 1 + \frac{3}{2} + \frac{3}{2} + \frac{9}{4} = \frac{25}{4} \end{aligned}$$

  Notice that for the square design region $R = [-1, 1] \times [-1, 1]$, the maximum occurs at $x_1 = \pm 1$ and $x_2 = \pm 1$.

- IV-Criterion: $average_{x \in R} N f^T(X^TX)^{-1} f(x) = average\ IV(\xi)$ where for the square design region R, the area $A = 4$.

$$\begin{aligned} IV(\xi) &= \frac{1}{A} \int_{-1}^{1} \int_{-1}^{1} N x^T (X^TX)^{-1} x \, \mathrm{d}x_1 \mathrm{d}x_2 \\ &= \frac{1}{4} \int_{-1}^{1} \int_{-1}^{1} 1 + \frac{3}{2}x_1^2 + \frac{3}{2}x_2^2 + \frac{9}{4}x_1^2x_2^2 \;\; \mathrm{d}x_1 \mathrm{d}x_2 = \frac{9}{4} \end{aligned}$$

- E-criterion: Maximum eigenvalue λ_{max} of $(X^TX)^{-1}$. Thus, $E(\xi) = \frac{1}{4}$.

- D_S-criterion. Consider the 3^2 design above for the interaction model. We would like to know if higher order (quadratic) terms are needed. To this end, we reorder the model terms yielding

$$E(y) = \beta_{11}x_1^2 + \beta_{22}x_2^2 + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

so that $f_1^T(x) = [x_1^2 \quad x_2^2]$ and $f_2^T(x) = [1 \quad x_1 \quad x_2 \quad x_1x_2]$. After appropriate reordering of the rows and columns of the model matrix X,

$$X^TX = \left[\begin{array}{cc|cccc} 6 & 4 & 6 & 0 & 0 & 0 \\ 4 & 6 & 6 & 0 & 0 & 0 \\ \hline 6 & 6 & 9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 6 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 \end{array}\right] = \left[\begin{array}{cc} M_{11}(\xi) & M_{12}(\xi) \\ M_{12}^T(\xi) & M_{22}(\xi) \end{array}\right] = M(\xi)$$

Notice here that $s = 2$ and so the upper left 2×2 matrix is $M_{11}(\xi)$ and the lower right 4×4 matrix is $M_{22}(\xi)$. The D_S criterion is

$$D_S(\xi) = \frac{|M(\xi)|}{|M_{22}(\xi)|} = 4.$$

4.9. Optimal Design Theory

By treating an experimental design as a probability measure, Kiefer (1959, 1960, 1961), Kiefer and Wolfowitz (1960) and Farrell, Kiefer, and Walbran (1968) pioneered the theoretical foundation for design optimality criteria. For a thorough discussion of optimal design theory, see Atkinson et al. (2007) and Pukelsheim (1993). The treatment of designs as probability measures is often referred to as *approximate theory* and the designs are called *approximate designs*. Approximate theory assigns a probability distribution to the points in the design space, $\mathcal{X}$. As a result, approximate theory does not require the number of trials at any design point to be an integer (i.e., the design is not necessarily implementable in practice). An experimental design with

trials at n distinct points in $\mathcal{X}$ can be summarized by

$$\xi = \left\{ \begin{array}{cccccc} x_1, & x_2, & . & . & . & , x_n \\ w_1, & w_2, & . & . & . & , w_n \end{array} \right\} \tag{2.27}$$

where x_i and w_i, $i = 1, .., n$, are, respectively, points of the design and the weights associated with these design points such that $\sum_i^n w_i = 1$. If $w_i = \frac{r_i}{N}$ for a particular $N-$point design, with $\sum_i^N r_i = N$ and r_i is a positive integer (that is the weights are rational), then the design is said to be **exact**. Otherwise, it is **continuous**. In practice, all designs are exact and, good N-point exact designs can often be found by using rational weights w_i that approximate the optimum w_i^* weights for the continuous measure ξ^*. The details of approximation rules are found in Pukelsheim and Reider (1992). For simple one-factor models with p parameters, there will be p support points with equal weights $\frac{1}{p}$ so that the exact design with $N = p$ is optimum. However, if the design weights are not rational, it is impossible to find an exact design ξ_N for any finite N that is identical to the continuous optimum design ξ^*. It must be noted, though, that when comparing exact designs, what matters are the corresponding values of the design criterion (Atkinson et al., 2007).

For the continuous design ξ, the information matrix associated with model parameter vector β in the linear model $y = X\beta + \epsilon$ is given by

$$\begin{aligned} M(\xi) &= \int_{\mathcal{X}} f(x) f^T(x) \xi \, \mathrm{d}x \\ &= \sum_{i=1}^{n} w_i f(x_i) f^T(x_i). \end{aligned} \tag{2.28}$$

For an N-trial exact design ξ_N, the information matrix $M(\xi_N)$ for β given in 2.28 is equivalently,

$$X^T X = \sum_{i=1}^{N} f(x_i) f^T(x_i) \tag{2.29}$$

where $f^T(x_i)$ is the ith row of the model matrix X (Atkinson et al., 2007). Note that $n \leq N$ with $n < N$ if any of the N design points is replicated, and $n = N$ if all design points are unique. Given the weights in the expression above and, summed over the n unique design points, the normalized version of the information matrix for the exact design, ξ_N, is given by

$$M(\xi_N) = \frac{X^T X}{N} \tag{2.30}$$

and the prediction variance function is

$$Var\left\{\hat{y}(x)\right\} = \sigma^2 f^T(x)(X^T X)^{-1} f(x). \tag{2.31}$$

For a continuous design, the standardized prediction variance function is given by

$$d(x, \xi) = f^T(x) M^{-1}(\xi) f(x)$$

which is clearly a function of both the design and the point in the design space at which prediction is made, but does not depend on any unknown model parameters. For an exact design, ξ_N, the standardized or scaled prediction variance function is given by

$$d(x, \xi_N) = f^T(x) M^{-1}(\xi_N) f(x). \tag{2.32}$$

4.9.1. The General Equivalence Theorem (GET): The General Equivalence Theorem (Kiefer, 1959) states that the following three conditions are equivalent if

$\Psi\{M(\xi)\}$ is the measure of imprecision and $\phi(x,\xi)$ is its derivative in the direction of a measure $\bar{\xi}$ which puts unit mass at a design point x:

1. The design ξ^* minimizes $\Psi\{M(\xi)\}$.
2. The design ξ^* maximizes the minimum over $\mathcal{X}$ of $\phi(x,\xi)$.
3. The minimum over $\mathcal{X}$ of $\phi(x,\xi^*) = 0$, and this minimum occurs at the points of support of the design.

For a detailed discussion of $\Psi\{M(\xi)\}$ and $\phi(x,\xi)$, see Atkinson et al. (2007). As a consequence of (3), a further condition obtained is:

4. For any non-optimum design, the minimum over $\mathcal{X}$ of $\phi(x,\xi) < 0$.

This theorem provides methods for the construction and checking of designs to see if they are optimal by some criteria. An important consequence of the theorem is the fact that for continuous designs, D-optimal designs are also G-optimal. For D-optimal designs, we have

$$\Psi\{M(\xi)\} = \log|M^{-1}(\xi)| = -\log|M(\xi)| \tag{2.33}$$

in which case the log determinant of the inverse matrix is minimized. The log is taken so that the resulting function is convex which guarantees that any minimum found is global rather than local. The D-optimal design minimizes $\Psi\{M(\xi)\}$ and by condition (2) of the theorem, it maximizes the minimum over $\mathcal{X}$ of the derivative function $\phi(x,\xi)$ given by

$$\phi(x,\xi) = p - d(x,\xi). \tag{2.34}$$

By condition (3) of the theorem, the minimum of $\phi(x,\xi^*)$ over $\mathcal{X}$ is 0, implying that

$$p - d(x,\xi^*) \geq 0 \tag{2.35}$$

for all $x \in \mathcal{X}$. Thus,

$$d(x, \xi^*) \leq p \tag{2.36}$$

which provides an upper bound for the standardized prediction variance. G-optimal designs minimize the maximum prediction variance over $\mathcal{X}$. By the general equivalence theorem, we observe that for a G-optimal design, the maximum prediction variance equals p, the number of parameters. Why are D-optimal and G-optimal continuous designs identical? To maximize the minimum over $\mathcal{X}$ of $\phi(x, \xi)$, we have to minimize the maximum of $d(x, \xi)$. This proves the equivalence. The general equivalence theorem holds for continuous designs, but in general, does not hold for exact designs. Another thing to note is that optimum designs are not necessarily unique.

A practical example of how the GET is used in checking the optimality of a proposed design for a quadratic regression model

$$E(y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 \tag{2.37}$$

$i = 1, ..., n$ is shown in Figure 2.2. The D-optimal design for the model is

$$\xi^* = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{array} \right\}. \tag{2.38}$$

The model has $p = 3$ parameters and by (2.36), the standardized prediction variance function $d(x, \xi^*) = 3$ at the points of support of the design as seen in Figure 2.2.

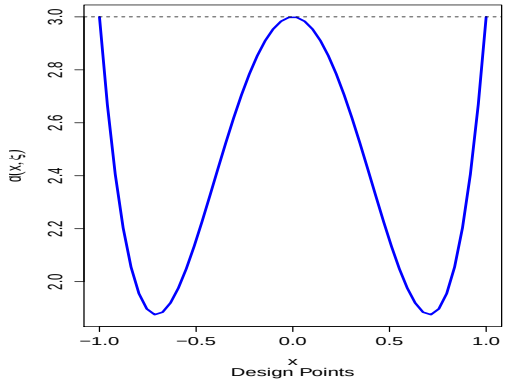

Figure 2.2: Standardized Prediction Variance Function of a D-optimal design for a Quadratic Regression Model $E(y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$.

5. Algorithms for Generating Optimal Designs

5.1. First-order Algorithms

The objective is to find a continuous design measure ξ that minimizes the measure of imprecision $\Psi\{M(\xi)\}$. The General Equivalence Theorem plays an important role in developing algorithms for obtaining the optimum (or near optimum) continuous experimental designs. Recall that condition (3) of the theorem states that the derivative or gradient function, $\phi(x, \xi)$, of $\Psi\{M(\xi)\}$ is non-negative for the optimum design. Therefore, we will expect the gradients to be negative away from the optimum design. Following Atkinson et al. (2007), let measure $\bar{\xi}_k$ put unit mass at the point x_k where x_k is chosen so that $\phi(x, \xi_k) < 0$, and ξ_k is an arbitrary starting design. For sufficiently small $\alpha > 0$, and for

$$\xi_{k+1} = (1 - \alpha)\xi_k + \alpha\bar{\xi}_k, \tag{2.39}$$

$\Psi\{M(\xi_{k+1})\} < \Psi\{M(\xi_k)\}$. Thus, an algorithm for generating the optimal design is a gradient descent method. The algorithms are called first-order because only the

first derivative is used. However, second-order algorithms can also be used which will converge faster than the first-order algorithms.

For D-optimum designs, which are the most widely used in industry, the standardized prediction variance function $d(x, \xi)$ can be used to generate the optimum design. Recall that $\phi(x, \xi) = p - d(x, \xi)$ for the D-optimal design, and define

$$\bar{d}(\xi) = \max_{x \in \mathcal{X}} d(x, \xi). \tag{2.40}$$

Then, the gradient descent (steepest descent) algorithm for D-optimality will successively add mass to the design measure corresponding to the point where $\bar{d}(\xi)$ is obtained (Atkinson et al., 2007). Using $\alpha_k = \frac{1}{k+1}$, the algorithm corresponds to the forward sequential algorithm implemented in *SAS* using `PROC OPTEX` (SAS/STAT, 2003).

5.2. Sequential Design Construction

A special case of the first-order algorithm can be used to construct D-optimal designs sequentially. Suppose the information matrix after N trials is $M(N)$ for an $N \times p$ model matrix X. Thus, $M(N) = X^T X$. Upon adding an additional point to the design matrix, the resulting information matrix, $M(N+1)$, can be written as follows: Let X^* be the resulting model matrix, with $f^T(x)$ being the $(N+1)$th row of X^*. Then,

$$X^* = \begin{bmatrix} X \\ f^T(x) \end{bmatrix} \quad \Rightarrow X^{*T} X^* = \begin{bmatrix} X^T X + f(x) f^T(x) \end{bmatrix}. \tag{2.41}$$

Consequently,

$$|M(N+1)| = |X^T X + f(x) f^T(x)|. \tag{2.42}$$

Using Rao (1973), this can be rewritten as a multiplicative update of $|M(N)|$:

$$\begin{aligned} |M(N+1)| &= |X^TX| \left\{1 + f^T(x)(X^TX)^{-1}f(x)\right\} \\ &= |M(N)| \left\{1 + \tfrac{d(x,\xi_N)}{N}\right\}. \end{aligned} \tag{2.43}$$

Thus, to maximize $|M(N+1)|$, trials are added to the design ξ_N to form ξ_{N+1}. Then, the trial that maximizes $d(x, \xi_{N+1})$ is retained to form the $N+1$-point sequential design. If the support points of the D-optimum design become clear through the sequential construction, the weights of the continuous design can be found by numerical optimization. Otherwise, a special algorithm by Silvey, Titterington, and Torsney (1978) can be used. Atkinson et al. (2007) suggest the following possibilities for finding the D-optimum design:

1. Use a numerical method to find an optimum continuous design with a starting point for the algorithm.

2. Use analytical optimization when feasible.

6. Nonlinear Models

6.1. Background

Researchers often consider models that are not linear in the parameters of interest. In fact, nonlinear models are prevalent in many areas of industry and scientific research such as manufacturing, pharmacokinetics, chemical kinetics, and engineering among others. Bates and Watts (1988) and Seber and Wild (1989) provide a thorough discussion of the subject of nonlinear regression analysis. Traditionally, when possible, researchers have resorted to transformations that linearize the nonlinear model in order to use estimation methods that are applicable to linear models. Common

among the transformations used is the natural logarithmic transformation, which has worked quite well in many cases. However, it is important to point out here that in cases where the errors are additive, rather than multiplicative, the use of the natural logarithmic transformation may not be advisable. Refer to Montgomery, Peck, and Vining (2006) and Johnson and Montgomery (2010) for a discussion of this issue. To emphasize the importance of using nonlinear estimation techniques even when a suitable transformation of the model can be done, consider the following interaction model in Johnson and Montgomery (2010):

$$y = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2) \times \epsilon. \tag{2.44}$$

In this case, the error is multiplicative. The response, viscosity, has been successfully modeled on the log scale and so a natural logarithmic transformation is appropriate. As a result, we have the linear model

$$\ln y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \ln \epsilon, \tag{2.45}$$

and parameter estimation can be done with ordinary least squares. Johnson and Montgomery (2010) point out the following drawbacks of using the linear model instead of the nonlinear model:

1. To make predictions of viscosity in the design region, an inverse transformation must be done. However, mean estimates of viscosity when inverse transformed are no longer means but medians and so are biased. In particular, if viscosity is positively (negatively) skewed, this means that the linear model will systematically underpredict (overpredict) mean viscosity.

2. Prediction intervals for the linear model may be substantially wider than those of the nonlinear model. This is a consequence of the transformations that are done to obtain the lower and upper prediction limits which does not preserve the equality of range across the design space.

Nonlinear models like the one in equation (2.44) are referred to as *transformably linear*. Others, called *conditionally linear*, are nonlinear models such that, for given values of certain nonlinear parameters, the resulting model is a linear model. Bates and Watts (1988) discuss many conditionally linear models.

6.1.1. Examples of Nonlinear Models: Atkinson et al. (2007) and Ford, Titterington, and Kitsos (1989) review several examples of nonlinear models. The following are several of their examples.

- *Exponential Decay.* In the study of pharmacokinetics and chemical kinetics, one of the simplest reactions is a first-order reaction in which a compound A is transformed into compound B at a constant rate, θ. The process is denoted by $A \longrightarrow B$ and is governed by the expectation model

$$\eta(t, \theta) = \theta_0 \exp(-\theta t), \quad t \geq 0, \theta > 0$$

  where θ_0 is the initial concentration of A and $\eta(t, \theta)$ the concentration of B at time t. This model is transformably linear using a natural logarithmic transformation. Also, the derivative of $\eta(t, \theta)$ with respect to θ_0 does not involve θ_0 and so it is conditionally linear. This seemingly trivial observation has important consequences for experimental design.

- *Inverse Polynomial Regression.* The first-order inverse polynomial is given by

$$\eta(t,\theta) = \frac{1}{1+\theta t}, \quad t \geq 0, \theta > 0 \tag{2.46}$$

This has very similar properties to the exponential decay model above and, when observations are made with error, it is difficult to distinguish between the model curves. Recall that distinguishing between two models is the purpose of T-optimal designs. The curves for $\theta_0 = 1$ and $\theta = 2$ are shown in Figure 2.3.

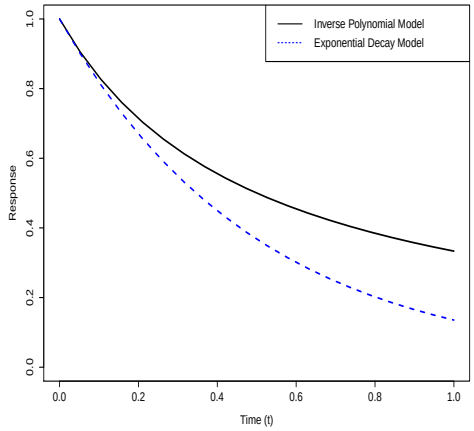

Figure 2.3: Two models for decay: inverse polynomial model (thick line) with $\theta = 2$ and exponential decay Model (dashed line) with $\theta_0 = 1$ and $\theta = 2$.

- *Two Consecutive First-order Reactions.* These consecutive reactions have the form $A \longrightarrow B \longrightarrow C$. In this case, A is transformed into B and then B is transformed into C at constant rates θ_1 and θ_2 respectively. The concentration of B at time t when the concentration of $A = 1$ is

$$\eta(t,\theta) = \frac{\theta_1}{\theta_1 - \theta_2}\left\{\exp(-\theta_2 t) - \exp(-\theta_1 t)\right\} \quad (t \geq 0) \tag{2.47}$$

provided that $\theta_1 > \theta_2 > 0$.

- Bates (1983) described a model in which the ith observation is given by

$$y_i = \theta_1\left\{\exp(-\theta_3 x_{i-1}) - \exp(-\theta_3 x_i)\right\} + \theta_2(x_i - x_{i-1}) + \epsilon_i. \tag{2.48}$$

In one application of this model, the response function is the concentration of a neurotransmitter released from rat-brain tissue immersed in a sequence of vials containing a buffer solution, x_i is the time from first immersion to transference from vial i to vial $i+1$, subject to a fixed total time for the whole experiment (Ford et al., 1989)

6.1.2. The Optimal Design Problem for Nonlinear Models: To illustrate the major design problem for nonlinear models, we will use the Michaelis-Menten model (Bates and Watts, 1988) for enzyme kinetics which relates the initial "velocity" of an enzymatic reaction to the substrate concentration x through the equation

$$f(x,\theta) = \frac{\theta_1 x}{\theta_2 + x} \tag{2.49}$$

The partial derivatives of the expectation function with respect to the parameters, called *parameter sensitivities*, (Atkinson et al., 2007), are

$$\frac{\partial f}{\partial \theta_1} = \frac{x}{\theta_2 + x} \tag{2.50}$$

$$\frac{\partial f}{\partial \theta_2} = \frac{-\theta_1 x}{(\theta_2 + x)^2} \tag{2.51}$$

The matrix of partial derivatives, evaluated at $x_1 = 1.10,$ and $x_2 = 0.22$, is

$$F = \begin{bmatrix} \frac{1.10}{\theta_2+1.10} & \frac{-1.10\theta_1}{(\theta_2+1.10)^2} \\ \frac{0.22}{\theta_2+0.22} & \frac{-0.22\theta_1}{(\theta_2+0.22)^2} \end{bmatrix}. \tag{2.52}$$

Hence,

$$F^T F = \begin{bmatrix} \sum_{i=1}^{2} \frac{x_i^2}{(\theta_2+x_i)^2} & -\sum_{i=1}^{2} \frac{\theta_1 x_i^2}{(\theta_2+x_i)^3} \\ -\sum_{i=1}^{2} \frac{\theta_1 x_i^2}{(\theta_2+x_i)^3} & \sum_{i=1}^{2} \frac{\theta_1^2 x_i^2}{(\theta_2+x_i)^4} \end{bmatrix}. \tag{2.53}$$

Notice that for a linear model, the partial derivatives of the expectation function do not depend on any unknown parameters, and hence, F will only be a function of x. The dependence of the parameter sensitivities, and hence the information matrix, on the unknown parameters for a nonlinear model constitutes a serious problem in the design of optimal experiments for nonlinear models. The implication of this dependence is that the optimal design for the model will depend on the unknown parameters. For example, it is impossible to obtain a D-optimal design for the Michaelis-Menten model without knowing the values of θ_1 and θ_2. The parameter dependence problem is not so much of an issue in cases where the nonlinear model is transformably linear. In such cases, Johnson and Montgomery (2010) have found that standard designs, such as 2^k factorial and 2^{k-p} fractional factorial designs, compare favorably with the optimal designs. Through a simulation study, they found that the D-efficiencies of the standard designs were comparable to those of the optimal designs. The main approaches suggested in the literature for finding optimal designs for nonlinear models are discussed below.

6.1.3. Locally Optimal Designs: In some cases, especially one-parameter models, where a reasonable guess of the parameter values can be made, a *locally optimal design* (Chernoff (1953); Box and Lucas (1959)), that is, a design that is optimal with

respect to a particular parameter value, has been proposed in the literature. Reviews of this subject are given in Ford et al. (1989) and Atkinson et al. (2007). Locally optimal designs will approximate the *true* optimal designs quite closely if reasonable guesses of the parameter values can be made prior to data collection. This obviously becomes more difficult in cases where there are multiple model parameters. Although in practice, the parameter, θ is rarely known, Ford et al. (1989) give some reasons why locally optimal designs are still of interest.

1. They provide a useful reference for other designs. For example, the usefulness of any design ξ is determined by computing its relative efficiency using a locally optimum design.

2. They are necessary for the construction of non-sequential designs based on efficiency and related criteria. Locally optimal designs are used to obtain designs based on expectation or minimax criteria.

3. Where experiments involve batches, the design for batch $(i + 1)$ might be a locally optimal design based on $\widehat{\theta}_i$.

4. Locally optimal designs may be stable over a range of θ values.

6.1.4. Sequential Designs: Atkinson (1982) stresses the importance of a sequential design scheme, with allowance for updating the parameter estimates, for nonlinear models given the parameter dependence problem. Sequential designs iterate the sequence of design formulation, experimentation and analysis of the experiments. In a sequential design, the design points are selected using a well-defined procedure which is outlined in Atkinson et al. (2007):

1. Start with a preliminary estimate or guess of the parameter vector θ,

2. Linearize the model by Taylor series expansion,

3. Find the optimal design for the linearized approximating model, then

4. One or several trials of the optimal design for the linearized model are executed and analysed. If the new estimate of θ is sufficiently precise, the process stops. Otherwise, step 2 is repeated for the new estimate and the process continued until sufficient precision is obtained or the experimental resources are exhausted.

6.1.5. Bayesian Optimal Designs: An effective approach to reducing the dependence of the design on specific parameter values is to use a Bayesian method. Chaloner and Verdinelli (1995) provide a thorough discussion of Bayesian experimental design. As mentioned above, locally optimal designs are optimal with respect to a particular value of the unknown parameter or parameter vector. In a Bayesian optimal design, a prior distribution $p(\theta)$ is assumed for the parameter θ. For example, the quantity that is maximized for a D-optimal design now becomes the *expectation* of the log of the determinant of the information matrix. That is, averaging occurs over many parameter values instead of plugging in just one. Atkinson et al. (2007), as well as other authors, define this as

$$\Phi(\xi) = E_\theta \log |M(\xi, \theta)| = \int_\theta \log |M(\xi, \theta)| p(\theta) \ \mathrm{d}\theta. \tag{2.54}$$

Similarly, for G-optimal designs, the expectation of the standardized prediction variance function is obtained:

$$d(x, \xi) = E_\theta d(x, \xi, \theta) = \int_\theta d(x, \xi, \theta) p(\theta) \ \mathrm{d}\theta, \tag{2.55}$$

again averaging over a range of parameter values.

6.1.6. Maximin and Minimax Designs: The parameter dependence problem can also be addressed using maximin designs (Pronzato and Walter, 1988) in which the parameter θ is assumed to belong to a set Θ. A maximin D-optimal design ξ^* satisfies (Atkinson et al., 2007)

$$\Phi(\xi^*) = \max_{\xi} \min_{\theta} \log |M(\xi, \theta)|. \tag{2.56}$$

Thus, the design ξ^* is found which maximizes $\log |M(\xi, \theta)|$ for that value of θ which minimizes the determinant. That is, the design is found which maximizes Fisher information for the parameter value which minimizes it. In a sense, this criterion is used to guard against the worst value of the log determinant (possible) given the range of θ. Similarly, a minimax design, ξ^* can be found which minimizes the log determinant of the inverse of the information matrix for that value of θ that maximizes it. Thus,

$$\Phi(\xi^*) = \min_{\xi} \max_{\theta} \log |M^{-1}(\xi, \theta)|. \tag{2.57}$$

6.1.7. Robust Designs: In addition to the aforementioned approaches to address the parameter dependence problem, Woods, Lewis, Eccleston, and Russell (2006), Dror and Steinberg (2006), and Ford, Torsney, and Wu (1992) have discussed approaches to the design problem by considering designs which are *robust* to a wide range of parameter values. Waterhouse, Eccleston, and Duffull (2009) introduced designs for both efficient parameter estimation and model discrimination in nonlinear models. Most of their work was concentrated in the area of pharmacokinetics and

pharmacodynamics which, as mentioned earlier, make use of a wide range of nonlinear models. In particular, they introduced *conditional* and *hybrid* designs which are jointly optimal with respect to both the D- and T-optimality criteria. However, for the most part, the issue of parameter-dependency has limited the amount of work done in the area of optimal experimental designs for nonlinear, and without loss of generality, generalized linear models. Moreover, most of the work done has focused primarily on D-optimal designs. Finding optimum designs in the area of nonlinear models remains the focus of many research endeavors.

6.2. Review of Graphical Methods

The literature, in terms of graphical methods for evaluating designs for nonlinear models, is rather sparse. Quantile Dispersion Graphs were used by Robinson and Khuri (2003) to compare designs for generalized linear models. In most of the examples of Atkinson et al. (2007), plots of the standardized prediction variance functions are made to compare designs. Generally, the methods discussed in the literature are not different from those already in use for linear models. By virtue of the fact that generalized linear models are a special case of nonlinear models, the graphical methods used to evaluate designs for such models can also be applied to nonlinear models. For a review of graphical methods, see Khuri and Lee (1998) and Ozol-Godfrey et al. (2008).

7. Stochastic Algorithms for Generating Optimal Designs

The problem of finding optimal designs is an optimization problem that is solved through the use of algorithms. Stochastic or probabilistic search methods have been used in the literature: Haines (1987) used the *simulated annealing* algorithm to obtain IV-, D- and G-optimal designs for various polynomial models. Atkinson (1992) used

a variant of the algorithm called *segmented annealing* which speeds the search for the design. Waterhouse et al. (2009) also use simulated annealing as the main search method to obtain their optimal designs. The algorithms used generally arise from the applications of combinatorial optimization. The use of stochastic search methods is motivated by the fact that they perform better than traditional exchange algorithms and are less likely to get trapped at local optima. The various design-generating algorithms are discussed below.

7.0.1. Simulated Annealing: Simulated annealing (SA) was introduced by Kirkpatrick, Gelatt, and Vecchi (1983) to find the global minimum of a cost function that may possess several local minima. Other papers, particularly, Corana, Marchesi, Martini, and Ridella (1987), have provided a straightforward implementation of the algorithm. It works on the principle that when a solid is allowed to cool, it eventually attains the minimum possible temperature. Cooling it too fast, however, will result in the solid attaining a less than optimum final temperature. The SA algorithm is designed to allow cooling to occur at a rate deemed optimal. The description of the algorithm given here is analogous to those of Goffe, Ferrier, and Rogers (1994), Waterhouse (2005), and Corana et al. (1987).

For a k-variable model, the essential starting parameters of the SA algorithm are the initial temperature T_0, the starting vector or matrix of algorithm parameters X that will maximize (or minimize) the criterion C; the step length vector or matrix V for X. To find an optimal experimental design, consider the design

$$\xi = \left\{ \begin{array}{cccccc} \mathbf{x_1}, & \mathbf{x_2}, & . & . & . & , \mathbf{x_m} \\ w_1, & w_2, & . & . & . & , w_m \end{array} \right\}$$

described earlier in equation 2.27 where each $\mathbf{x_i}$ is $k \times 1$. The support points and the weights are the SA elements of X, that is, the vector or matrix of parameters. Without loss of generality, let $X = \xi$. For simplicity, we treat the design as an $m \times n$ matrix with elements ξ_{ij}, with $i = 1, .., m; j = 1, .., n$. We also have matrices $\boldsymbol{L}$ and $\boldsymbol{U}$ whose elements are the lower and upper bounds of the support points and the weights. Let ξ_{ij}^{p-1} be ξ_{ij} after the $(p-1)st$ iteration. At the pth iteration of the algorithm, each element of ξ_{ij}^{p-1} is perturbed in sequence to give ξ_{ij}^{k} using the corresponding step or perturbation quantity, $v_{ij} \in V$, drawn from a uniform $[-1, 1]$ distribution. Thus,

$$\xi_{ij}^{k} = \xi_{ij}^{k-1} + v_{ij}. \tag{2.58}$$

Therefore, each iteration involves finding new mn designs. Designs that violate the bounds are promptly rejected. For each design generated, the criterion C is calculated and compared to the existing *best* design criterion. If the objective of the optimization is to find a global maximum, the following steps are executed, assuming $\xi^{'}$ and ξ^{*} are the current new and *best* designs, respectively, and at current temperature T:

1. Compute $\delta C = C(\xi^{'}) - C(\xi^{*})$.

2. If $\delta C \geq 0$, then accept the new design as the best design. That is, $\xi^{*} = \xi^{'}$.

3. Else if $(\delta C < 0)$, then accept the new design with probability $p = \exp(\delta C/T)$.

Notice that for negative values of δC, a metropolis criterion determines acceptance or rejection of the design. Generally, the probability of a downhill move (accepting inferior designs) is reduced at lower temperatures, whereas at higher temperatures, this probability is larger. Also, larger jumps in the criterion's value decrease the probability of downhill moves. Cooling is done geometrically, that is, $T_{k+1} = \alpha T_k$, where $0.85 \leq \alpha < 1$ appear in the published literature. The optimization procedure

ends when the step sizes are less than some tolerance at which point the optimal design is the current ξ^*.

7.0.2. Cross Entropy: The following description is based on Waterhouse (2005). The cross-entropy method (Rubinstein and Kroese, 2004) was originally developed as an algorithm for estimating the probabilities of rare events in stochastic networks. The algorithm can be adapted to continuous multi-extremal optimization as shown in Kroese, Porotsky, and Rubinstein (2006)

The design to be considered here is characterized as a vector of length mn. The upper- and lower bound matrices $\boldsymbol{U}$ and $\boldsymbol{L}$ are transformed appropriately to conform to the design. The elements of ξ are associated with independent truncated normal distributions with lower and upper limits defined by the elements of $\boldsymbol{L}$ and $\boldsymbol{U}$. The means and variances of these distributions are $\boldsymbol{\mu} = (\mu_1, ..., \mu_{mn})^T$ and $\boldsymbol{\sigma}^2 = (\sigma_1^2, ..., \sigma_{mn}^2)^T$. Each element of ξ can then be written as a random deviate from a truncated normal distribution, that is, $\xi_j \sim N(\mu_j, \sigma_j^2, \boldsymbol{L}_j, \boldsymbol{U}_j)$, for $j = 1, ..., mn$. The vector of interest is μ^* which is the optimal design ξ^*. Rubinstein and Kroese (2004) provide the steps of the CE algorithm for continuous optimization:

1. **Initialize**: Choose initial estimates $\boldsymbol{\mu}_0$ and $\boldsymbol{\sigma}_0^2$. Set $t = 1$.

2. **Draw**: At the kth iteration, generate random samples, $\xi^1, \xi^2, ..., \xi^N$ from $N(\mu_{k-1}, \sigma_{k-1}^2, \boldsymbol{L}, \boldsymbol{U})$

3. **Select**: Let $\mathcal{I}$ be the indices of the $N^{elite} = \rho N$ samples, (typically $\rho = 0.1$).

   Update: For all $j = 1, .., mn$, let

$$\tilde{\mu}_{kj} = \sum_{i \in \mathcal{I}} \xi_j^i / N^{elite}, \quad \tilde{\sigma}_{kj}^2 = \sum_{i \in \mathcal{I}} (\xi_j^i - \tilde{\mu}_{kj})^2 / N^{elite} \tag{2.59}$$

and

$$\widehat{\boldsymbol{\mu}}_k = \alpha_1 \tilde{\boldsymbol{\mu}}_k + (1-\alpha_1)\widehat{\boldsymbol{\mu}}_{k-1}, \;\; \widehat{\boldsymbol{\sigma}}_k^2 = \alpha_2 \tilde{\boldsymbol{\sigma}}_k^2 + (1-\alpha_2)\widehat{\boldsymbol{\sigma}}_{k-1}^2. \tag{2.60}$$

4. If $\max_j \{\widehat{\sigma}_{kj}\} < tol$, where tol is a small positive real number tolerance, the algorithm converges. Otherwise, increment t by 1 and return to step 2.

The *injection method* (Botev and Kroese, 2004) is used such that every time the stopping criterion is met, the standard deviations are inflated by adding

$$|C_k^* - C_{k-1}^*|h \tag{2.61}$$

where h is between 0.1 and 10, and C_k^* is the best value of the criterion obtained at the kth iteration. This is used to avoid stopping at local optima.

7.1. Genetic Algorithms

Genetic algorithms (GAs) (Holland, 1975) are evolutionary stochastic search strategies based on the principles of genetics and natural selection. The foundations of GAs were developed by John Holland in the 1960s and were popularized by David Goldberg (Goldberg, 1989). A genetic algorithm (GA) maintains a population of potential solutions, called chromosomes, and then uses the processes of *selection*, *reproduction* and *mutation* to select the solutions that seem to work well for the optimization problem. GAs have provided solutions to complex optimization problems like wire routing, job scheduling, machine learning, transportation and optimal control problems.

GAs are attractive not only because they are relatively easy to implement but also for the following reasons based on Michalewicz (1992), Haupt and Haupt (2004) and Borkowski (2003):

1. GAs maintain a population of solutions.
2. GAs allow optimization with continuous or discrete variables.
3. GAs do not require derivative information.
4. GAs can deal with a large number of variables.
5. GAs can simultaneously search multiple wide samplings of the solution space.
6. GAs are well-suited to parallel computing.

An in-depth introduction to GAs can be found in Michalewicz (1992) and Haupt and Haupt (2004). The brief description given below is mostly based on Borkowski (2003) and Michalewicz (1992).

7.1.1. Structure of a GA: In order to optimize an objective function F, the components of a genetic algorithm (as an evolutionary program) are enumerated below.

1. **Genetic Representation**. For any GA, the construction of a chromosome is required. A chromosome represents a potential solution to the problem of interest and traditionally is represented by a string of *genes* that are either *binary encoded* or *real-number encoded*. In a binary encoding, the genes are encoded 0 or 1, while in a real representation the genes are encoded with real numbers. The type of encoding used results in a binary or real GA. In the context of optimal experimental designs, a chromosome is simply an experimental design as defined in equation 2.27.

2. **Objective Function** F. An objective function assesses a chromosome's superiority or inferiority in its environment by rating it in terms of its *fitness*. F takes a chromosome as input and outputs an objective function value. For example, if the optimization problem is to generate a D-optimal design, the objective function is the D-optimality criterion function.

3. **Genetic Operators**. These operators modify existing chromosomes to produce new chromosomes called offspring. The processes of selection, reproduction and mutation are made possible by genetic operators. The selection phase of the GA involves choosing pairs of parents for reproduction. This can be done in many different ways, with the population size being one thing to consider. In this dissertation, the selection scheme used is random pairing. Thus, two parents are randomly chosen for reproduction irrespective of their fitness. Other selection strategies as outlined in Haupt and Haupt (2004) are rank weighting, cost weighting and tournament selection. Rank and cost weighting select chromosomes for reproduction based on the rank and cost (that is, a measure of chromosome undesirability) of the chromosomes respectively. Tournament selection is often used for large population sizes.

Once parents have been selected, the process of reproduction can take place through the use of reproduction operators. It must be noted that the purpose of reproduction is not only to create offspring but introduce variability into the population in order to enhance the search for the most desirable chromosome (or experimental design). Offspring can be produced through blending, a widely used reproduction operator. Blending is simply a linear combination of the genes (or support points) of the two parents and it results in the production of two offspring. A new reproduction operator which modifies the concept of blending is proposed in Chapter 3. Other

reproduction operators will exchange (or swap) support points at randomly selected positions on the parent chromosomes while others will exchange whole sections of the parent chromosomes. Haupt and Haupt (2004) and Michalewicz (1992) discuss other reproduction operators.

Mutation is also an evolutionary process by which the variability of chromosomes in the population is increased, resulting in a more efficient search of the problem space. It involves altering at least one of the genes in a chromosome. In uniform mutation, a randomly selected gene on a chromosome is replaced (or altered) by a random deviate from a uniform distribution. Non-uniform mutation perturbs a randomly selected gene about its current position (Michalewicz, 1992). Gaussian mutation, used in this dissertation, delivers a similar functionality as non-uniform mutation. In this type of mutation, a gene is replaced with a draw from a truncated normal distribution with mean equal to the gene (or support point) and a standard deviation σ usually set to one.

It must be emphasized that the optimal combination of selection strategies, reproduction and mutation operators is problem-specific. In addition, the GA parameters must also be tuned for the problem.

4. **GA Parameters**. Selection, reproduction and mutation happen probabilistically and therefore, the associated probabilities must be defined beforehand. In addition, parameters such as the population size and the number of generations the population evolves (i.e, number of iterations) must also be specified.

To put these concepts in the context of optimal experimental designs, assume that the objective is to obtain a D-optimal design for some model (linear or nonlinear). During iteration (generation) t, the GA maintains a population of potential solutions $P(t) = \{x_1^t, ..., x_n^t\}$ to the optimization problem. These solutions are potential D-

optimal designs for the model of interest. Each x_i^t is evaluated under F (D-optimality criterion) to give its measure of fitness (D-optimality). Then a new population is formed by selecting the more fit chromosomes (that is, designs that are closer to D-optimal designs) to reproduce generating offspring (new potentially D-optimal designs), and also to mutate. While reproduction leads to the creation of new designs based on D-optimality, mutation introduces extra variability into the gene pool. In the selection process, the GA (generally) selects chromosomes with superior fitness relative to the existing population so that their good traits can be passed on to future generations of chromosomes. In other words, the success of the GA is based on the *survival of the fittest* biological imperative. Michalewicz (1992) and others give examples of many different selection, reproduction and mutation methods. Some GAs incorporate the concept of **elitism** where the most fit chromosomes do not participate in the reproduction and mutation processes. For example, Borkowski (2003) retains the top two elite chromosomes (designs). The fitnesses of the new population are evaluated and the best D-optimality-based designs become the elite designs for the next generation. The processes of selection, reproduction and mutation are iterated until a stopping criterion is met. The most elite (or best) D-optimal design is reported as the optimal (or near-optimal) design. A flowchart of how a basic GA works is given in Figure 2.4.

7.2. Application to Optimal Designs

In spite of the advantages of GAs, a review of the optimal design literature suggests that their use is minimal at best. In general, derivative-based and exchange algorithms dominate the literature. Fortunately, GAs are gradually making their way into the area of optimal design. Borkowski (2003) showed how a GA can be used to generate small exact response surface designs that were superior to designs generated by other

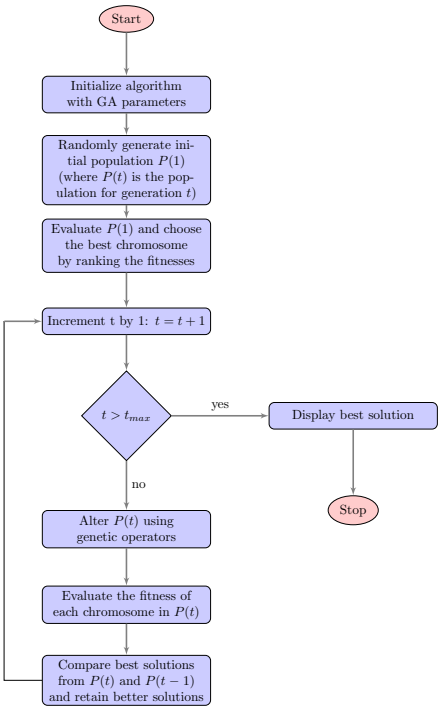

Figure 2.4: Flowchart of a Genetic Algorithm with t_{max} Generations.

existing algorithms. Also, Heredia-Langner et al. (2003) demomstrate the use of a GA in the construction of D-optimal designs. More recently, Limmun, Borkowski, and Chomtee (2012) have used a GA to generate D-optimal designs in situations where the experimental region is an irregularly shaped polyhedral region for mixture experiments. While the foregoing examples have focused mainly on optimal designs for linear models, GAs have not been exploited in the context of design generation for nonlinear models. Preliminary results in this research, however, show that optimal designs for nonlinear models based on GAs perform better than, if not as well as, those based on existing algorithms.

GAs, given their robustness and ease of implementation, can be used to obtain designs (that are close to optimum) which can be further improved by the simulated annealing or cross entropy algorithms. Preliminary work shows that this two-stage approach to optimization produces much better designs with respect to D- and T-optimality, for example.

CHAPTER 3

DESIGNS ROBUST TO MULTIPLE PRIOR DISTRIBUTIONS

1. Introduction

Nonlinear models are widely used in many scientific fields, and in particular, the physical and biological sciences. These models have the property that their Fisher information matrices depend on any unknown model parameters. Thus, unlike linear models whose information matrices are only functions of the design points up to a proportionality constant, designing optimal experiments (e.g, to estimate model parameters) for nonlinear models is non-trivial.

A common topic in the literature on nonlinear experimental design is locally optimal designs, including local optimality criteria and Bayesian approaches to design construction and evaluation. Locally optimal designs (Box and Wilson, 1951) are designs that are optimal conditional on a particular value or guess of the model parameter vector. These have been widely used in practice and actually provide the basis for design comparisons in terms of efficiency (Ford et al., 1989). However, as the dimension of the parameter vector increases, making reliable guesses becomes difficult and therefore obtaining reasonable locally optimal designs becomes a difficult problem. Further, if the guesses are not within a relatively small neighborhood of the true parameter values, the resulting design will be sub-optimal. More importantly, consulting several well-trained experts for reasonable guesses of the parameter values necessitates the requirement of a design that performs efficiently across the different experts' guesses. In particular, local optimality can be thought of as concentrating unit mass at a particular parameter vector value which is inconsistent with having multiple parameter vectors across the experts.

Some Bayesian approaches to the nonlinear design problem have been proposed. These, in a broad sense, generalize the idea of local optimality by using non-degenerate prior distributions. Thus, Bayesian optimal designs are obtained by averaging over the prior distribution. However, when considering the issue of multiple prior distributions elicited from multiple experts, the robustness of design to different priors is a more practical concern than optimality - with respect to a single prior - in the case of nonlinear models.

In this chapter, new robust design criteria, aimed at achieving designs that are robust to multiple prior distributions, are presented for nonlinear models. In addition, novel algorithmic approaches to implementing the methods are introduced.

2. Bayesian Optimality Criteria

Most of the Bayesian design approaches can be put in a decision-theoretic framework. Chaloner and Verdinelli (1995) indicate that the general principle used is averaging over what is unknown (in this case, the sample space). Following Clyde (1993), a prior distribution $p(\theta)$ is defined on the parameter space Θ; a real-valued loss function $L(\theta, a, \xi)$ is defined on the product space $\Theta \times \mathcal{A} \times \Xi$, where Ξ is the set of all probability measures or experimental designs defined on $\mathcal{X}$, and $\mathcal{A}$ is the action space. The action space is determined by the experimental objectives and consists of estimation, prediction and model discrimination to name a few. Specifically, $L(\theta, a, \xi)$ quantifies the loss in observing data y based on design ξ and taking action a. Given the uncertainty in θ, a natural procedure in this case is to choose an *optimal* decision to minimize expected loss (Berger, 1985). Averaging over the unobserved sample using a Bayes decision rule $\delta^{\epsilon}(y)$ yields as Bayes risk

$$r(\xi) = \iint L(\theta, \delta^{\epsilon}(y), \xi)\, p(\theta|y)p(y)\, \mathrm{d}\theta\, \mathrm{d}y. \tag{3.1}$$

The expression in (3.1) can be re-arranged using Bayes rule as

$$r(\xi) = \iint L(\theta, \delta^{\epsilon}(y), \xi)\, p(y|\theta)p(\theta)\, \mathrm{d}y\, \mathrm{d}\theta \tag{3.2}$$

which is the pre-posterior expected loss of using design ξ based on prior distribution $p(\theta)$. Thus, the optimal decision is to choose the design ξ^* that minimizes $r(\xi)$. The dependence of this framework on $p(\theta)$ is hereby noted. For example, the choice of design ξ^* will be sub-optimal if $p(\theta)$ is misspecified or if there is a class of possible prior distributions. The contention here is that ξ^* is sensitive to the choice of $p(\theta)$ as pointed out by Toman and Notz (1991). The issue of multiple prior distributions is analogous to that of multiple models investigated by Lauter (1976).

To put things in perspective, $-L(\theta, a, \xi)$ is a utility function. Thus, the optimization problem becomes that of maximization of expected utility functions instead of minimization of expected loss functions. Bayesian equivalents to some classical optimality criteria can be obtained by specifying the appropriate utility function. For example, the use of Shannon information (Shannon, 1948) results in the Bayesian version of the D-optimality criterion. This is given by

$$\Phi(\xi) = \int \log |M(\xi, \theta)|\, p(\theta)\, \mathrm{d}\theta. \tag{3.3}$$

where $|M(\xi, \theta)|$ is the determinant of the information matrix at θ.

The Bayesian approach to nonlinear design is particularly intuitive due to the uncertainty in θ. However, as observed above, there is still the issue of robustness

that needs to be addressed. The next section introduces new approaches to obtaining robust designs for nonlinear models.

3. New Robust Criteria for Nonlinear Models

The previous section emphasized the fact that the main issue with the Bayesian design paradigm is the sensitivity of the design to the prior distribution which is not uncommon in Bayesian analysis. As DasGupta and Studden (1988) point out, a single prior distribution is approximate at best and so it makes sense to think in terms of a class or family of prior distributions (instead of a single one) in a robustness framework. In other cases, like in nonlinear models, the uncertainty in the location or spread of the model parameters requires that these quantities be elicited from experts.

The purpose of the elicitation process is to identify an underlying probability distribution that captures an expert's beliefs about θ. An expert's beliefs are usually in the form of summaries of their *ideal* probability distribution in the form of quantiles and moments, for example. An analyst, or decision maker, in Bayesian parlance, obtains a density function that adequately fits the summaries. A problem, among others, generic to all elicitation is the extent to which the density function represents the expert's actual beliefs (Oakley and O'Hagan, 2007). In a sense, the elicited prior distribution (or density function) is only approximate. That is an approximate representation of the researcher's uncertainty in parameter values. There is a substantial amount of literature on prior elicitation and the methods used. Kadane and Wolfson (1998) provide a useful discussion of the subject. Elicitation from multiple experts invariably results in different prior distributions and accounting for this uncertainty in the prior distribution at the design stage is worthwhile. Thus, a family of priors is used instead of a single prior.

DasGupta and Studden (1991) consider this problem for linear models. In their approach they assume a favored prior distribution and proceed such that the resulting design minimizes Bayes risk with respect to the favored prior subject to being robust to a class of priors. Thus, they solve a constrained optimization problem given that the design is optimized according to the favored prior conditional on being robust to a set of priors. The new approaches discussed here differ from the previous approaches in that a favored prior is not assumed, but the entire set of priors is used and the design problem is unconstrained. More specifically, the context is extended to nonlinear models.

Following DasGupta and Studden (1988), suppose we have a class Γ of Normal prior distributions where

$$\Gamma = \left\{ p(\theta) : \theta \sim N(\mu, \sigma^2\Sigma);\ \mu \in C \text{ and } \Sigma \in \Sigma_\theta \right\} \tag{3.4}$$

where Σ_θ is a class of positive definite matrices up to a proportionality constant. For the purposes of this section, μ may be fixed, in which case C is a singleton, implying complete confidence in the location of θ. Otherwise, C is a set in $\mathcal{R}^p$ where $p = \dim(\mu)$. As DasGupta and Studden (1991) observe, μ is usually fixed since an approximate location of θ can be easily elicited compared to the higher moments and strength of correlations which are more difficult to elicit. In the following, $\phi(\xi, \theta)$ is a utility function or equivalently, any of the alphabetic design optimality criteria.

3.1. Maximin Criterion

In a situation where more than one prior is plausible, it is reasonable to obtain a design that maximizes the minimum expected utility over Γ. Thus, a maximin

criterion of the form

$$\eta_1(\xi) = \min_{p_i(\theta)\in\Gamma} E_i\{\phi(\xi,\theta)\} \tag{3.5}$$

where E_i, the expected utility under the ith prior, is desirable, $1 \leq i \leq k$ for k priors in Γ. The design ξ^* that maximizes equation (3.5) is a robust design in the sense that it maximizes the minimum expected utility over all the priors in Γ.

3.2. Product Criterion

Another functional that is based on the idea of product optimality (Atkinson and Cox, 1974) is also considered here. This is defined as

$$\eta_2(\xi) = \prod_{i=1}^{k} E_i\{\phi(\xi,\theta)\} \tag{3.6}$$

where E_i is as previously defined. The objective of this criterion is to obtain a design that maximizes the product of expected utilities under the different priors in Γ. The choice of a product is made here so that the resulting design performs efficiently for a wide range of parameter values. On the log scale, we can think of the robust design as the one that maximizes the sum, over all priors in Γ, of the logarithms of the expected utilities.

3.3. Weighted Product Criterion

The product criterion in (3.6) can be modified to incorporate varying *a priori* weights for the prior distributions. This can be useful when the confidence in each prior is not the same. Thus, we have a weighted product criterion of the form

$$\eta_3(\xi) = \prod_{i=1}^{k} E_i\{\phi(\xi,\theta)\}^{w_i} \tag{3.7}$$

where w_i are non-negative and $\sum_{i=1}^{k} w_i = 1$. Practically, the weights reflect the credibility of the experts from whom each prior is elicited. Intuitively, the robust design maximizes expected utility under the prior with the largest weight relative to the other priors. On the log scale, the criterion is the sum of weighted log expected utilities. Like $\eta_2(\xi)$ in (3.6), $k = 1$ for $\eta_3(\xi)$ in (3.7) yields the criterion in (3.3) for D-optimality.

3.4. Geometric Criterion

Intuitively, the geometric mean of expected utilities provides a better "compromise" than their arithmetic mean. Also, in cases when expected utilities tend to be small, the resulting product in (3.6) will be near 0. The situation is made worse when the expected utilities cannot be precisely represented due to computer precision limitations resulting in the criterion being assigned a value of zero. A reasonable fix for this is to take the kth root of the criterion in (3.6). This results in a criterion which can be interpreted as the geometric mean of expected utilities. Hence a new criterion

$$\eta_4(\xi) = \left\{ \prod_{i=1}^{k} E_i\{\phi(\xi, \theta)\} \right\}^{\frac{1}{k}} \tag{3.8}$$

is proposed. Similarly, the geometric mean of (3.7) can be used as a criterion. An interesting idea as far as (3.7) is concerned is to simultaneously optimize the design and the weights. This is particularly reasonable if no *a priori* assumptions can be made about the prior weights. Admittedly, this comes at a computational cost.

4. Aggregating Probability Distributions

Another possible solution to the nonlinear design problem, when more than one prior distribution for the parameters can be specified, is to find a reasonable approach

to aggregating or pooling the different prior distributions into a single consensus prior distribution that can, for example, be used in equation (3.3). This naturally leads to opinion pooling (or group decision making), a subject that has recently received attention in the Bayesian literature. Numerous articles can be found in the literature pertaining to group decision making and methods for combining probability distributions with Genest and Zidek (1986) giving a thorough review and annotated bibliography. The rationale behind opinion pooling is to find one consensus probability distribution $p(\theta)$ that sufficiently reflects the opinions of all the individuals from whom elicitation was done. In this section, some methods for opinion pooling in the literature are introduced in the light of their applicability to experimental design.

4.1. Linear Opinion Pooling

Suppose k experts make subjective probability judgments about an unknown parameter vector $\theta \in \Theta \subset \mathcal{R}^p$ resulting in a set of k probability densities $\{p_1(\theta), ..., p_k(\theta)\}$. The assumption is made here that the experts are independent and so the probability densities are independent. The $p_i(\theta)$'s can be probability mass functions for discrete θ without any loss of generality.

The probability density obtained through a linear opinion pool of the expert opinions is

$$p(\theta) = \sum_{i=1}^{k} w_i p_i(\theta) \tag{3.9}$$

where $w_i \geq 0$ and $\sum w_i = 1$. This was proposed by Stone (1961) and is attributed to Laplace by Bacharach (1979). This is simply a mixture distribution and $p(\theta)$ is a valid probability density by construction. The appeal of this method is in its simplicity, and from a computational standpoint, it is also easy to sample from $p(\theta)$ especially when the $p_i(\theta)$ have the same form. Genest and Zidek (1986) note that (3.9) is usually

multimodal which means that there is no clear-cut consensus for a jointly preferred action (in decision making situations). They also point out the fact that it is not sensitive to the expert weights. Depending on the context, this may be advantageous or disadvantageous.

4.1.1. Independent and Logarithmic Opinion Pooling: Using the same premise as before, when the information sources are independent, a method called independent opinion pooling results in an overall prior distribution

$$p(\theta) = g \prod p_i(\theta) \tag{3.10}$$

where

$$\frac{1}{g} = \int \prod_{i=1}^{k} p_i(\theta) \, \mathrm{d}\theta \tag{3.11}$$

is a normalizing constant. The multiplication of prior distributions here is similar to the combination of independent likelihoods from statistical experiments (Berger, 1985). A generalization of (3.10) is logarithmic opinion pooling which results in the prior distribution

$$p(\theta) = \frac{\prod_{i=1}^{k} [p_i(\theta)]^{w_i}}{\int \prod_{i=1}^{k} [p_i(\theta)]^{w_i} \, \mathrm{d}\theta}, \tag{3.12}$$

$w_i \geq 0$ and $\sum w_i = 1$. Logarithmic pooling overcomes some of the issues associated with linear pooling. The resulting distribution is typically unimodal and less dispersed and it is externally Bayesian (Givens and Roback, 1999). The unimodality implies that it is more likely to indicate consensual values when decisions must be made. External Bayesianity, proved by Genest (1984), in the words of Givens and Roback (1999) is that "the pooling process produces the same result from combining all expert priors into a single aggregate prior and then updating with a likelihood as

from updating each expert's prior and then merging the resulting individual posterior distributions into a single group posterior." A weakness of this approach is that it has the characteristic that if any of the experts assigns a zero probability to some values of θ, the aggregated distribution summarily assigns a zero probability to those values regardless of the opinions of the other experts.

4.2. Supra-Bayesian Approach

The Supra-Bayesian approach uses Bayes' rule as a pooling operator. It is based on the assumption that there is a fictitious decision maker or overall expert who is the "synthetic personality" of the group of experts. In other cases, this supra-Bayesian is one to whom the group reports to. This person treats the opinions of the individual experts as data and updates his/her own prior using Bayes' Theorem. Thus, the posterior distribution represents the group consensus probability density (Genest and Zidek, 1986)

$$p(\theta|p_1, ..., p_k) \propto L(p_1, ..., p_k|\theta)p_s(\theta) \tag{3.13}$$

where $L : \theta \to [0, \infty)$ is the supra-Bayesian's likelihood function and $p_s(\theta)$ is his/her prior distribution. This method of combining priors is favored by Lindley (1983) and others and it has the intuitive Bayesian appeal in the sense that prior knowledge is updated using a likelihood. However, Genest and Zidek (1986) point out that in situations where the supra-Bayesian is only virtual, the choice of an appropriate likelihood falls on the group. In addition, they also observe that the supra-Bayesian's prior would have to be the object of consensus. From a computational standpoint, the implementation of this method is also non-trivial except perhaps in low-dimensional cases.

5. Implementation

The new strategies for designing robust experiments for nonlinear models discussed earlier are implemented using genetic algorithms. The GA uses genetic operators that mimic the processes of natural evolution like recombination, mutation and reproduction. The GA has been used to improve existing designs in the published literature for both linear and nonlinear models. Details of some of these improvements are given in Chapter 4.

GAs are well-suited to design problems in the sense that they search the design space using a population of designs. Theoretically, this ensures that the search is exhaustive if the GA is not terminated after an insufficient number of generations. Typically, a GA produces offspring for every set of parents. This choice has a bearing on how much time it takes to generate the optimal design because it impacts how exhaustively the design space of possible "chromosomes" or designs is searched. In problems where the optimal design is supported on only a few points, this may not be an issue. However, convergence to the optimal (or robust) design is slow when the design is supported on more than just a few points. In particular, implementing the classical GA for such problems requires computers with substantial resources. This is the case with nonlinear models and Bayesian designs where the number of support points increases with the variance of the parameters. The optimization problems solved using the GA are of the form

$$\max_{\xi \in \Xi} f(\xi) \tag{3.14}$$

where $f(\xi)$ is any of the functionals in (3.5) through (3.8). In this research, a new and efficient reproduction operator has been developed to reduce the amount of time needed to search for designs.

5.1. New GA Reproduction Operator

An efficient reproduction operator is developed to minimize the time to convergence to the robust designs based on the functionals proposed. A description of this operator is given here. Consider two designs

$$\xi^1 = \left\{ \begin{array}{cccccc} x_{11}, & x_{21}, & . & . & . & , x_{n1} \\ w_{11}, & w_{21}, & . & . & . & , w_{n1} \end{array} \right\} \text{ and } \quad \xi^2 = \left\{ \begin{array}{cccccc} x_{12}, & x_{22}, & . & . & . & , x_{n2} \\ w_{12}, & w_{22}, & . & . & . & , w_{n2} \end{array} \right\} \tag{3.15}$$

randomly selected for the reproduction process with a probability $p \in (0, 1)$ where x_{ij} is the ith point in design $j = 1, 2$. Similarly, w_{ij} is the ith weight in design j. Two offspring result from this genetic operation so that we have

$$\xi^{11} = \left\{ \begin{array}{cccccc} x^*_{11}, & x^*_{21}, & . & . & . & , x^*_{n1} \\ w^*_{11}, & w^*_{21}, & . & . & . & , w^*_{n1} \end{array} \right\} \text{ and } \quad \xi^{22} = \left\{ \begin{array}{cccccc} x^*_{12}, & x^*_{22}, & . & . & . & , x^*_{n2} \\ w^*_{12}, & w^*_{22}, & . & . & . & , w^*_{n2} \end{array} \right\} \tag{3.16}$$

The support points of the offspring will be a linear combination of two points chosen at random from parents ξ^1 and ξ^2, respectively, if blending is the reproduction operator used. That is, for example, if x_{m1} and x_{r2} are randomly selected from ξ^1 and ξ^2, respectively, then x^*_{11} in ξ^2 is

$$x^*_{11} = a x_{m1} + (1 - a) x_{r2} \tag{3.17}$$

for $a \in (0, 1)$, and $m, r \in \{1, 2, ..., n\}$. To improve the efficiency of the algorithm, the number of offspring can be tripled for each parent to give a total of six (instead of two) at the end of the reproduction operation. By taking advantage of the fact that both the support points and weights are optimized, the blending of the support points and the weights can be alternated. Thus, ξ^1 can have two offspring by fixing its support

points and blending its weights with those of ξ^2; or by blending its support points with those of ξ^2 and fixing the weights. The third offspring is produced through blending both support points and weights (with those of ξ^2) at the same time. Thus, ξ^1 can have 3 offspring, denoted by ξ^{a1}, ξ^{b1}, and ξ^{c1} where

1. ξ^{a1} results from fixing points, but blending weights,

2. ξ^{b1} results from blending points, but fixing weights and,

3. ξ^{c1} results from blending both points and weights.

Generating Bayesian optimal designs requires the evaluation of a multidimensional integral in most practical situations. Monte Carlo integration is used to approximate the integral in the GA. For example, approximation of the integral in (3.3) is

$$\Phi(\xi) = \int \log|M(\xi,\theta)|\, p(\theta)\, \mathrm{d}\theta \approx \frac{1}{N}\sum \log|M(\xi,\theta)| \tag{3.18}$$

where N, usually very large, is the number of random draws taken from $p(\theta)$.

The simulated annealing (SA) algorithm is a competitor to the GA in most optimization problems. The GA, though, has the advantage of searching the design space faster than the SA algorithm. However, the advantage of the SA algorithm lies in its ability to probabilistically improve a single design and move it towards the optimum. Therefore, the designs generated using the GA can be fed into a SA algorithm in order to improve it, if possible. This is useful because it serves as a check in cases where the design obtained by the GA does not converge sufficiently due to time constraints.

CHAPTER 4

DESIGN IMPROVEMENTS USING GENETIC ALGORITHMS

1. Overview

This section presents improvements over existing designs in the literature for some linear and nonlinear models. Improvements in the designs may be the result of changes to the weights w_i and/or design points x_i. The improvements in existing designs seen here motivate the use of genetic algorithms as the optimization mechanism for all of the work hereafter.

1.1. Bayesian T-optimal Designs

Ponce De Leon and Atkinson (1991) consider a design that discriminates between two competing (or rival) linear regression models in the presence of prior information about the model parameters over the $[-1, 1]$ design space. The models used by the authors had previously been presented in Atkinson and Fedorov (1975) where it was assumed one of the models is true and a locally T-optimal design was found. In Ponce De Leon and Atkinson (1991), there is a specified prior probability that each model is true and prior distributions for the model parameters in each of the models are also specified conditional on the prior probabilities. The aim of a Bayesian T-optimal design is to maximize the expected noncentrality parameter of the false model with the expectation taken over the models and the prior distributions. That is, a T-optimal design is one which provides the most powerful F-test for lack-of-fit of the false model. The true model η_t is one of two known functions $\eta_1(x, \theta_1)$ and $\eta_2(x, \theta_2)$ with respective prior probabilities π_{01} and $\pi_{02} = 1 - \pi_{01}$. The parameters $\theta_j,\ j = 1, 2$, are of dimension m_j and have prior distributions $p_{0j}(\theta_j)$ and $\Theta_j \in \mathcal{R}^{m_j}$

is the parameter space of θ_j. The criterion they maximize is

$$\Gamma(\xi) = \pi_{01} E_{\theta_1}\{\Delta_2(\xi, \theta_1)\} + \pi_{01} E_{\theta_2}\{\Delta_1(\xi, \theta_2)\} \tag{4.1}$$

and

$$\Delta_1(\xi, \theta_2) = \inf_{\theta_1 \in \Theta_1} \int \{\eta_2(x, \theta_2) - \eta_1(x, \theta_1)\}^2 \, \xi(\mathrm{d}x)$$
$$\Delta_2(\xi, \theta_1) = \inf_{\theta_2 \in \Theta_2} \int \{\eta_1(x, \theta_1) - \eta_2(x, \theta_2)\}^2 \, \xi(\mathrm{d}x)$$

are the noncentrality parameters for η_1 when η_2 is true and vice versa (Ponce De Leon and Atkinson, 1991). The models are

$$\eta_1(x, \theta_1) = \theta_{10} + \theta_{11} e^x + \theta_{12} e^{-x} \text{ and } \eta_2(x, \theta_2) = \theta_{20} + \theta_{21} x + \theta_{22} x^2. \tag{4.2}$$

The Bayesian T-optimal design the authors found given $\pi_{01} = 0.6$ and $\pi_{02} = 0.4$ is

$$\xi_P = \left\{ \begin{array}{ccccc} -1.0000 & -0.6634 & 0.1624 & 0.8466 & 1.0000 \\ 0.2438 & 0.4265 & 0.2535 & 0.0206 & 0.0556 \end{array} \right\} \tag{4.3}$$

using the discrete prior distribution in Table 4.1 with $\Gamma(\xi_P) = 0.8481 \times 10^{-3}$. Using the genetic algorithm, the Bayesian T-optimal design found with $\Gamma(\xi_{GA}) = 0.8994 \times 10^{-3}$ is

$$\xi_{GA} = \left\{ \begin{array}{ccccc} -1.0000 & -0.6614 & 0.1612 & 0.1700 & 1.0000 \\ 0.2486 & 0.4317 & 0.1236 & 0.1246 & 0.0716 \end{array} \right\}. \tag{4.4}$$

$\Gamma(\xi_{GA}) > \Gamma(\xi_P)$ implies the design generated by the GA is superior to the design presented by Ponce De Leon and Atkinson (1991), underscoring the efficiency and

performance of the GA compared to other optimization routines. The two designs are also similar with respect to certain weights and design points. For example, the design point 1.000 in ξ_{GA} is weighted more than that in ξ_P because of the absence of 0.8466 in ξ_{GA}.

$\pi_{01} = 0.6$				$\pi_{02} = 0.4$			
θ_{10}	θ_{11}	θ_{12}	$p_{01}(\theta_1)$	θ_{20}	θ_{21}	θ_{22}	$p_{02}(\theta_2)$
4.5	-1.5	-2.0	0.25	1.0	0.5	-2.0	0.23
4.0	-1.0	-2.0	0.14	0.8	0.4	-2.0	0.33
4.0	-1.0	-1.5	0.11	1.0	0.6	-1.5	0.17
5.0	-1.5	-1.5	0.06	1.2	0.5	-1.5	0.15
4.0	-2.0	-1.0	0.05	0.8	0.6	-1.0	0.12
4.5	-1.5	-1.5	0.08				
4.0	-1.5	-2.0	0.05				
4.0	-2.0	-2.0	0.12				
4.5	-2.0	-2.0	0.07				
5.0	-1.5	-2.0	0.07				

Table 4.1: Prior probabilities for the true model and prior distributions for θ_1 and θ_2.

1.2. Designs for a Compartmental Model

Atkinson et al. (1993) present designs for the properties of a compartmental model including the area under the curve (AUC), the time to maximum concentration (t_{max}) and the maximum concentration. They considered both locally- and Bayesian optimal designs. The model considered was

$$y = \theta_3(e^{-\theta_1 t} - e^{-\theta_2 t}) + \epsilon, \quad for \;\; t \geq 0 \text{ and } \theta_2 > \theta_1. \tag{4.5}$$

The observed errors ϵ are taken to be independent and identically distributed normal random variables with mean zero and variance σ^2. The model was fitted to data realized from an experiment in which six horses each received 15 mg/kg of theophylline

as aminophylline by intragastric administration. The analysis of data from horse number 3 provided prior parameter values

$$\theta_1 = 0.0589, \quad \theta_2 = 4.290, \quad \theta_3 = 21.80 \tag{4.6}$$

that were used to generate a locally c-optimal design for the AUC defined as

$$AUC = \int_0^\infty \eta(t, \theta)\, \mathrm{d}t = \frac{\theta_3}{\theta_2} - \frac{\theta_3}{\theta_1} = g(\theta). \tag{4.7}$$

where $\eta(t, \theta)$ is the expectation of the response function in equation 4.5. Interest is in minimizing the variance of $g(\hat{\theta})$. A design criterion used in this situation is the c-optimality criterion. The c-optimal design minimizes

$$Varg(\hat{\theta}) = Var(c^T\hat{\theta}) = c^T M^{-1}(\xi)c \tag{4.8}$$

where $c_i(\theta) = \dfrac{\partial g(\theta)}{\partial \theta_i}$ and $M^{-1}(\xi)$ is the inverse of the information matrix. The dependence of both c and M on θ is removed because θ is assumed to be known. The locally c-optimal designs obtained for the AUC by Atkinson et al. (1993) is

$$\xi_A = \left\{ \begin{array}{cc} 0.2331 & 17.6322 \\ 0.0135 & 0.9865 \end{array} \right\} \tag{4.9}$$

with a c-criterion value of 2193. Using the GA, the design obtained is

$$\xi_{GA} = \left\{ \begin{array}{cc} 0.2329 & 17.6179 \\ 0.0135 & 0.9865 \end{array} \right\} \tag{4.10}$$

with a c-criterion value of 2190 and thus is marginally more efficient for estimating AUC. The similarities in the two designs cannot be over-emphasized. In this case, the difference between the two designs is in the support points. An important thing to note about c-optimal designs is that they usually result in singular information matrices as can be seen from the fact that both ξ_A and ξ_G have $n = 2$ support points compared to $p = 3$ model parameters. In practice, the information matrices are *regularized* by adding some random quantity to the diagonal entries.

Further, a Bayesian D-optimal design was obtained by Atkinson et al. (1993) using uniform priors for θ_1 and θ_2. It should be recalled that θ_3 enters the model linearly and so does not have an impact on the design. The priors are $\theta_1 \sim [0.04884, 0.06884]$ and $\theta_2 \sim [3.298, 5.298]$ and the Bayesian D-optimal design obtained is

$$\xi_A = \left\{ \begin{array}{ccc} 0.2288 & 1.4170 & 18.4513 \\ 0.3333 & 0.3334 & 0.3333 \end{array} \right\} \tag{4.11}$$

and a D-criterion value of 7.3760. The design

$$\xi_{GA} = \left\{ \begin{array}{ccc} 0.2428 & 1.4514 & 18.0698 \\ 0.3287 & 0.3524 & 0.3189 \end{array} \right\} \tag{4.12}$$

with a D-criterion value of 7.4953 is obtained by the GA using the proposed new reproduction operator. The designs may be similar in terms of support points but vary more in terms of where they concentrate experimental effort, that is, the weights.

CHAPTER 5

APPLICATIONS

1. Pharmacokinetics: One-Compartment Open Model

In drug development research, the processes of Absorption, Distribution, Metabolism and Excretion (ADME) or Absorption, Distribution and Elimination (ADE) are of critical importance. When a dose D of a drug is administered, there is a site of action where the drug will have its effect. Concentrations of the drug at this location cannot be directly measured and are determined by ADME. In general, a pharmacologist is interested in keeping the concentration of the drug high enough to achieve a desirable response and low enough to avoid toxicity. Understanding ADME allows manipulation of concentrations through different dosing strategies. Although concentrations at the site of action are not directly measurable, concentrations of the drug in the blood, plasma or serum reflect those at the site. Consequently, information about ADME is gained by measuring blood concentrations over time. This study is what is usually called a pharmacokinetics (PK) study. Of particular interest is the optimal choice of time points after drug administration at which to measure the concentration of the drug in the blood.

Compartmental models are used in PK studies to represent the body of an individual subject. A schematic of a one-compartment model is shown in Figure 5.1. This involves two parameters which quantify the rate at which the drug is absorbed into the body (or compartment) κ_a and the rate at which the drug is eliminated κ_e.

The one-compartment model, although quite simplistic, is used in drug development. It relates the amount X of the drug in the blood (or plasma) to the time after

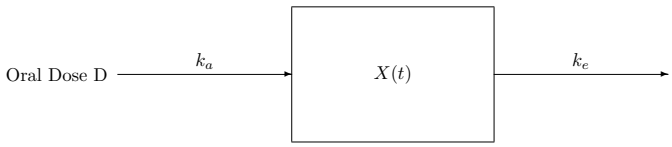

Figure 5.1: Schematic of a One-Compartment Open Model.

drug administration t:

$$E(X|t) = \frac{\kappa_a}{\kappa_a - \kappa_e}\{\exp(-\kappa_e t) - \exp(-\kappa_a t)\}, \quad t > 0. \tag{5.1}$$

Knowledge of the PK parameters enable the pharmacologist to predict the concentration that would be achieved by a subject at any time following the dose. Thus, it is important that the PK parameters are estimated efficiently by deliberating choosing the times at which to measure the amount of the drug in the blood. A plot of the one-compartment model is shown in Figure 5.2 for $\kappa_a = 0.05$ and $\kappa_e = 0.005$.

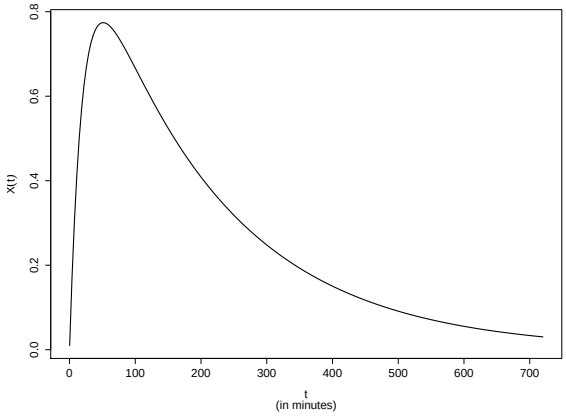

Figure 5.2: An example plot of a One-Compartment Open Model with $\kappa_a = 0.05$ and $\kappa_e = 0.005$.

2. Dose-Response Studies: Four-Parameter Logistic Model

Dose-response studies involve the use of bioassays to determine the potency or toxicity of a chemical substance on an organism. A bioassay (or biological assay) is simply a type of scientific experiment. In a typical bioassay, a specified amount of a chemical substance is administered and a response measured. In the vast majority of dose-response studies, the goal of investigators is the inhibitory concentration (IC_{50}), that is, the dose amount that inhibits 50% of a biological process.

The assumed relationship between the response and the logarithm of the dose in these studies is nonlinear, specifically, sigmoidal. A wide variety of models approximate this relationship. For example, when the response is quantal, a logistic regression model is used for analysis. For continuous responses, the four-parameter logistic (4PL) model is used. The mean response given a dose amount x is

$$\eta(x, \theta) = \theta_3 + \frac{\theta_4 - \theta_3}{1 + (x/\theta_1)^{\theta_2}}. \tag{5.2}$$

θ_3 and θ_4 are the minimal and maximal responses respectively, thus, $\theta_4 > \theta_3 > 0$. θ_1 denotes ED_{50} (or LD_{50}) and θ_2 is a slope parameter. Plots of the 4PL model for $\theta_1 = 15.03$, $\theta_2 = 1.31, -1.31$, $\theta_3 = 530$ and $\theta_4 = 1587$ are shown in Figure 5.3. Experimental design in bioassays is important because optimal choices of dose concentrations lead to precise estimation of the model parameters.

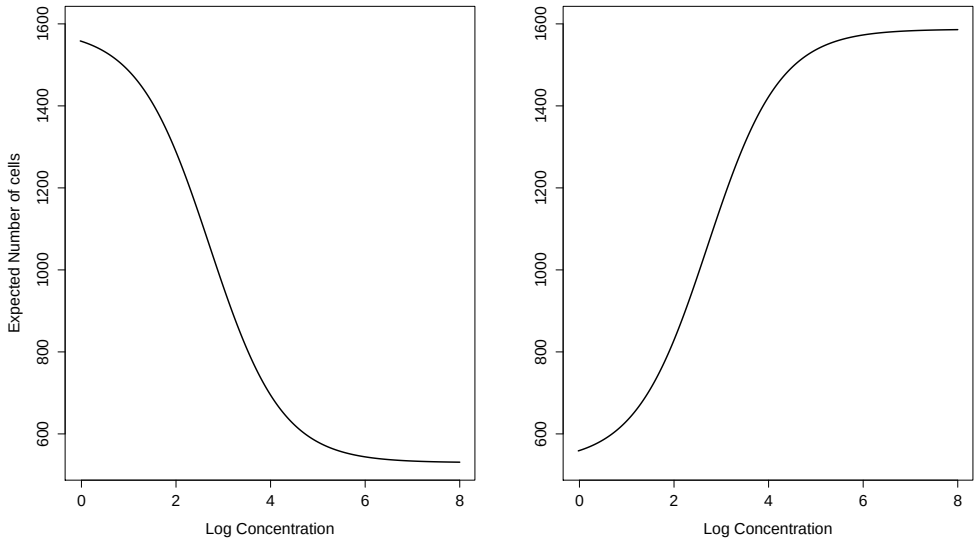

Figure 5.3: Plots of the 4PL model with $\theta_1 = 15.03$, $\theta_3 = 530$, $\theta_4 = 1587$ for slope parameters $\theta_2 = 1.31$ (left) and $\theta_2 = -1.31$ (right).

3. Enzyme Kinetics: Michaelis-Menten Model

The primary function of enzymes is to enhance rates of reactions so that they are compatible with the needs of an organism. Knowledge of the properties of enzymes leads to an in-depth understanding of the way the body functions in health and can provide useful information about how problems arise when the normal processes are impacted by disease, trauma or environmental agents (Matthews and Allcock, 2004).

A reaction produces a *product* from a *substrate*. For many enzymes, the rate of catalysis v, defined as the number of moles of product formed per second, is a function of the substrate concentration s (Berg et al., 2002). The Michaelis-Menten (MM) equation governs the kinetics of many such reactions. It relates mean v to s through

$$E(v|s) = \frac{V_{max}s}{K_M + s}. \tag{5.3}$$

Here $V_{max} > 0$ is the maximum rate at which substrate can be turned into product. The parameter $K_M > 0$ is the Michaelis parameter and it is the substrate concentration at which the reaction proceeds at half its maximum rate.

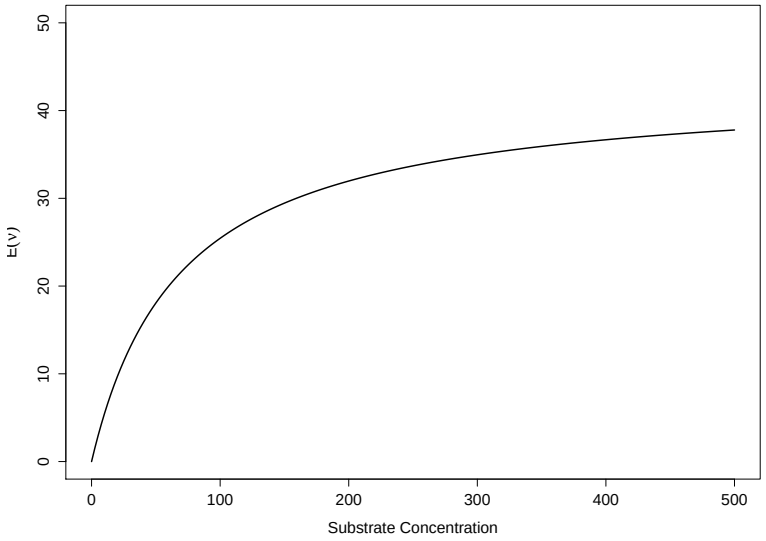

Figure 5.4: An example plot of the Michaelis-Menten model with $V_{max} = 43$ and $K_M = 69$.

Designs will be obtained for the one-compartment and Michaelis-Menten models in this chapter. Appendix C has results for the 4PL model.

4. Designs for the Compartmental Model

Robust designs are generated here for the one-compartment model shown in Figure 5.2 based on the new strategies proposed in the previous chapter. As mentioned above, this is an important model in drug development and designing an experiment to estimate the model parameters requires prior knowledge of them. The assumption is made here that prior distributions for the model parameters are elicited from

four independent experts, for example, and thus four prior distributions are used. For the compartmental model the biological parameters κ_a and κ_e are both positive with $\kappa_a > \kappa_e$. It is important to take this information into account when thinking about plausible prior distributions for these parameters. For example, skewed rather than symmetric distributions could be biologically more plausible given the nature of the parameters. Consequently, the distributions considered here are the lognormal, triangular and normal distributions which reflect right-skewed, left-skewed, and symmetric distributions respectively. The normal distribution, although not ideal biologically, can be parameterized reasonably to reflect the relationship between the model parameters.

To estimate the multidimensional integral in equation (3.18) on page 66, draws need to be made from the prior distributions. Given that $\kappa_a > \kappa_e$, draws for κ_a are made conditional on draws for κ_e. For the lognormal prior, for example, κ_e is defined as $\kappa_e \sim LN(\mu_e = -5.64, \sigma_e = 0.83)$. The lognormal distribution parameter values are based on Rodda et al. (1975). To ensure $\kappa_a > \kappa_e$, define $\delta \sim LN(\mu_\delta = -3.16, \sigma_\delta = 0.58)$ such that

$$\kappa_a = \kappa_e + \delta. \tag{5.4}$$

It is important to note that κ_a does not have a lognormal distribution although it is the sum of two lognormal random variables. In particular, its distribution has no closed form but it can be approximated by a lognormal distribution. This fact becomes useful later when designs based on aggregated prior distributions are considered. Plots of the distributions of κ_a and κ_e (based on $N = 10000$ draws each) are shown in Figure 5.5.

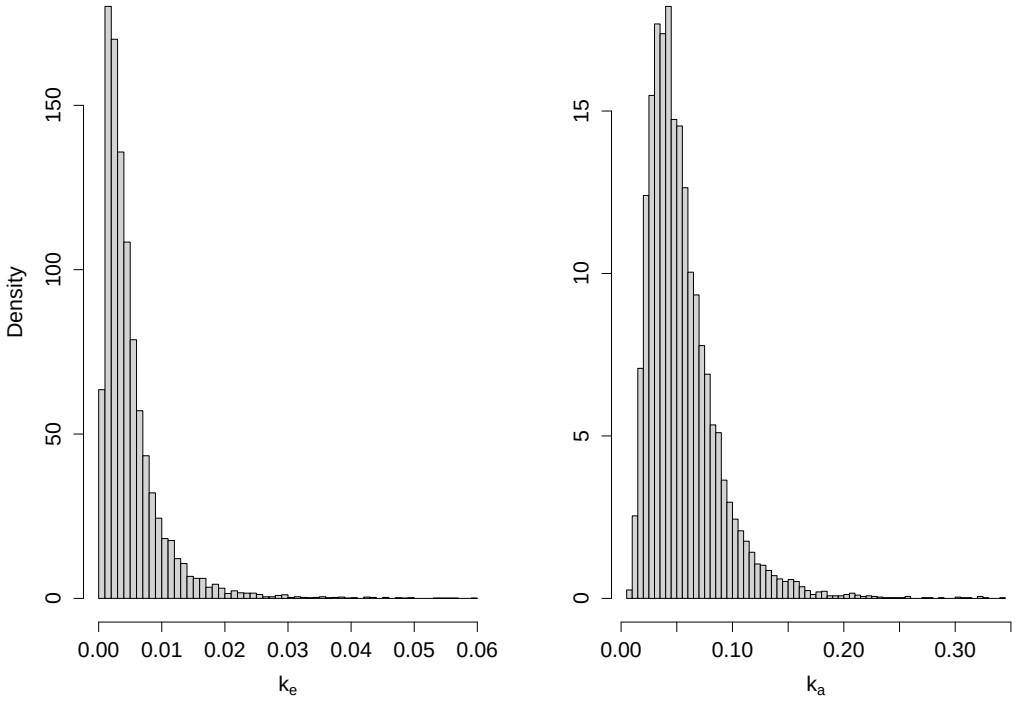

Figure 5.5: Left: Distribution of κ_e based on 10000 draws from $LN(\mu_e, \sigma_e)$. Right: Distribution of κ_a drawn conditionally on 10000 draws of κ_e.

The probability density function of a triangular distribution in terms of minimum a, mode c and maximum b is given in Equation (5.5). For the triangular prior, the draws for κ_a are made similarly with $\kappa_e \sim T(a = 0.0001, c = 0.008, b = 0.01)$ and $\delta \sim T(a = 0.01, c = 0.1, b = 0.1)$. The distribution of κ_a is once again not available in closed form. Plots of the distributions of κ_e and κ_a are shown in Figure 5.6. It is worth mentioning that generating draws this way implicitly incorporates a dependence between the parameters.

$$f(\theta|a,b,c) = \begin{cases} 0 & \text{for } \theta < a, \\ \frac{2(\theta-a)}{(b-a)(c-a)} & \text{for } a \leq \theta \leq c, \\ \frac{2(b-\theta)}{(b-a)(b-c)} & \text{for } c < \theta \leq b, \\ 0 & \text{for } b < \theta. \end{cases} \tag{5.5}$$

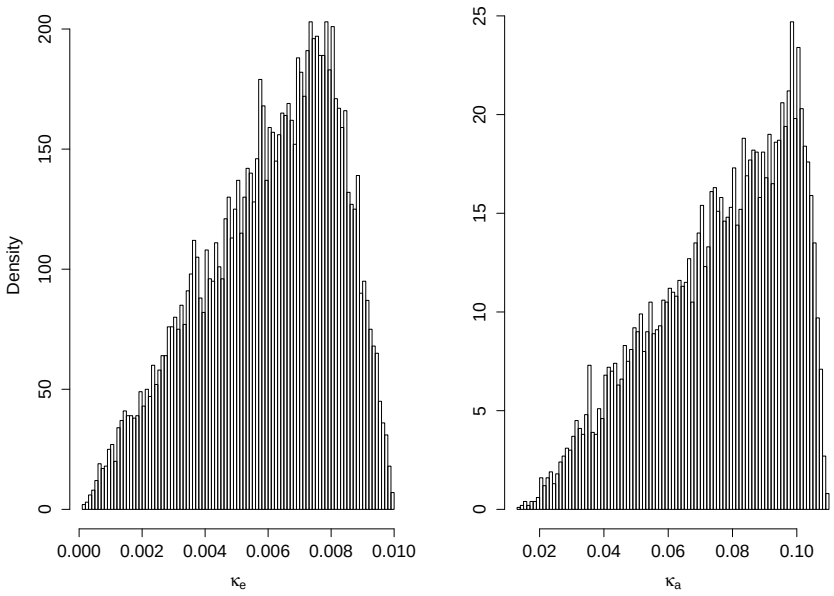

Figure 5.6: Left: Distribution of κ_e based on 10000 draws from $T(a = 0.0001, c = 0.008, b = 0.01)$. Right: Distribution of κ_a drawn conditionally on 10000 draws of κ_e.

Bivariate normal priors with the same mean but different covariance matrices are used. The first of these priors is more informative (that is, more precise) with

$$\begin{pmatrix} \kappa_a \\ \kappa_e \end{pmatrix} \sim \mathcal{MVN}\left(\begin{pmatrix} 0.07 \\ 0.009 \end{pmatrix}, \quad \begin{pmatrix} 0.005^2 & 0 \\ 0 & 0.0005^2 \end{pmatrix} \right). \tag{5.6}$$

The second bivariate normal prior is more variable (or vague) with

$$\begin{pmatrix} \kappa_a \\ \kappa_e \end{pmatrix} \sim \mathcal{MVN}\left(\begin{pmatrix} 0.07 \\ 0.009 \end{pmatrix}, \quad \begin{pmatrix} 0.01^2 & 0 \\ 0 & 0.001^2 \end{pmatrix} \right). \tag{5.7}$$

These priors are in Figure 5.7 with the informative prior on the left and the vague prior on the right.

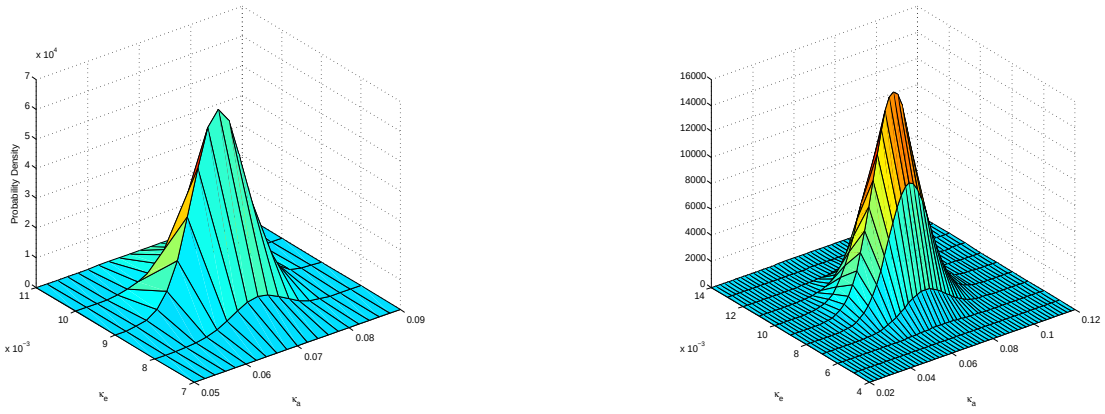

Figure 5.7: Left: Three-dimensional plot of the *informative* bivariate normal distribution. Right: Three-dimensional plot of the *vague* bivariate normal prior distribution.

The designs in this section will be generated assuming parameter estimation is of interest. For example, the PK parameters κ_a and κ_e must be efficiently estimated for

the pharmacologist to recommend dosing strategies. Thus, for the maximin design to be obtained, the proposed criterion in equation (3.5) is maximized using a GA and $E_i\{\phi(\xi,\theta)\}$ is the expected D-criterion value with respect to the ith prior distribution.

4.1. Maximin Design

The maximin design for the one-compartment model is a 5-point design

$$\xi_M = \left\{ \begin{array}{ccccc} 13.01 & 120.51 & 124.18 & 127.07 & 576.07 \\ 0.4963 & 0.1365 & 0.2015 & 0.1643 & 0.0014 \end{array} \right\}. \tag{5.8}$$

Generally, a design that spreads its points over the design space is a good design. This, in particular, is a characteristic of good designs for linear models. Designs for nonlinear models, however, do not always have this property. For a model like the one-compartment model, it is imperative for data collection to be done in such a way that the processes of absorption, distribution and elimination can be adequately captured representing regions of the curve with different and changing slopes. A design that samples only at the extremes or only in the distribution phase may not be efficient for estimating the model parameters. The maximin design in equation (5.8) concentrates about half of its mass near 13 minutes and the remaining half of the observations at 120, 124 and 127 minutes after the drug is administered with a minimal sampling effort at about 576 minutes. The closeness of the middle three design points may indicate that a 3-point design and not necessarily a 5-point design may suffice for this model. In practice, the design points 120, 124 and 127 may be replaced by their average (or median). A 3-point design is admissible for the one-compartment model because it is a two-parameter model. The 5-point design, though, may be preferred

for the purpose of checking lack of fit. The plot of the one-compartment model with the points of support of the maximin design are shown in Figure 5.8.

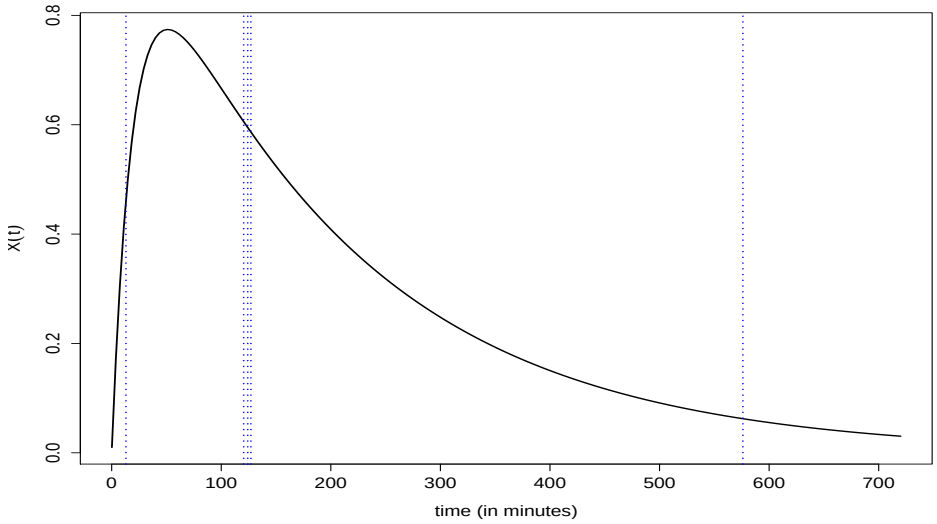

Figure 5.8: The one-compartment open model with the points of support (dashed vertical lines) of the maximin design.

The maximin design ξ_M is compared to Bayesian D-optimal designs based on the lognormal, triangular and two bivariate normal distributions. First, the Bayesian optimal designs obtained by maximizing the criterion in equation (3.3) are shown in Table 5.1. These designs are based on $N = 2000$ draws from each of the respective prior distributions. The designs essentially exhibit the same property as the maximin design in that most of the mass is concentrated on earlier time points. In fact, this property is common to almost all designs for compartmental models found in the literature.

Possible comparisons that can be made in order to demonstrate the robustness of the maximin D-optimal design are enumerated below. These include:

1. Compare ξ_M to the Bayesian D-optimal designs in terms of the minimum D-criterion values,

2. Obtain distributions of D-criterion using ξ_M and the Bayesian D-optimal designs over $N = 10000$ draws from each of the prior distributions,

3. Obtain the the distribution of efficiency of ξ_M relative to locally optimal designs with respect to each of the prior distributions and,

4. Compare the Bayesian D-optimal designs in terms of their D-criterion values.

Prior Distribution	Bayesian Optimal Design
Lognormal	$\xi_L = \left\{ \begin{array}{ccccc} 16.54 & 154.19 & 296.10 & 370.32 & 435.40 \\ 0.4943 & 0.1982 & 0.0781 & 0.2162 & 0.0133 \end{array} \right\}$
Triangular	$\xi_T = \left\{ \begin{array}{ccccc} 12.51 & 115.09 & 309.09 & 648.34 & 682.50 \\ 0.4815 & 0.4149 & 0.0858 & 0.0053 & 0.0126 \end{array} \right\}$
MVN (Informative)	$\xi_{Nf} = \left\{ \begin{array}{ccccc} 13.04 & 152.82 & 290.81 & 342.20 & 499.10 \\ 0.4952 & 0.4897 & 0.0056 & 0.0064 & 0.0031 \end{array} \right\}$
MVN (Vague)	$\xi_{Nv} = \left\{ \begin{array}{ccccc} 13.68 & 145.38 & 170.26 & 339.28 & 389.94 \\ 0.4865 & 0.4737 & 0.0230 & 0.0058 & 0.0110 \end{array} \right\}$

Table 5.1: The Bayesian D-optimal designs based on each of the four prior distributions.

The first of the four possible comparisons is given in Table 5.2. The minima in terms of D-criterion values were found using $N = 10000$ draws from each of the four prior distributions.

Prior Distribution	Minimum D-criterion Value	
	Bayesian	Maximin
Lognormal	-12.3672	-6.2035
Triangular	7.7921	7.9209
MVN (Informative)	8.3627	8.5050
MVN (Vague)	7.3435	7.5968

Table 5.2: Comparison of the minimum D-criterion values across the four prior distributions to the minimum D-criterion value of the maximin D-optimal design across the same priors. This is based on $N = 10000$ draws from the each of the priors.

The robustness of the maximin D-optimal design (ξ_M) can be seen from Table 5.2. It must be noted that in calculating D-criterion, the logarithm of the determinant of the information matrix is used. It maximizes the minimum D-criterion value for each of the four prior distributions. The implication of this is that it is reasonable to collect data based on the time points in ξ_M because it maximizes the least information about the PK parameters that can be gained by using any one of the prior distributions. In particular, ξ_M maximizes the determinant of the least informative information matrix. In other words, it improves the worst possible variance-covariance matrix of the estimated model parameters. Further, ξ_M improves the worst possible confidence ellipsoid of the estimated parameters.

The second comparison as outlined above is to evaluate ξ_M over $N = 10000$ draws from each of the prior distributions. Each Bayesian D-optimal design is also evaluated over the same set of draws and comparisons made. Boxplots of the D-criterion values obtained are shown in Figure 5.9.

The information from the plots is summarized in Table 5.3 with the percentiles based on ξ_M in parentheses. The plots in Figure 5.9 give a strong indication that ξ_M is a reasonable compromise (or robust) design. With the exception of the lognormal

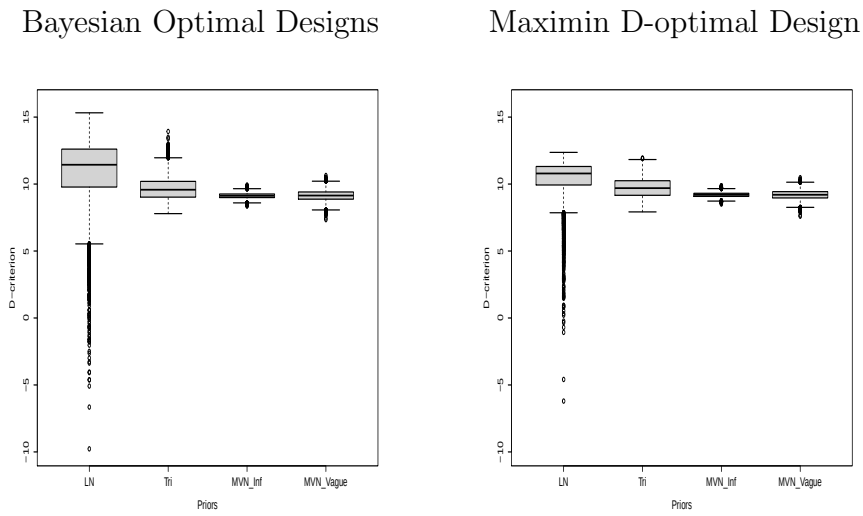

Figure 5.9: Left: Boxplots of D-criterion values of the Bayesian D-optimal designs evaluated over $N = 10000$ draws from the respective prior distributions. Right: Boxplots of the D-criterion values of the maximin D-optimal design evaluated over $N = 10000$ draws from each of the prior distributions.

prior, the percentiles of ξ_M seem to closely approximate those from the triangular and normal distributions as can be seen in the density plots below. Thus, in the case of prior ambiguity, ξ_M is a good compromise design. It must also be noted that although the design based on the lognormal prior has higher D-criterion values, it also has the lowest minimum D-criterion value as observed in Table 5.2. As a result ξ_M is preferred to any one of the Bayesian designs.

Further, the distribution of efficiency of ξ_M relative to locally optimal designs over the priors is also investigated. A design is said to be locally optimal if its optimality is conditional on a particular value of the unknown parameter vector. Locally optimal designs will be denoted as ξ_θ. Two designs can be compared by calculating relative efficiency. The relative D-efficiency of a design ξ_1 compared to another design ξ_2 at a

Prior Distribution	Percentiles			
	$25th$	$50th$	$75th$	$90th$
Lognormal	9.78 (9.93)	11.44 (10.79)	12.61 (11.31)	13.33 (11.60)
Triangular	9.02 (9.16)	9.58 (9.70)	10.20 (10.25)	10.86 (10.75)
MVN-Inf	8.99 (9.09)	9.12 (9.19)	9.26 (9.31)	9.38 (9.42)
MVN-Vague	8.87 (8.96)	9.14 (9.20)	9.41 (9.43)	9.65 (9.64)

Table 5.3: Comparison of the percentiles of D-criterion values of ξ_M (in parentheses) and the Bayesian D-optimal designs based on $N = 10000$ draws from the respective prior distributions.

particular parameter value θ is

$$D_{rel-eff}(\theta) = \left\{ \frac{|M(\xi_1, \theta)|}{|M(\xi_2, \theta)|} \right\}^{1/p} \tag{5.9}$$

where $|M(\xi, \theta)|$ is the determinant of the information matrix M using a design ξ and evaluated at $\theta \in \mathcal{R}^{p \times 1}$, where p is the number of parameters in the model. For the one-compartment model $p = 2$. $D_{rel-eff}(\theta) > 1$ implies that ξ_1 is more efficient than ξ_2 for the particular θ. To compare the maximin D-optimal design to locally optimal designs based on the lognormal prior, for example, the relative D-efficiencies will be of the form

$$D_{rel-eff}(\theta) = \left\{ \frac{|M(\xi_M, \theta)|}{|M(\xi_\theta, \theta)|} \right\}^{1/p} \tag{5.10}$$

where the θ are drawn from the lognormal prior distribution. The distribution of $D_{rel-eff}$ of ξ_M over the lognormal and triangular prior distributions are given in Figure 5.11.

In making the plots in Figure 5.11, relative efficiencies greater than 7 were considered extreme and therefore, excluded. These large relative efficiencies could be a result of the maximin D-optimal design being highly efficient or just a consequence of the

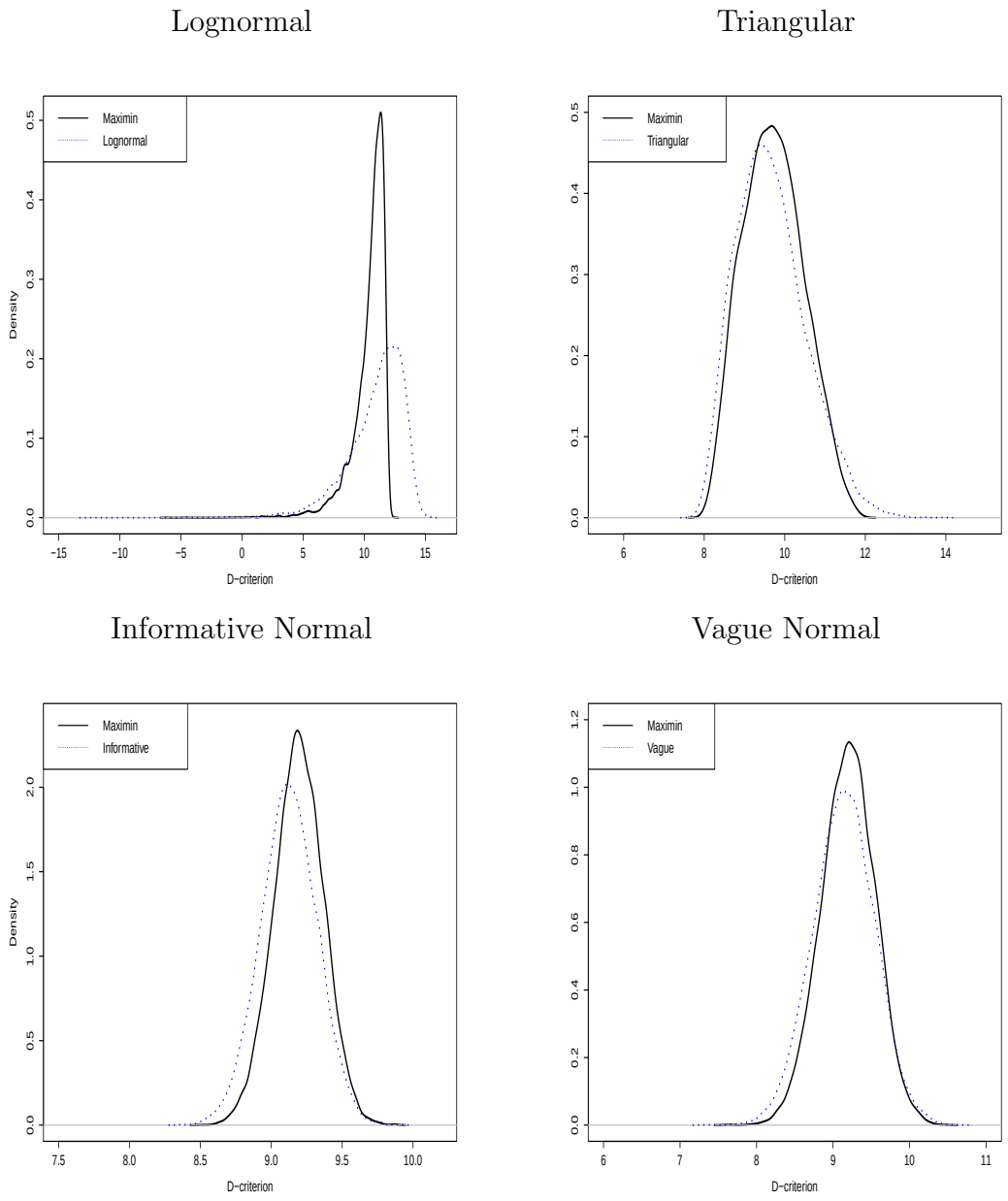

Figure 5.10: Density plots comparing the distribution of D-criterion values of the Bayesian designs to the maximin D-optimal design.

GA not converging to the optimal local D-criterion values. Both of the distributions are skewed towards larger relative efficiency values but focus should not necessarily be on large values of $D_{rel-eff}$. Also evident from the plots is the possibility of low

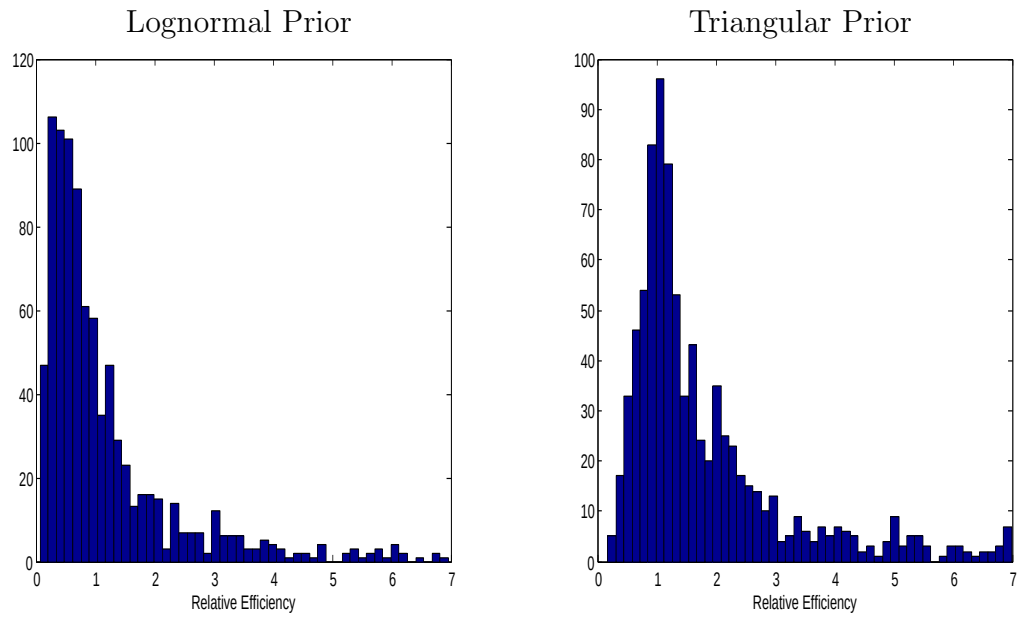

Figure 5.11: Left: Distribution of $D_{rel-eff}$ of ξ_M relative to $N = 1000$ locally optimal designs based on the lognormal prior distribution. Right: Distribution of $D_{rel-eff}$ of ξ_M relative to $N = 1000$ locally optimal designs based on the triangular prior distribution.

relative efficiencies. These low relative efficiencies are not unexpected because ξ_M is not maximin D-efficient but maximin D-optimal. In obtaining a maximin D-efficient design, the optimization is done by taking the locally optimal designs into account. This is what is done in Pronzato and Walter (1988) and is advisable when p is at most 2 and the parameters belong to small intervals. This is obviously not the case where prior distributions and not small intervals are being considered. At the very minimum, it is computationally prohibitive. Table 5.4 summarizes the information in Figures 5.11 and 5.12 which present summaries of the distribution of relative D-efficiency of the maximin design.

It is noted that although ξ_M has poor minimum efficiencies for the lognormal and triangular priors, the median efficiencies are quite high for all priors. This may

Summary Statistic	Relative D-efficiency			
	Lognormal	Triangular	MVN-Inf	MVN-Vague
Minimum	0.0704	0.1730	0.9915	0.8808
Q_1	0.4342	0.9626	1.2844	1.3680
Median	0.7448	1.2965	1.9347	2.0187
Q_3	1.3783	2.1917	3.1970	3.1444

Table 5.4: Summary statistics based on the relative efficiency plots in Figure 5.11.

suggest perhaps that the maximin D-optimal design could be a good substitute for the maximin D-efficient designs of Pronzato and Walter (1988).

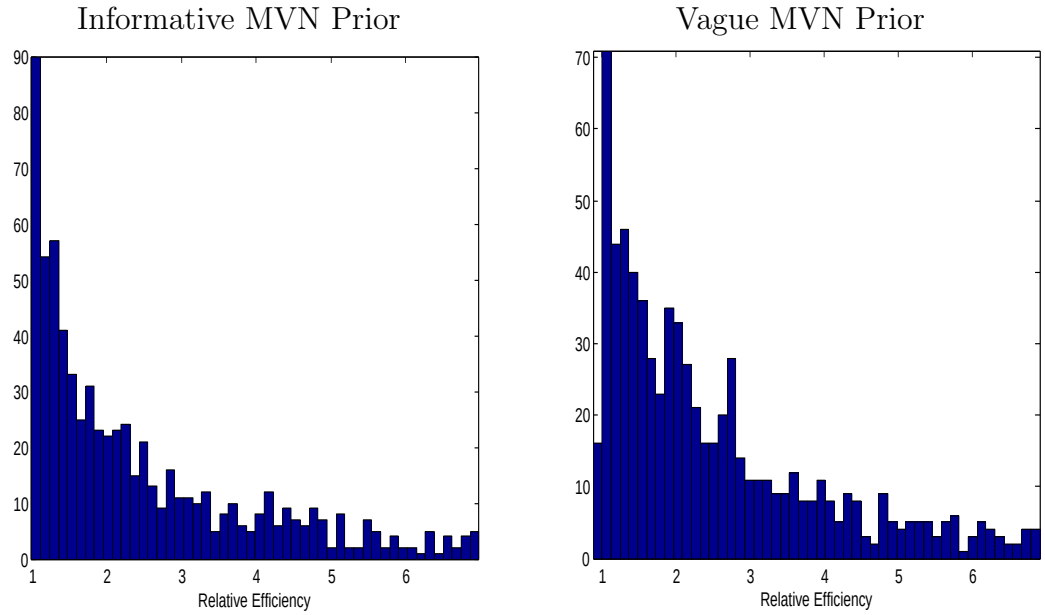

Figure 5.12: Left: Distribution of $D_{rel-eff}$ of ξ_M relative to $N = 1000$ locally optimal designs based on the informative MVN prior distribution. Right: Distribution of $D_{rel-eff}$ of ξ_M relative to $N = 1000$ locally optimal designs based on the vague MVN prior distribution.

It can be observed from the boxplots in Figure 5.9 that for the lognormal prior distribution, both the Bayesian D-optimal and maximin D-optimal designs have a considerable proportion of low D-criterion values. It is of interest to have information about the parameter values for which these D-criterion values occur. Color maps of the D-criterion values are made for each of the four Bayesian D-optimal designs and the maximin D-optimal design to investigate this in Figure 5.13.

It is quite evident from the plots that the low D-criterion values occur at high values of the parameters for all the priors. In particular, the lognormal priors have larger parameter values compared to the other priors because of the long right-tail. The parameter values in the tails have low probability values from the perspective of the expert from whom they were elicited. This notwithstanding, it is still imperative to find a design that is good for such unlikely parameter values because the true parameter values are unknown.

4.2. Product Design

This section presents a design based on the new product criterion proposed in the previous chapter. Recall that this design maximizes the product of expected utilities across the prior distributions. The product design for the compartmental model is

$$\xi_P = \left\{ \begin{array}{ccccc} 13.53 & 152.75 & 283.39 & 437.04 & 505.73 \\ 0.5008 & 0.4941 & 0.0011 & 0.0031 & 0.0009 \end{array} \right\}, \tag{5.11}$$

and the support points are shown in Figure 5.14. The product design in equation (5.11) is similar to the maximin in terms of the concentration of sampling effort in the sense that the vast majority of observations are to be made at earlier time points (13.53 to 152.75 minutes after the dose has been administered). They are different,

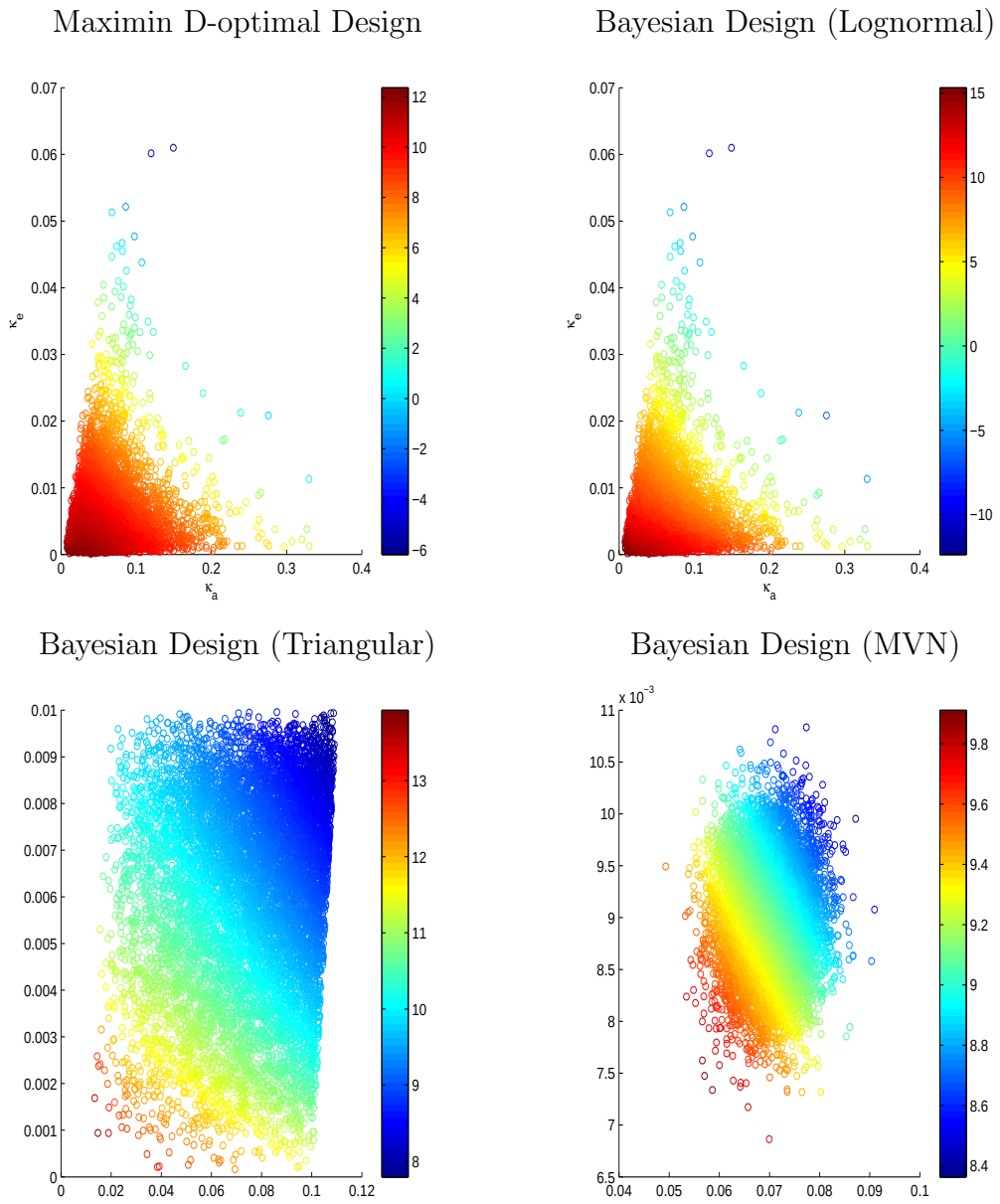

Figure 5.13: Topleft: Heatmap of D-criterion values of ξ_M based on $N = 10000$ from the lognormal prior distribution. Topright: Heatmap of D-criterion values of ξ_L based on $N = 10000$ from the lognormal prior distribution. Bottom left: Heatmap of D-criterion values of ξ_T based on $N = 10000$ from the triangular prior distribution. Bottom right: Heatmap of D-criterion values of ξ_{Nf} based on $N = 10000$ from the informative multivariate normal prior distribution.

however, in terms of spread of the support points as seen in Figure 5.14. Comparisons similar to those made earlier will be done to demonstrate the robustness of ξ_P. Table 5.5 compares the minimum D-criterion values of ξ_P to those of the Bayesian D-optimal designs. The minimum D-criterion values for the Bayesian designs remain unchanged since the same set of $N = 10000$ draws are used.

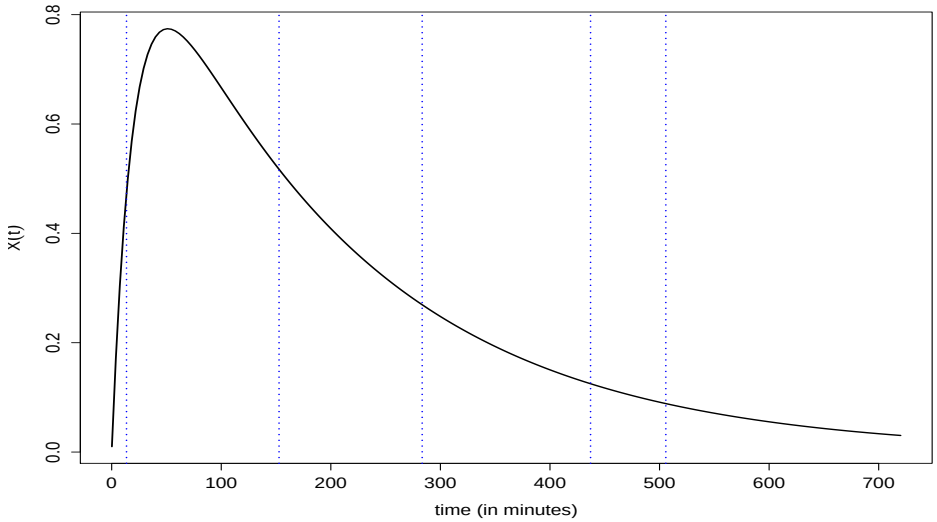

Figure 5.14: The one-compartment open model with the points of support (dashed vertical lines) of the product design.

It can be seen, though, that the minimum D-criterion value of the product design is greater than the smallest D-criterion value for the Bayesian designs, that is, -9.5757 compared to -12.3672. This minimum D-criterion value is less than what what was attained by the maximin design. This is not unexpected because the product design is not a maximin D-optimal design. Thus, the maximin D-optimal design was expected to have a larger minimum D-criterion value across the priors compared to the product design.

Prior Distribution	Minimum D-criterion Value	
	Bayesian	Product
Lognormal	-12.3672	-9.5757
Triangular	7.7921	7.7405
MVN (Informative)	8.3627	8.3587
MVN (Vague)	7.3435	7.3241

Table 5.5: Comparison of the minimum D-criterion values across the four prior distributions to the minimum D-criterion value of the product design across the same priors. This is based on $N = 10000$ draws from the each of the priors.

Density plots comparing the distribution of D-criterion values of the Bayesian designs to those of the product design are in Figure 5.15. It can be seen that the density plots comparing the D-criterion values based on the lognormal distribution to those of the product design are similar to the corresponding plots in Figure 5.10. The density plots of D-criterion values based on the triangular distribution and the product design are also similar to those in the previous figure. The density plots for the informative and vague multivariate normal distributions indicate that the product design is a good approximation to the Bayesian designs based on these priors. It can be observed that the product design more closely approximates the designs based on the latter prior distributions than the maximin D-optimal design. This indicates that the product design may be a reasonable compromise design in the face of prior ambiguity than the maximin design.

Comparison of the product design to $N = 1000$ locally D-optimal designs across each of the prior distributions is done using relative efficiency as defined in equation (5.9). Summaries of the distribution of D-efficiency of the product design relative to the four prior distributions is given in Table 5.6.

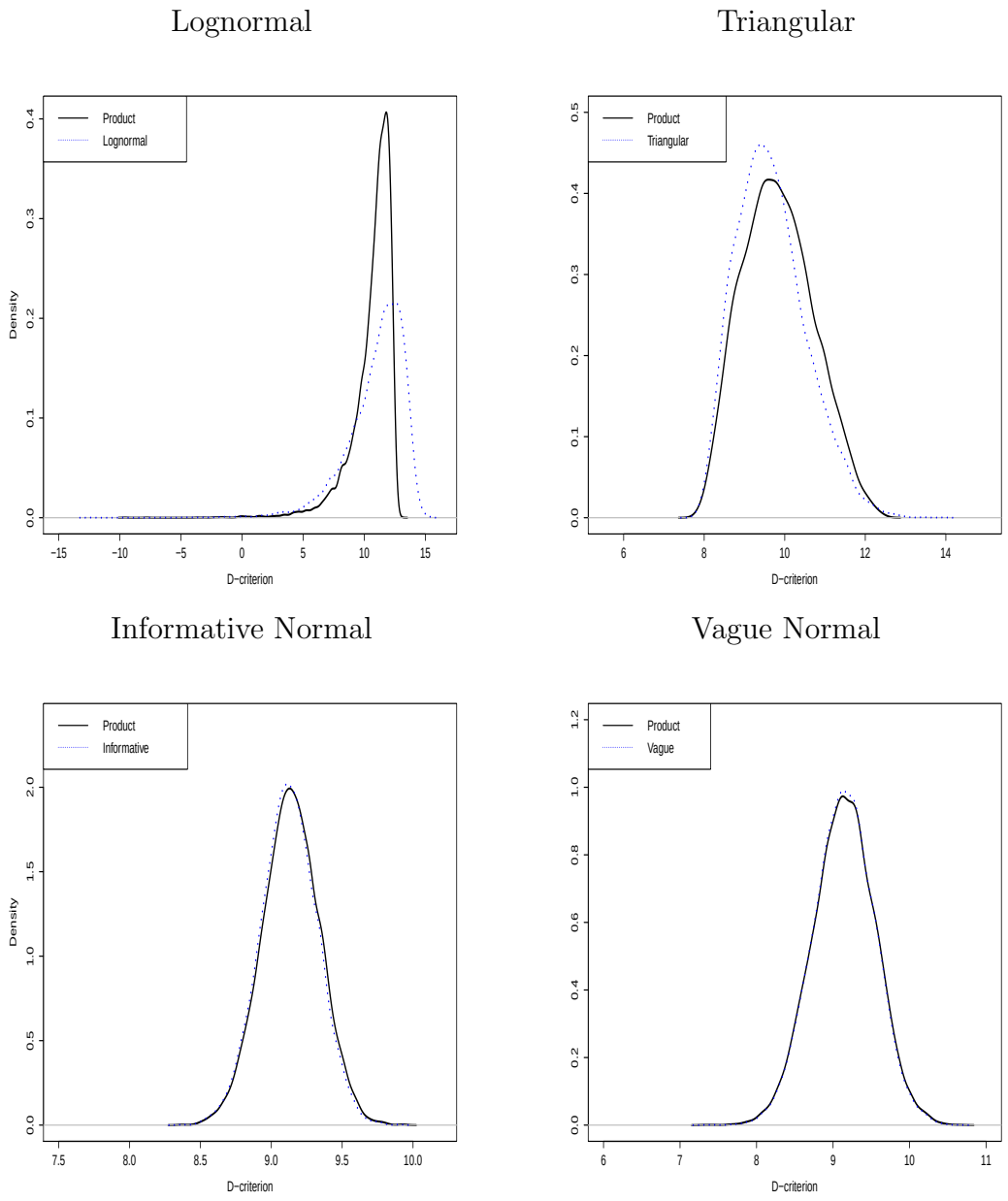

Figure 5.15: Density plots comparing the distribution of D-criterion values of the Bayesian designs to the product design.

A quick comparison of the values in Tables 5.4 and 5.6 indicates that the latter has consistently larger values. It is important to emphasize that this difference in relative efficiency between the product design and the maximin D-optimal design is

Summary Statistic	Relative D-efficiency			
	Lognormal	Triangular	MVN-Inf	MVN-Vague
Minimum	0.0979	0.2351	0.9726	0.9187
Q_1	0.5863	1.0616	1.4585	1.5448
Median	0.9714	1.6104	2.8631	2.7455
Q_3	2.1612	3.3963	8.8100	9.2923

Table 5.6: Summary statistics of the distribution of relative efficiency of the product design.

as expected given that the maximin D-optimal design addresses a different robustness objective than the product design. In particular, the maximin criterion is a "worst-case" criterion.

4.3. Weighted Product Design

The product design can be modified to incorporate varying *a priori* weights in prior distributional assumptions as mentioned in Chapter 3. In situations where the experts have different levels of training (or experience), it is a reasonable practice to weight their assessments (or assumptions) about the model parameters to reflect their "credibility". The appropriate criterion is proposed in equation (3.7).

A weighted product design was generated with weights $w = (0.52, 0.16, 0.16, 0.16)$ assigned respectively, to the lognormal, triangular, and the normal priors. Thus, the lognormal prior is considered more plausible than the other priors. This results in the design

$$\xi_W = \left\{ \begin{array}{ccccc} 15.78 & 137.91 & 218.33 & 310.46 & 383.93 \\ 0.4840 & 0.1837 & 0.1812 & 0.1000 & 0.0511 \end{array} \right\}. \tag{5.12}$$

Noticeable in the design is the fact that most of the experimental effort is not concentrated on just the first two support points. There is also an appreciable amount

of weight on the latter three points. This pattern is not consistent with what has been seen previously in the Bayesian designs based on the triangular and normal distributions but quite similar (in terms of weight distribution) to that seen in the design based on the lognormal distribution in Table 5.1. This similarity in weights is a consequence of putting more weight on the lognormal prior than the others. It can be observed also that there is some kind of a shift in terms of the support points to the left, thus favoring earlier time points compared to ξ_L in Table 5.1.

Improving minima in terms of D-criterion values across the priors is an important robustness consideration. Table 5.7 compares the minimum D-criterion for the weighted product design ξ_W to those of the Bayesian designs. While ξ_W improves the minimum D-criterion for the lognormal prior, it does not do the same for the other priors. Its minima are smaller than those for the maximin and product designs. Density plots are shown below in Figure 5.16, as before, comparing the weighted

Prior	Minimum D-criterion Value	
Distribution	Bayesian	Product
Lognormal	-12.3672	-10.0974
Triangular	7.7921	7.0455
MVN (Informative)	8.3627	7.7771
MVN (Vague)	7.3435	6.5732

Table 5.7: Comparison of the minimum D-criterion values across the four prior distributions to the minimum D-criterion value of the weighted product design across the same priors. This is based on $N = 10000$ draws from the each of the priors.

product design to the Bayesian designs. It can be quickly seen that the weighted product design seems to approximate the design based on the lognormal prior to a reasonable extent which makes sense given the weight on the lognormal prior. The

design does not approximate the designs based on the triangular and normal priors as closely. This could also be an artifact of the weights used in generating the design as observed earlier.

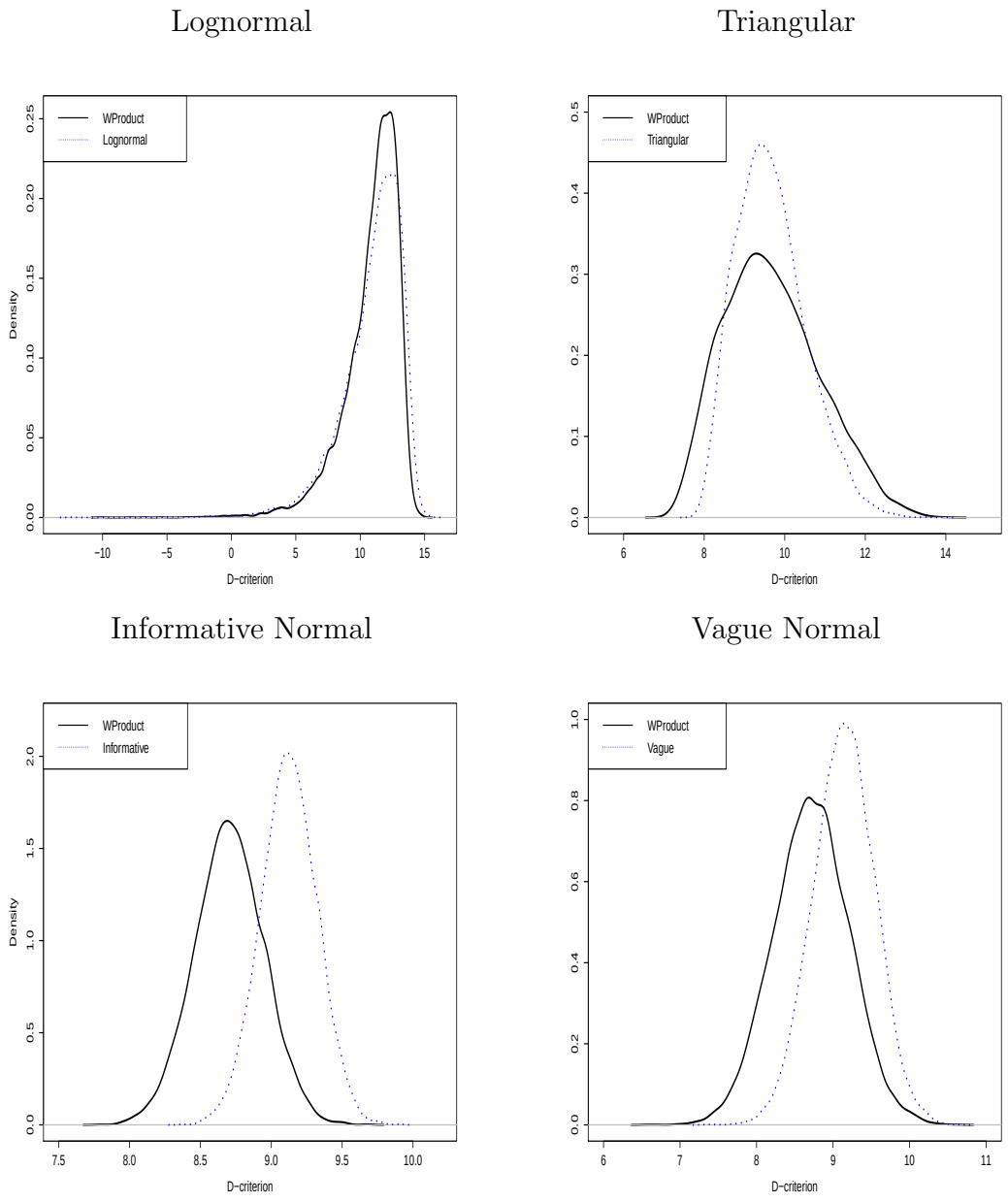

Figure 5.16: Density plots comparing the distribution of D-criterion values of the Bayesian designs to the weighted product design.

Further, comparing the values in Tables 5.4, 5.6 and 5.8 (shown below), it can be observed that ξ_W improves the relative minimum efficiency with respect to the lognormal prior in particular. There are also improvements in the relative efficiencies for the triangular prior. However, for ξ_W, the minimum relative efficiencies decrease for the normal priors. The minimum relative efficiencies, though, for the normal priors based on ξ_W are very high for all practical purposes. Thus, although there is a loss in relative efficiency for the normal priors by using ξ_W, this is reasonably compensated for by the increase in minimum relative efficiency for the lognormal prior. This increase in minimum relative efficiency is a direct result of the weighting scheme used in design generation.

Summary	Relative D-efficiency			
Statistic	Lognormal	Triangular	MVN-Inf	MVN-Vague
Minimum	0.2231	0.4013	0.7469	0.6897
Q_1	0.7570	0.9814	1.1722	1.2681
Median	1.0381	1.4534	2.3194	2.2467
Q_3	2.0733	3.0906	6.9989	7.5859

Table 5.8: Summary statistics of the distribution of relative efficiency of the weighted product design.

5. Designs Based on Aggregated Prior Distributions

Recall that in Chapter 3 some methods for aggregating prior distributions in the Bayesian literature were introduced. It is reiterated here that the rationale for combining (or pooling) prior distributions is to arrive at a consensus prior distribution for the purpose of obtaining a Bayesian experimental design. In this section, new applications making use of pooling methods are introduced in the context of experimental design. An important aspect of design based on pooling of priors is the ability to

obtain draws from the consensus (or composite) prior. In particular, focus is on the independent and logarithmic pooling operators.

5.1. Sampling from Composite Prior

The functions of interest are those in equations (3.10) and (3.12) repeated here for convenience. The consensus probability distributions arising from independent and logarithmic opinion pooling are respectively

$$p(\theta) = g \prod p_i(\theta)$$

where $\frac{1}{g} = \int \prod_{i=1}^{k} p_i(\theta)\, \mathrm{d}\theta$ is a normalizing constant and,

$$p(\theta) = \frac{\prod_{i=1}^{k} [p_i(\theta)]^{w_i}}{\int \prod_{i=1}^{k} [p_i(\theta)]^{w_i}\ \mathrm{d}\theta},$$

$w_i \geq 0$ and $\sum w_i = 1$. The consensus probability distributions are generally not in any recognizable parametric form and so Markov-Chain Monte Carlo (MCMC) methods are used to obtain draws from them. For example, in this application, lognormal, triangular and normal distributions are going to be combined and the resulting distribution is not recognizable. Focus will be on independent pooling initially and later, geometric pooling.

A random-walk Metropolis-Hastings (M-H) algorithm is used to draw from the composite probability distribution. A bivariate normal proposal density is used and so the M-H ratio is of the form

$$r = \frac{p(\theta^*)}{p(\theta^{cur})} \tag{5.13}$$

where θ^* is the proposed parameter vector value and θ^{cur} is the current state of the Markov chain. Tuning the algorithm is important to ensure that acceptance rates

are not too high (as a result of the chain stagnating) or too low (as a result of rejecting too many proposals). This can be accomplished by "properly" choosing the variance of the proposal distribution. In the present case, the values of $\sigma_{ka} = 0.005$ and $\sigma_{ke} = 0.001$ produced satisfactory results. It is also important to note that the prior aggregation methods used here are such that they summarily assign a zero probability to any parameter value if at least one expert believes that that value is not admissible. An intuition for this property can be gained by recognizing the fact that $p(\theta)$, the consensus distribution, is a normalized product of the individual densities. As a result, even if a particular parameter value has high probability in a particular prior but zero or low probability in the other priors, it will be rejected as a proposal with very high probability. Thus, the resulting distribution is expected to consist of values that are "consensual". For example, the extreme values in the lognormal prior distribution are completely eliminated from the consensus distribution $p(\theta)$. Three chains were used with acceptance rates of 33.71%, 33.32% and 33.80% respectively.

The algorithm is run for 200000 iterations after which a burn-in of 150000 is used. Sample path history plots for both parameters based on the three chains are shown in Figure 5.17. Histograms of the samples are also shown in Figure 5.18.

The plots in Figure 5.17 are used to help assess convergence of the chains to $p(\theta)$ for both parameters. They suggest that the chains have converged to the target (or consensus) distribution $p(\theta)$. Thus, samples obtained are actually from $p(\theta)$. Samples, based on the three chains, are shown in Figure 5.18. It should be recalled that independent pooling usually produces a unimodal distribution as demonstrated by the histograms in Figure 5.18.

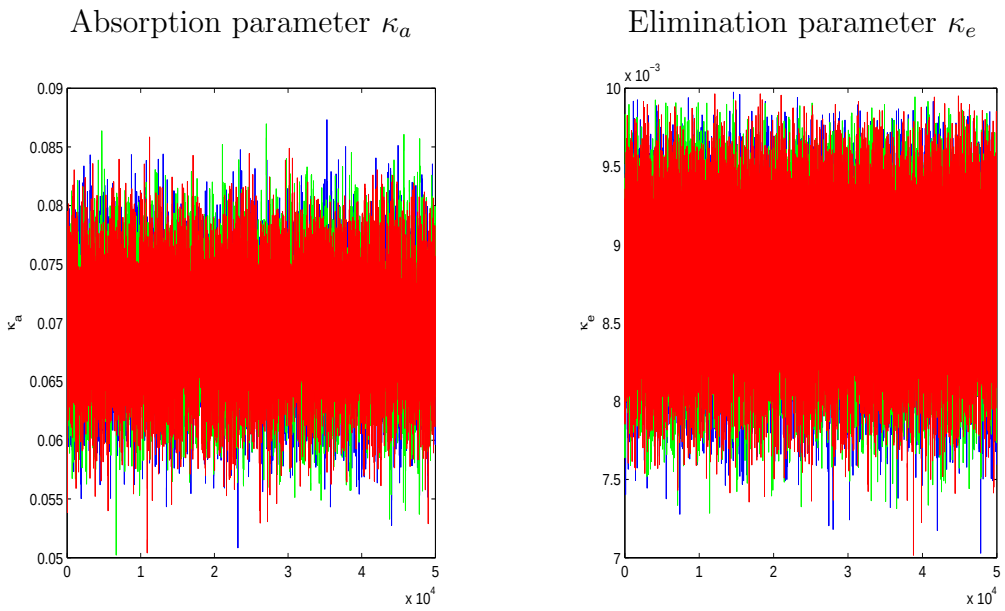

Figure 5.17: Sample path history plots for the random-walk M-H algorithm for each of the parameters after a burn-in of 150000.

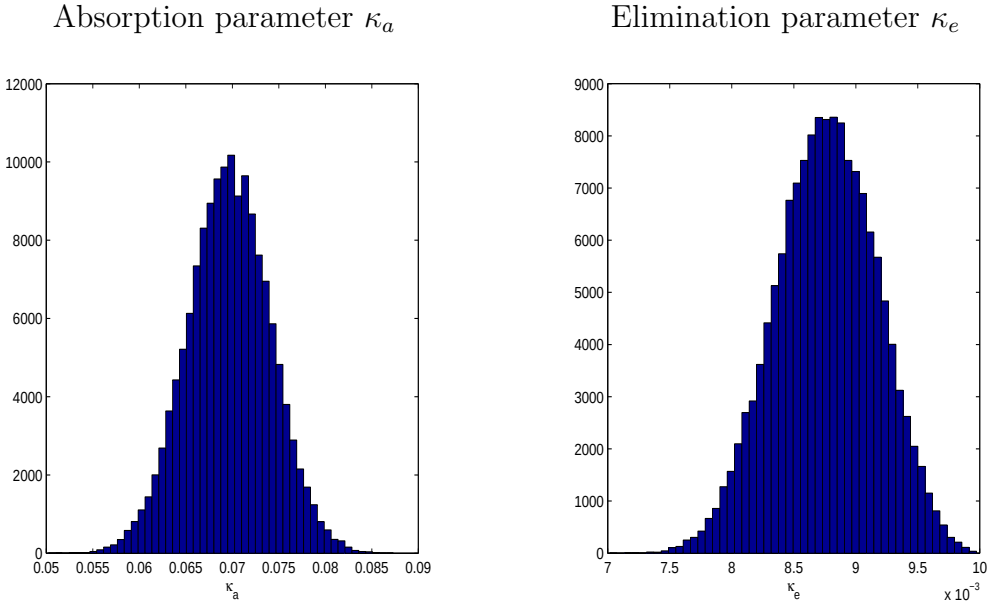

Figure 5.18: Approximate distributions of the parameters based on the three chains.

A Bayesian D-optimal design is generated based on the draws for κ_a and κ_e shown in Figure 5.18. The design obtained is

$$\xi_{IP} = \left\{ \begin{array}{ccccc} 13.23 & 129.45 & 220.20 & 377.54 & 635.23 \\ 0.4456 & 0.5371 & 0.0098 & 0.0025 & 0.0049 \end{array} \right\}. \tag{5.14}$$

It is seen in ξ_{IP}, like in the previous designs, that a large amount of sampling is done at earlier time points after dose administration. In this case, 44.56% of the observations are taken near 13 minutes. Very little sampling is done at later time points as seen in the weights on $220.20, 377.54$ and 635.23. This pattern is quite clear in the designs presented so far and an explanation will be presented for it later.

Comparing ξ_{IP}, the Bayesian design based on the consensus prior $p(\theta)$, to the individual Bayesian designs in Table 5.1 is a reasonable comparison. This will be done by evaluating ξ_{IP} and the Bayesian designs over the same set of $N = 10000$ parameter values. First, density plots of the D-criterion values are shown in Figure 5.19 followed by plots of efficiencies of ξ_{IP} relative to each of the Bayesian designs in Figure 5.20. Relative efficiencies in excess of 6.0 were excluded for the lognormal prior distribution in making the plot in the topleft corner of Figure 5.20. This removes the influence of a few extremely high relative efficiencies which will otherwise distort the information in the plot.

The density plots of D-criterion values of the informative and vague normal priors are similar in shape to the respective density plots based on ξ_{IP}. In addition, the near-symmetry in the density plots for these priors is a result of the symmetry in the normal priors. The negative skewness in the density plots of D-criterion values for ξ_{IP} and the design based on the lognormal distribution is not unexpected. This is caused by the large low-probability parameter values in the lognormal distribution. Similar

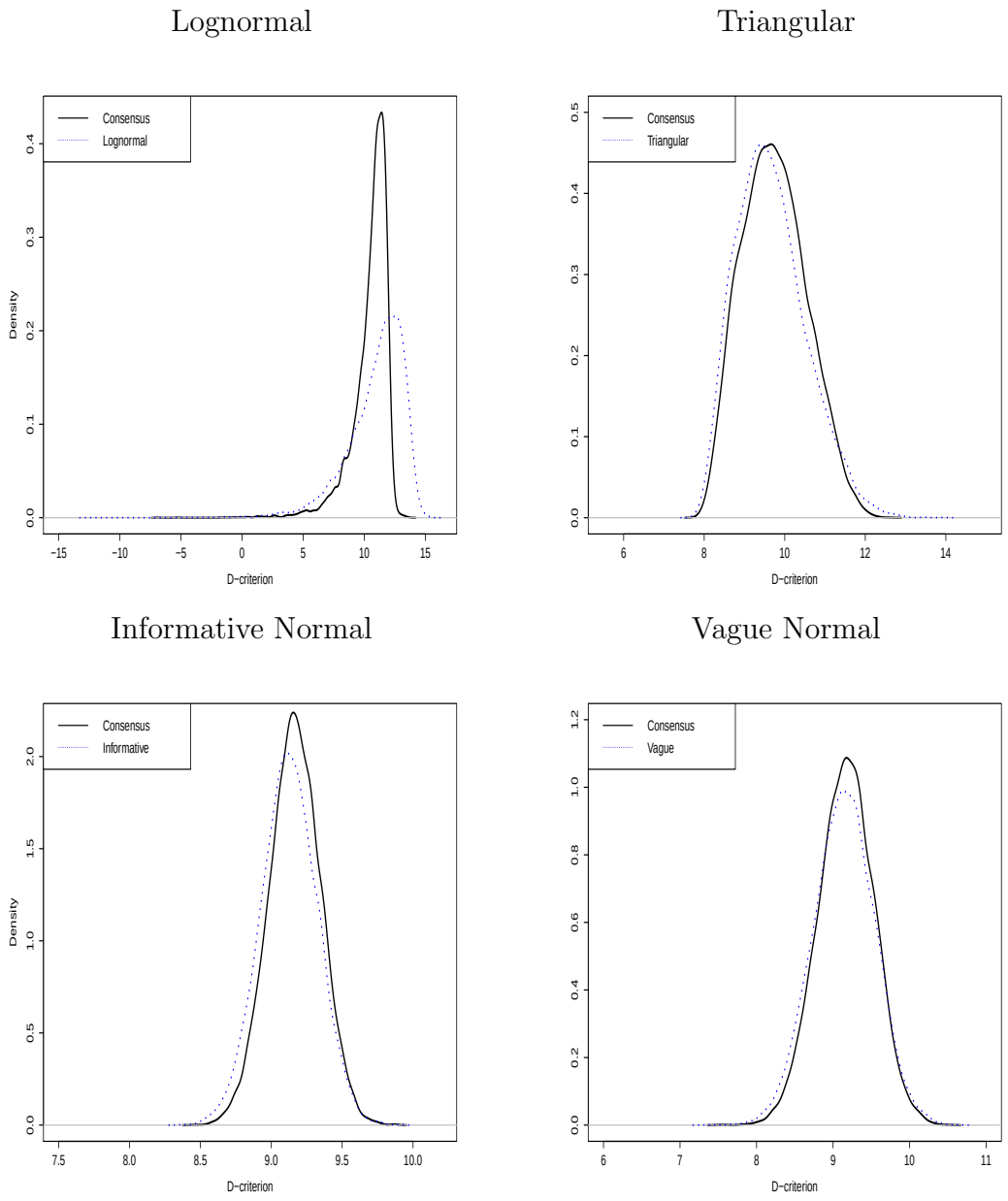

Figure 5.19: Plots showing the efficiencies of the Bayesian design ξ_{IP} relative to the Bayesian designs based on the individual priors.

comments can be made about the density plots for ξ_{IP} and ξ_T for the triangular prior distribution.

The distributions of relative efficiencies of ξ_{IP} in Figure 5.20 follow directly from the density plots. In fact, the information provided by the two figures is equivalent. The closeness in approximation of the density plots of D-criterion values for the normal priors by ξ_{IP} implies that relative efficiencies for these priors will be close to 1.0 as seen in the corresponding plots in Figure 5.20. The positive skewness in the distribution of relative efficiency of ξ_{IP} for the lognormal prior is also directly related to the positive skewness in the lognormal distribution as previously remarked.

The question may then be asked: "why might a practitioner use ξ_{IP} instead of any of the Bayesian designs?". To answer this question, assume that a practitioner uses the Bayesian D-optimal design based on the lognormal prior, ξ_L, instead of ξ_{IP}. Suppose also that θ is in the domain of the normal prior distributions. The performance of ξ_L relative to the designs based on the normal priors, ξ_{Nf} and ξ_{Nv} respectively, for the same sets of parameter values is shown in Figure 5.21.

Looking at the distributions, it is clear that all the relative efficiencies are to the left of 1.0. This means that if θ is in the domain of the normal priors, using the design based on the lognormal prior (and ignoring the normal priors) will be sub-optimal for making inference about θ. On the other hand, the plots in Figure 5.20 show that ξ_{IP} is preferred to ξ_L. Performance of ξ_{IP} relative to locally optimal designs was investigated but the details are excluded for brevity. The results showed the robustness of ξ_{IP}.

Using logarithmic pooling, draws are obtained from the consensus distribution using a random-walk M-H algorithm (as described above) using a weight of 0.25 for each prior. Thus, the resulting distribution is a geometric mean of the prior distributions. Sample path history plots and draws are shown in Figure A.1 in Appendix A. Acceptance rates were approximately 56% across the three chains. Immediately noticeable in the distribution of draws for both parameters is the larger spread in the distri-

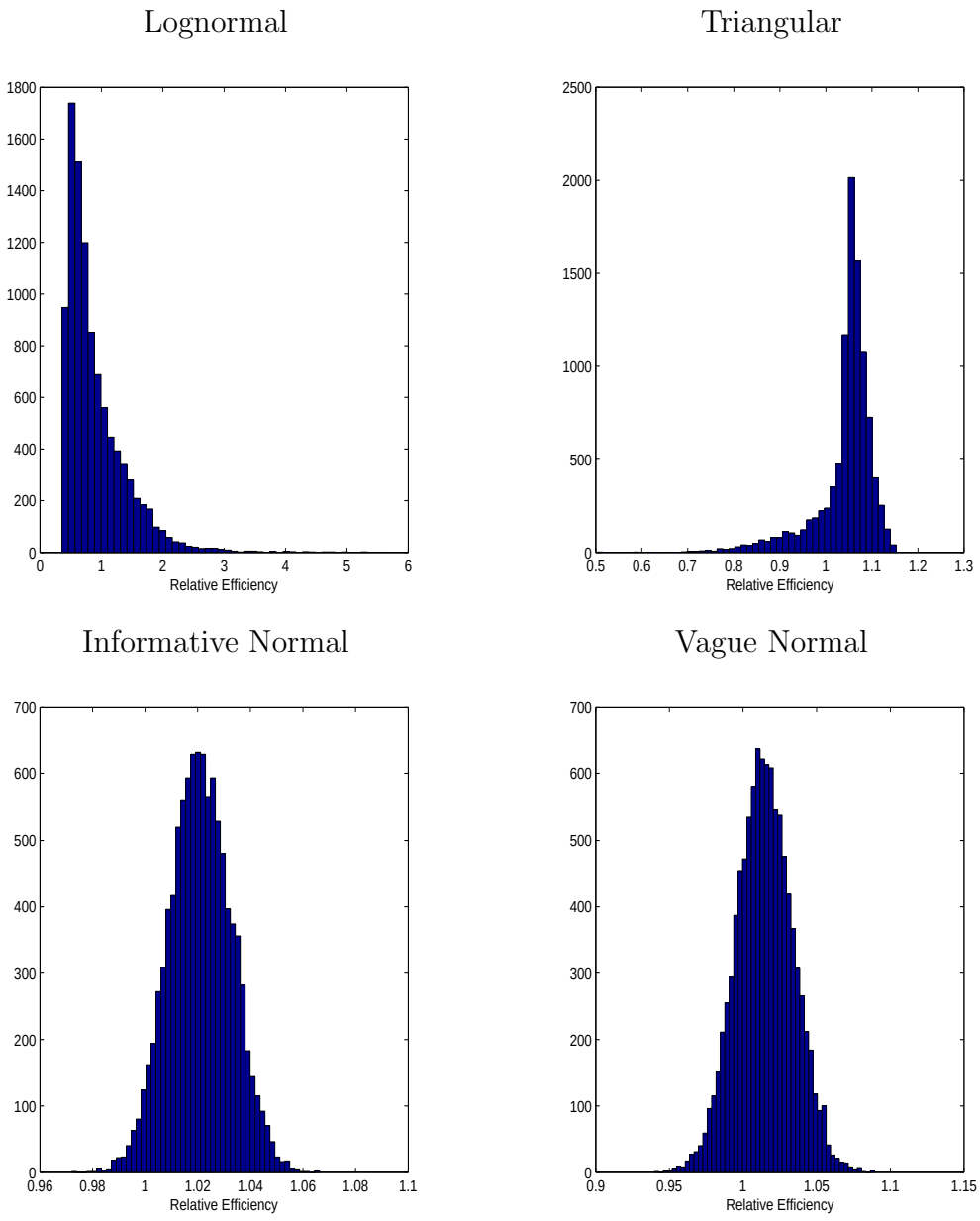

Figure 5.20: Plots showing the efficiencies of the Bayesian design ξ_{IP} relative to the Bayesian designs based on the individual priors.

butions compared to the distribution of draws obtained using independent pooling. This spread is attributable to the prior weights and thus a consequence of logarithmic pooling. For example, increasing the weight on the lognormal distribution toward 1.0

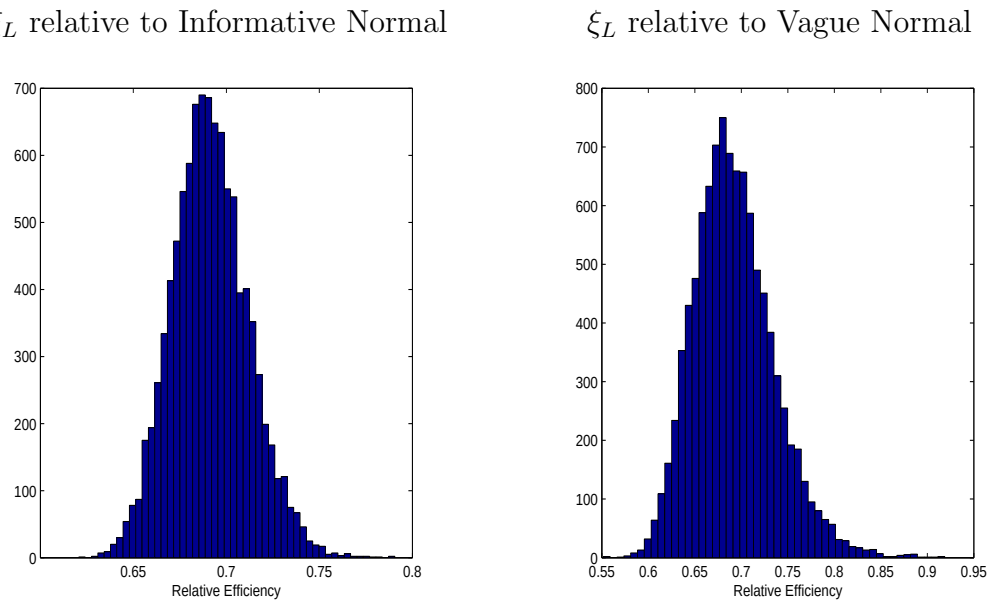

Figure 5.21: Distributions of the performance of ξ_L relative to the Bayesian designs, ξ_{Nf} and ξ_{Nv} for the normal priors.

results in a consensus distribution that has a slightly longer right tail. A Bayesian D-optimal design generated based on these draws is

$$\xi_{LP} = \left\{ \begin{array}{ccccc} 12.95 & 130.64 & 364.02 & 546.79 & 677.48 \\ 0.4949 & 0.4915 & 0.0016 & 0.0058 & 0.0062 \end{array} \right\}. \tag{5.15}$$

The distribution of weights and, to some extent, the support points in ξ_{LP} are similar to those of previous designs: earlier time points are weighted considerably more than latter time points. The performance of ξ_{LP} relative to the Bayesian designs is shown in Figure 5.22. It can be seen that the performance of ξ_{LP} is very similar to that of ξ_{IP} across the prior distributions. Table 5.9 compares the performances of ξ_{IP} and ξ_{LP}. The results in Table 5.9 show how close the two designs are. It is worth pointing out that ξ_{LP} improves the minimum relative efficiency over the triangular prior distribution.

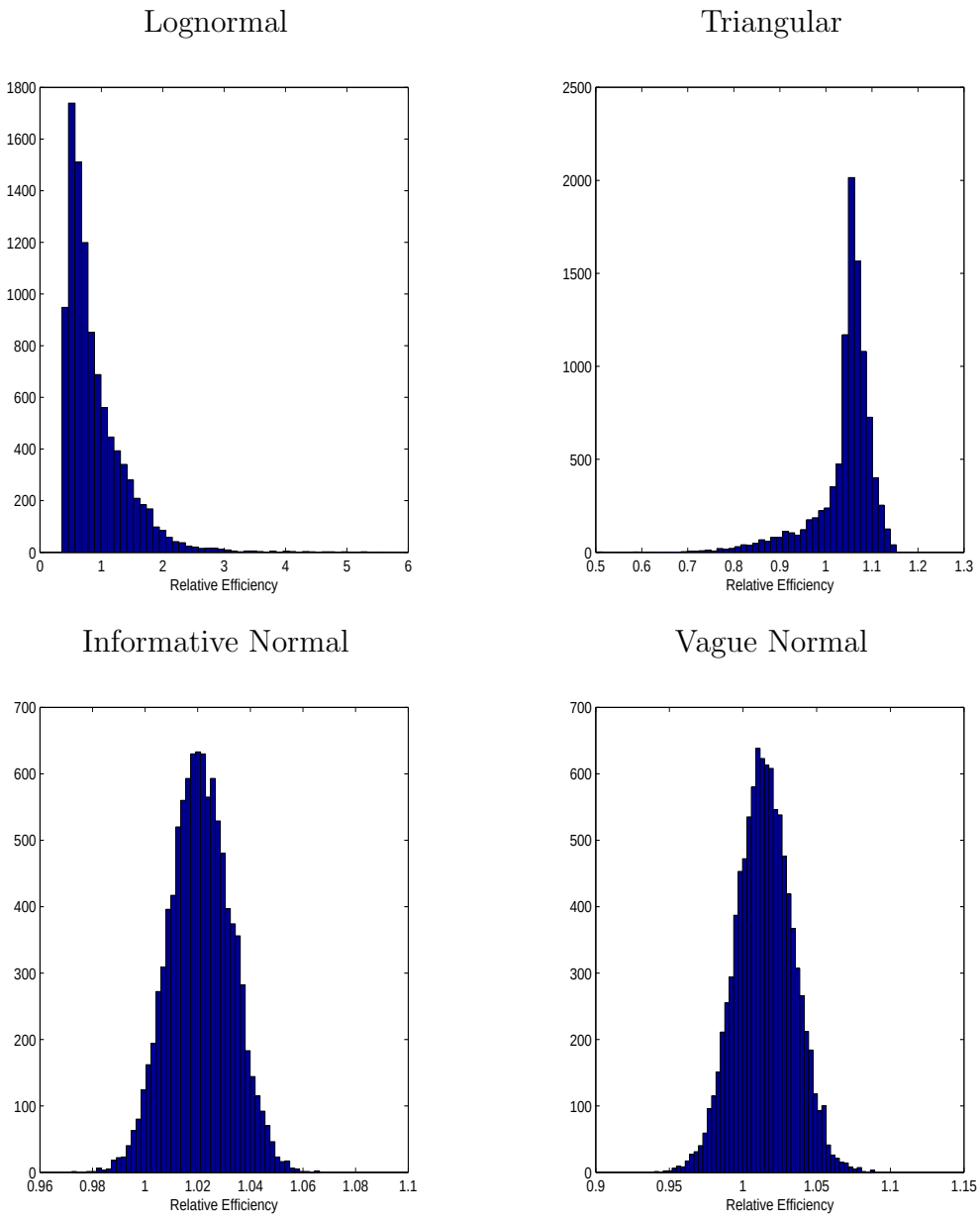

Figure 5.22: Relative efficiency plots showing the performance of ξ_{LP} relative to the Bayesian designs based on the individual priors.

The similarity in design performance begs the question, "how sensitive are the designs to the pooling weights?". A design is generated with pooling weights $w = (0.70, 0.10, 0.10, 0.10)$ on the lognormal, triangular, informative and vague normal

Summary	ξ_{IP}				ξ_{LP}			
Statistic	LN	TL	MVN-Inf	MVN-Vag	LN	TL	MVN-Inf	MVN-Vag
Minimum	0.36	0.50	0.97	0.94	0.39	0.57	0.97	0.93
Q_1	0.56	1.03	1.01	1.00	0.57	1.04	1.02	1.00
Median	0.74	1.06	1.02	1.01	0.75	1.06	1.02	1.02
Q_3	1.10	1.08	1.03	1.03	1.10	1.09	1.03	1.03

Table 5.9: Comparison of the performances of ξ_{IP} and ξ_{LP} relative to the prior distributions.

priors in that order. Acceptance rates across the three chains were about 70% which is related to the weights. For example, it is expected that larger parameter values from the lognormal distribution that were previously rejected would be now accepted with a reasonably high probability. This is seen in the larger spread in the consensus distributions of the model parameters in Figure A.1 in Appendix A. The Bayesian D-optimal design obtained

$$\xi_{WP} = \left\{ \begin{array}{ccccc} 13.38 & 134.84 & 137.52 & 348.27 & 350.16 \\ 0.5001 & 0.3514 & 0.1393 & 0.0014 & 0.0077 \end{array} \right\} \tag{5.16}$$

concentrates almost all sampling effort on earlier time points. Also noticeable is the fact that unlike previous designs where at least 40% of the sampling effort is concentrated on the support point after the earliest time point, about 49% of the weight is split between 134.84 and 137.52. The closeness of these support points, in light of the structure of previous designs, suggests a convergence of the points to one or the other of the points or perhaps the average. The closeness of the latter support points, 348.27 and 350.16, is also noted. Thus, for all intents and purposes, ξ_{WP} may be a 3-point design. Some of this can be explained by the weights on the prior distributions. It is conceivable that by increasing the weight on a particular prior, the

lognormal in this case, the variability in the distribution of model parameter values decreases and thus fewer support points are needed for the design. On the other hand, it is also known that the number of support points of Bayesian optimal designs increases with the spread in the prior distribution. Thus, the spread in the consensus prior should result in more design points. Thus, the weights and the spread in the consensus distributions could possibly lead to the structure of ξ_{WP}. It suffices to say that more work must be done to see the full impact of the prior weights and also the parametric forms of the priors on the resulting experimental design. Whether a practitioner decides to use ξ_{WP} as a 3-point or 5-point design may depend on the costs of running one versus the other as well as other experimental objectives.

6. Discussion

Designs based on new design criteria for nonlinear models were presented for the one-compartment model. It was seen that these designs for several examples were good approximations to the Bayesian designs in terms of the distribution of D-criterion values. The designs also serve as good compromise designs given their performance relative to locally optimal designs across the prior distributions considered. The performance of the Bayesian designs relative to each other - for example, the design based on the lognormal design relative to those based on the normal distributions - indicated the sub-optimality in design that occurs when all the priors are not taken into account when generating a design.

The designs generated, both based on new robust criteria and Bayesian, followed a similar pattern: they concentrate half of the sampling effort at a very early time point, usually between 12 to 13 minutes after drug administration, a substantial amount on time points between 120 and 140 followed by very little sampling at latter time points.

This observed pattern makes sense and is intuitive given a compartmental model. The absorption phase of the compartmental model in Figure 5.2 occurs before the "hump" or turning point while the elimination phase is generally in the region where the curve is decaying. Thus, to efficiently estimate or make any kind of inference about the absorption parameter, it is imperative to sample soon after drug administration. The spread in the design points afterward ensures precise estimation of the elimination parameter. In fact, a design that concentrates all sampling effort away from early time points will be very inefficient.

Four prior distributions were considered here, with the lognormal prior having more support for larger parameter values. In such situations, designs are required that are efficient in the tails of the distribution. The poor performance of the Bayesian designs in the tails is quite illustrative of the need for robustness. The maximin design based on the new criterion performed quite efficiently by maximizing the minimum D-criterion value in the right tail. This makes it reasonable to recommend the maximin design for situations where the priors are skewed or have some extreme values. The maximin design protects the experimenter against sub-optimality if the true parameter values occur somewhere in the tails in this case.

Designs based on independent and logarithmic pooling were also presented as alternative robust designs. These designs followed the same structure as those based on the new criteria - maximin, product and weighted product criteria and geometric - and can be given the classical Bayesian optimal design interpretation because they are obtained by averaging over a single prior distribution. Thus, they are essentially Bayesian optimal designs. It is important to emphasize here that pooling may not always be an option. For example, when there is little overlap of the supports of the prior distributions, the two pooling methods considered cannot be used. Thus, it is imperative that the supports overlap in order for some kind of a consensus to be

possible. Questions about the sensitivity of designs to the pooling weights have yet to be fully answered.

7. Discretization of Continuous Designs

Recall that a continuous design assigns weights w_i to a design point x_i and consequently directs the experimenter to take a fraction w_i of the total number observations, N, at experimental condition x_i. A practical concern an experimenter may have is how to obtain an implementable design, that is an exact design, given a continuous design. A commonly used method of discretizing a continuous design is presented with an application the product design in equation (5.11) restated below assuming total sample sizes of $N = 1000, 100,$ and 10.

$$\xi_P = \left\{ \begin{array}{ccccc} 13.53 & 152.75 & 283.39 & 437.04 & 505.73 \\ 0.5008 & 0.4941 & 0.0011 & 0.0031 & 0.0009 \end{array} \right\}.$$

The approach used here calculates Nw_i and rounds it to an integer N_i^*. For the product design, if the Nw_i are rounded to the largest integer smaller than Nw_i, then the resulting design is

$$\xi_P^{N_1} = \left\{ \begin{array}{ccccc} 13.53 & 152.75 & 283.39 & 437.04 & 505.73 \\ 500 & 494 & 1 & 3 & 0 \end{array} \right\} \tag{5.17}$$

while rounding to the smallest integer larger than Nw_i results in

$$\xi_P^{N_2} = \left\{ \begin{array}{ccccc} 13.53 & 152.75 & 283.39 & 437.04 & 505.73 \\ 501 & 495 & 2 & 4 & 1 \end{array} \right\}. \tag{5.18}$$

In each case, the N_i^* do not sum to $N = 1000$. The inclusion of the sampling time of 505.73 minutes depends on how the discretization is done. A decision must be made in this case whether to use $\xi_P^{N_1}$ which excludes 505.73 or $\xi_P^{N_2}$ which includes it. For example, the experimenter may want to determine if it makes any sense to collect data that far out in time after drug administration. That is, does it make sense biologically to even collect data this far out? If not, then using $\xi_P^{N_1}$ will suffice. It must be noted here that taking $N = 1000$ observations is even not something that is likely to happen in practice. Another important observation is that the appearance of the large sampling times is a result of the range $[20, 720]$ of times that was used in generating the designs. The maximum sampling time of 720 minutes is equal to 12 hours and it might not make sense to collect data in a 12-hour period in a biopharmaceutical context. Limiting the range to something that is more biologically meaningful will potentially minimize the extent to which judgment must be exercised in design implementation.

For $N = 100$, it can be seen that only the first two sampling times will have any importance in the resulting exact design. For $N = 10$, the same exact design results given the near zero weights assigned to latter sampling times. Once again the experimenter will be called upon to make some judgments. An important lesson that can be learned from here is that a lot of thought must be put into specifying the space of sampling times used in generating the continuous design.

Pukelsheim and Reider (1992) and others propose more complicated methods of rounding off continuous designs into exact designs with a minimum loss in relative efficiency. The discussion presented here focuses on a less mathematical but intuitive approach.

8. Designs for the Michaelis-Menten Model

Robust designs for the Michaelis-Menten model are presented in this section. Before the designs are presented, it is important to emphasize a benefit that can be derived from the perspective of design due to the conditionally linear property of the model discussed in Chapter 2. In particular, the maximum rate at which the substrate is converted into product V_{max} is a linear parameter, and hence, the robust (or optimal) designs will not depend on it. Thus, if the assumption can be made that the Michaelis constant K_M is independent of V_{max}, as is the case in this section, then it will not be necessary to specify a prior distribution for V_{max}.

Consequently, the designs in this section are generated under the assumption that two lognormal prior distributions are plausible for K_M. These distributions are shown in Figure 5.23 with the distribution on the left having non-zero probabilities for values of K_M near zero. The lognormal distributions in Figure 5.23 were generated with the same variance but different means. The rationale behind this is to allow for sensitivity checks. The range of substrate concentrations used in generating the designs is arbitrarily $(0, 50]$. The designs associated with the various robust criteria will be presented in the same order as was done for the compartmental model.

8.1. Maximin Design

The maximin design for the Michaelis-Menten model is

$$\xi_{MM} = \left\{ \begin{array}{ccccc} 2.00 & 49.79 & 49.87 & 49.95 & 50.00 \\ 0.4808 & 0.0891 & 0.1780 & 0.0807 & 0.1714 \end{array} \right\}. \tag{5.19}$$

The design points are shown on the plot of the model with $V_{max} = 10$ and $K_M = 1.025$, the average of the two prior means. It is clear from equation (5.17) and the plot in

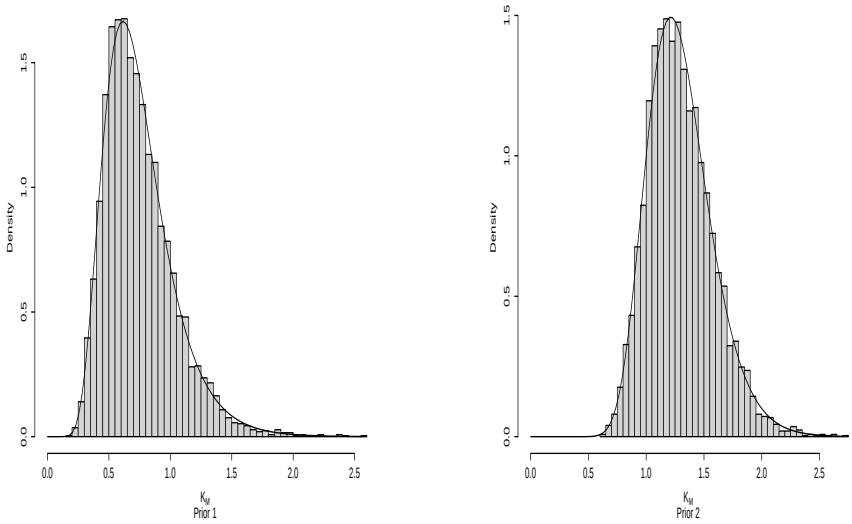

Figure 5.23: Left: Prior distribution of K_M based on a lognormal distribution with parameters $\mu_1 = -0.354, \sigma_1 = 0.365$. Right: Prior distribution of K_M based on a lognormal distribution with parameters $\mu_2 = 0.239, \sigma_2 = 0.215$.

Figure 5.24 that the design has effectively two support points: one support point near zero and the other near or at the maximum allowable substrate concentration, and with both weights close to 0.50. It will be shown later that the structure of ξ_{MM} is actually a characteristic of designs for the Michaelis-Menten model.

The Bayesian D-optimal designs based on the two priors are also given in Table 5.10. The structure of the designs is similar to that of ξ_{MM} and agrees with theoretical results about the designs for the Michaelis-Menten model. Comparisons similar to those made for the compartmental model will be made. Table 5.11 compares the minima for the maximin design ξ_{MM} to the two Bayesian designs. These minima are based on evaluating both ξ_{MM} and the two Bayesian designs over a set of $N = 5000$ draws from $LN(\mu_1, \sigma_1)$ and $LN(\mu_2, \sigma_2)$ respectively. The robustness of ξ_{MM} can be seen from Table 5.11 in that it improves the minima across the two

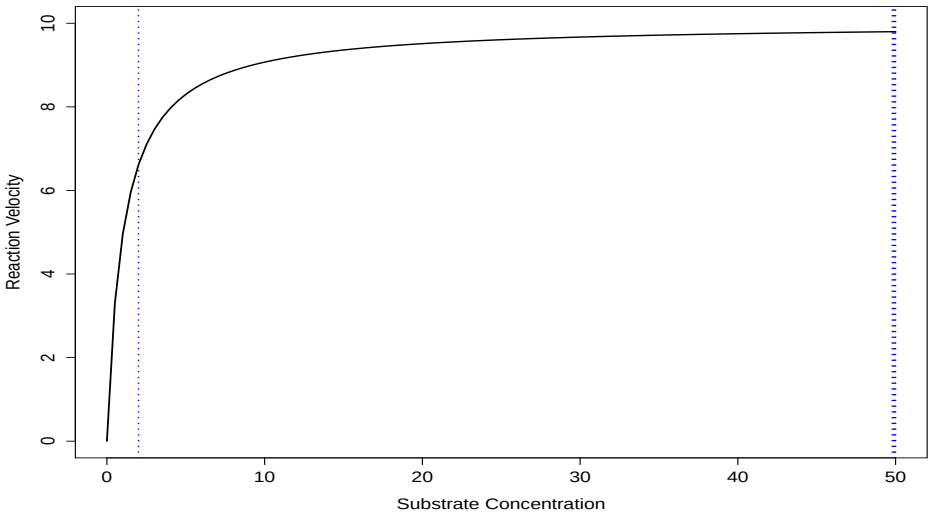

Figure 5.24: The Michaelis-Menten model with the points of support (dashed vertical lines) of the maximin design ξ_{MM}.

Prior Distribution	Bayesian Optimal Design
$LN(\mu_1, \sigma_1)$	$\xi_L = \left\{ \begin{array}{ccccc} 0.54 & 49.82 & 49.85 & 49.86 & 49.95 \\ 0.5017 & 0.0198 & 0.0773 & 0.2398 & 0.1614 \end{array} \right\}$
$LN(\mu_2, \sigma_2)$	$\xi_T = \left\{ \begin{array}{ccccc} 1.07 & 1.09 & 1.12 & 49.84 & 49.85 \\ 0.0735 & 0.2001 & 0.2252 & 0.0987 & 0.4025 \end{array} \right\}$

Table 5.10: The Bayesian D-optimal designs based on each of the lognormal prior distributions.

priors. It is reiterated here that maximizing the worst D-optimality criterion value across both priors is the objective of the maximin design ξ_{MM}.

The distribution of D-criterion values based on ξ_{MM} and the Bayesian D-optimal designs is shown in the boxplots in Figure 5.25. It can be seen that ξ_{MM} considerably improves the minimum D-criterion values with respect to each of the priors. In fact,

Prior Distribution	Minimum D-criterion Value	
	Bayesian	Maximin
$LN(\mu_1, \sigma_1)$	-2.2556	-1.8127
$LN(\mu_2, \sigma_2)$	-3.4886	-2.1036

Table 5.11: Comparison of the minimum D-criterion values across the two prior distributions to the minimum D-criterion values based on ξ_{MM}.

the 10th percentile of the D-criterion values based on ξ_{MM} are -0.1017 and -0.8060 compared to -0.1264 and -0.8347 for $LN(\mu_1, \sigma_1)$ and $LN(\mu_2, \sigma_2)$ respectively. Thus, for a set of 500 parameter values from each prior, ξ_{MM} has larger D-criterion values. Because the actual value of K_M is unknown, it is particularly important to guard against worst-case scenarios.

It must be mentioned that although the maximin design ξ_{MM} improves the minimum considerably, it is consistently out-performed by the Bayesian designs when higher percentiles are considered as the boxplots in Figure 5.25 show. This is unlike the behavior of the maximin design for the compartmental model. This is not surprising since the objective of ξ_{MM} is to maximize a minimum. One reason for the differences in the two designs in terms of their behavior when higher percentiles are considered is because the two models are different. For the compartmental model, the maximin design not only improves minimum, but as Table 5.3 shows, it also compares favorably to the Bayesian designs at higher percentiles. Thus, a researcher interested not only in a design that improves minima but performs as efficiently as the Bayesian designs - when higher percentiles are considered - can use the maximin design as a robust design for the compartmental model. In the case of the Michaelis-Menten model, the maximin design can be used if the researcher is primarily interested in guarding against the worst-case scenario, but if interest goes beyond that, the prod-

uct or weighted product designs could be used. Thus, the questions: "which robust criterion", or equivalently "which design should a researcher use given a nonlinear model?" may be answered. In essence, the "right" design depends on the model (and possibly on the prior distributions used).

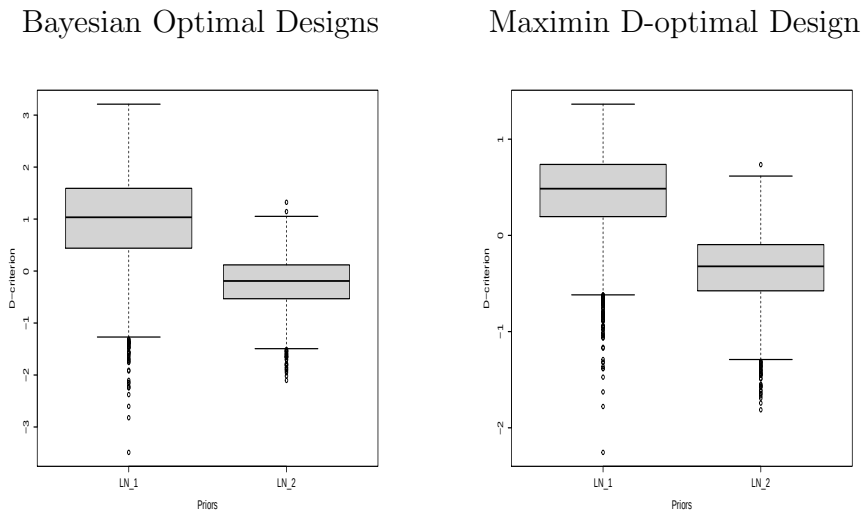

Figure 5.25: Left: Boxplots of D-criterion values of the Bayesian D-optimal designs evaluated over $N = 5000$ draws from the respective prior distributions. Right: Boxplots of the D-criterion values of the maximin D-optimal design evaluated over $N = 5000$ draws from each of the priors distributions.

Density plots of the D-criterion values are given in Figure 5.26. These plots put in perspective the foregoing argument about the behavior of the maximin design ξ_{MM} compared to the Bayesian designs at low and high percentiles. It shows, as already mentioned, that ξ_{MM} does not perform as well as the Bayesian designs at higher percentiles.

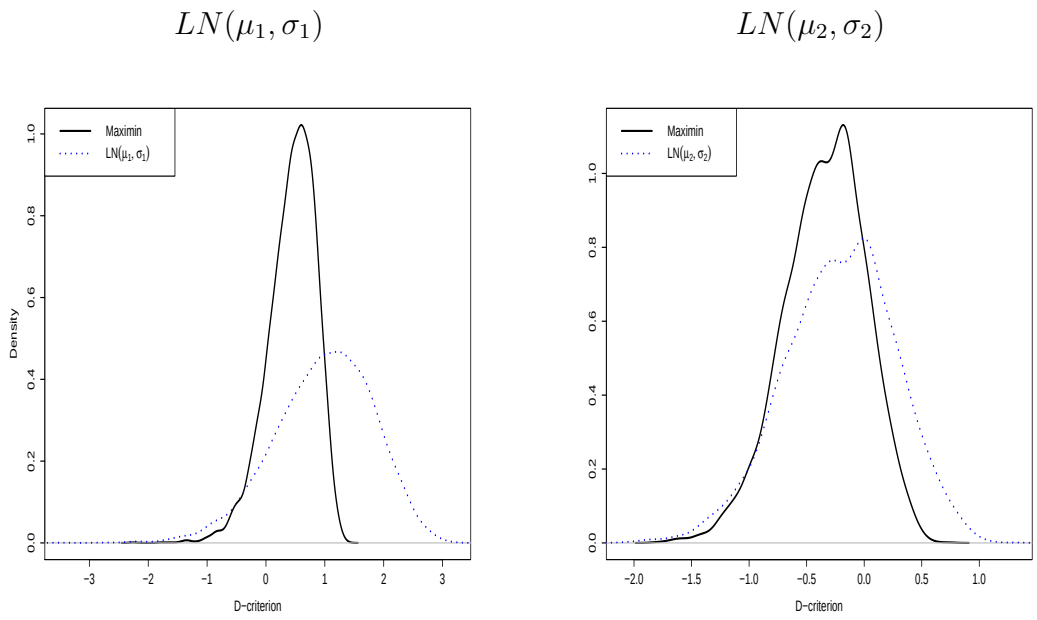

Figure 5.26: Density plots comparing the distribution of D-criterion values of the Bayesian designs to the maximin design.

Relative efficiency plots are also given in Figure 5.27 based on $N = 2000$ locally optimal designs for each prior distribution. The information in the plots are summarized in Table 5.12. The summary statistics in Table 5.12 show that ξ_{MM} is quite efficient as a compromise design given the high median efficiencies relative to locally optimal designs with respect to each of the priors.

Summary Statistic	Relative D-efficiency	
	$LN(\mu_1, \sigma_1)$	$LN(\mu_2, \sigma_2)$
Minimum	0.4750	0.8131
Q_1	0.9194	1.0148
Median	1.0164	1.0254
Q_3	1.0316	1.0419

Table 5.12: Summary statistics based on the relative efficiency plots in Figure 5.27.

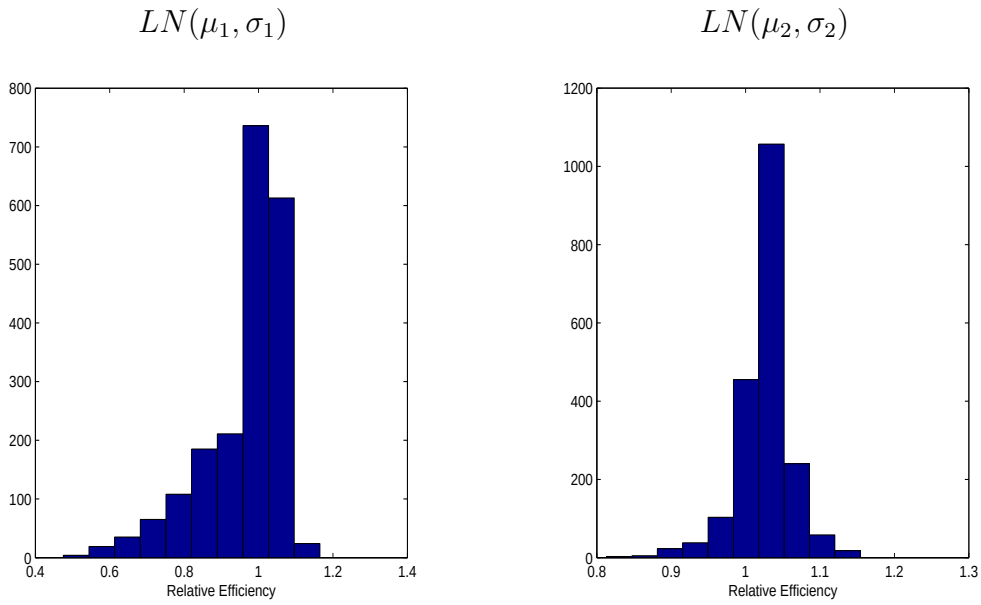

Figure 5.27: Left: Distribution of $D_{rel-eff}$ of ξ_M relative to $N = 2000$ locally optimal designs based on the $LN(\mu_1, \sigma_1)$ prior distribution. Right: Distribution of $D_{rel-eff}$ of ξ_{MM} relative to $N = 2000$ locally optimal designs based on the $LN(\mu_2, \sigma_2)$ prior distribution.

8.2. Product Design

The product design that maximizes the product of expected utilities for the Michaelis-Menten model is

$$\xi_{PM} = \left\{ \begin{array}{ccccc} 0.78 & 49.38 & 49.43 & 49.77 & 49.93 \\ 0.5015 & 0.0069 & 0.0342 & 0.1062 & 0.3512 \end{array} \right\}. \tag{5.20}$$

Noticeable is the similarity in support points between the product design ξ_{PM} and the maximin design ξ_{MM}. The density plots comparing the distribution of D-criterion for the product design to the Bayesian designs are given in Figure 5.28. The plots in Figure 5.28 contrasted with those for the maximin design in Figure 5.26 suggest that the product density is a much better "fit" overall to the Bayesian designs than

the maximin design. Thus, it is a better compromise design for the Michaelis-Menten model than the maximin design for these particular prior distributions.

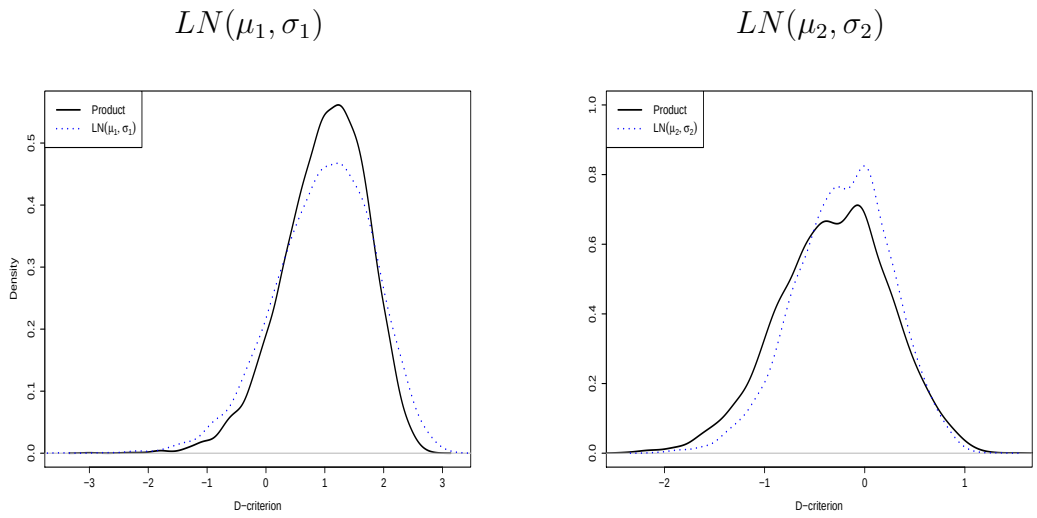

Figure 5.28: Density plots comparing the distribution of D-criterion values of the Bayesian designs to the product design ξ_{PM}.

Distribution of $D_{rel-eff}$ of the product design ξ_{PM} based on $N = 2000$ locally optimal designs for each prior distribution is in Figure 5.29. The plots and the summary in Table 5.13 emphasize that ξ_{PM} performs efficiently relative to locally optimal designs for each of the priors and thus, it can be used instead of either Bayesian designs if parameter estimation is of interest.

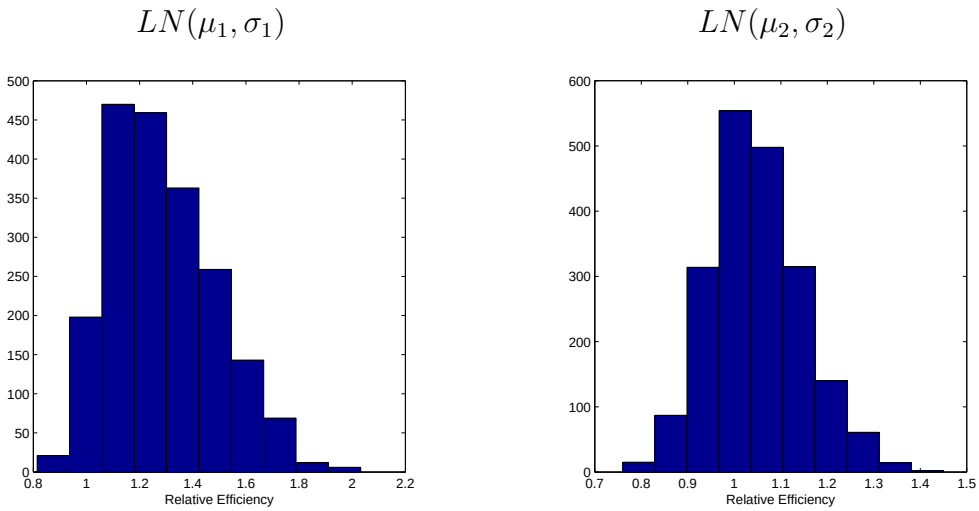

Figure 5.29: Left: Distribution of $D_{rel-eff}$ of ξ_M relative to $N = 2000$ locally optimal designs based on the $LN(\mu_1, \sigma_1)$. Right: Distribution of $D_{rel-eff}$ of ξ_{PM} relative to $N = 2000$ locally optimal designs based on the $LN(\mu_2, \sigma_2)$.

Summary Statistic	Relative D-efficiency	
	$LN(\mu_1, \sigma_1)$	$LN(\mu_2, \sigma_2)$
Minimum	0.8144	0.7605
Q_1	1.1326	0.9785
Median	1.2600	1.0395
Q_3	1.4177	1.1107

Table 5.13: Summary statistics based on the relative efficiency plots in Figure 5.29.

8.3. Weighted Product Design

A weighted product design assuming equal confidence in the two priors is also generated for the Michaelis-Menten model. This is given by

$$\xi_{WM} = \left\{ \begin{array}{ccccc} 0.61 & 46.89 & 47.30 & 47.40 & 48.83 \\ 0.5029 & 0.1673 & 0.1348 & 0.0690 & 0.1260 \end{array} \right\}. \tag{5.21}$$

The weighted product design ξ_{WM} is similar to the designs seen so far except that the latter design points are not as close to the maximum allowable concentration. Nonetheless, the structure is consistent with the designs presented so far.

Density plots comparing the distribution of the D-criterion values of the weighted product design ξ_{WM} to the Bayesian designs are shown in Figure 5.30. The plots suggest that ξ_{WM} "fits" the Bayesian design based on $LN(\mu_1, \sigma_1)$ better than that based on $LN(\mu_2, \sigma_2)$ although the priors were equally weighted. Relative efficiency plots of ξ_{WM} are also shown in Figure 5.31 with Table 5.14 giving summaries. These indicate that the product and weighted product designs both improve the minimum relative efficiency with respect to $LN(\mu_1, \sigma_1)$. Thus, from a design perspective, it will make more sense to use either of those designs than the maximin design if parameter estimation in the Michaelis-Menten model is of interest.

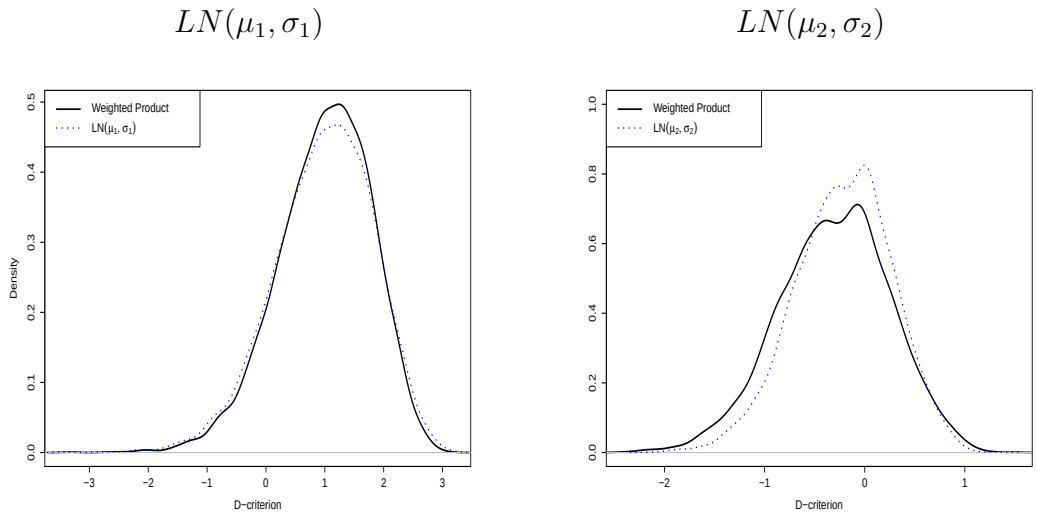

Figure 5.30: Density plots comparing the distribution of D-criterion values of the Bayesian designs to the weighted product design ξ_{WM}.

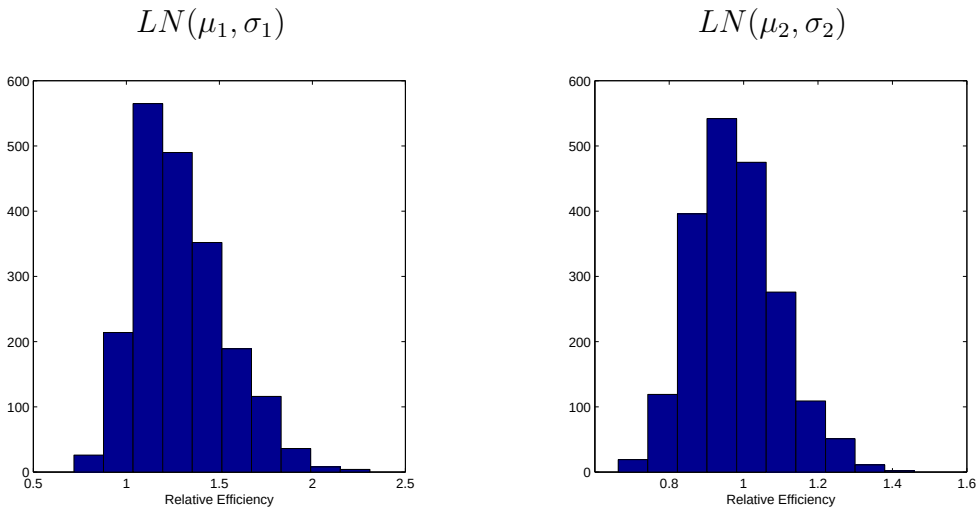

Figure 5.31: Left: Distribution of $D_{rel-eff}$ of ξ_M relative to $N = 2000$ locally optimal designs based on $LN(\mu_1, \sigma_1)$. Right: Distribution of $D_{rel-eff}$ of ξ_{PM} relative to $N = 2000$ locally optimal designs based on $LN(\mu_2, \sigma_2)$.

Summary Statistic	Relative D-efficiency	
	$LN(\mu_1, \sigma_1)$	$LN(\mu_2, \sigma_2)$
Minimum	0.7181	0.6627
Q_1	1.1113	0.8972
Median	1.2538	0.9698
Q_3	1.4378	1.0513

Table 5.14: Summary statistics based on the relative efficiency plots in Figure 5.31.

8.4. Reason for Design Structure

It was observed that all of the D-optimal designs for the Michaelis-Menten model contained one design point that was near zero and the other close to or at the maximum concentration x_{max}. The designs contain points so that the parameters K_M, the Michaelis constant, and the asymptote V_{max} can be precisely estimated. Bates and Watts (1988) show that the D-optimal design for the Michaelis-Menten model

has design points

$$x_1 = \frac{K_M}{1 + 2(K_M/x_{max})} \qquad \text{and} \qquad x_2 = x_{max}. \tag{5.22}$$

It is immediately seen from equation (5.20) that the D-optimal design has as a support point the maximum allowable concentration, thus, $x_2 = x_{max}$. This support point is free of $\theta = (K_M, V_{max})$ because it is for efficient estimation of V_{max}, the conditionally linear parameter. The other design point is a function of K_M, the nonlinear parameter. Thus, if a distribution for K_M can be specified, then a distribution of possible values of x_1 can be obtained. Figure 5.32 shows two distributions of x_1 based on the two lognormal prior distributions.

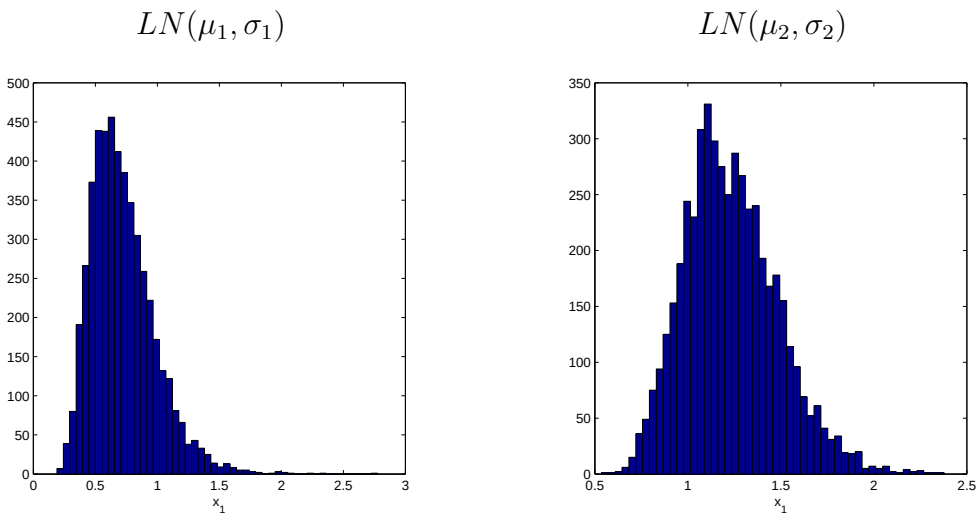

Figure 5.32: Left: Distribution of x_1 based on $LN(\mu_1, \sigma_1)$. Right: Distribution of x_1 based on $LN(\mu_2, \sigma_2)$.

The settings of x_1 in all the designs presented for the Michaelis-Menten model make sense given the distributions in Figure 5.32. Also noticeable is the similarity

between the corresponding plots in Figures 5.23 and 5.32. It must be noted that the amount of variability in x_1 is approximately equal to the amount of variability in K_M for sufficiently large x_{max}. Another consequence of equation (5.20) is that x_1 approaches K_M if the maximum allowable concentration increases.

8.5. Sensitivity Analysis

A sensitivity analysis is carried out for the product designs for the Michaelis-Menten model. The three scenarios considered are enumerated below.

1. $L1$: Fix prior means and vary prior variances.

2. $L2$: Vary prior means and fix prior variances.

3. $L3$: Vary prior means and vary prior variances.

The objective is to ascertain the impact on the product design of (1) through (3). Recall that the product design in equation (5.18) is based on (2) as well as the Bayesian designs in Table 5.10. Differences between the Bayesian designs and the product design under the enumerated scenarios are also of interest. The prior distributions for $L1$ and $L3$ are shown in Figure 5.33. The priors for $L2$ are given in Figure 5.23. $L1$ has mean of 1.0 and variances of 0.05 and 0.09, whereas $L3$ has means and variances 0.75 and 0.03, and 1.3 and 0.06 respectively.

The product designs are given in Table 5.15. It is obvious that at least two of the product designs in Table 5.15 are essentially 2-point designs. The product design based on $L3$, that is varying both prior means and variances, has two support points close to zero. Thus, it is slightly different from the designs based on $L1$ and $L2$. The designs based on $L1$ and $L2$ are quite similar in terms of the support distribution. Given that they are basically 2-point designs, does the similarity in designs mean $L1$

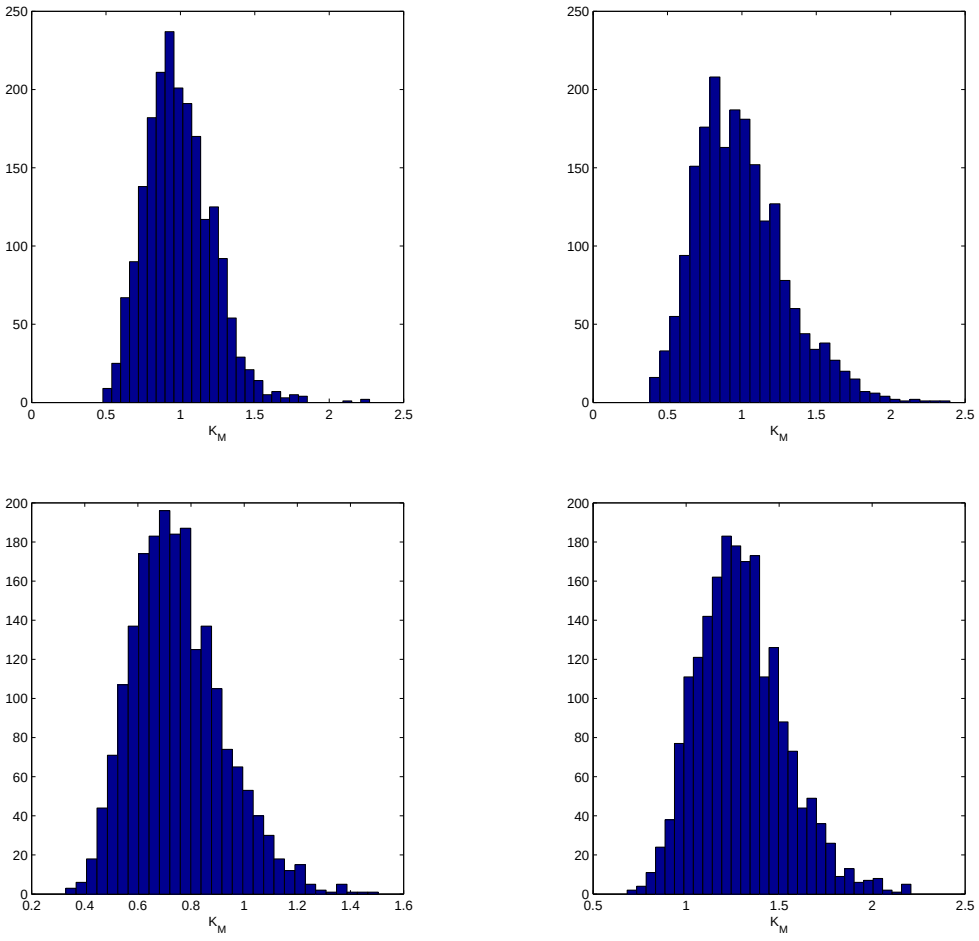

Figure 5.33: Top row: Lognormal distributions of K_M with the same mean and different variances ($L1$). Bottom: Lognormal distributions of K_M with different means and variances ($L3$)

and $L2$ do not impact the product design in any noticeable manner? Is it just an artifact of designs for the Michaelis-Menten model? What effect does the settings (or values) of the prior means and variances as well as their relative magnitudes have? It is thus not straightforward to conclude that $L1$ and $L2$ have no impact on the product design because multiple factors are at play. The difference between the design based on $L3$ compared to those based on $L1$ and $L2$ may weakly suggest that the product design may, at least be impacted - albeit weakly - by $L3$.

Scenario	Product Design
$L1$	$\xi_L = \left\{ \begin{array}{cccccc} 0.80 & 49.26 & 49.73 & 49.78 & 49.88 & 49.91 \\ 0.4957 & 0.0177 & 0.0566 & 0.0246 & 0.3208 & 0.0846 \end{array} \right\}$
$L2$	$\xi_T = \left\{ \begin{array}{ccccc} 0.78 & 49.38 & 49.43 & 49.77 & 49.93 \\ 0.5015 & 0.0069 & 0.0342 & 0.1062 & 0.3512 \end{array} \right\}$
$L3$	$\xi_T = \left\{ \begin{array}{cccccc} 0.84 & 0.86 & 49.56 & 49.77 & 49.87 & 49.90 \\ 0.2252 & 0.2732 & 0.1934 & 0.0901 & 0.1176 & 0.1005 \end{array} \right\}$

Table 5.15: The product designs based on each of the three scenarios $L1$, $L2$ and $L3$.

The Bayesian designs based on the three scenarios are given Table 5.16. The differences in the Bayesian designs in Table 5.16 reasonably suggest that the scenarios have an impact on the Bayesian designs. In particular, it can be observed that the settings of the low levels of concentration are quite variable across the Bayesian designs. Comparing the product design and the Bayesian designs based on $L1$, it is seen that the low level of concentration for the product design, $x_1 = 0.80$, appears to be situated between the low concentration levels of the Bayesian designs. A similar observation can be made for the product and Bayesian designs based on $L2$ and $L3$. For a model like the Michaelis-Menten model, where it is known that one of the support points is the maximum allowable concentration, could this mean that the low concentration of the product design is merely an average of the prior means for K_M? The results of the sensitivity study may give some credence to this.

Scenario	Bayesian Design
$L1$	$\xi_L = \left\{ \begin{array}{cccccc} 0.84 & 0.85 & 49.29 & 49.67 & 49.77 & 49.92 \\ 0.3713 & 0.1295 & 0.0950 & 0.0988 & 0.1165 & 0.1888 \end{array} \right\}$
	$\xi_T = \left\{ \begin{array}{ccccc} 0.77 & 49.46 & 49.76 & 49.84 & 49.87 \\ 0.4978 & 0.0754 & 0.1308 & 0.0924 & 0.2036 \end{array} \right\}$
$L2$	$\xi_L = \left\{ \begin{array}{ccccc} 0.54 & 49.82 & 49.85 & 49.86 & 49.95 \\ 0.5017 & 0.0198 & 0.0773 & 0.2398 & 0.1614 \end{array} \right\}$
	$\xi_T = \left\{ \begin{array}{ccccc} 1.07 & 1.09 & 1.12 & 49.84 & 49.85 \\ 0.0735 & 0.2001 & 0.2252 & 0.0987 & 0.4025 \end{array} \right\}$
$L3$	$\xi_T = \left\{ \begin{array}{ccccc} 0.64 & 49.67 & 49.75 & 49.77 & 49.88 \\ 0.5018 & 0.1938 & 0.2079 & 0.0607 & 0.0357 \end{array} \right\}$
	$\xi_T = \left\{ \begin{array}{ccccc} 1.15 & 1.20 & 49.59 & 49.90 & 50.00 \\ 0.2402 & 0.2614 & 0.0817 & 0.1558 & 0.2608 \end{array} \right\}$

Table 5.16: The Bayesian designs based on each of the three scenarios $L1$, $L2$ and $L3$.

CHAPTER 6

CONCLUSION AND FUTURE WORK

New experiments are often conducted to confirm the results of previous experiments. The Bayesian paradigm is particularly suited to such situations because information gained from previous experiments can be used to suggest prior distributions going forward. When there have been multiple previous studies, more than one prior distribution for the model parameters will usually result. Also, where prior distributions are elicited from a group of subject matter experts, multiple prior distributions will invariably result. In each of these cases, it is reasonable to assume that the priors will belong to a class of prior distributions. Bayesian experimental designs that are based on a single prior distribution are quite sensitive to the particular prior. It is important, therefore, to find experimental designs that work efficiently across the class of prior distributions. That is, a design that is robust to the class of priors.

The issue of multiple prior distributions has been examined, to some extent, for linear models but not for nonlinear models. The overall aim of this dissertation was two-fold:

1. Extend work to nonlinear models by introducing new robust design criteria - maximin, product, weighted product and geometric criteria.

2. Develop a reusable algorithm for generating designs based on the new design criteria.

Thus, the robust design problem is accompanied by the design generation problem which is a computationally intensive task. The design problem is addressed by the new design criteria introduced in Chapter 3. The design generation problem is tackled with genetic algorithms using a new genetic operator that is suited for the design problem

and is also in Chapter 3. The usefulness of genetic algorithms in conjunction with the new genetic operator is shown in Chapter 4. Improvements to existing designs for nonlinear models in the literature are presented.

The robust criteria introduced in this dissertation extend, in some sense, the Bayesian paradigm of design. The objective of the maximin design, based on the maximin criterion, is to maximize the minimum value of an optimality criterion over the class of prior distributions as shown in the examples. In practice this means protecting the experimenter against the worst-case scenario, that is, maximizing the minimum amount of information over the class or set of prior distributions. In a sense, this is designing an experiment with a cautious or somewhat pessimistic perspective. The maximin idea is not new in the statistical literature but the application in this dissertation is new.

Maximizing the product criterion results in a design that maximizes the product of expected utilities across the set of priors. The choice of product is made so that the resulting design is robust in the sense that it performs satisfactorily for a wide range of parameter values. Weighting the prior distributions in the product criterion results in the weighted product criterion. This is useful in cases where there is varying *a priori* weights in prior distributional assumptions. For example, if experts from whom priors are elicited have different amounts of experience and/or training, the weighted product criterion is recommended. The geometric criterion is based on the idea that the geometric mean, and not the arithmetic mean, of expected utilities provides a compromise. Thus, the design based on the geometric criterion is intended to provide a compromise between the Bayesian designs based on the prior distributions. It is noted here that if the priors are equally weighted, then the design based on the weighted product criterion is identical to that based on the geometric

criterion. Common among the proposed criteria is the fact that the resulting designs are functions of all the prior distributions or prior information matrices.

The new criteria were applied to the compartmental, Michaelis-Menten and four-parameter logistic models as shown in Chapter 5 and Appendix B. The designs were found to be satisfactory approximations to the Bayesian designs in some cases. Thus, they are recommended in cases of prior ambiguity. The choice of which design is appropriate given a nonlinear model is not straightforward as it may depend also on the prior distributions. Performance of the robust designs relative to locally optimal designs based on the priors is one way to address the question. Invariably, more work must be done to more completely address this.

Axiomatic prior pooling methods were also used to address the issue of multiple priors. The objective was to use these methods to pool or aggregated the prior distributions into a consensus prior distribution and obtain a Bayesian design based on this prior. It was observed in Chapter 5 that the independent pooling operator cannot be used if the supports of the priors do not overlap. This drawback is obviously not a concern for the proposed criteria. The Bayesian designs obtained using the consensus prior distributions were found to perform efficiently relative to the Bayesian designs and also locally optimal designs. Thus, this approach to the problem may also be an alternative solution.

It may be of interest in some cases to assess the differences between the robust design and the Bayesian designs if, for example, two priors with the same mean but different covariance matrices are available for design. In a broader sense, how sensitive are the robust designs to the fixing and/or varying the means and covariances of the prior distributions? Sensitivity analyses are complicated by the number of nonlinear parameters in a model. The Michaelis-Menten model, with only one nonlinear parameter, lends itself to such sensitivity analysis. Although some minor differences

could be seen between the product design and the corresponding Bayesian designs, not very much was learned about the sensitivity of the product design to fixing and/or varying prior means and variances. This could be a result of the nature of designs for the Michaelis-Menten model and/or the settings of the parameters of the priors. Consequently, more work in the future is required to fully assess sensitivity.

The limited use of algorithmic-based experimental designs occurs because of a lack of accessible and/or reusable algorithms to generate designs. In cases where algorithms are readily available, they may not be well-tuned to the problem and so convergence to the robust or optimal designs may take longer than necessary. This dissertation takes advantage of the power of genetic algorithms and introduces a new genetic operator which takes advantage of the structure of the problem to speed the search for the robust designs. The designs obtained using the genetic algorithm may be improved using the simulated annealing algorithm.

Additional statistical properties of the proposed criteria could be the subject of future research. In addition, it will be interesting to examine what class of prior distributions are sensitive to the weights used in the weighted product criterion. Other functionals upon which design criteria could be based are the minimum range and norm. Further work is required to assess the usefulness of these. Aggregating locally optimal designs into a compromise design using an appropriate clustering algorithm could also be pursued in the future.

REFERENCES CITED

Atkinson, A. C., 1982. Optimum biased coin designs for sequential clinical trials with prognostic factors. Biometrika 69 (1), 61–67.

Atkinson, A. C., 1992. A segmented algorithm for simulated annealing. Statistics and Computing 2, 221–230.

Atkinson, A. C., Chaloner, K., Herzberg, A. M., Juritz, J., 1993. Optimum experimental designs for properties of a compartmental model. Biometrics 49 (2), pp. 325–337.

Atkinson, A. C., Cox, D. R., 1974. Planning experiments for discriminating between models. Journal of the Royal Statistical Society. Series B (Methodological) 36 (3), pp. 321–348.

Atkinson, A. C., Donev, A. N., Tobias, R., 2007. Optimum Experimental Designs, With SAS. Oxford Statistical Science Series. Oxford University Press.

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

Bacharach, M., 1979. Normal bayesian dialogues. Journal of the American Statistical Association 74 (368), pp. 837–846.

Bates, D., 1983. The derivative of $|x^{'}x|$ and its uses. Technometrics 25 (4).

Bates, D., Watts, D., 1988. Nonlinear regression analysis and its applications. Wiley Series in Probability and Statistics. John Wiley & Sons.

Belloto, R., Dean, A., Moustafa, M., Molokhia, A., Gouda, M., Sokoloski, T., 1985. Statistical techniques applied to solubility predictions and pharmaceutical formulations: an approach to problem solving using mixture response surface methodology. International Journal of Pharmaceutics 23 (2), 195 – 207.

Berg, J., Tymoczko, J., Stryer, L., 2002. Biochemistry, 5th Edition. New York: W H Freeman, section 8.4, The Michaelis-Menten Model Accounts for the Kinetic Properties of Many Enzymes.

Berger, J., 1985. Statistical Decision Theory and Bayesian Analysis. Springer Series in Statistics. Springer.

Biles, W., 1975. A response surface method for experimental optimization of multi-response processes. Industrial and Engineering Chemistry Process Design and Development 14.

Borkowski, J. J., 2003. Using a genetic algorithm to generate small exact response surface designs. Journal of Probability and Statistical Science 1, 65–88.

Botev, Z., Kroese, D., 2004. Global likelihood optimization via the cross-entropy method with an application to mixture models. In: Simulation Conference, 2004. Proceedings of the 2004 Winter. Vol. 1. p. 535.

Box, G. E. P., 1957. Evolutionary operation: A method for increasing industrial productivity. Journal of the Royal Statistical Society. Series C (Applied Statistics) 6 (2), 81–101.

Box, G. E. P., Behnken, D. W., 1960. Some new three-level designs for the study of quantitative variables. Technometrics 2, 455–475.

Box, G. E. P., Lucas, H. L., 1959. Design of experiments in nonlinear situations. Biometrika 46, 77–90.

Box, G. E. P., Wilson, K. B., 1951. On the experimental attainment of optimum conditions. Journal of the Royal Statistical Society, Series B 13, 1–45.

Carter, W., Wampler, G., Stablein, D., 1983. Regression Analysis of Survival Data in Cancer Chemotherapy. Marcel Dekker.

Chaloner, K., Verdinelli, I., 1995. Bayesian experimental design: A review. Statistical Science 10 (3), 273–304.

Chernoff, H., 1953. Locally optimal designs for estimating parameters. The Annals of Mathematical Statistics 24 (4), 586–602.

Clyde, M. A., 1993. Bayesian optimal designs for approximate normality. Ph.D. thesis, Univ. Minnesota.

Corana, A., Marchesi, M., Martini, C., Ridella, S., 1987. Minimizing multimodal functions of continuous variables with the "simulated annealing" algorithm. ACM Transactions on Mathematical Software 13, 262–280.

DasGupta, A., Studden, W., 1988. Robust bayesian analysis and optimal experimental designs in normal linear models with many parameters. Tech. rep., Dept. of Statistics, Purdue Univ.

DasGupta, A., Studden, W. J., 1991. Robust bayesian experimental designs in normal linear models. The Annals of Statistics 19 (3), pp. 1244–1256.

Dincer, S., Ozdurmus, S., 1977. Mathematical model for enteric film coating of tablets. Journal of Pharmaceutical Sciences 66, 1070–1073.

Dror, H. A., Steinberg, D. M., 2006. Robust experimental design for multivariate generalized linear models. Technometrics 48 (4), 520–529.

Farrell, R., Kiefer, J., Walbran, A., 1968. Optimum multivariate designs. In: Proceedings of the 5th Berkeley Symposium. Vol. 1. University of California Press, Berkeley, CA, pp. 113–138.

Finney, D. J., 1943. The fractional replication of factorial arrangements. Annals of Eugenics 12 (1), 291–301.

Ford, I., Titterington, D. M., Kitsos, C. P., 1989. Recent advances in nonlinear experimental design. Technometrics 31 (1), 49–60.

Ford, I., Torsney, B., Wu, C. F. J., 1992. The use of a canonical form in the construction of locally optimal designs for non-linear problems. Journal of the Royal Statistical Society. Series B (Methodological) 54 (2), 569–583.

Genest, C., 1984. A characterization theorem for externally bayesian groups. The Annals of Statistics 12 (3), pp. 1100–1105.

Genest, C., Zidek, J. V., 1986. Combining probability distributions: A critique and an annotated bibliography. Statistical Science 1 (1), pp. 114–135.

Givens, G. H., Roback, P. J., 1999. Logarithmic pooling of priors linked by a deterministic simulation model. Journal of Computational and Graphical Statistics 30, 8–457.

Goffe, W. L., Ferrier, G. D., Rogers, J., 1994. Global optimization of statistical functions with simulated annealing. Journal of Econometrics 60, 65–99.

Goldberg, D., 1989. Genetic algorithms in search, optimization, and machine learning. Artificial Intelligence. Addison-Wesley.

Haines, L. M., 1987. The application of the annealing algorithm to the construction of exact optimal designs for linear regression models. Technometrics 29 (4), 439–447.

Haupt, R., Haupt, S., 2004. Practical Genetic Algorithms. Wiley.

Heredia-Langner, A., Carlyle, W. M., Montgomery, D. C., Borror, C. M., Runger, G. C., 2003. Genetic algorithms for the construction of d-optimal designs. Journal Name: Journal of Quality Technology, 35(1):28-46.

Hill, W. J., Hunter, W. G., 1974. Design of experiments for subsets of parameters. Technometrics 16 (3), pp. 425–434.

Holland, J. H., 1975. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. University of Michigan Press.

Huck, P. M., Murphy, K. L., Reed, C., LeClair, B. P., 1977. Optimization of polymer flocculation of heavy metal hydroxides. Journal (Water Pollution Control Federation) 49 (12), 2411–2418.

Johnson, R. T., Montgomery, D. C., 2010. Designing experiments for nonlinear modelsan introduction. Quality and Reliability Engineering International 26 (5), 431–441.

Kadane, J., Wolfson, L. J., 1998. Experiences in elicitation. Journal of the Royal Statistical Society: Series D (The Statistician) 47 (1), 3–19.

Khuri, A. I., Lee, J., 1998. A graphical approach for evaluating and comparing designs for nonlinear models. Computational Statistics & Data Analysis 27 (4), 433 – 444.

Kiefer, J., 1959. Optimum experimental designs (with discussion). Journal of the Royal Statistical Society, Series B 21, 272–319.

Kiefer, J., 1960. Optimum experimental designs v. In: Proc. Fourth Berkeley Symp. Math. Statist. Vol. 1. Univ. of California Press, pp. 381–405.

Kiefer, J., 1961. Optimum designs in regression problems, ii. The Annals of Mathematical Statistics 32 (1), pp. 298–325.

Kiefer, J., Wolfowitz, J., 1959. Optimum designs in regression problems. The Annals of Mathematical Statistics 30 (2), pp. 271–294.

Kiefer, J., Wolfowitz, J., 1960. The equivalence of two extremum problems. Canadian Journal of Mathematics 12, 363–366.

Kirkpatrick, S., Gelatt, C. D., Vecchi, M. P., 1983. Optimization by simulated annealing. Science 220, 671–680.

Kroese, D. P., Porotsky, S., Rubinstein, R. Y., 2006. The cross-entropy method for continuous multi-extremal optimization. Methodology and Computing in Applied Probability 8, 383–407.

Lauter, E., 1976. Optimal multipurpose designs for regression models. Mathematische Operationsforschung Statistik 7 (1), 51–68.

Limmun, W., Borkowski, J. J., Chomtee, B., 2012. Using a genetic algorithm to generate D —optimal designs for mixture experiments. Quality and Reliability Engineering International.

Lindley, D., 1983. Reconciliation of probability distributions. Operations Research 31 (5), pp. 866–880.

Maddox, I. S., Richert, S. H., 1977. Use of response surface methodology for the rapid optimization of microbiological media. Journal of Applied Microbiology 43 (2), 197–204.

Mager, P., 1982a. Multivariate response surface optimization. Pharmazie 37, 658–660.

Mager, P., 1982b. Structure-neurotoxicity relationships applied to organphosphorus pesticides. Toxicology Letters 11, 67–71.

Matthews, J. N. S., Allcock, G. C., 2004. Optimal designs for michaelis–menten kinetic studies. Statistics in Medicine 23 (3), 477–491.

Michalewicz, Z., 1992. Genetic Algorithms + Data Structures = Evolution Programs. Springer-Verlag.

Montgomery, D. C., Bettencourt, V. M., 1977. Multiple response surface methods in computer simulation. Simulation 29, 113–121.

Montgomery, D. C., Peck, E. A., Vining, G. G., 2006. Introduction to Linear Regression Analysis, 4th Edition. John Wiley & Sons.

Myers, R. H., Khuri, A. I., Carter, W. H., 1989. Response surface methodology:1966-1988. Technometrics 3, 137–157.

Myers, R. H., Montgomery, D. C., Anderson-Cook, C. M., 2009. Response Surface Methodology: Process and Product Optimization Using Designed Experiments. Wiley.

Oakley, J. E., O'Hagan, A., 2007. Uncertainty in prior elicitations: A nonparametric approach. Biometrika 94 (2), pp. 427–441.

Ozol-Godfrey, A., Anderson-Cook, C. M., Robinson, T. J., 2008. Fraction of design space plots for generalized linear models. Journal of Statistical Planning and Inference 138 (1), 203 – 219.

Plackett, R., Burman, J., 1946. The design of optimum multifactorial experiments. Biometrika 33, 305–325.

Ponce De Leon, A. C., Atkinson, A. C., 1991. Optimum experimental design for discriminating between two rival models in the presence of prior information. Biometrika 78 (3), 601–608.

Pronzato, L., Walter, E., 1988. Robust experiment design for nonlinear regression models. In: Model-Oriented Data Analysis. Vol. 297 of Lecture Notes in Economics and Mathematical Systems. Springer, Berlin, pp. 77–86.

Pukelsheim, F., 1993. Optimal Design of Experiments. Wiley, New York.

Pukelsheim, F., Reider, S., 1992. Efficient rounding of approximate designs. Biometrika 79, 763–770.

Rao, C., 1973. Linear Statistical Inference and its Applications, 2nd Edition. Wiley, New York.

Robinson, K. S., Khuri, A. I., 2003. Quantile dispersion graphs for evaluating and comparing designs for logistic regression models. Comput. Stat. Data Anal. 43 (1), 47–62.

Rodda, B. E., Sampson, C. B., Smith, D. W., 1975. The one-compartment open model: Some statistical aspects of parameter estimation. Journal of the Royal Statistical Society. Series C (Applied Statistics) 24 (3), pp. 309–318.

Rubinstein, R. Y., Kroese, D. P., 2004. The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation, and Machine Learning. Springer-Verlag.

SAS/STAT, 2003. SAS/STAT Software, Version 9.1. Cary, NC.

Seber, G., Wild, C., 1989. Nonlinear Regression. Wiley, New York.

Shannon, C. E., 1948. A mathematical theory of communication. Bell System Technical Journal 27, 379–243, 623–656.

Shechter, M., Heady, E., 1970. Response surface analysis and simulation models in policy choices. American Journal of Agricultural Economics 52, 41–50.

Shek, E., Ghani, M., Jones, R., 1980. Simplex search in optimization of capsule formulation. Journal of Pharmaceutical Sciences 69, 1135–1142.

Silvey, S. D., Titterington, D. H., Torsney, B., 1978. An algorithm for optimal designs on a design space. Communications in Statistics - Theory and Methods 7 (14), 1379–1389.

Smith, K., 1918. On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. Biometrika 12 (1/2), pp. 1–85.

Stone, M., 1961. The opinion pool. Annals of Mathematical Statistics (32), 1339–42.

Toman, B., 1992. Bayesian robust experimental designs for the one-way analysis of variance. Statistics & Probability Letters 15 (5), 395 – 400.

Toman, B., Gastwirth, J. L., 1993. Robust bayesian experimental design and estimation for analysis of variance models using a class of normal mixtures. Journal of Statistical Planning and Inference 35 (3), 383 – 398.

Toman, B., Notz, W. I., 1991. Bayesian optimal experimental design for treatment - control comparisons in the presence of two-way heterogeneity. Journal of Statistical Planning and Inference 27 (1), 51 – 63.

Wallis, R., 1978. Some uses of multifactorial response surface analysis in temperature tolerance studies. Australian Journal of Marine and Freshwater Research 27, 487–498.

Waterhouse, T. H., September 2005. Optimal experimental design for nonlinear and generalised linear models. Ph.D. thesis, University of Queensland.

Waterhouse, T. H., Eccleston, J. A., Duffull, S. B., 2009. Optimal design criteria for discrimination and estimation in nonlinear models. Journal of Biopharmaceutical Statistics 19 (2), 386–402.

Woods, D. C., Lewis, S. M., Eccleston, J. A., Russell, K. G., 2006. Designs for generalized linear models with several variables and model uncertainty. Technometrics 48 (2), 284–292.

APPENDICES

APPENDIX A

PLOTS

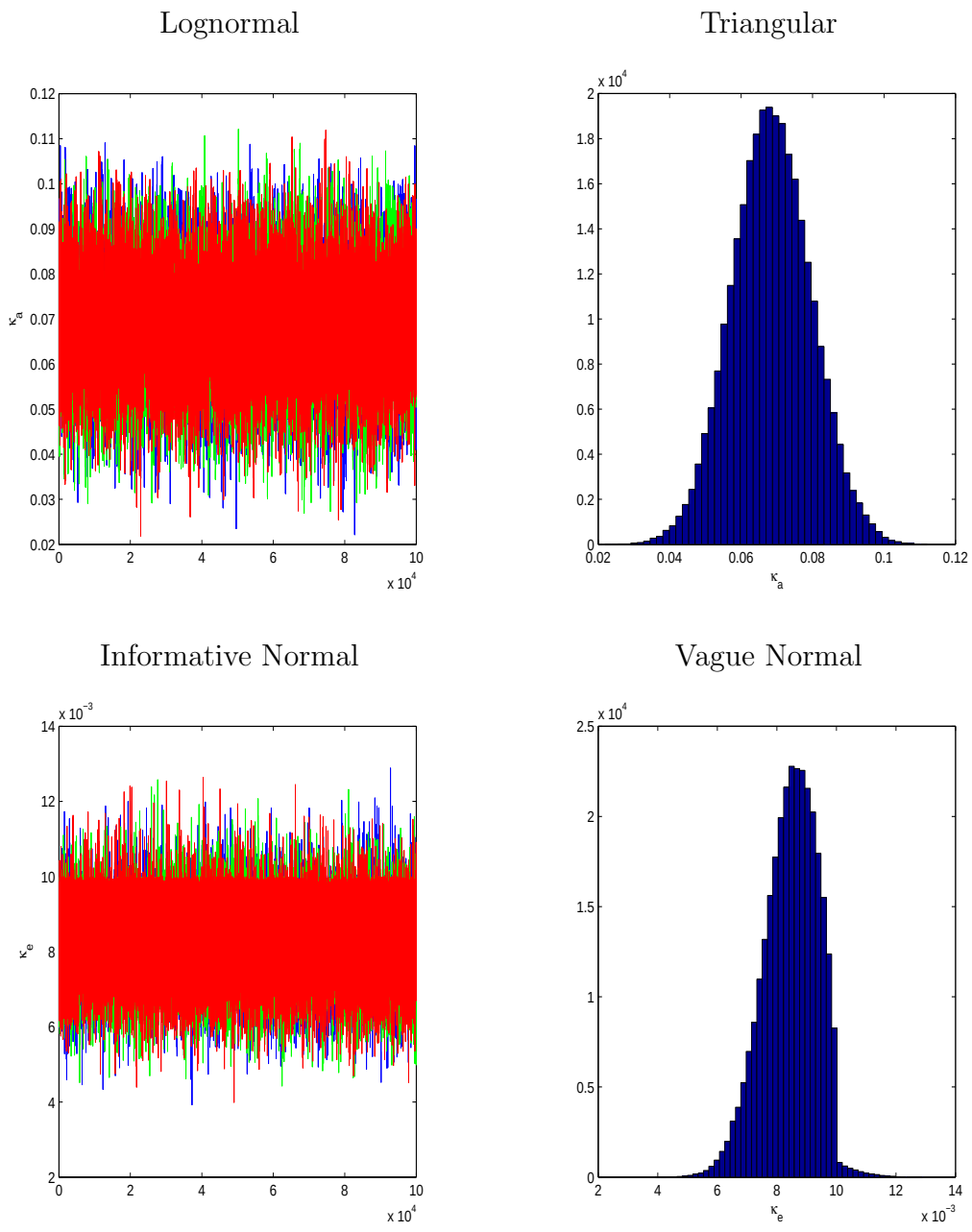

Figure A.1: Sample path history plots and distributions of draws of the parameters based on logarithmic pooling using weights of 0.25 for each prior.

APPENDIX B

DESIGN FOR FOUR-PARAMETER LOGISTIC MODEL

Biological assays are methods that investigate the biological properties of a compound (e.g., a drug) by the analysis of its effects on living matter. In a typical bioassay, a stimulus (e.g., a dose of drug) is applied to a subject yielding a change in a measurable characteristic (or response) of the subject. In drug development research, the relationship between the dose of a drug and a clinical endpoint (response) is of paramount interest. Consequently, estimating the parameters of the model describing the dose-response relationship is critical. In most pharmacological studies, the four-parameter logistic (4PL) model has been found to adequately model this relationship. In designing an experiment that will optimally estimate the model parameters, suppose that prior elicitation results in two multivariate Normal distributions $p_1(\theta)$ and $p_2(\theta)$ with means and covariance matrices $\mu_{\mathbf{1}}, \mathbf{V_1}$ and $\mu_{\mathbf{2}}, \mathbf{V_2}$ respectively, where

$$\begin{aligned} \mu_{\mathbf{1}} &= (15.03, 1.31, 530, 1587), \qquad \mathbf{V_1} = diag(1.00, 0.01, 1, 0.50) \qquad \text{and} \\ \mu_{\mathbf{2}} &= (5.01, 0.44, 177, 529), \qquad \mathbf{V_2} = diag(2.00, 0.02, 2.00, 1.00). \end{aligned}$$

It is insightful to look at the distribution of logistic curves under these two prior distributions. Figure B.1 contains logistic curves based on a random sample of 200 sets of parameter values from each of the two prior distributions. The plots show that there is a large number of different profiles (or shapes) that the 4PL curve can assume. The goal is to find a design that performs sufficiently well, for estimation purposes, for example, across these different profiles.

The distribution of curves under $\mu_{\mathbf{2}}, \mathbf{V_2}$ are more variable compared to those under $\mu_{\mathbf{1}}, \mathbf{V_1}$ due to the relatively larger variability in the second prior. The objective here is is to show that a design that is a function of the two information (or precision) matrices is more desirable than one that is based on exactly one of the prior distributions.

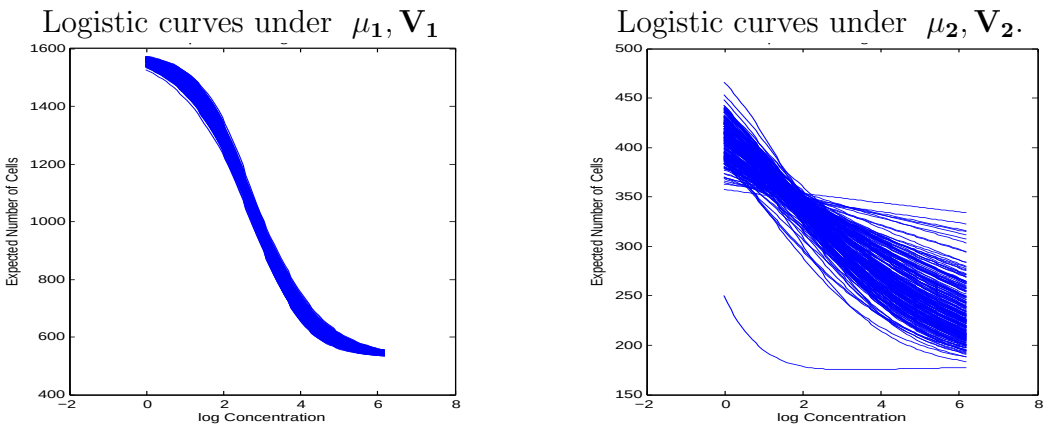

Figure B.1: Distribution of a random sample of logistic curves under two prior parameter distributions.

The following Bayesian D-optimal designs are obtained using priors $p_1(\theta)$ and $p_2(\theta)$:

$$\xi_{D1} = \left\{ \begin{array}{cccc} 0.0338 & 1.9794 & 3.5168 & 6.1215 \\ 0.2726 & 0.2611 & 0.1943 & 0.2719 \end{array} \right\} \tag{B.1}$$

and

$$\xi_{D2} = \left\{ \begin{array}{cccc} 0.0379 & 1.5555 & 3.7501 & 6.1409 \\ 0.2152 & 0.1721 & 0.3496 & 0.2631 \end{array} \right\} \tag{B.2}$$

respectively. The equi-weighted product design obtained is

$$\xi_C = \left\{ \begin{array}{cccc} -0.0138 & 1.8378 & 3.5650 & 6.1625 \\ 0.2416 & 0.2431 & 0.2801 & 0.2352 \end{array} \right\}. \tag{B.3}$$

This design is a weighted product design as well as a geometric design because $w_1 = w_2 = \frac{1}{2}$. The three designs are similar with respect to the latter two design points but quite different in terms of the earlier two design points. Differences can also be seen in the weight distributions of the designs.

To evaluate the performance of the robust design, relative efficiency is used. Distributions of relative efficiency of the composite design ξ_C are given in Figure B.2.

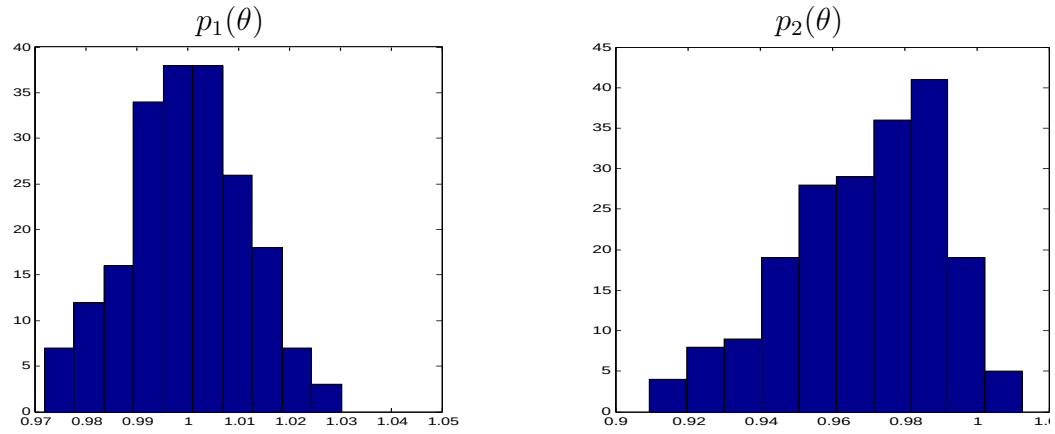

Figure B.2: Empirical distribution of relative efficiencies of the robust design, ξ_C across two prior distributions.

Plots of the distribution of relative efficiency of ξ_C (based on a random sample of 200 sets of parameter values from each of the two prior distributions) across the two prior distributions in Figure B.2. Numerical summaries of the these plots are also given in Table B.1. The relative efficiencies are generally greater than 1 or within a small neighborhood of it as seen in Table B.1. This suggests the robustness of ξ_C to the two prior distributions.

The essential features of the distributions in Figure B.3 are summarized in Table B.2 below where it is worth noting that the relative efficiencies are all less than 1. This is indicative of the sub-optimality of the ξ_{D1} for $p_2(\theta)$ and also ξ_{D2} for $p_1(\theta)$.

Summary Statistic	Relative D-efficiencies	
	$\mu_1, \mathbf{V_1}$	$\mu_2, \mathbf{V_2}$
Minimum	0.972	0.910
Maximum	1.030	1.013
Median	1.000	0.972
Mean	1.000	0.970

Table B.1: Numerical summaries of the empirical distribution of the relative frequency of the robust (or composite) design across the two priors.

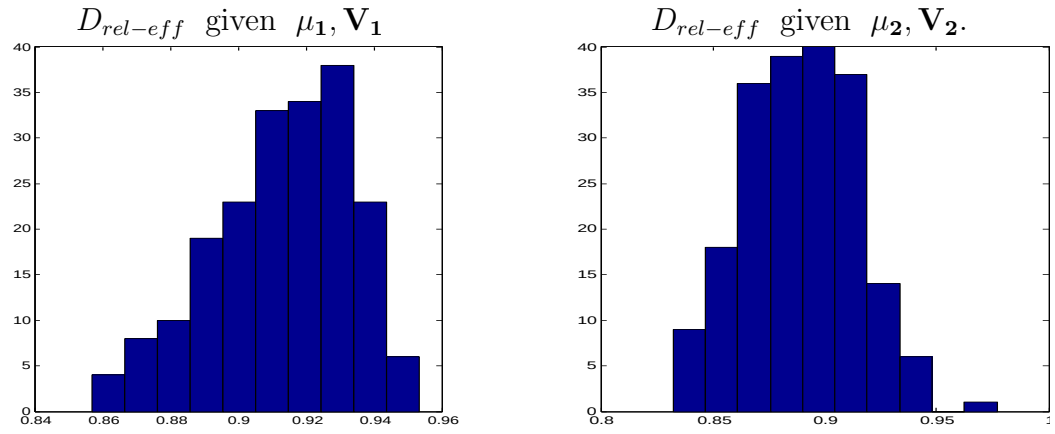

Figure B.3: Left: Distribution of the efficiency of the Bayesian optimal design ξ_{D1} relative to locally optimal designs based on $p_2(\theta)$. Right: Distribution of the efficiency of the Bayesian optimal design ξ_{D2} relative to locally optimal designs based on $p_1(\theta)$.

Summary Statistic	Relative D-efficiencies	
	ξ_{D1} on μ_2, V_2	ξ_{D2} on μ_1, V_1
Minimum	0.857	0.832
Maximum	0.954	0.977
Median	0.916	0.890
Mean	0.913	0.889

Table B.2: Numerical summaries of the distribution of relative efficiencies of the Bayesian optimal designs ξ_{D1} and ξ_{D2} on $p_2(\theta)$ and $p_1(\theta)$ respectively.

APPENDIX C

MATLAB CODE

```
function [bestD, bestcrit, R] = gaMaxMin_Mate(npoints, popsize, PrDist, Ndraws, Ngen, pc, pm)
%This function is the main function that calls all other functions.
%It differs from the gaMaxMin() function in that it uses the proposed
%mating operator. Its arguments are identical to those of gaMaxMin

%Use an odd population size
    npairs = (popsize-1)/2;

%number of parameters (in model)
    [~, nparms, ~] = size(PrDist);

%Allocate space for robust criterion value at each generation.
R = zeros(Ngen,1);

    %Robust criterion vector
    Rcrit = zeros(1, popsize);

acc = zeros(5,2); %Accumulator

%matrix of offspring in each generation.
%npairs is doubled because each pair of parents
%produces two offspring.
    offmat = zeros(2*npairs, 2*npoints+1);    %last column is for saving row index of parent

    %Generate initial population
    pop = InitialPop(popsize, npoints);

% Compute fitness of initial population
    for i = 1:popsize
         Xs = check_ind2(pop(i,:), npoints, nparms);
```

```
          %Calcuate fitness
          Rcrit(i) = fitMaxMin(Xs, Ndraws, PrDist);

          %Replace ith chromosome with its healthier version
          pop(i,:)= Xs;
    end

%Sort criterion values in order to rank them.
%Sorting is done from smallest to largest.
    [Rcrit, ind] = sort(Rcrit);
    bestcrit = Rcrit(popsize);
    bestD = pop(ind(popsize),:);

clc;
    tic;
    h = waitbar(0,'Initializing waitbar...');

%Perform genetic operations
    for k = 1:Ngen
        perc = k/Ngen;
        waitbar(perc,h,sprintf('%d%% percent...',perc*100));

      for i = 1:npairs
       %sample without replacement from the first (popsize-1) elements of ind
        pa = randsample(ind(1:popsize-1),2);
        [off1, off2, acc] = mateMaxMin(pop(pa(1),:), pop(pa(2),:), npoints, nparms, acc, ......
        pc, Ndraws, PrDist);

        %Replace parents in population by fitter offspring
         pop(pa(1),:)= off1;
         pop(pa(2),:)= off2;
```

```
end

  %%Check to make sure population is healthy before next
  %%generation(Not necessary since mate function checks health)
  for j = 1:popsize
      pop(j,:) = check_ind2(pop(j,:), npoints, nparms);
      Rcrit(j) = fitMaxMin(pop(j,:), Ndraws, PrDist);
  end

  %Sort fitnesses
  [Rcrit, ind] = sort(Rcrit);

  if Rcrit(popsize) > bestcrit
      bestcrit = Rcrit(popsize);
      bestD = pop(ind(popsize),:);
  end;


 %Perform Gaussian mutation
 for i = 1:popsize
     if i ~= ind(popsize)     %don't mutate best chromosome
         pop(i,:) = mutation(pop(i,:),npoints,nparms, pm);
     end
      Rcrit(i) = fitMaxMin(pop(i,:), Ndraws, PrDist);
 end

 %Sort fitnesses
  [Rcrit, ind] = sort(Rcrit);
  if Rcrit(popsize) > bestcrit
      bestcrit = Rcrit(popsize);
      bestD = pop(ind(popsize),:);
```

```
        end;

        R(k) = bestcrit;
        disp(bestcrit);
        disp(bestD);

        save('MaximinMate.txt', 'bestcrit', 'bestD', '-ascii', '-append');

    end

    close(h)
    toc


function D  = fitMaxMin(design, Ndraws, PrDist)
%function [log_dets, thetaMat, F, D]  = fitMaxMin(design, Ndraws, PrDist)
%This function returns the min D-criterion value for k prior distributions
%using monte carlo integration. The result is stored in D.

%PrDist is a 3-dimensional matrix of the prior distns.
%It is 2 X 3 X K where K is the number of prior distns.

%Number of draws from each prior distn.
[~,m,k] = size(PrDist);

N = Ndraws;

%Extract design points from design
num_dpts = length(design)/2;
design_pts = design(1:num_dpts);
```

```
%Allocate space for the derivative matrices
%F = zeros(num_dpts, m, k);

%Allocate space for the log determinants
log_dets = zeros(N, k);

%Create weight matrix
W = diag(design(num_dpts + 1: 2*num_dpts));



%Column 1 of PrDist
%contains the prior means
%Get parameters for distns
%mu = reshape(PrDist(:,1,:), m, k)'; %Get the means
%sigma = PrDist(:,2:n,:);   %Get the variances

for i = 1:N
%Get draws from prior distributions
%thetaMat is k X m.
%thetaMat = mvnrnd(mu, sigma);  %Make draws
    thetaMat = vec2mat(reshape(PrDist(i,:,:), m*k,1)', m);

%This is passed to CalDets
%Create derivative matrices and compute determinants
    % for j = 1:k
        % F(:,:,j) = DerMatrix(design_pts, thetaMat(j,:));
% log_dets(i,j) = log(det(F(:,:,j)'*W*F(:,:,j)));
    % end

log_dets(i,:) = CalDets(num_dpts, m, thetaMat, W, design_pts);
end
```

```
%Sum N log determinants for each prior distn and divide by number
%of draws (Monte Carlo Integration)
D = min(sum(log_dets)/N);

function M = CalDets(num_dpts, m, thetaMat, W, design_pts)

%Allocate space for the derivative matrices
[k, ~] = size(thetaMat);
F = zeros(num_dpts, m, k);

%Allocate space for determinants
M = zeros(1,k);

%Create derivative matrices and compute determinants
    for j = 1:k
        F(:,:,j) = DerMatrix(design_pts, thetaMat(j,:));
M(1,j) = log(det(F(:,:,j)'*W*F(:,:,j)));
    end

function F = DerMatrix(x, theta)

%This function generates the derivative matrix F
%for a one-compartment open model. The number of
%rows of the matrix is equal to the length of indv

n = length(x);

%Create matrix F (of zeros)
F = zeros(n, length(theta));
```

```
%Get the parameters: absorption and
%elimination constants
ka = theta(1); ke = theta(2);

%Fill out derivative matrix
F(:,1) = ((ka*(ka-ke).*x + ke).*exp(-ka.*x) - ke.*exp(-ke.*x))/(ka-ke).^2;

F(:,2) = ((-ka.*(ka-ke).*x + ka).*exp(-ke.*x) - ka.*exp(-ka.*x))/(ka-ke).^2;



function [cA,cB, acc] = mateMaxMin(A, B, npoints, nparms, acc, pc, Ndraws, PrDist)
%(pop(pa(1),:),pop(pa(2),:),npoints, acc, pc, theta, cov)
%This function produces several offspring given 2 parents and outputs the
%'fitter' of the bunch.

%Initial offpsring of both parents
off1A = A; off2A = A;
off1B = B; off2B = B;

%offspring matrices
fmatA = zeros(4,length(A)); fmatA(4,:)= A;
fmatB = zeros(4,length(B)); fmatB(4,:)= B;

if rand < pc
   b = rand;
   %Fix weights, vary (blend) support points
   off1A(1:npoints) = b*A(1:npoints)+(1-b)*B(1:npoints);
   off1B(1:npoints) = b*B(1:npoints)+(1-b)*A(1:npoints);
   fmatA(1,:)= off1A; fmatB(1,:)= off1B;

   %Fix support points, vary (blend) weights
```

```
b = rand;
off2A(npoints+1:length(A))= b*A(npoints+1:2*npoints)+(1-b)*B(npoints+1:2*npoints);
off2B(npoints+1:length(B))= b*B(npoints+1:2*npoints)+(1-b)*A(npoints+1:2*npoints);
fmatA(2,:)= off2A; fmatB(2,:)=off2B;

%Vary support points, vary (blend) weights
b = rand;
off3A = b*A + (1-b)*B;
off3B = b*B + (1-b)*A;
fmatA(3,:)= off3A; fmatB(3,:)= off3B;

%First check health of the offspring
%and then calculate fitness (d-optimality)
dcritA = zeros(1,4);
dcritB = dcritA;

for i = 1:length(dcritA);
    fmatA(i,:) = check_ind2(fmatA(i,:),npoints,nparms);
    fmatB(i,:) = check_ind2(fmatB(i,:),npoints,nparms);
    dcritA(i) = fitMaxMin(fmatA(i,:), Ndraws, PrDist);
    dcritB(i) = fitMaxMin(fmatB(i,:), Ndraws, PrDist);
end

%sort fitnesses
[~,indA] = sort(dcritA);
[~,indB] = sort(dcritB);
cA = fmatA(indA(4),:);
cB = fmatB(indB(4),:);


%Increment counters appropriately (for A and B)
```

```
    if indA(4)==1
        acc(1,1) = acc(1,1)+1;
    end

    if indA(4)==2
        acc(2,1)=acc(2,1)+1;
    end

    if indA(4)==3
        acc(3,1)=acc(3,1)+1;
    else
        acc(4,1)=acc(4,1)+1;
    end

    if indB(4)==1
        acc(1,2) = acc(1,2)+1;
    end

    if indB(4)==2
        acc(2,2)=acc(2,2)+1;
    end

    if indB(4)==3
        acc(3,2)=acc(3,2)+1;
    else
        acc(4,2)=acc(4,2)+1;
    end
else
     cA = A;
     cB = B;
     acc(5,1)= acc(5,1)+1;
```

```
end

%Auxiliary Functions
function new_ind = check_ind2(ind, npoints, nparms)

xs = ind;
%Check if support points are within the design region
spts = checkpoints(ind(1:npoints),npoints);

%Make sure number of unique support points are greater or equal to
%number of parameters in model
while length(unique(spts)) < nparms
    spts = spts + 0.001.*rand(1,npoints);
    spts = checkpoints(spts,npoints);
end

xs(1:npoints)=spts;

wts = ind(npoints+1:2*npoints);

%Make sure all weights are non-negative
for i = 1:length(wts)
    if wts(i)<0
        wts(i)=0;
    end
end

%Make sure no weight is identically zero
wts = wts/sum(wts);
ind2 = find(wts==0);
if ~isempty(ind2)
```

```
    wts = checkweights(wts);
end

xs(npoints+1:2*npoints)= wts;
new_ind = xs;

function pts = checkpoints(pts,npoints)

%Make sure support points are within the design space

lx = 2; hx = 720;
    for i = 1:npoints
        if pts(i) < lx
            pts(i)= lx;
        else
            if pts(i) > hx
                pts(i) = hx;
            end
        end
    end

 function W = checkweights(w)

ind3 = find(w <= 0);

for i=ind3
    w(i) = w(i) + 0.0001;
end

W = w/sum(w);
```

```
function pop = InitialPop(popsize, npoints)

pop = [720*lhsdesign(popsize, npoints) rand(popsize,npoints)];

function mut = mutation(ind, npoints,p1,pm)

%mutation of support points
%first randomly draw a support point

l = 2; u = 720;
sigma = 0.5;

if rand < pm
    m = randsample(1:npoints,1);
    ind(m) = normt_rnd(ind(m),sigma,l,u);
    m2 = randsample(npoints+1:2*npoints,1); %Now weights
    ind(m2) = normt_rnd(ind(m2),0.1*sigma,0,1);
    mut = check_ind2(ind,npoints,p1);
else
    mut = check_ind2(ind, npoints,p1);
end

function result = normt_rnd(mu,sigma2,left,right)
% PURPOSE: random draws from a normal truncated to (left,right) interval
% ---------------------------------------------------
% USAGE: y = normt_rnd(mu,sigma2,left,right)
% where: mu = mean (nobs x 1)
% sigma2 = variance (nobs x 1)
% left = left truncation points (nobs x 1)
% right = right truncation points (nobs x 1)
% ---------------------------------------------------
```

```
% RETURNS: y = (nobs x 1) vector
% -------------------------------------------------------
% NOTES: use y = normt_rnd(mu,sigma2,left,mu+5*sigma2)
% to produce a left-truncated draw
% use y = normt_rnd(mu,sigma2,mu-5*sigma2,right)
% to produce a right-truncated draw
% -------------------------------------------------------
% SEE ALSO: normlt_rnd (left truncated draws), normrt_rnd (right truncated)
%
% adopted from Bayes Toolbox by
% James P. LeSage, Dept of Economics
% University of Toledo
% 2801 W. Bancroft St,
% Toledo, OH 43606
% jpl@jpl.econ.utoledo.edu
% For information on the Bayes Toolbox see:
% Ordinal Data Modeling by Valen Johnson and James Albert
% Springer-Verlag, New York, 1999.
if nargin ~= 4
    error('normt_rnd: wrong # of input arguments');
end;
std = sqrt(sigma2);
% Calculate bounds on probabilities
lowerProb = Phi((left-mu)./std);
upperProb = Phi((right-mu)./std);
% Draw uniform from within (lowerProb,upperProb)
u = lowerProb+(upperProb-lowerProb).*rand(size(mu));
% Find needed quantiles

result = mu + Phiinv(u).*std;
function val=Phiinv(x)
```

```
% Computes the standard normal quantile function of the vector x, 0<x<1.
 val=sqrt(2)*erfinv(2*x-1);


function y = Phi(x)
% Phi computes the standard normal distribution function value at x
y = .5*(1+erf(x/sqrt(2)));
```