Stochastic Optimization and Subgroup Selection

Bob Salim

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

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Lurdes Y.T. Inoue, Chair

Scott Emerson

Kathleen Kerr

Marcia Ciol, GSR

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Stochastic Optimization and Subgroup Selection

Bob Salim

Chair of the Supervisory Committee:
Professor Lurdes Y.T. Inoue
Biostatistics

An important area in statistics is that of experimental or study design. The most typical problem is that of finding the required sample size that meets specific goals such as controlling type I and II error rates at given levels. In most experiments, additional constraints may be imposed from both practical and technical perspectives. From the practical perspective, it is expensive, both in monetary and time scales, or even impossible, to perform experiments over all possible values that the design variables can take. Some approaches have been developed to design an experiment to achieve maximum information given the restrictions in sample size, known as 'optimal designs'. In an optimal design, the design points are selected to maximize a pre-selected optimality criterion, and it can be done using optimization methods. In this work, we propose a method for stochastic optimization called "forward slice" and evaluate its performance relative to other optimization methods. We also demonstrate the use of our method in design problems. Specifically, our simulation studies indicate that the "forward slice" selects the global optimum more often than other optimization methods. Further, when applying the "forward slice" to design problems, our method performs well when obtaining locally D-optimal design points and achieves a higher median overall D-optimality compared to the design points obtained using an alternative algorithm. In addition to experimental design, optimality considerations also arise post-design and post-experiment. One such example is in the problem of subgroup selection in a clinical

trial. Most studies are designed to address only the primary inferential questions. However, in many studies, it is also of interest to assess differential associations in subpopulations. We propose a decision-theoretic approach to subgroup analysis consisting of two stages: model selection and subgroup reporting. We assess and compare the performance of our proposed method with some traditional approaches for subgroup selection under different scenarios. Our simulation studies show that the performance of the decision-theoretic method is similar to that of testing an interaction followed by stratified analyses based on the results of the interaction test. In the selection of the subgroups, the proposed method favors reporting subgroups that exhibit larger treatment effect, are larger, and are simpler (less complex). We also observe a trade-off where approaches that tend to have a larger power for detecting a subpopulation may perform more poorly when the effect is in fact homogeneous in the overall population. In addition, the proposed method allows for incorporation of prior information in both the model selection and subgroup reporting stages which may increase power while keeping the type-I error controlled.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank Professor Lurdes Inoue for her guidance, direction, support, and help throughout my time at the University of Washington. I would also like to thank Professors Scott Emerson, Kathleen Kerr, and Marcia Ciol for their helpful comments and for their participation on my Supervisory Committee. I thank the Department of Biostatistics for this privilege and opportunity to pursue my study in the field of Biostatistics. Finally, I thank my family and friends for their support and encouragement throughout my education.

Chapter 1

INTRODUCTION

An important area in statistics is that of experimental or study design. The most typical problem is that of finding the required sample size that meets specific goals such as control of type I and II error rates at given levels. However, a broader range of goals may also need to be addressed to fully specify an experiment or study. Specifically, in experimental designs, the choice of values for the variables controlled by the investigator may be of great interest. This, in turn, arises from practical and technical difficulties. From the practical perspective, it is expensive, both in monetary and time scales, or even impossible, to perform experiments over all possible values that the design variables can take. From the technical perspective, we face the issue that the precision for estimating model parameters largely depends on the design.

Some approaches have been developed to design an experiment that would obtain maximum information given the restrictions in sample size, known as 'optimal design'. For standard linear models (i.e. continuous response), tools for design selection, estimation and evaluation of the model and response surface based on data generated by the design have been developed in the area of response surface methodology (RSM). However, in many experiments, such as in clinical or epidemiological experiments, the response of interest may be non-continuous in nature (e.g. binary or count data), and generalized linear models (GLM) may be more appropriate. Unfortunately, the area of 'optimal designs' is less developed for generalized linear models due to the dependence of the design on the unknown parameters of the GLM model [38].

In addition to experimental design, optimality considerations also arise post-design and post-experiment. One such example is in the problem of subgroup selection in a clinical

trial. Most studies are designed to address only the primary inferential questions. However, in many studies, it is also of interest to assess differential associations in the subpopulations level. However, given the sample size, most trials have low power to detect differential subgroup effects. Furthermore, improper conduct of subgroup analysis inflate the false positive rate excessively, resulting in misleading conclusions. Thus, it is of great importance that the probability that a treatment truly works in a reported subgroup be optimal.

The overarching goal of this dissertation is to develop methods for optimal design. Specifically, this dissertation tackles two problems. First, we propose a stochastic optimization method. We evaluate the method in comparison to standard optimization methods and apply it to design problems. Second, we develop a method for subgroup selection.

This dissertation is organized as follows. In Chapter 1, we provide some background on optimal design and optimization methods. In Chapter 2, we propose and evaluate a method for stochastic optimization. In Chapter 3, we adapt and apply the method developed in Chapter 2 to experimental design problems, such as dose-finding studies and multifactor experiments. In Chapter 4, we develop a method for subgroup selection under a Bayesian-decision theoretic framework, aided by the method proposed in Chapter 2. Finally, in Chapter 5, we offer some conclusions and discuss potential future directions based on our findings.

Chapter 2
BACKGROUND

We start our work with a review of the general model formulation and criteria for optimal design in specific applications in Section 2.1. Next, we briefly review some of the commonly used optimization methods in Section 2.2. In Section 2.3 we provide a discussion.

2.1 Optimization in Experimental Design Problems

Experiments are studies wherein the values of the predictor variables can be controlled by the investigators. Experimental designs have been used in various areas, including chemistry, engineering, physics, industrial research, and to some extent, early phases of clinical trials. The goal of most experiments is to obtain information and assess the relationship between the predictors and the response of interest. In many experiments, however, the amount of information that can be collected is limited by the sample size restrictions, which arises due to constraints in time, money, etc. Hence, given the restrictions on the number of samples that can be obtained, it is of great importance that the samples are selected carefully so that the experiment is "optimal", yielding the maximum amount of information of interest.

In this section, we discuss some of the approaches for designing these experiments in both general experimental designs and designs of clinical trials, more specifically, dose finding studies.

2.1.1 General Experiments

In most experiments, the relationship between the predictors and the response is usually estimated by some model. One class of parametric models that have been used widely in modeling these relationships is the generalized linear models (GLM). Here we discuss briefly

the properties of GLM.

2.1.1.1 Generalized Linear Models

Generalized linear models (GLMs) have been widely used in applied statistics on a wide range of applications to address questions about association or prediction. There are three components for GLMs.

(i) The response y is identically independently distributed with a distribution belonging to the exponential family. The probability mass (density) function of y is given by

$$f(y, \theta, \phi) = \exp\left(\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\right) \tag{2.1}$$

where $a(\cdot), b(\cdot)$, and $c(\cdot)$ are known functions, θ is the canonical parameter, and ϕ is a dispersion parameter.

(ii) The linear predictor is of the form

$$\eta(\mathbf{x}) = \mathbf{f}^T(\mathbf{x})\boldsymbol{\beta} \tag{2.2}$$

where $\mathbf{f}^T(\mathbf{x})$ is a known p-component function of $\mathbf{x}$, and $\boldsymbol{\beta}$ is an unknown parameter vector of length p.

(iii) A link function $g(\cdot)$ relates η to the mean response $\mu(\mathbf{x})$, so that $\eta(\mathbf{x}) = g(\mu(\mathbf{x}))$. The function $g(\cdot)$ is monotone and differentiable.

Given that we have a class of models that describe the relationship between predictors and outcome, we can now define a measure for the 'information' of interest that can be obtained from the experiment. These 'measures' are commonly referred to as 'optimality criteria'. Here we discuss some common optimality criteria that have been used in designing an experiment.

2.1.1.2 Frequentist Optimality Criteria

In order to select an optimal or efficient design, one needs to define a criterion. We list below some of the commonly used optimality criteria, usually referred to as 'alphabetical' optimality criteria, defined by Kiefer (1959) [39]. For a given model with parameter vector $\boldsymbol{\beta}$ and design point d, denote the Fisher's information matrix for that model by $\mathbf{I}(\boldsymbol{\beta}; d)$. Then, for a given model, with the set of all allowable designs $\mathcal{D}$, the optimal designs based on the various optimality criteria are given below.

(i) A-optimality criterion [39]. An efficient design based on the A-optimality criterion is the design d_A such that

$$d_A = \underset{d\in\mathcal{D}}{\operatorname{argmin}}[\operatorname{trace}(\mathbf{I}(\boldsymbol{\beta}; d)^{-1})].$$

In other words, an A-optimal design is the design which minimizes the trace of the inverse of the information matrix. Kiefer (1959) showed that an A-optimal design minimizes the average variance of the best linear estimators of $\boldsymbol{\beta}$ [39].

(ii) c-optimality criterion [22]. The c-optimality criterion was first defined by Elfving (1952) [22], where the inference of interest is for a linear combination of $\boldsymbol{\beta}$, that is, we are interested in the estimation of $\mathbf{c}^T\boldsymbol{\beta}$. An efficient design based on the c-optimality criterion is the design d_c such that

$$d_c = \underset{d\in\mathcal{D}}{\operatorname{argmin}}[\mathbf{c}^T(\mathbf{I}(\boldsymbol{\beta}; d)^{-1})\mathbf{c}].$$

(iii) D-optimality criterion [39]. An efficient design based on the D-optimality criterion is the design d_D such that

$$d_D = \underset{d\in\mathcal{D}}{\operatorname{argmax}}[\det(\mathbf{I}(\boldsymbol{\beta}; d))].$$

In other words, a D-optimal design is the design which maximizes the determinant of the information matrix. Kiefer (1959) showed that a D-optimal design minimizes the generalized variance of the best linear estimators of $\boldsymbol{\beta}$ [39]. Kiefer also showed that the D-optimality criterion possesses an invariant property that is not possessed by other optimality criteria [39]. Suppose that $\boldsymbol{\beta}'$ is related to $\boldsymbol{\beta}$ by a non-singular linear transformation. Then, a D-optimal design d^* for $\boldsymbol{\beta}$ is also D-optimal for $\boldsymbol{\beta}'$ [39].

Often, prior to data collection, that is, at the start of the experiments, decisions have to be made as to which design points should be sampled/tested. In most experiments, existing data or knowledge is available prior to experimentation. In fact, existing data or knowledge often motivates the experimentation. Hence, a Bayesian approach is natural in this context, and can play an important role. Chaloner and Verdinelli (1995) discussed the Bayesian approaches to optimal designs and reviewed the Bayesian analogs of the 'alphabetical' optimality criteria discussed previously [13].

2.1.1.3 Bayesian Optimality Criteria

Lindley (1972) proposed a decision-theoretic approach to experimental design [43]. Under this approach, a Bayesian solution to the experimental design problem is a design that maximizes a given utility function, $U(\boldsymbol{\beta}, d)$, averaged over the prior distributions, $\pi(\boldsymbol{\beta})$. Subsequently, Chaloner and Verdinelli (1995) discussed utility functions that provide the Bayesian analog to the 'alphabetical' optimality criteria. Specifically:

(i) Bayesian A-optimality criterion [13]. The Bayesian A-optimality criterion is obtained when the utility function is given by $U(\boldsymbol{\beta}, d) = -(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$ [13]. Chaloner and Verdinelli (1995) showed that maximizing the expected utility is equivalent to maximizing

$$\phi(d) = -\int \text{trace}(\mathbf{I}(\boldsymbol{\beta}; d))d\pi(\boldsymbol{\beta}).$$

(ii) Bayesian c-optimality criterion [13]. The Bayesian c-optimality criterion is obtained

when the utility function is given by $U(\boldsymbol{\beta}, d) = -(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \mathbf{c}(\boldsymbol{\beta})\mathbf{c}(\boldsymbol{\beta})^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$ [13]. Chaloner and Verdinelli (1995) showed that maximizing the expected utility is equivalent to maximizing

$$\phi(d) = -\int \mathbf{c}(\boldsymbol{\beta})^T(\mathbf{I}(\boldsymbol{\beta}; d))^{-1}\mathbf{c}(\boldsymbol{\beta})d\pi(\boldsymbol{\beta}).$$

(iii) Bayesian D-optimality criterion [12] [13]. The Bayesian D-optimality criterion is obtained when the utility function is the Shannon information, as suggested by Lindley (1956), DeGroot(1962, 1986), Stone (1959), and Bernardo (1979) [42] [19] [18] [63] [4] [13]. Chaloner and Verdinelli (1995) showed that maximizing the expected utility is equivalent to maximizing

$$\phi(d) = \int \log(|\mathbf{I}(\boldsymbol{\beta}; d)|)d\pi(\boldsymbol{\beta}).$$

2.1.1.4 Locally Optimal Designs

As seen previously, most commonly used optimality criteria, including Frequentist and Bayesian criteria, maximize some function of the information matrix. The information matrix in generalized linear models depends on the model parameters and the predictors. Thus, the optimal design depends on the value of the parameters in the model. In other words, given different parameter values, a different optimal design would be obtained. However, in all experiments and studies, the true value of the parameters ($\boldsymbol{\beta}$) are unknown. In fact, most experiments and studies are conducted to obtain estimates of such parameters (or equivalently to obtain estimates of the predictor-response relationship). Hence, we have a problem where on the one hand we are trying to obtain an efficient estimate of the parameters for which we need an optimal design, and on the other hand, we are trying to obtain an optimal design for which we need the value of the parameters. A common approach to this problem is to obtain a design that is optimal for the current 'best guess' of the parameter values.

Chernoff [15] introduced the term 'locally optimal' design, which refers to a design that is optimal for the 'best guess' of the parameters. Naturally, the 'best guess' of the parameters before the start of the experiment is represented by the prior distribution on the parameters under the Bayesian framework. Next, we review some existing approaches to obtaining a locally optimal design.

2.1.1.5 Locally Optimal Quantal Response Design with Binary Outcomes

Quantal response experiments are described by a function that relates the level of a predictor to the probability of the response. An example of a quantal response experiment is a dose-response study with a binary outcome in which case we may represent the probability of response as

$$P[Y = 1|\mu, \sigma, X] = f\left(\frac{X - \mu}{\sigma}\right) \tag{2.3}$$

where Y is the outcome variable, X is the experimental factor (design variable), μ and σ are the parameters of interest, and f is a given function. For example, for a simple logistic regression model using the usual parameterization, we have $P[Y = 1|\beta_0, \beta_1, X] = expit(\beta_0 + \beta_1 X)$, where *expit* is the inverse of the logistic function, and β_0 and β_1 are parameters of interest. The $\mu-\sigma$ parameterization shown in (2.3) is a one-to-one transformation of the usual parameterization with f being the expit function and $\beta_0 = -\frac{\mu}{\sigma}$ and $\beta_1 = \frac{1}{\sigma}$. In such studies, investigators are interested in characterizing the quantal response $P[Y = 1|\mu, \sigma, X = x]$ as a function of x. In particular, often, quantiles x_p for some given p and such that $P[Y = 1|\mu, \sigma, X = x_p] = p$ are the focus of the study.

There are many approaches to optimal designs for quantal response settings. We review some of such approaches, but focusing on dose-response studies.

Abdelbasit and Plackett [1] found that using a D-optimality criterion, for a two-point symmetric design about μ using a logistic model, the covariates x that give a probability of response of 0.824 (and similarly 0.176 due to symmetry) would give a locally D-optimal design. However, due to the parameters being unknown, they proposed a sequential design.

At each step, the parameter of interest is estimated by the maximum likelihood estimator (MLE), and the next design point is chosen as the point that maximizes the D-optimality criterion based on the current estimate of the parameter.

Neyer [51] proposed a design formulated in three parts. The first part of the design is used to focus on the region of interest. The second part is designed to efficiently estimate the parameter of interest. In the third part of the design, the parameter estimates are refined continuously once unique estimates have been found. In computing the MLE, Neyer also proposed a way to handle cases where we have unusual estimates from the MLE, such as when the estimated σ (in the $\mu - \sigma$ parameterization) from the MLE is 0. This is done by restricting the parameter values at each step and having "guesses" of the parameters prior to the experiment as the estimate when the estimated σ is 0.

Robbins and Monro [55] proposed nonparametric designs that focused on estimating the quantile x_p associated with the probability of response p. The proposed design by Robbins and Monro assumed that the function for the response is unknown, and hence used a stochastic approximation to obtain an estimate of the quantile x_p. Wu (1985) proposed an improved version of the Robbins-Monro method by using all current data to estimate logit-MLE using a logistic regression model, and choosing the next design point based on the fitted logistic model [70].

2.1.1.6 Locally Optimal Multifactor Designs

Chaudhuri and Mykland [14] proposed a sequential approach to obtain a locally optimal multifactor design. They showed that given certain conditions that their design is asymptotically D-optimal. The algorithm for the method by Chaudhuri and Mykland is given in Appendix A.2.

Dror and Steinberg [21] proposed a method for obtaining a D-optimal multifactor design. Their method relies on approximations to the Bayesian D-optimality criterion to improve computational efficiency and time by using a discretization of the prior distribution $\pi(\boldsymbol{\beta})$. Specifically, let $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_N$ be the samples from the prior distribution for large N (e.g.

$N = 10,000$). Then, the likelihood $L(\boldsymbol{\beta}_u)$ for each of those vectors can be quickly computed at any stage of the experiment. Define the weights r_u based on these likelihoods by $r_u = L(\boldsymbol{\beta}_u)/\sum_{v=1}^{N} L(\boldsymbol{\beta}_v)$. The posterior for $\boldsymbol{\beta}$ is given by $\sum r_u \boldsymbol{\beta}_u$. Let $\tilde{\boldsymbol{\beta}}$ denote the posterior median of $\boldsymbol{\beta}$. Then, the approximate criteria are based on:

$$\phi_1(d) = \sum_{u=1}^{N} r_u \log(|\mathbf{I}(\boldsymbol{\beta}; d)|) \tag{2.4}$$

$$\phi_2(\tilde{\boldsymbol{\beta}}; d) = \log(|\mathbf{I}(\tilde{\boldsymbol{\beta}}; d)|) \tag{2.5}$$

The authors proposed a notion called augmentation horizon. The augmentation horizon (m) is determined at the start of the experiment as the number of observations needed for a highly efficient design with high D-efficiency at the prior median of the parameters ($\boldsymbol{\beta}^{(0)}$).

Furthermore, the D-efficiency of a design d is defined as follows:

$$\text{D-efficiency} = (\phi_2(\boldsymbol{\beta}_T; d)/\phi_2(\boldsymbol{\beta}_T; d^*))^{1/p} \tag{2.6}$$

where $\phi_2(\boldsymbol{\beta}_T; d) = \log(|\mathbf{I}(\boldsymbol{\beta}_T; d)|)$ as defined above, $\boldsymbol{\beta}_T$ is the true parameter vector, p is the number of parameters, and d^* is the n-sample D-optimal design that would have been obtained using Fedorov's exchange algorithm if the true values of the parameters ($\boldsymbol{\beta}_T$) were known in the beginning of the experiment.

The algorithm for obtaining the D-optimal design, starting with the null design, is provided in Appendix A.2. with more details.

2.1.2 Designs of Phase I Clinical Trials

An important application of designs for quantal response is in dose-finding studies. These are usually conducted in phase I of a clinical trial, where the dose-response relationship between the dose of the drug with the potential toxicity that might arise is explored.

The common assumption made in a dose-finding study is that the efficacy of the drug rises monotonically with the dose, but so does the probability of a toxic event. Thus, in

a dose-finding study, the goal is to estimate the dose for which the most efficacy can be obtained without having a too high probability of a toxic event, which is usually called the maximum tolerated dose (MTD). In other words, the MTD is the maximum dose under which the expected probability of toxicity is not unacceptable.

Dose-finding studies are typically small in size. O'Quigley, Pepe, and Fisher [52] also noted that in many of the oncology trials, the patients are of high risk of short-term fatality. These factors underline the importance of designing an optimal dose-finding study that would efficiently estimate the MTD. However, in addition to efficiency or optimality, many other considerations, such as ethical concerns, need to be considered when designing such trials, where 'experiments' are done to human subjects. (These are the 'special' considerations that in general need not be made for general experiments, for example, in a chemistry experiment aimed to find the limit of detection of a particular compound). In dose-finding studies, generally it is undesirable and unethical to expose patients to extremely high doses that would cause extreme toxicity, even though it might yield a design that is 'optimal' in estimating the MTD.

We review in more detail the approaches commonly used to find the MTD in dose-finding studies, including the continual reassessment method (CRM) [52], the escalation with overdose control (EWOC) [3], and the decision-theoretic approach [68].

2.1.2.1 Continual Reassessment Method (CRM)

O'Quigley, Pepe, and Fisher [52] developed the method for phase I oncology clinical trials as an alternative when conventional approaches would not work due to the slow dose-escalation and the high short-term risk of fatality for the patients. The CRM, on the other hand, continually updates the estimates as more information is collected, and treats the next patient at the MTD based on the currently available evidence.

The first assumption of the CRM is that the probability of a toxic response increases monotonically with the treatment dose [52]. The CRM assumes an a-priori dose-response relationship, which is updated over time as observations of toxicities become available. Suppose

that the dose levels x_i $(i = 1, 2, ..., k)$ are chosen for experimentation from the dose range x. Let Y_j be a binary random variable indicating the event of severe toxicity for the j^{th} patient. Suppose that θ is the probability of a dose-limiting toxicity (DLT) of interest, and there is a dose x^* (not necessarily one of $x_1, ..., x_k$) for which the probability of toxicity is θ. Denote the dose–response function by $\psi(x_i, a), a \in \mathcal{A}$. The only assumption for the function ψ is that it is monotonic in both x and a and that, for some a_0, $\psi(x^*, a_0) = \theta$. O'Quigley et al proposed a model that is flexible enough such that for any dose $\tilde{x}$ and response probability $\tilde{\theta}$, there exists a parameter value $\tilde{a}$ such that $\psi(\tilde{x}, \tilde{a}) = \tilde{\theta}$.

Let $\Omega_j = \{y_1, ..., y_{j-1}\}$ be the data collected up until patient $j - 1$ and $f(a, \Omega_j)$ be a nonnegative function summarizing the information about the parameter a. The function $f(a, \Omega_j)$ is then the prior distribution for a before we obtain the information for patient j. For $\mathcal{A} = (0, \infty)$, we have

$$\int_0^\infty f(a, \Omega_j) da \quad = \quad 1$$

for $j = 1, ..., n$. We can also estimate the probability of a toxic response at dose level i for patient j given the available information up until patient $j - 1$, denoted by θ_{ij}.

$$\theta_{ij} \quad = \quad \int_0^\infty \psi(x_i, a) f(a, \Omega_j) da$$

for $i = 1, ..., k$. O'Quigley et al proposed that, compared to calculating the θ_{ij} above every time, it is preferable to work with the expected value of a over $\mathcal{A}$. The alternative estimate of the toxic response probability is then $\theta'_{ij} = \psi(x_i, \mu_a(j))$, where $\mu_a(j) = \int_0^\infty a f(a, \Omega_j) da$. The expression for the likelihood of toxic events, having chosen the dose level for patient j (denoted by $x(j)$), is assumed to be the usual Bernoulli likelihood

$$\phi(x(j), y_j, a) \quad = \quad \psi(x(j), a)^{y_j} (1 - \psi(x(j), a))^{1 - y_j}.$$

Using Bayes' theorem, given the response for the j^{th} patient, the information on a can

be updated by obtaining $f(a, \Omega_{j+1})$ from $f(a, \Omega_j)$.

$$f(a, \Omega_{j+1}) = \frac{\phi(x(j), y_j, a) f(a, \Omega_j)}{\int_0^\infty \phi(x(j), y_j, u) f(u, \Omega_j) du}.$$

The algorithm for CRM is given in Appendix A.3.

2.1.2.2 Escalation with Overdose Control (EWOC)

The escalation with overdose control (EWOC) is a method proposed by Babb, Rogatko, and Zacks [3] that aims to estimate the maximum tolerated dose (MTD) while controlling the proportion of people subject to overdose. Suppose that α is the maximum allowable proportion of people subject to overdose. The idea of EWOC is that the dose level is selected so that the predicted proportion of people subject to overdose is equal to the pre-specified α by utilizing the posterior cumulative distribution function (CDF) of the MTD. Formally, for the k^{th} dose assignment, the posterior CDF of the MTD is given by

$$\pi_k(\gamma) = P(\gamma \geq \text{MTD}|\text{data}).$$

Based on this, the EWOC would assign to the k^{th} patient the dose x_k for which $\pi_k(x_k) = \alpha$. In other words, the dose is selected such that the probability that it exceeds the MTD is α. Let X_{min} and X_{max} denote the minimum and maximum dose levels for the study. The dose-response relationship is modeled by

$$P(\text{Toxicity}|\text{Dose} = \text{x}) = F(\beta_0 + \beta_1 x),$$

where F is a specified function. For example, F is the inverse logit (expit) function for a logistic regression. Let x_i denote the dose administered to the i^{th} subject, y_i denote the binary indicator of toxicity for the i^{th} subject, and Ω_k denote the data collected until the $(k-1)^{th}$ subject. Let $h(\beta_0, \beta_1)$ be the prior on the parameters β_0 and β_1. Using Bayes

theorem, the posterior distribution of the parameters given the data Ω_k is given by

$$P(\beta_0, \beta_1|\Omega_k) = \tau^{-1} Lh(\beta_0, \beta_1),$$

where

$L = \Pi_{i=1}^{k-1} F(\beta_0 + \beta_1 x_i)^{y_i}(1 - F(\beta_0 + \beta_1 x_i))^{1-y_i}$ is the likelihood and

$\tau = \int\int Lh(\beta_0, \beta_1) d\beta_0 d\beta_1$ is the normalizing constant.

The marginal CDF of the MTD can be obtained by first transforming the parameters through the transformation $T(\beta_0, \beta_1) = (\rho_0, \gamma)$. The inverse transformation is given by

$$T^{-1}(\rho_0, \gamma) = (f_1(\rho_0, \gamma), f_2(\rho_0, \gamma))$$

where f_1 and f_2 are defined by

$$f_1(\rho_0, \gamma) = \frac{\gamma F^{-1}(\rho_0) - X_{min} F^{-1}(\theta)}{\gamma - X_{min}} \quad \text{and} \quad f_2(\rho_0, \gamma) = \frac{F^{-1}(\theta) - F^{-1}(\rho_0)}{\gamma - X_{min}}.$$

The posterior distribution of the transformed parameters (ρ_0, γ) given the data Ω_k can then be written as

$$P(\rho_0, \gamma|\Omega_k) = \tau^{-1} Lg(\rho_0, \gamma),$$

where $g(\rho_0, \gamma)$ is the prior induced for (ρ_0, γ). The marginal posterior CDF of the MTD given the data Ω_k can then be obtained as

$$\pi_k(z) = \int_{X_{min}}^{z} \int P(\rho_0, \gamma|\Omega_k) d\rho_0 d\gamma.$$

The algorithm for EWOC is given in Appendix A.3.

2.1.2.3 Decision-Theoretic Method

A decision-theoretic approach for dose-finding studies was introduced by Whitehead and Bruner (1995) [68]. The dose to be administered to the next patient was considered as the

decision that is to be made upon observing the data so far. Whitehead and Bruner specified that the 'ingredients' for the decision-theoretic method are: (1) the model for the data F, which then specifies the joint density function $f(\mathbf{x};\beta)$ for which the parameter β is unknown, (2) a prior distribution for β, $h_0(\beta)$, (3) a set of possible actions $A_1, A_2, ...$ which are specified and under the investigator's control, and (4) a gain function, $G_i(\beta)$, which represents the gain made if action A_i was taken.

The posterior subjective distribution of β can then be obtained based on the data so far ($\mathbf{x}$) using Bayes theorem as follows

$$h(\beta|\mathbf{x}) = \frac{f(\mathbf{x};\beta)h_0(\beta)}{\int f(\mathbf{x};\beta)h_0(\beta)d\beta},$$

and the action A is taken to maximize the posterior expected gain G defined by

$$G = \int G(\beta)h(\beta|\mathbf{x})d\beta.$$

Whitehead and Bruner claimed that the decision theoretic method is the general method encompassing the CRM. In other words, CRM is a special case of the decision theoretic method [68]. Whitehead and Bruner (1995) showed that given a specific model F, prior h_0, and gain function $G(\theta)$, the decision theoretic method yields the CRM utilized in O'Quigley, Pepe, and Fisher (1990) [68] [52].

As discussed in this section, a optimal (or locally optimal) design maximizes a specific optimality criterion. The next section reviews some of the existing methods for optimization.

2.2 Optimization Methods

In this section we review some of the commonly used optimization methods that have been developed to date. Formally, let the objective function be $f : \mathbb{R}^n \to \mathbb{R}$, that is, with domain $\mathbb{R}^n$ and codomain $\mathbb{R}$. The goal of such methods is to find $x \in \mathbb{R}^n$ that maximizes (or minimizes) f. The optimization methods are broadly classified as deterministic or stochastic.

Some of these methods are implemented in R with the optim() function. We will use the methods available in R to compare with our proposed method in the next Chapter 3.

2.2.1 Deterministic Optimization Methods

Deterministic optimization is a branch in mathematics which encompasses classical methods for optimization. Deterministic optimization methods, in general, rely on the gradient or the Hessian of the objective function. In this section we briefly review some of the existing methods for deterministic optimization. For a select subset of the methods, details of the corresponding algorithms are provided in Appendix A.

2.2.1.1 Conjugate Gradient Method

The conjugate gradient (CG) method is a method that performs minimization along conjugate directions [69]. Specifically, we say that a set of non-zero vectors $\{p_1, ..., p_l\}$ is *conjugate* for a symmetric positive-definite vector A if $p_i^T A p_j = 0$, $\forall i \neq j$. The idea underlying CG is that minimization of the target function f can be done by successively minimizing f along the individual directions in the conjugate set.

2.2.1.2 BFGS Method

The BFGS method, named after those who proposed the method (Broyden, Fletcher, Goldfarb, and Shanno), is a quasi-Newton method for minimizing a given function f [69]. Define a quadratic model associated with the given function f at the current iterate x_k as

$$m_k(p) = f_k + \nabla f_k^T p + \frac{1}{2} p^T B_k p,$$

where B_k is a $n \times n$ symmetric positive definite matrix that is updated at every iteration. The minimizer of this convex quadratic equation is

$$p_k = -B_k^{-1} \nabla f_k.$$

The resulting p_k is then used as the search direction for the next iterate given by $x_{k+1} = x_k + \alpha_k p_k$, where α_k is the step length. The step length α_k is chosen such that it satisfies the Wolfe condition, which is informally "the condition that there is sufficient decrease and curvature". The Wolfe condition, mathematically, is such that

$$\begin{aligned} f(x_k + \alpha_k p_k) &\leq f(x_k) + c_1 \alpha_k \nabla f_k^T p_k \\ \nabla f(x_k + \alpha_k p_k)^T &\geq c_2 \nabla f_k^T p_k \end{aligned}$$

where $0 < c_1 < c_2 < 1$.

2.2.1.3 Nelder-Mead Method

The Nelder-Mead method is a derivative-free optimization method that uses a simplex in $\mathbb{R}^n$ at each iteration [69]. At any iteration, the $n+1$ points in $\mathbb{R}^n$ that are kept track of, denoted by $\{x_1, ..., x_{n+1}\}$, are the vertices of the current simplex. The vertices $\{x_1, ..., x_{n+1}\}$ are ordered such that $f(x_1) \leq ... \leq f(x_{n+1})$. The centroid of the best n points at that iteration is given by $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$. The points along the line joining $\bar{x}$ and the worst vertex x_{n+1} is given by the function $\tilde{x}(t) = \bar{x} + t(x_{n+1} - \bar{x})$.

The Nelder-Mead algorithm is an iterative algorithm where the iterations continue until the stopping criterion is met. Nelder and Mead suggested stopping when the 'standard error' of the responses for the points in the simplex ($\sqrt{\Sigma_{i=1}^{n}(f(x_i) - \frac{1}{n}\Sigma_{i=1}^{n} f(x_i))^2/n}$) is less than some pre-set value [50].

2.2.2 Stochastic Optimization Methods

Deterministic optimization methods are usually sensitive to the initial values and may not converge depending on the initial value and the shape of the objective function. Furthermore, the solution from the deterministic optimization methods may converge to a local extremum but not a global extremum. Stochastic optimization methods attempt to address some of these issues. Stochastic optimization methods refer to optimization methods that utilize

iterates that are 'random'. Next, we review some of the existing stochastic optimization methods. For a select subset of the methods, details of the corresponding algorithms are provided in Appendix A.

2.2.2.1 Simulated Annealing Method

Simulated annealing is a local search algorithm used for optimization problems that is capable of escaping from a local optimum by allowing a chance to select inferior points at a given iteration [32]. The probability of selecting an inferior point depends on the "temperature" parameter (t_k) at each iteration (k). One condition for the temperature parameter is that $t_k>0\ \forall k$ and $\lim_{k\to\infty} t_k = 0$.

2.2.2.2 Genetic Algorithms

The genetic algorithm (GA), first introduced by Holland (1975) [34], is a method for solving large optimization problems. The algorithm draws an analogy with the genetic structure of a chromosome and hence the name genetic algorithm. The goal of the genetic algorithm is to maximize a given function $f(\mathbf{x})$ of the vector $\mathbf{x} = (x_1, x_2, ..., x_p)$ where each x_i is binary. The x_i's are 'stacked' together to form a binary string (bitstring), resembling a chromosome. The algorithm then proceeds to find the optimum via a process analogous to procreation, mutation, and natural selection.

Define g to be the 'fitness' of the bitstring, which is proportional to the probability of a bitstring to live long enough to reproduce. The function g is pre-selected and is a monotone increasing function of f. The algorithm starts with randomly generating an even number n of bitstrings of length p. The n 'parents' for the next generation are then chosen with replacement from the existing n strings, where the probability of choosing string j is proportional to its fitness (g_j).

The 'parents' are then considered in pairs, and for each pair, a crossover is performed with a pre-specified probability p_c. If crossover is performed, a random integer k between 1

and $p-1$ is selected and the last $(p-k)$ elements of the strings are exchanged between the two parents.

After crossover, mutation is performed for each element, with pre-specified probability p_m. If mutation is performed, an element is switched from 0 to 1 or vice versa, independently for each element. The algorithm then is allowed to repeat for a specified number of generations. At the end of the algorithm, the string with the highest f is selected as the optimum.

2.3 Discussion

In this chapter, we have reviewed the idea of optimal experiments, some of the existing methods for obtaining optimal experiments, and the existing optimization methods.

The dose-finding experiment can be viewed as an optimal design problem as well, for which our objective function is to obtain an efficient estimate of MTD given the small sample size, penalizing for overdose. The CRM and EWOC procedures have been shown to converge to the unknown MTD [56].

Many of the optimization methods that have been developed, including traditional optimization methods (conjugate gradient, BFGS) and stochastic methods (local search, Nelder-Mead) suffer from the problem that the search may be stuck in a local optimum. Some stochastic methods have been developed, including the simulated annealing and genetic algorithms, that would enable 'escape' from the local optimum. However, the simulated annealing method does not guarantee an 'escape' from the local optimum and the probability of escape highly depends on the 'temperature' parameter. The method of genetic algorithms is naturally used for optimizing a combination of binary predictors. Extensions to genetic algorithm for real-valued predictors have been attempted, which involves converting the value to a base-2 binary number [28] [31]. However, converting a real-valued predictor to a binary number may require a long binary string, and the precision is highly sensitive to the length of the string. Furthermore, the chance of 'escaping' from a local mode is highly dependent on the pre-specified parameters (p_c and p_m). In general, we would expect that analogous to living organisms, the chance of escape from a local mode is quite low, just as it is highly

unlikely that a mutation and recombination would yield a much superior organism.

In this dissertation we address some of the above limitations with the standard optimization methods by proposing a new method for stochastic optimization. We develop the method in the next Chapter 3 and apply it to some design problems in Chapter 4.

Chapter 3

FORWARD SLICE

3.1 Introduction

In this chapter we propose a method for stochastic optimization which we will later apply to design problems in Chapter 4. At its core, our method is based on the procedure that Neal (2003) called the 'slice sampling' procedure [49] [17] [48], originally developed as a Markov chain sampling procedure to draw samples from a target distribution. The slice sampling method relies on an auxiliary variable which defines a level at which we slice the target density to obtain regions from which we draw samples of the target distribution. Similar to Neal's method, our procedure uses an auxiliary variable for stochastic optimization that also defines slices, but of an objective function to be maximized (or minimized). Moreover, unlike with Neal's method, the auxiliary variable in our approach is not sampled and takes on non-decreasing values in the sequential iterations of the procedure so that, for a given pre–specified tolerance, at the end of the procedure we attain the maxima and the argument of the maxima (or close values given the selected tolerance level).

This chapter is organized as follows. In Section 3.2 we review the method developed by Radford Neal [49]. Next, in Section 3.3 we build on the method and describe our proposed *forward slice procedure for stochastic optimization*. In Section 3.4 we show that the method selects the global maxima of any given objective function that satisfies some conditions. In Section 3.5, we use simulation studies to investigate the performance of our method and constrast it with those from the alternative methods reviewed in Chapter 2. Finally, in Section 3.6 we provide a discussion.

3.2 Posterior Sampling via Slice: a review

The slice sampling was a method originally developed by Radford Neal [49] as an alternative to some standard MCMC procedures such as the Gibbs sampling and Metropolis–Hastings algorithm for drawing samples from a target posterior distribution. Specifically, the slice sampling allows one to draw samples $x \in \mathbb{R}^n$ from the target $f(x)$. The method relies on obtaining these samples by drawing uniformly from the region that lies under $f(x)$. Formally, consider an auxiliary variable y and define the joint distribution over x and y that is uniform over the region $U = \{(x, y) : 0<y<f(x)\}$. We can obtain samples of x by sampling over the joint distribution (x, y) and ignoring the variable y. Since obtaining uniform independent points from U is non-trivial, Neal proposed using a MCMC procedure that will converge to the target uniform distribution. This is done by first sampling from the conditional distribution of y given the current x, that is, uniformly sample from the interval $(0, f(x))$. Next, sample from the conditional distribution of x given the current y, which is uniform over the region $S = \{x : f(x) \geq y\}$. Neal called this region S the horizontal 'slice'.

3.2.1 Univariate slice sampling

The algorithm for the univariate slice sampling for obtaining a new value x_1 given the current value x_0 is as follows:

1. Obtain a real value y drawn uniformly from the interval $(0, f(x_0))$. This defines the horizontal 'slice' $S = \{x : f(x) \geq y\}$, which is depicted in Figure 3.1 by the full horizontal red lines on the x-axis.

2. Find an interval $I = (L, R)$ around x_0 that contains at least a big part of the slice. This is done by first defining a step size w. First, we obtain an interval of length w around x_0, by obtaining a number u sampled uniformly from $(0, w)$. The interval of length w around x_0 is then the interval $(x_0 - u, x_0 + w - u)$. We then step out to the left and right of the initial interval with step size w until the ends of the interval are

not part of the slice S. The updated interval is $I = (L, R)$ shown in the Figure 3.2.

3. The new point x_1 is then sampled uniformly from the 'slice' part of the interval (i.e. $S \cap I$).

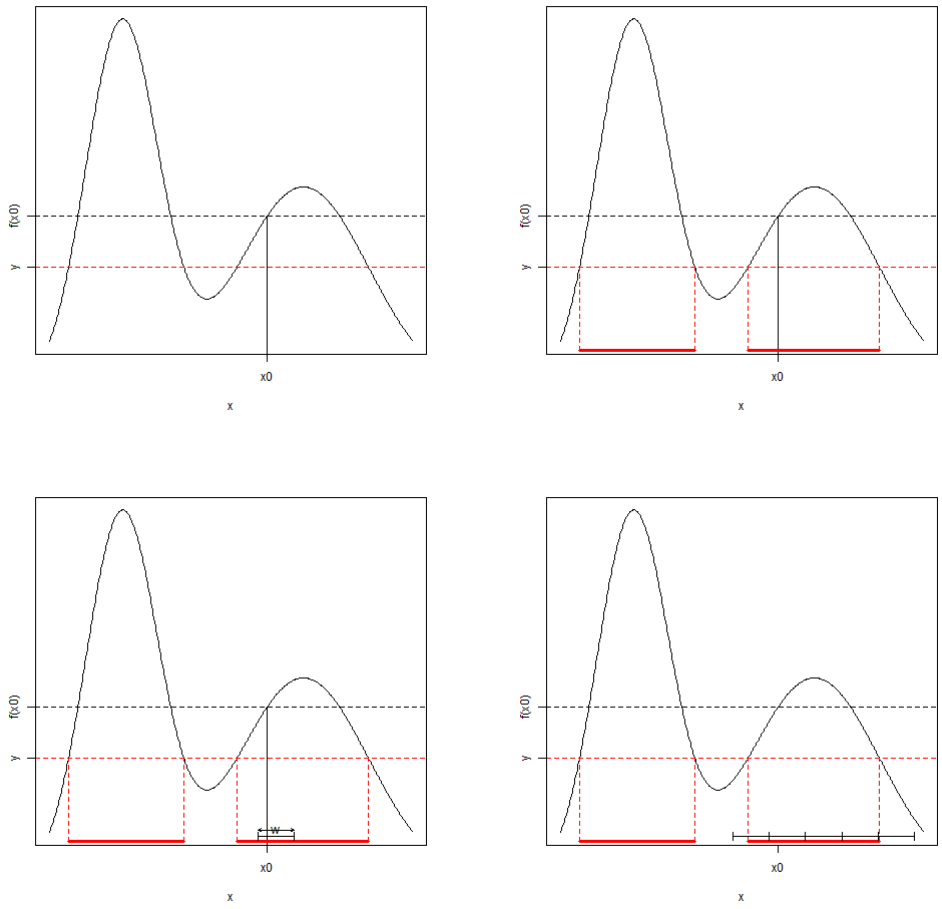


Figure 3.1: Illustration of slice sampling. The top left panel shows the y sampled from the interval $(0, f(x_0))$. The top right panel shows the horizontal 'slice', depicted by the full horizontal red lines. The bottom left panel shows the first interval of length w around x_0. The bottom right panel shows the stepping to the left and right until the ends are outside the slice.

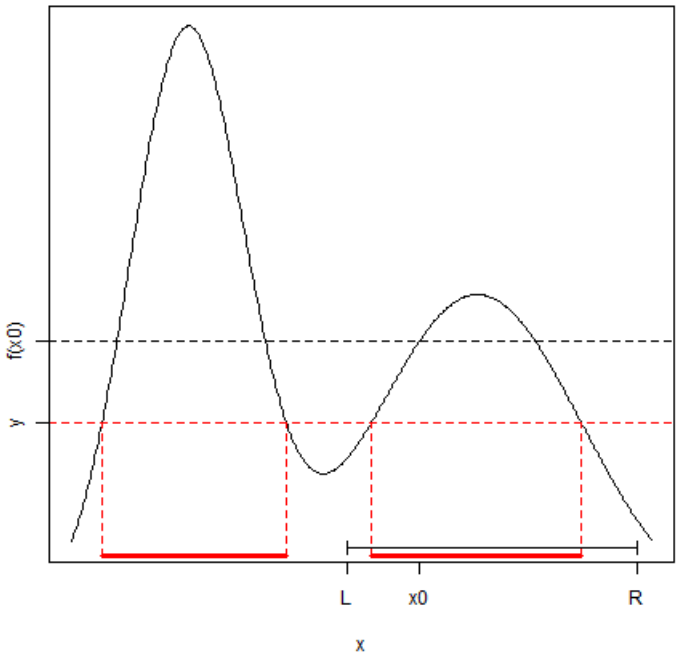


Figure 3.2: The obtained interval $I = (L, R)$

3.2.2 Multivariate Slice Sampling

Neal [49] proposed an extension of slice sampling for sampling directly from a multivariate distribution. The idea is extended to sampling from the multivariate distribution by replacing the interval $I = (L, R)$ by an axis-aligned hyperrectangle $H = \{x_i : L_i \leq x_i \leq R_i \text{ for all } i = 1, ..., n\}$, where L_i and R_i are the extent of the hyperrectangle along the axis of the variable x_i. The procedure of finding $x_1 = (x_{1,1}, ..., x_{1,n})$ from the current $x_0 = (x_{0,1}, ..., x_{0,n})$ is similar to the univariate procedure. Specifically:

1. Obtain a real valued y uniformly sampled from $(0, f(x_0))$, which defines the slice $S = \{x : f(x) \geq y\}$.

2. Find the hyperrectangle, $H = (L_1, R_1) \times \cdots \times (L_n, R_n)$ around x_0, which preferably

should contain a big part of the slice.

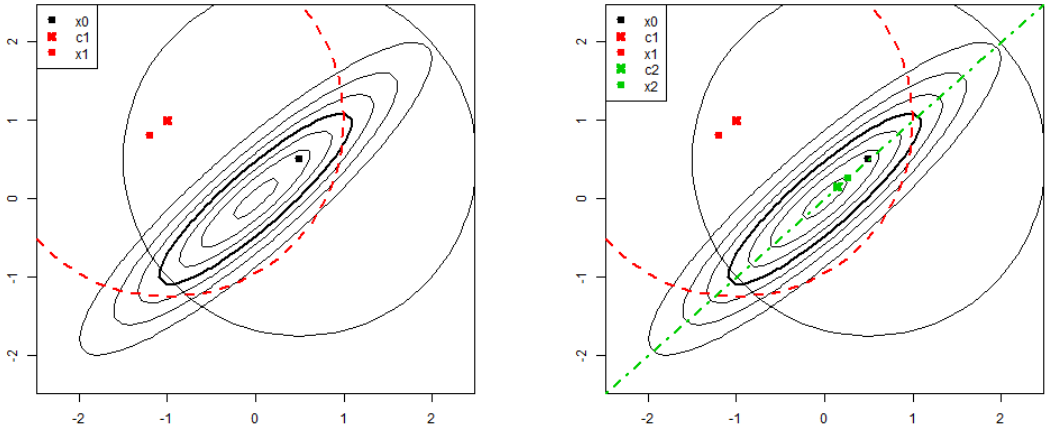
3. Obtain x_1 by uniformly sampling from the slice within the hyperrectangle (i.e. $S \cap H$)

3.2.3 Multivariate Slice Sampling with Shrinking Rank

When the parameters in the target distribution are highly correlated, many MCMC methods mix slowly. Neal [49] proposed an idea called the crumb framework for updating x_0 more efficiently, by leaving "crumbs" to guide where to sample and where not to sample. First, based on a given initial point x_0, a crumb is sampled from a known distribution. Given the crumb, the next state x_1 is proposed under a distribution that could have generated the crumb. If the proposed x_1 is in the slice, then we accept that as the new state, otherwise, new crumbs and proposals are drawn until the proposal is accepted (within the slice).

Thompson and Neal [66] developed an adaptive slice sampling method for multivariate distributions, called the shrinking-rank method, which extends upon the crumb framework to obtain new states more efficiently. They proposed that the crumbs are spherical Gaussian random variables that are centered at the current state, with an initial standard deviation (σ_c), which is a tuning parameter specified by the user. The distribution of the proposal x_1 based on the first crumb c_1 is a spherical Gaussian centered at c_1 with standard deviation σ_c. The idea of the shrinking rank method is that when a proposal x is outside of the slice (i.e. not accepted), the method adapts based on the information of the rejected proposal for drawing the next crumb and proposal. Let J be the nullspace of the subspace from which the next crumb is to be drawn, with unit-length and orthogonal columns. Let g^* be the projection of the gradient of the log density at the rejected proposal onto the nullspace of J. If g^* has a "large angle" with the gradient of the log density at the rejected proposal, then there will be no adaptation based on it, since the nullspace of J and the gradient is already nearly orthogonal. On the other hand, when the angle is small, the adaptation is done by appending the column of J by $g^*/\|g^*\|$. Thompson and Neal defined large angle as larger than 60^o, but claimed that the exact value is not crucial. An example of the shrinking-rank

method is given below in Figure 3.3.



(a) (b)

Figure 3.3: Adaptive slice sampling. The two-dimensional distribution is represented with the contour plots. The initial state is x_0. The bold ellipse represents the boundary of the slice. (a) The first crumb c_1 is drawn from the spherical Gaussian centered at x_0 (the solid circle). The first proposal x_1 is drawn from a spherical Gaussian centered at c_1 (the red dashed circle). Here, it shows that the first proposal is rejected since it is outside the slice. (b) The adaptation based on the first rejected slice results in the one-dimensional line (the green dashed-and-dotted line) as the space from which the next crumb to be sampled. Shown here are the next crumb c_2 and the next proposal x_2, which is accepted as it is in the slice.

3.3 Stochastic Optimization via Forward Slice: a proposal

Let $f : \mathcal{X} \to \mathbb{R}$ be a function with domain $\mathcal{X}$ and codomain $\mathbb{R}$. We note that we alternatively may refer to $\mathcal{X}$ as the design space. Our goal is to find values $x \in \mathcal{X}$ that maximize $f(\cdot)$. We build on the slice sampling idea to derive a procedure for stochastic optimization.

3.3.1 Forward Slice: Uni–dimensional case

In our method we obtain slices with increasing function level y at each iteration. We will call this the *forward slice* procedure. For a given $x_0 \in \mathcal{X}$, instead of sampling y from $(0, f(x_0))$,

we take $y = f(x_0)$. The forward slice procedure selects the next point from the slice region $S = \{x : f(x) \geq y\}$, more specifically, the slice region within the design space, which we denote as S^X (i.e. $S^X = S \cap \mathcal{X}$). The forward slice procedure guarantees that the next sample generated from the forward slice (x_1) follows $f(x_1) \geq f(x_0)$ because x_1 is sampled from S, a region where the function cannot take on decreasing values.

Note that the set S may not be completely known. Regardless, the sampling of x_1 from the set S can be done using a simple rejection sampling scheme. Specifically, first, we sample an initial proposal candidate (call it p_0) by uniformly sampling from the entire design space $\mathcal{X}$. If $f(p_0) \geq f(x_0)$, then we take $x_1 = p_0$. Otherwise, we reduce the space from which we sample a proposal based on p_0. Specifically, since $x_0 \in S$ (because $f(x_0) \geq f(x_0)$), then we can obtain the revised sampling space as $\mathcal{X}^{(1)} = \mathcal{X}^{(0)} \cap (-\infty, p_0]$ if $p_0 > x_0$ or $\mathcal{X}^{(1)} = \mathcal{X}^{(0)} \cap [p_0, \infty)$ if $p_0 < x_0$. Then, we sample a new proposal p_1 from $\mathcal{X}^{(1)}$. Similarly, if $f(p_1) \geq f(x_0)$, then we take $x_1 = p_1$. Otherwise, we reduce the sampling space again based on p_1 as $\mathcal{X}^{(2)} = \mathcal{X}^{(1)} \cap (-\infty, p_1]$ if $p_1 > x_0$ or $\mathcal{X}^{(2)} = \mathcal{X}^{(1)} \cap [p_1, \infty)$ if $p_1 < x_0$. We continue to sample proposal candidates and to reduce the sampling space until we sample p_j such that $f(p_j) \geq f(x_0)$, for which we take $x_1 = p_j$.

There is no guarantee that the next selected point x_1 maximizes the function $f(x)$. Thus, we define the forward slice as an iterative procedure, using the obtained x_1 from the previous iteration as the next input in the next iteration. We use the following notation: x_0 denotes the initial point input to the forward slice, while x_k and S_k denote, respectively, the point and the slice region after the k^{th} iteration where $S_k = \{x : f(x) \geq f(x_k)\}$, and S_k^X denotes the portion of the slice region S_k within the design space $\mathcal{X}$, i.e. $S_k^X = S_k \cap \mathcal{X}$. In other words, S_k^X is a set of points in $\mathcal{X}$ that is inside the slice region after the k^{th} iteration, i.e. $S_k^X = \{x : x \in \mathcal{X}, f(x) \geq f(x_k)\}$. Furthermore, we use the following notation for the proposal candidates for one slice iteration: p_0 denotes the initial proposal candidate, $\mathcal{X}^{(1)}$ denotes the space from which the next proposal (p_1, if needed) is drawn; and more generally p_j denotes the proposal candidate after j previously-rejected proposals, and $\mathcal{X}^{(j+1)}$ denotes the space from which the next proposal candidate (p_{j+1}) is to be drawn if p_j is also rejected.

By convention, $\mathcal{X}^{(0)} = \mathcal{X}$. Given a user-specified tolerance for convergence ϵ, the forward slice procedure is given as follows in Algorithm 1.

Algorithm 1 Univariate forward slice

Set $\epsilon>0$ as convergence tolerance, initial point x_0, and design space $\mathcal{X}$.
Set $k = 0$
repeat
    Define $S_k = \{x : f(x) \geq f(x_k)\}$
    Set $j = 0$. Sample x_{k+1} uniformly from $S_k^X = S_k \cap \mathcal{X}$ as follows:
    repeat
        Sample p_j uniformly from $\mathcal{X}^{(j)}$
        if $f(p_j) \geq f(x_k)$ **then**
            $x_{k+1} \leftarrow p_j$
            break
        else
            if $p_j > x_k$ **then**
                $\mathcal{X}^{(j+1)} \leftarrow \mathcal{X}^{(j)} \cap (-\infty, p_j]$
            else
                $\mathcal{X}^{(j+1)} \leftarrow \mathcal{X}^{(j)} \cap [p_j, \infty)$
            end if
            $j \leftarrow j + 1$
        end if
    until $f(p_j) \geq f(x_k)$
    Calculate $D = |(f(x_{k+1}) - f(x_k))/f(x_k)|$
    $k \leftarrow k + 1$
until $D \leq \epsilon$

3.3.2 Forward Slice: Multi–dimensional case

We can extend the forward slice method to the multi–dimensional case in analogy to the adaptive slice sampling by Thompson and Neal [66] that utilizes sampled "crumbs" from a known auxiliary distribution. Neal [49] proposed these "crumbs" as a trail of points in the space $\mathcal{X}$ that guide the drawing of the next sample given the initial point. Assume the domain space $\mathcal{X}$ has dimension p. Let J denote the matrix whose nullspace represents the subspace of $\mathcal{X}$ from which the next crumb is to be drawn. This space is informed by the rejected samples obtained from the crumbs. Similarly to the unidimensional case, for a given

$\mathbf{x}_0 \in \mathcal{X}$, we take $y = f(\mathbf{x}_0)$. Let S denote the horizontal slice as defined by Neal, which is the region defined as $S = \{\mathbf{x} : f(\mathbf{x}) \geq y\}$. The sampling from the slice S is done also by rejection sampling similarly to the unidimensional case. We start by obtaining a proposal sample $\mathbf{x}_p$ from the design space $\mathcal{X}$. If $f(\mathbf{x}_p) > f(\mathbf{x}_0)$, then we take $\mathbf{x}_1 = \mathbf{x}_p$. Otherwise, the proposal $\mathbf{x}_p$ is rejected, and the subspace from which the next proposal sample is taken from is guided by $\mathbf{x}_p$. For a rejected proposal $\mathbf{x}_p$, let g^* be the projection of $\nabla f(\mathbf{x}_p)$ onto the nullspace of J, where ∇f denotes the gradient of the function f. If g^* has a "large angle" with the gradient of the function at the rejected proposal, then the nullspace of J and the gradient is "nearly" orthogonal. On the other hand, when the angle is small, then we adapt the sampling subspace by appending the column of J by the normalized g^*. Thompson and Neal defined θ to be the threshold angle for "non-orthogonality", and they proposed 60^o as the threshold [66]. Then, for a rejected proposal $\mathbf{x}_p$, a new proposal sample $\mathbf{x}_p$ is obtained by sampling from $\mathcal{N}(J) \cap \mathcal{X}$, where $\mathcal{N}(J)$ denotes the nullspace of J. This can be done by sampling $\mathbf{x}_p^0$ from $\mathcal{X}$ and taking its projection onto the nullspace of J, i.e. $\mathbf{x}_p = (I - JJ^T)\mathbf{x}_p^0$. This continues until a proposal $\mathbf{x}_p$ such that $f(\mathbf{x}_p) > f(\mathbf{x}_0)$ is obtained. Thus, adapting the shrinking rank approach by Thompson and Neal [66] for stochastic optimization via the iterative forward slice procedure we attain the following Algorithm 2.

Algorithm 2 Multivariate forward slice

Set the the tolerance ϵ, initial design point x_0, the maximum number of candidates per slice n_c, the shrinking multiplier $\phi \in (0, 1)$, and the angle θ.
repeat
  Obtain $y \leftarrow f(x_0)$
  Define $S = \{x : f(x) \geq y\}$
  Initialize $J = [\,]$ as an empty-matrix.
  Initialize $\mathcal{S} = \mathcal{X}$.
  Sample x_1 from $S^X = S \cap \mathcal{X}$ as follows
  repeat
    Obtain a random sample z_k from $\mathcal{S}$
    if $J = [\,]$ **then**
      $\tilde{x}_k \leftarrow x_0 + z_k$
    else
      $\tilde{x}_k \leftarrow x_0 + z_k - JJ^T z_k$
    end if
    if $f(\tilde{x}_k) \geq y$ **then**
      $x_1 \leftarrow \tilde{x}_k$
      break
    else
      if the number of columns of $J <p-1$ **then**
        Calculate g^* as follows
        if $J = [\,]$ **then**
          $g^* = \nabla f(\tilde{x}_k)$
        else
          $g^* = \nabla f(\tilde{x}_k) - JJ^T \nabla f(\tilde{x}_k)$
        end if
        if $\frac{g^{*T} \nabla f(\tilde{x}_k)}{\|g^{*T}\| \|\nabla f(\tilde{x}_k)\|} > \cos(\theta)$ **then**
          Update $J = [J \quad g^*/\|g^*\|]$
        else
          Update $\mathcal{S} \leftarrow \mathcal{S} \times \phi$
        end if
      else
        Obtain n_c candidate points from the sampling subspace.
        Call this candidate set $\mathcal{C}$.
        if $\exists x_c \in \mathcal{C}$ such that $f(x_c) \geq y$ **then**
          $x_1 \leftarrow \underset{c \in \mathcal{C}}{\text{argmax}}[f(c)]$
          break
        else
          Reset $J = [\,]$
        end if
      end if
    end if
  until $f(x_1) \geq y$
  Calculate $D = |(f(x_1) - f(x_0))/f(x_0)|$
  $x_0 \leftarrow x_1$
until $D \leq \epsilon$

3.4 Heuristics for Stochastic Optimization via Forward Slice

The forward slicing method should select a point that gives a global maximum of a function $f(\cdot)$ under the following conditions:

(1). The function f is continuous, smooth, and differentiable in the space $\mathcal{X}$.

(2). There is no *local plateau* in $\mathcal{X}$. We define a local plateau as follows: a local plateau exists in $\mathcal{X}$ if $\exists$ an interval $P \subset \mathcal{X}$ such that $f(x_i) = f(x_j)\ \forall x_i, x_j \in P$ and $\exists x_k \in \mathcal{X}$ such that $f(x_k) > f(x)\ \forall x \in P$.

(3). The design space $\mathcal{X}$ is compact.

Under the above conditions, in the Appendix B we show that the iterates converge to a global maximum. For practical implementation of the algorithm, however, we set a tolerance $\epsilon > 0$ to restrict the number of iterations.

A graphical illustration of the forward slice procedure is shown in Figure 3.4. Consider the case where we have the initial point x_0 closer to one of the local modes that is not the global maximum.

The upper left panel of Figure 3.4 shows that the next sampled point based on the first iteration is at the second mode or very close to it. Since the difference in the response between $f(x_1)$ and $f(x_0)$ is larger than the set tolerance, we perform another iteration of the forward slice, as shown in the upper right panel of Figure 3.4. Given the new slice level, the slice S contains two intervals: one interval for the first mode and another for the second mode. Since the first interval is much larger than the second one, we are more likely to sample the next x_1 from the slice portion containing the first mode, which is what happened in this case.

Given that we have not reached the tolerance for convergence, we perform another forward slice step, as shown in the lower left panel of the Figure 3.4. We observe that the slice level now excludes the second mode entirely, ensuring that we will obtain the maximum that is

in the first mode if we continue running the forward slice until convergence. In this case, convergence is achieved after 17 iterations which is shown in the lower right panel of Figure 3.4, and we observe that the x obtained gives the global maximum for $f(x)$.

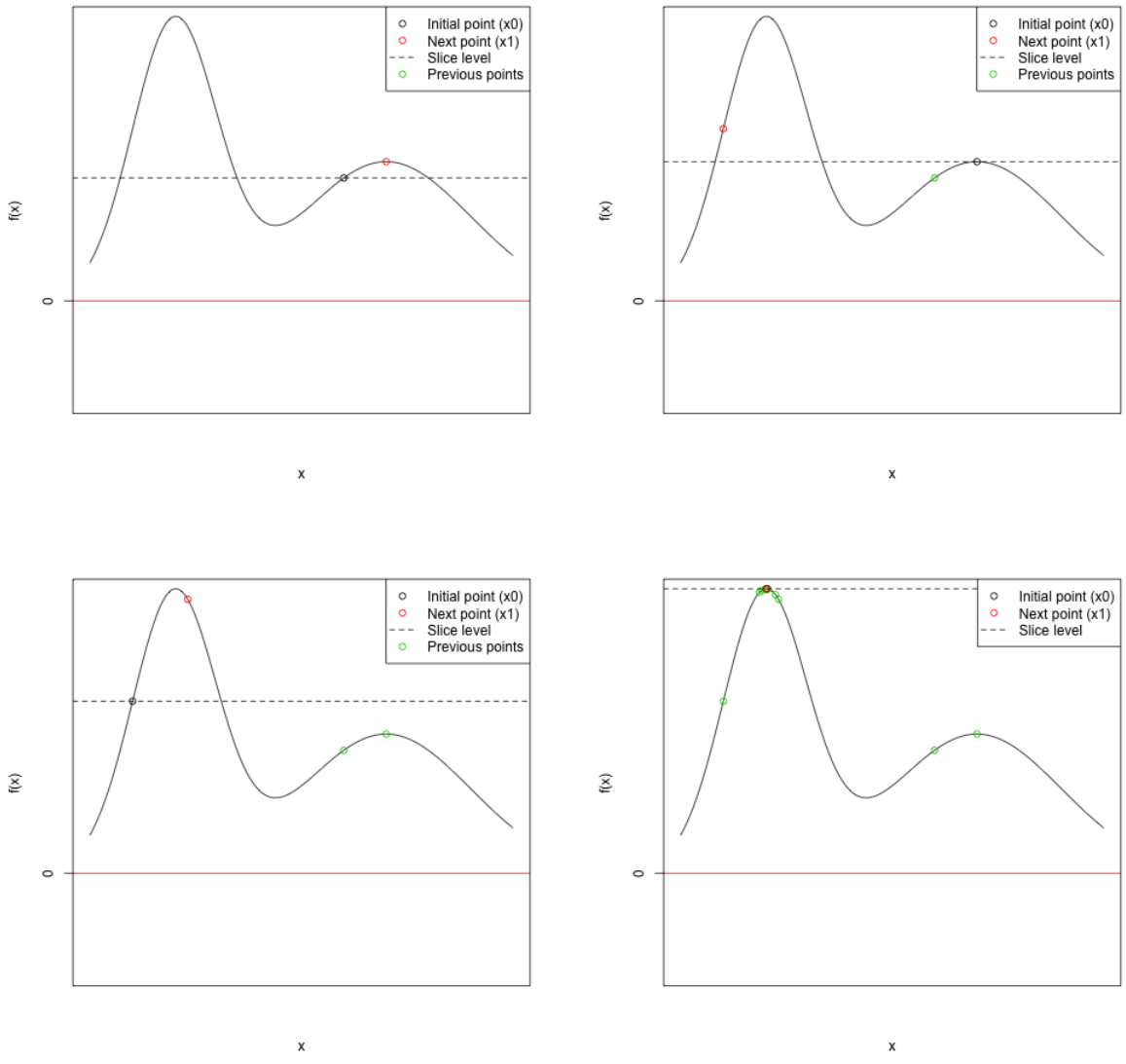


Figure 3.4: An example run of the forward slice with tolerance $\epsilon =$ 1E-8.

3.5 Comparison with alternative optimization methods

We utilized simulation studies to assess the performance of the forward slice method compared to alternative optimization methods, namely, the conjugate gradient (CG) method, the Nelder-Mead method (NM), the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method, the bounded BFGS (L-BFGS-B) method, simulated annealing (SANN), and the Brent method, all of which available with the optim() function in R.

3.5.1 Uni–dimensional case

We considered three different scenarios of the response function (described below), and for each scenario, we assessed the methods over 1000 replications by comparing their resulting points x that maximize the function $f(x)$.

- Scenario 1 – Bimodal function:

$$f(x) = \sqrt{\frac{2}{\pi}}\exp(-2x^2) + \frac{1}{2\sqrt{2\pi}}\exp(-\frac{(x-2.5)^2}{2})$$

  with the boundary of [-1,4] on x. For the application of all methods, for each replication, the initial point was randomly chosen from a Unif(-1,4) distribution. All of the methods use the same initial point for each replication.

- Scenario 2 – Trimodal function:

$$f(x) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{(x+4)^2}{2}) + \frac{1}{\sqrt{3.92\pi}}\exp(-\frac{x^2}{3.92}) + \frac{1}{2\sqrt{2\pi}}\exp(-\frac{(x+4)^2}{2})$$

  with the boundary of [-7,7] on x. For each replication, the initial point was randomly chosen from a Unif(-7,7) distribution and all of the methods use the same initial point for each replication.

- Scenario 3 – No local/unique maximum:

$$f(x) = (x/10)^2 + 0.05$$

  with the boundary of [-7,7] on x. For each replication, the initial point is randomly chosen from a Unif(-7,7) distribution and all of the methods use the same initial point for each replication.

- Scenario 4 – Bimodal function with a narrow global peak and a wider local peak:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x-2)^2}{2}) + \sqrt{\frac{2}{\pi}} \exp(-50x^2)$$

  with the boundary of [-1,4] on x. For each replication, the initial point was randomly chosen from a Unif(-7,7) distribution and all of the methods use the same initial point for each replication.

Figure 3.5 shows that the NM, BFGS, L-BFGS-B, CG, and SANN methods select from the first mode around 50% of the time. On the other hand, the multiple forward slice method correctly selects the optimum from the first mode 96% of the time. A similar performance is seen in Figure 3.6 which shows that the NM, BFGS, L-BFGS-B, CG, and SANN methods correctly select from the highest mode around 50% of the time. On the other hand, the multiple forward slice method correctly selects the optimum from the highest mode 98% of the time. Figure 3.7, shows that all the methods correctly select from either extremes of the design space around half of the time. Figure 3.8 shows that the multiple forward slice method correctly selects the highest mode 57.3% of the time, whereas the other methods select the highest mode around 20% of the time.

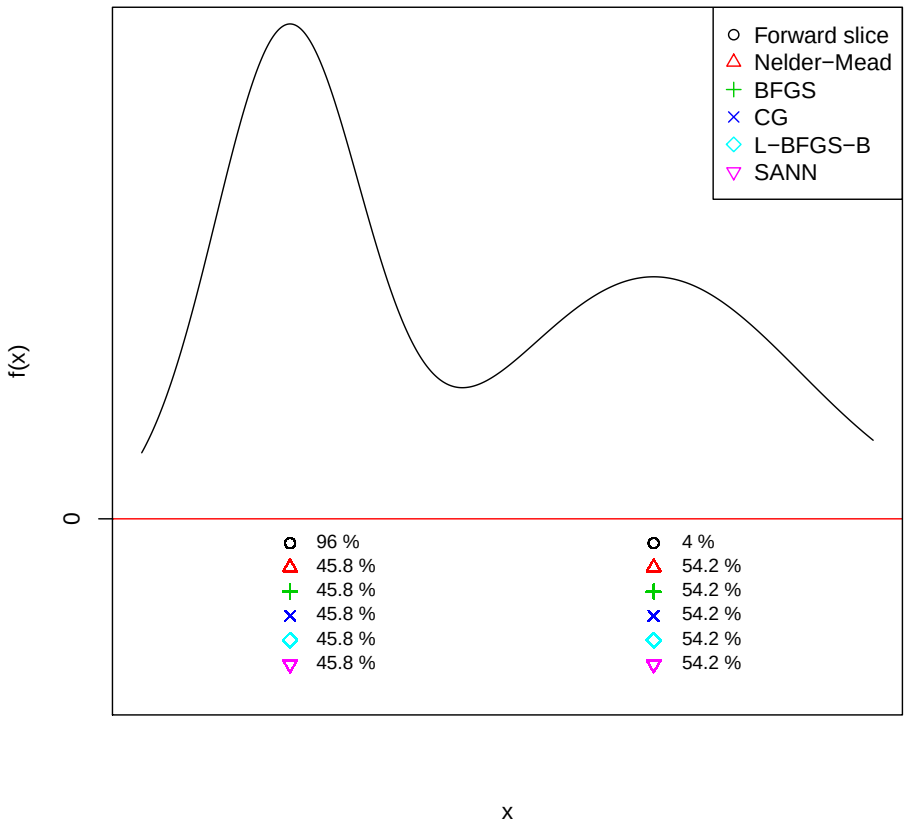


Figure 3.5: Scenario 1: Comparison of methods. The points represent where the optimum was selected for one replication. Reported percentages are the proportion of points selected to be the maximum for each local maximum.

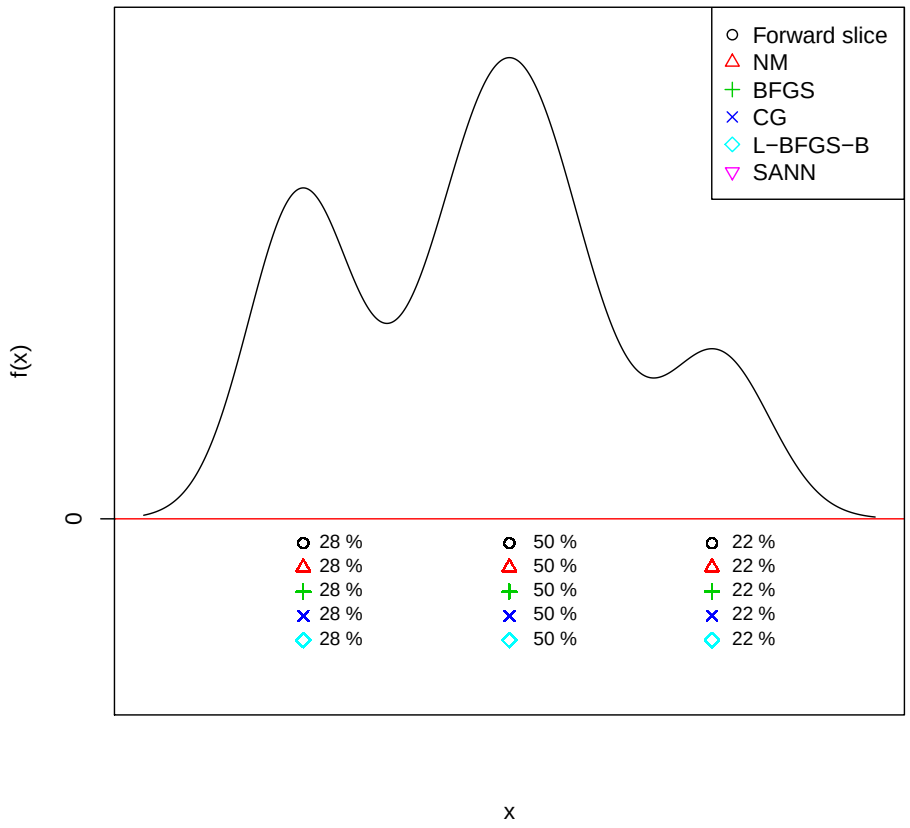


Figure 3.6: Scenario 2: Comparison of methods. The points represent where the optimum was selected for one replication. Reported percentages are the proportion of points selected to be the maximum for each local maximum.

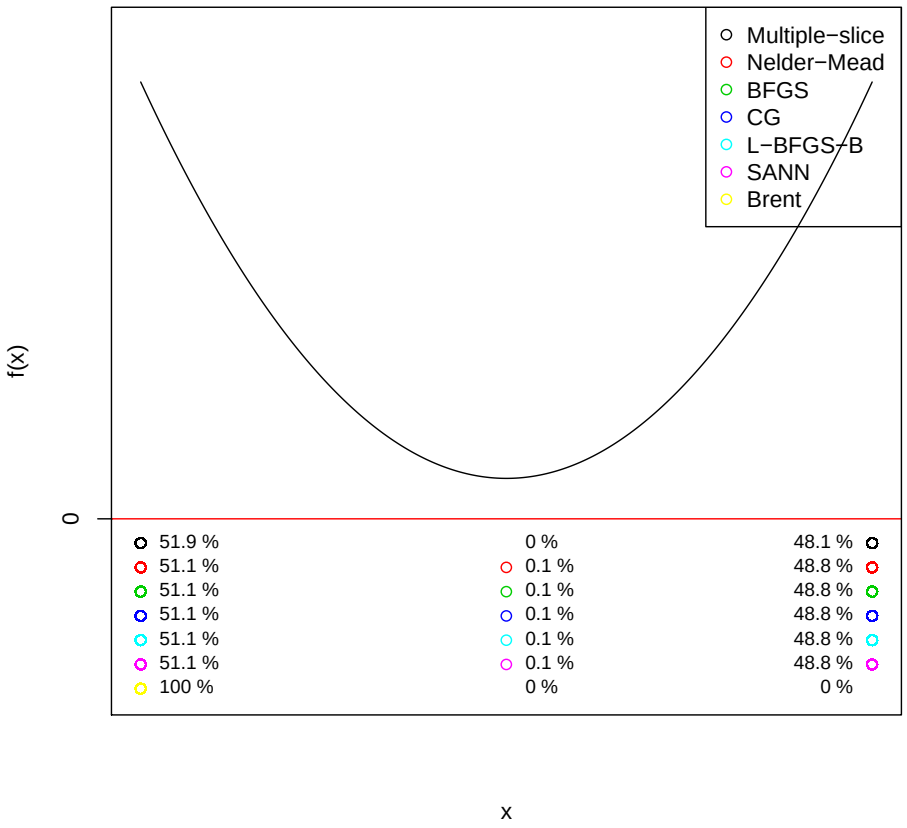


Figure 3.7: Scenario 3: Comparison of methods. The points represent where the optimum was selected for one replication. Reported percentages are the proportion of points selected to be the maximum for each local maximum.

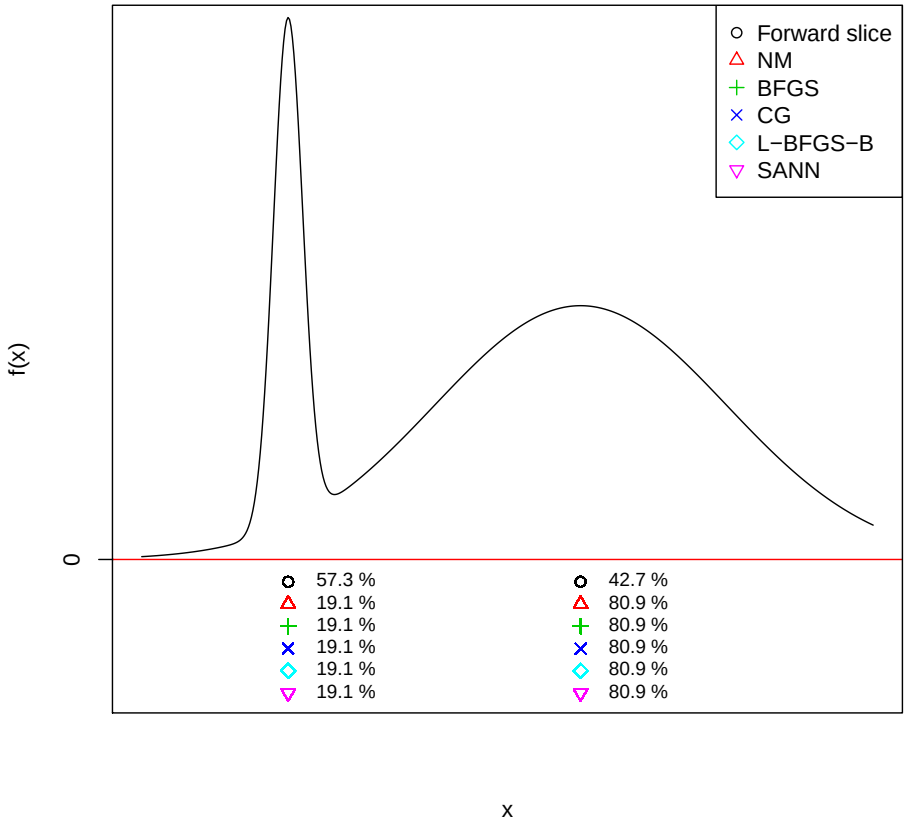


Figure 3.8: Scenario 3: Comparison of methods. The points represent where the optimum was selected for one replication. Reported percentages are the proportion of points selected to be the maximum for each local maximum.

3.5.2 Multi–dimensional Case

Similarly, we considered an additional set of three scenarios of the response function (described below), and for each scenario, we assessed the methods over 1000 replications by comparing their resulting points x that maximize the function $f(x)$.

- Scenario 1 – Bimodal function:

$$f(x) = BVN\left(\begin{pmatrix} 0 \\ 1.8 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) + BVN\left(\begin{pmatrix} 1.8 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}\right)$$

  where $BVN()$ is a bivariate normal distribution function, and with the boundary of [-3,5] on each of the variables. For each replication, the initial point is randomly chosen from a Unif(-3,5)×Unif(-3,5) distribution and all of the methods use the same initial point for each replication.

- Scenario 2 – Hexamodal function:

$$\begin{aligned} f(x) \;=\; & 1.2 * BVN\left(\begin{pmatrix} -4 \\ 4 \end{pmatrix}, \begin{pmatrix} 1 & -0.6 \\ -0.6 & 1 \end{pmatrix}\right) + BVN\left(\begin{pmatrix} -3 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \\ & + BVN\left(\begin{pmatrix} 0 \\ 4 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) + 0.7 * BVN\left(\begin{pmatrix} -1.5 \\ 0 \end{pmatrix}, \begin{pmatrix} 4/9 & 0 \\ 0 & 4/9 \end{pmatrix}\right) \\ & + BVN\left(\begin{pmatrix} 5 \\ 3 \end{pmatrix}, \begin{pmatrix} 0.25 & 0 \\ 0 & 0.25 \end{pmatrix}\right) + BVN\left(\begin{pmatrix} 3 \\ -3 \end{pmatrix}, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}\right) \end{aligned}$$

  where $BVN()$ is a bivariate normal distribution function, and with the boundary of [-3,5] on each of the variables. For each replication, the initial point was randomly chosen from a Unif(-7,6)×Unif(-7,6) distribution. All of the optimization methods utilized the same initial point for each replication.

Figures 3.9–Figure 3.10 show that the multivariate forward slice procedure chooses the correct mode more frequently compared to the alternative methods. The average maxima is also higher and closer to the true maxima under the multivariate forward slice compared to the average from the alternative methods.

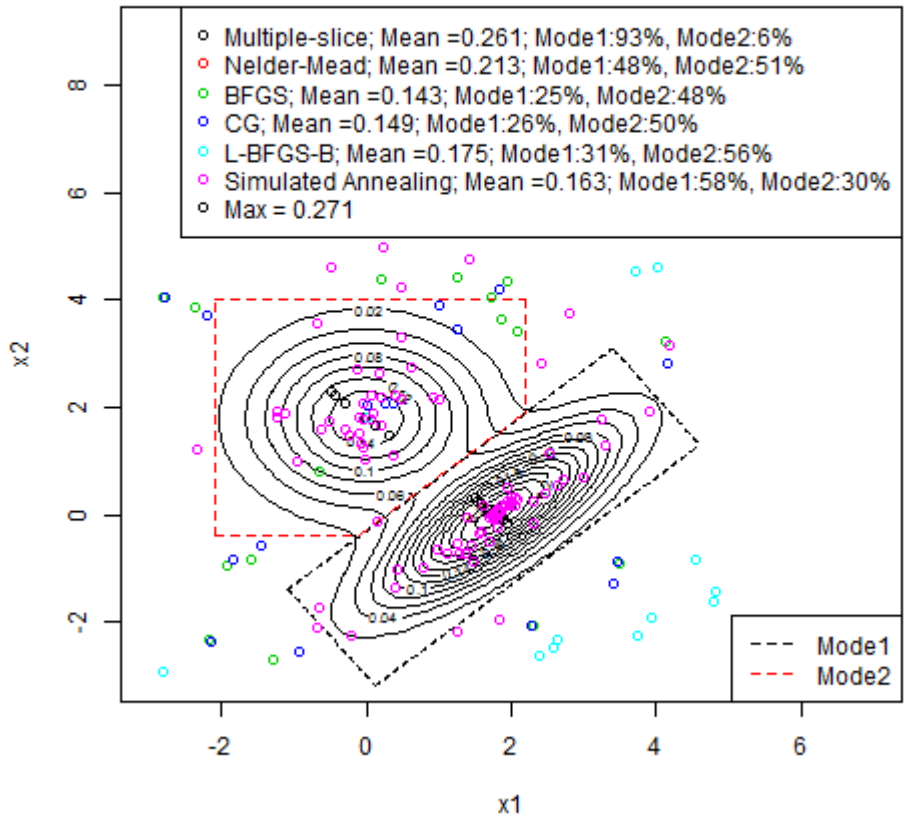


Figure 3.9: Scenario 1: Comparison of methods. The hollow round points represent where the optimum was selected for one replication. Reported percentages are the proportion of points selected to be the maximum for each local maximum.

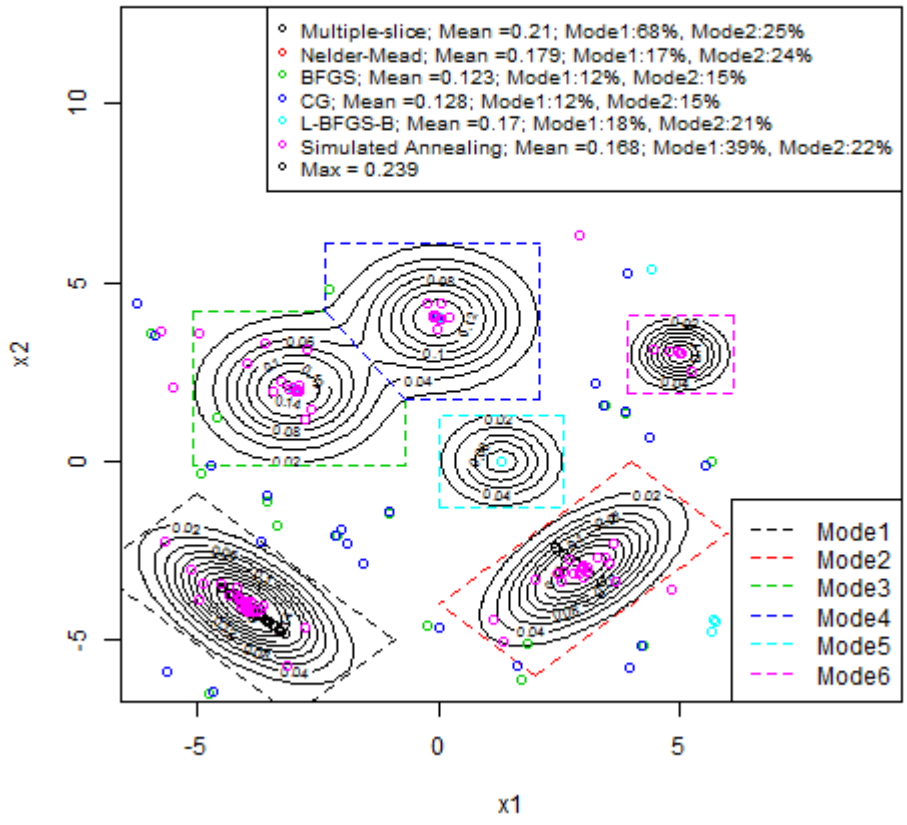


Figure 3.10: Scenario 3: Comparison of methods. The hollow round points represent where the optimum was selected for one replication. Reported percentages are the proportion of points selected to be the maximum for each local maximum.

3.6 Discussion

The forward slice procedure is a stochastic optimization procedure that selects points with response closer to the optimum at each iteration. The points obtained from the forward slice procedure theoretically converge to the global maximum within a constraint of the design space $\mathcal{X}$. The forward slice procedure allows escape from a local optimum by not restricting the search for the next point to the local search. Rather, the forward slice procedure samples from $S^X = S \cap \mathcal{X}$, that is, the slice within the design space.

The forward slice procedure selects the highest mode (or maxima) most of the time in univariate optimization, as shown in Figures 3.5, 3.6, and 3.7, as well as in multivariate optimization, as shown in Figures 3.9 and 3.10 and outperforms the selection obtained with alternative optimization methods (conjugate gradient, BFGS, L-BFGS-B, Nelder-Mead, and simulated annealing).

One limitation for the forward slice procedure is that the probability of sampling from the highest mode compared to a local lower mode depends on the relative size of the 'support' of the modes at each slice level. For example, for a function with narrow high mode and wider lower mode, as shown in Figure 3.8, we observe that the forward slice procedure only selects the highest mode around 57% of the time. We observe, however, that the forward slice procedure still outperforms the alternative methods which only selected the highest mode around 20% of the time.

Another limitation to the forward slice procedure is that in theory there is no upper bound to the amount of computational time needed for the forward slice procedure to converge. In our simulations, the forward slice procedure took on average, around 5 times the computing time required for the simulated annealing method. To explain this increase in computation time, take for an example the one–dimensional optimization. The alternative methods may stop after finding a local mode. The forward slice procedure, however, may continue sampling if there is support for other modes in a given slice. In multi–dimensional setting the increase in computational time can further be explained by the adaptive procedure for sampling from

the slice. Thus, we observe a performance–computing time trade–off with respect to our proposed method versus existing alternatives.

Chapter 4

FORWARD SLICE PROCEDURE IN DESIGN PROBLEMS

4.1 Introduction

The proposed stochastic optimization method via forward slice presented in Chapter 3 was originally motivated by specific design problems. In this chapter we examine the application of the forward slice procedure in two design problems. First, in Section 4.2 we apply the forward slice method to determine the maximum tolerated dose (MTD) in a dose–finding study and compare the results to those from the standard continuous reassessment method (CRM). Second, in Section 4.3 we apply our method to determine the optimal selection of covariate values and compare the results to those obtained by the method in Dror and Steinberg (2008). Finally, in Section 4.4 we provide a discussion.

4.2 Uni–dimensional design problem: dose–finding

Recall from Chapter 2 that the standard CRM treats the next patient at the dose level closest to the target level by directly inverting the currently estimated dose-response curve. We propose a modification of the standard CRM to incorporate our forward slice procedure for sampling the treatment dose for the next patient via maximization of an objective function. The objective function reaches its maximum at the dose with probability of toxicity equal to the target. The simplest form for such a function could be $\tilde{f}(x) = 1 - |\psi(x) - p_{target}|$, where $\psi(x)$ is the dose-response function and p_{target} is the target toxicity probability. However, since it is undesirable to expose patients to unnecessarily high doses, we impose a penalty on higher doses with the objective function f shown below:

$$f(x) = (1 - \text{pen} \times I_{\{\psi(x) > p_{target}\}}) \times (1 - |\psi(x) - p_{target}|)$$

where pen>0 is the term associated with the penalty. The maximum of this function is at x for which $\psi(x) = p_{target}$, since $f(x) = 1$ when $\psi(x) = p_{target}$ and $f(x)<1$ when $\psi(x) \neq p_{target}$. Given that the forward slice procedure gives the value of x that maximizes $f(x)$, we expect the selected doses to be the same as the ones selected by the standard CRM. We conducted a couple of simulation studies to compare the performance of the standard CRM versus the modified CRM via forward slice.

4.2.1 Simulation study 1

We conducted a simulation study to assess and compare the performance of the CRM with forward slice versus that from the standard CRM. Our first simulation considers the situation in which a drug can potentially be tested at 13 dose levels with the true probability of toxicity shown in the Table 4.1. Suppose that the probability of toxicity that is of interest (MTD) is 0.30. We will use the logistic regression model for the dose-response relationship, that is, $\psi(x, a) = \frac{\exp(3+\exp(a)*x)}{1+\exp(3+\exp(a)*x)}$. The 'slope' parameter in the model is given by $\exp(a)$, which is strictly positive. The prior on a is a Normal distribution with mean 0 and variance 1.34. Thus, the 'slope' $\exp(a)$ has a lognormal distribution with $\mu = 0$ and $\sigma = \sqrt{1.34}$, which is quite heavily skewed towards smaller values of $\exp(a)$, but still has a considerable tail for larger values. The maximum sample size is $N = 40$. The results from 1000 replications using the standard CRM and the CRM with forward slice are shown in Table 4.1. Tables 4.1 and 4.2 show that the methods perform similarly.

Table 4.1: Simulation study 1: Dose finding study. The true probability of toxicity for each dose, and the percent of times each dose is selected as MTD under each method across 1000 replications. We assume pen = 0.8.

		% selected as MTD	
Dose number	True probability of toxicity	Standard CRM	CRM with forward slice
1	0.05	0	0.3
2	0.08	1.4	0.4
3	0.12	4.8	6.5
4	0.18	16.6	18.8
5	0.27	33.5	31.5
6	0.38	30.5	29.5
7	0.50	12.6	12.2
8	0.62	0.6	0.8
9	0.73	0	0
10	0.82	0	0
11	0.88	0	0
12	0.92	0	0
13	0.95	0	0

Table 4.2: Simulation study 1: Dose finding study. Sample size and toxicity for both approaches

	Standard CRM	CRM with forward slice
Average sample size	40	40
Proportion of people with toxicity	30.4%	30.5%

We note that our application of the forward slice to CRM requires the specification of an objective function with a penalty term, pen. Thus, we considered additional simulations to assess the sensitivity of the performance to the choice of a penalty term pen. The results are given in Table 4.3.

Table 4.3: Simulation study 1: Dose finding study. Results assessing the sensitivity to the choice of a penalty term pen.

		% selected as MTD			
Dose	True probability of toxicity	pen = 0	pen = 0.1	pen = 0.2	pen = 0.3
1	0.05	0	0	0.2	0
2	0.08	1.1	0.7	0.4	1.1
3	0.12	6.8	6.3	4	5.3
4	0.18	20.6	18.9	18.6	20.7
5	0.27	34.1	35.3	36.8	36.2
6	0.38	27.1	29.5	28.6	26.7
7	0.5	8.7	8.3	10.2	8.7
8	0.62	1.6	0.9	1	1.3
9	0.73	0	0.1	0.2	0
10	0.82	0	0	0	0
11	0.88	0	0	0	0
12	0.92	0	0	0	0
13	0.95	0	0	0	0
	Average Sample Size	40	40	40	40
	Prop. toxicity	0.30715	0.306225	0.303425	0.30515

From Table 4.3, we observe that the proportion selected as MTD for each dose, the average sample size, and the proportion of people who had toxicities appear to be comparable for different values of pen. Hence, it appears that the method is quite robust to the choice of the penalty term pen.

4.2.2 Simulation study 2

Our second simulation considers the situation in which a drug can potentially be tested at 16 dose levels with the true probability of toxicity shown in the Table 4.4 and with the probability of toxicity that is of interest (MTD) equal to 0.20. The dose–response is described by $\psi(x, a) = \frac{\exp(3+\exp(a)*x)}{1+\exp(3+\exp(a)*x)}$. The 'slope' $\exp(a)$ also has a lognormal distribution with $\mu = 0$ and $\sigma = \sqrt{1.34}$. The maximum sample size is $N = 30$. The results from 1000 replications using both the standard CRM and the CRM via forward slice sampling are shown below. Tables 4.4 and Table 4.5 again show similar performances of both methods.

Table 4.4: Simulation study 2: Dose finding study. The true probability of toxicity for each dose, and the percent of times each dose is selected as MTD under each method across 1000 replications. We assume pen = 0.8.

		% selected as MTD	
Dose number	True probability of toxicity	CRM	CRM with slice design procedure
1	0.054	4.4	5.9
2	0.071	2.8	3.1
3	0.092	2.8	2.1
4	0.119	23.2	23.1
5	0.153	18.8	21.3
6	0.193	26.1	26.5
7	0.242	13.6	11.9
8	0.298	4.4	3.7
9	0.361	1.7	1.5
10	0.429	1.8	0.6
11	0.500	0.4	0.3
12	0.571	0	0
13	0.639	0	0
14	0.702	0	0
15	0.758	0	0
16	0.807	0	0

Table 4.5: Simulation study 2: Dose finding study. Sample size and toxicity for both approaches

	CRM	CRM with slice design procedure
Average sample size	20.02	19.80
Proportion of people with toxicity	24.3%	24.3%

4.3 Multi–dimensional design problem: comparison of the forward slice procedure with the method by Dror and Steinberg (2008)

In this section we consider the multi–dimensional design problem where the goal is to select design points meeting the D-optimality criterion. Recall from Chapter 2 that Dror and Steinberg (2008) approached this problem using the Fedorov exchange algorithm to obtain

the next design point. Alternatively, we propose to select the next design point using the forward slice procedure. We note that the D-optimality criterion specifically define the objective function for the forward slice. Specifically, we utilize the function

$$f(x) \;=\; \phi_2(\tilde{\boldsymbol{\beta}}; d) \;=\; \phi_2\left(\tilde{\boldsymbol{\beta}}; \begin{pmatrix} X \\ x \end{pmatrix}\right) \;=\; \log(|\mathbf{I}(\tilde{\boldsymbol{\beta}}; d)|).$$

Thus, in the above, $d = \begin{pmatrix} X \\ x \end{pmatrix}$, where X is the current design matrix and x is the next (multivariate) design point. For completeness, we re-state the forward slice procedure for the D-optimality criterion using the above function. We note that the initial design points are obtained using the algorithm by Dror and Steinberg to guarantee that the information matrix is of full rank.

4.3.1 Forward Slice under the D–Optimality Criterion

Let n denote the current number of samples, and N denote the total sample size.

1. Obtain the first n_0 design points defining a full–rank current design matrix X_0 using the method by Dror and Steinberg using the Fedorov's exchange algorithm. Define the tolerance ϵ.

2. Given that i design points have been observed, the $(i+1)th$ design point is selected as follows. Determine x_{0i}, the input for our slice procedure, as the last observed design point, i.e. the last row of X_0, where X_0 is the currently observed design matrix.

3. Define $d_i = \begin{pmatrix} X_0 \\ x_{0i} \end{pmatrix}$, with x_{0i} chosen in the previous step. Let $y = f(x_{0i}) = \phi_2(\tilde{\boldsymbol{\beta}}; d_i)$ define the level of the horizontal 'slice'. Note that f is a function of the next design point, x, to be appended as follows: $f(x) = \phi_2\left(\tilde{\boldsymbol{\beta}}; \begin{pmatrix} X_0 \\ x \end{pmatrix}\right)$.

4. Perform the multivariate forward slice procedure using the defined slice to obtain the temporary design point x_1.

5. Calculate the relative difference $D = |(f(x_1) - f(x_0))/f(x_0)|$.

   - If $D \leq \epsilon$, then report x_1 as the next design point that maximizes f
   - Otherwise, take $x_0 = x_1$ and repeat steps 2-5.

4.3.2 Simulation Study – Choosing the Next Point

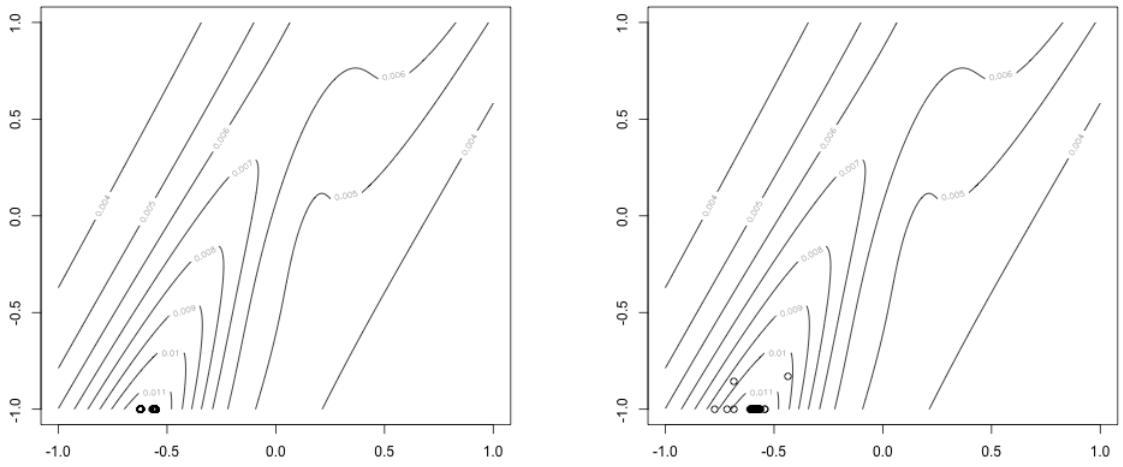
Simulation with 2 variables. We performed a simulation study with 1000 replications to assess the performance of the slice design procedure compared to Fedorov's exchange algorithm (proposed by Dror and Steinberg) for selecting a single next design point. We used a logistic regression model with 2 variables with true regression parameters $(0, 7, -3)^T$. The initial design matrix is given by $\mathbf{X} =$

$$\begin{pmatrix} 1 & -0.262 & -1 \\ 1 & 0.259 & 1 \\ 1 & 0.754 & 1 \\ 1 & 0.108 & -1 \end{pmatrix}$$

This initial design matrix $\mathbf{X}$ is based on running the algorithm by Dror and Steinberg up until the information matrix given the current design points is non-singular. Given that there are only 2 variables, we can obtain the contour plot of the D-optimality criterion upon addition of one design point (x_1, x_2) as a function of x_1 and x_2. With 1000 replications, the scatter of the next design point given the current design $\mathbf{X}$ for both methods is given below.

Figure 4.1(a) shows that the obtained design point is close to the optimal using the Fedorov's exchange algorithm, whereas in Figure 4.1(b), we observe that the obtained design point using the slice procedure also clusters around the optimal. The mean log D-optimality for Fedorov's is 0.01189, whereas the mean log D-optimality for the slice procedure is 0.01188. The two methods appear to be comparable in selecting the next design point in terms of the

D-optimality criterion.



(a) (b)

Figure 4.1: Contour plot of the D-optimality criterion upon addition of one design point (x_1, x_2). Scatter of next optimal design points using the Federov's exchange algorithm (panel a) or the forward slice procedure (panel b).

Simulation with 4 variables. Next, we considered a logistic regression model with 4 variables with true regression parameters $(0, 7, 8, -3, 0.5)^T$. Moreover, the initial design matrix is given by $\mathbf{X} =$

$$\begin{pmatrix} 1 & 0.134 & -0.124 & 0 & 0 \\ 1 & -0.146 & -0.118 & -0.000192 & 0.000455 \\ 1 & 0.751 & -1 & -1 & -1 \\ 1 & -0.156 & -0.091 & 0.786 & 1 \\ 1 & -0.376 & 1 & 1 & 1 \end{pmatrix}$$

The mean log D-optimality for Fedorov's is -14.27, whereas the mean log D-optimality for the slice procedure is -11.97, and for 89.6% of the experiments, the log D-optimality from the design point picked by the slice procedure is higher than the log D-optimality from the

design point picked by Fedorov algorithm. Hence, it appears that the forward slice procedure picks design points that would give higher D-optimality when appended to the current design point $\mathbf{X}$.

4.3.3 Simulation Study – Complete Design

Simulation studies are then performed to assess the performance of the forward slice method throughout a study instead of selecting a single next design point to sample. We assess the performance of the forward slice method in Dror and Steinberg's sequential D-optimal design with binary response variable. As discussed in Section 2.1.1.6, Dror and Steinberg's sequential D-optimal design re-estimates the model parameter estimates after the response under a new design point is observed. Based on the current parameter estimates, the next design point is chosen to maximize the estimate of the D-optimality criterion using the Fedorov Exchange Algorithm. This procedure is repeated until the maximum sample size. In our simulation studies, we only replace the Fedorov Exchange Algorithm with the forward slice procedure, but do not change the parameter estimation procedure or stopping criterion.

We performed a simulation study with 100 replications to assess the performance of the slice design procedure compared to the Fedorov's exchange algorithm using a logistic regression model with 4 variables with the maximum sample size of n=30.

Table 4.6 shows the true parameter values and the prior distribution for each parameter along with corresponding simulation results. We note that in this set of simulations the prior distributions are flat, but centered at the true parameter values.

Table 4.6: Simulation results comparing the performance of the forward slice to the Fedorov's exchange algorithm in the context of the method by Dror and Steinberg when the priors are flat, but centered around the true parameter value. We assume n=30.

			Mean parameter (Mean 95% range)	
Parameter	True value	Prior dist.	Fedorov	Forward-slice
β_0	0	Unif(-10,10)	-0.233 (2.246)	-0.049 (1.98)
β_1	7	Unif(-3,17)	11.84 (6.64)	12.20 (7.06)
β_2	8	Unif(-2,18)	13.56 (7.46)	13.30 (7.91)
β_3	-3	Unif(-13,7)	-5.33 (3.80)	-4.75 (2.87)
β_4	0.5	Unif(-9.5,10.5)	0.79 (2.61)	0.79 (2.19)
Median D-Efficiency			0.52	0.56

From Table 4.6 above, we see that the median D-efficiency is slightly higher using the forward-slice procedure compared to using the Fedorov Exchange Algorithm. The bias and uncertainty of the parameter estimates are comparable between the two methods.

We further explored the methods to assess sensitivity to the the assumed priors with a simulation with 100 replications, but this time assuming that the priors are flat and centered around 0 instead of centered around the true parameter values, and with maximum sample size of $n = 40$ per replication.

Table 4.7: Simulation results comparing the performance of the forward slice to the Fedorov's exchange algorithm in the context of the method by Dror and Steinberg when the priors are flat, but centered around 0. We assume $n = 40$.

			Mean parameter (Mean 95% range)	
Parameter	True value	Prior dist.	Fedorov	Forward-slice
β_0	0	Unif(-10,10)	0.124 (1.528)	0.404 (1.25)
β_1	7	Unif(-10,10)	6.77 (4.05)	7.88 (3.60)
β_2	8	Unif(-10,10)	7.92 (3.29)	8.28 (3.04)
β_3	-3	Unif(-10,10)	-3.37 (2.09)	-3.03 (1.70)
β_4	0.5	Unif(-10,10)	-0.23 (2.23)	-0.39 (2.46)
Median D-Efficiency			0.46	0.50

From Table 4.7 above, we see that the median D-efficiency is slightly higher using the forward-slice procedure compared to using Fedorov Exchange Algorithm. The bias and uncertainty of the parameter estimates are comparable between the two methods.

4.4 Discussion

In this chapter, we presented some applications of the forward slice procedure in design problems: univariate dose-finding and multifactor experiments. For the univariate dose-finding design problem, we viewed the search for the maximum tolerated dose as optimizing a form of utility function, for which the utility is the highest when the probability of toxicity is the desired maximum probability of toxicity, and utilized the forward slice procedure to obtain the optimal dose. For the multifactor experiment, we used the design proposed by Dror and Steinberg and utilized the forward slice procedure instead of the Federov Exchange Algorithm to obtain the design point that would maximize the D-optimality criterion when appended to the current design matrix.

One distinction between the results presented here relative to those presented in Chapter 3 is that, in this chapter, the parameters defining the true objective function are unknown. Thus, the objective function changes upon accrual of new information. As a consequence, the optimum design points are selected based on the current estimate of the objective function (by assuming a given model and based on the current estimates of the model parameters). From the univariate dose-finding simulations, the performance of the modified CRM using the forward slice procedure is comparable to that of the traditional CRM, as shown in Tables 4.1 and 4.4. For the multifactor experiments, we found that the forward slice procedure obtains design points with slightly higher D-efficiency compared to the design points obtained with the Fedorov Exchange Algorithm.

One limitation of the forward slice procedure in design applications is again the computational time which is longer compared to that under the existing methods. However, its performance is comparable or slightly improved relative to that attained under existing methods.

Chapter 5

BAYESIAN SUBGROUP SELECTION

5.1 Preliminaries

Randomized clinical trials are experiments conducted on human subjects to assess the characteristics of a proposed treatment. Biologically, it is possible for a given treatment to be more effective in some populations and less effective in others. This has led to increasing interests in examining and searching for subgroups from the data collected in clinical trials. The search for subgroups are commonly done either prospectively or retrospectively [44] [45]. Prospective search of subgroups are usually performed after a Phase II trial. The subgroup(s) found, if any, are then taken into consideration when designing the Phase III trial. Retrospective search of subgroups are conducted using the Phase III trial. If the trial outcome is positive in the trial, retrospective search is done to ascertain potential subgroup(s) which may be harmed by the treatment. If such subgroup(s) are found, they can be excluded from the indication of the treatment. On the other hand, if the trial outcome is negative, search is done for potential subgroup(s) that may benefit from the treatment.

Subgroups are typically defined by one or more baseline covariates or biomarkers measured in the trial. While it might be important or of great interest to conduct subgroup analyses, there have also been concerns about the current practice. The main concerns include potential for data dredging, multiple comparisons, inflation of false positives, lack of power, and interpretability of the results.

Assmann, Pocock, Enos, and Kasten (2000) [2] conducted a survey of clinical trial reports published in four major medical journals and found that among the 50 trials surveyed, two-third of the trials reported results from subgroup analysis. Most of these trials reported statistical inference based on subgroup-specific treatment effect, and no statistical test on

the interaction effects.

Pocock, Assmann, Enos, and Kasten (2002) [54] also outlined some issues based on the same survey regarding subgroup analysis and its current practice of reporting in publications. Pocock et al noted similarly that most of the reported subgroup analysis were done inappropriately. Pocock et al also noted that most of the trials were powered only to detect the overall treatment effect, and there is lower power to detect the differential effects based on the subgroups due to the size of the trial. This problem is further worsened if subgroup analysis were conducted not based on the pre-planned analysis, allowing for data dredging, which highly inflates the rate of false positives.

Brookes et al [9] [10] noted similar problems and they had conducted simulations to quantify the increased rate of false positives due to inappropriately conducting subgroup analysis. They found that the magnitude of the damage may differ based on trial characteristics, but the false positive rate may be inflated up to 64% based on simulations. They also noted that most of the trials that are planned with 80% power to detect overall effect have only around 29% power to detect differential subgroup effects based on an interaction test.

Many methods and approaches have been developed in subgroup selection. Some current approaches developed in subgroup selection have been in conjuction with methods from the field of causal inference and machine learning, and are data-driven approaches to subgroup search. Foster et al (2011) [27] proposed a method called "Virtual Twins", modeling the potential outcomes using random forests [6], and then using a regression tree to search for subgroup(s). Su et al (2009) [64] proposed a subgroup selection approach by recursive partitioning, extending the method of Classification and Regression Trees (CART) by Breiman et al (1984) [7]. The recursive partitioning method seeks to identify subgroups by splitting the domain space of the variables in a tree-like manner into nodes that are more or less homogeneous with respect to the outcome. Lipkovich et al (2011) [45] proposed a similar approach called "Subgroup Identification based on Differential Effect Search (SIDES)" using a tree-like split and recursive partitioning, adding the 'complexity control' and 'multiplicity control' to control the size of the tree and the false positive rate.

Dixon and Simon (1991) [20] provided an approach for subgroup analysis in the Bayesian framework. They considered the subgroups defined by baseline variables, and they used a generalized linear model framework with one-way baseline variable treatment interaction terms in the model. On their paper, they assume the baseline variables to be binary. A strong assumption in their approach is the exchangeability between interaction effects, that is, no specific baseline variable is assumed to be more or less likely to have a higher or lower interaction effect than all the others. They proposed using noninformative priors for the main-effect parameters in the model, a Normal prior for the treatment effect, and a modification of a standard Jeffrey's prior as the hyperprior on the parameter for the interaction term. Their result shows that the exchangeability leads to the shrinking of the posterior estimates of the interaction parameters towards 0.

Jones et al (2011) [35] proposed an extension to the model proposed by Dixon and Simon (1991). The extension is that they allow for higher-level interactions of the baseline variables in the model. The prior for each level of the interaction terms is taken to be a Normal distribution with mean zero and a variance component. Each level of the interaction term, thus, have one variance component in the prior. They illustrated their method using a model with three binary baseline variables, hence 8 different subgroups are assumed from the model. Hence, there are three different parameters describing the variances of the normal distributions, one for each of the level of the interactions.

Sivaganesan et al (2011) [62] proposed a Bayesian model selection approach to subgroup selection. They considered the enumeration of all possible scenarios of the subgroup effects, considering whether there are effects within each subgroup and the number of unique magnitudes of effects on subgroups that have any. The set of all possible models are then considered, with each model corresponding to one effect combination from all of the possible enumerations. They assigned mixture g-priors for the subgroup effects, and noninformative priors for all the other parameters. The subgroup effect is then determined based on a Bayesian model selection approach.

In this dissertation we examine the problem of subgroup selection under a Bayesian

decision–theoretic paradigm. In Section 5.2, we discuss our proposed method for subgroup selection using a utility formulation. Our method focuses on the decision of which subgroup(s) would be selected, with the subgroup(s) defined by pre-selected variables. In the proceeding sections, we discuss the application of this method for various types of outcomes (continuous, binary, and censored).

5.2 Proposed Method

We propose a method for prospective subgroup selection of subpopulation(s), if any, which would be recommended for further testing in Phase III trial. We assume that the model for analysis is of the GLM family. Let T denote the treatment indicator and let $\mathbf{X}$ denote the variables defining the subgroups. For the subgroup analysis problem, we propose using a model with one-way interaction between each variable defining the subgroups with the treatment variable. We assume the covariates $\mathbf{X}$ are binary or categorical variables, and we assume the model

$$g(\mu(X,T)) \quad = \quad \alpha_0 + \beta_T T + \boldsymbol{\beta}_X \mathbf{X} + \boldsymbol{\gamma} \mathbf{X} T \tag{5.1}$$

where $g(\cdot)$ denotes the link function, and $\mu(\cdot,\cdot)$ denotes the mean. Note that the magnitude of any differential treatment effect depends on the link function. A treatment that provides a non-zero homogeneous effect on one scale (e.g. additive or multiplicative) will necessarily have differential effect with another scale (link function). Here, we assume that the investigators have in mind the scale in which a treatment effect is of interest. Let $\boldsymbol{\theta} = (\alpha_0, \beta_T, \boldsymbol{\beta}_X, \boldsymbol{\gamma})$ denote the model parameters.

5.2.1 Subgroup Notation

We define the notation for the subgroups as follows. Let $\mathbf{X} = \{X_1, ..., X_p\}$ denote the covariates in the model. Let each variable X_i be a categorical variable with h_i categories.

For each variable, there are (h_i+1) possible ways it is included in the subgroup definition, one of them being that the variable is not included as part of the subgroup definition. There are $K = \prod_{i=1}^{p}(h_i + 1)$ number of possible subgroups based on the combination of the variables. We enumerate the distinct subgroups, with indexes $0, 1, 2, ..., K-1$. By convention, we define "subgroup 0" to be no subgroups (the overall population). The enumeration can be done by utilizing the expand.grid() function in R.

We illustrate this using an example with three-variables: X_1 having two levels (0 and 1), X_2 having two levels (0 and 1), and X_3 having three levels (1, 2, and 3). In this case, $K = (3)(3)(4) = 36$, and the different subgroups are given below in Table 5.1.

Table 5.1: Enumeration of subgroups based on three variables.

Subgroup	(Sub)Population
0	Overall
1	X_1=0
2	X_1=1
3	X_2=0
4	X_2=1
5	X_1=0,X_2=0
6	X_1=1,X_2=0
7	X_1=0,X_2=1
8	X_1=1,X_2=1
9	X_3=1
10	X_3=2
11	X_3=3
12	X_1=0,X_3=1
13	X_1=1,X_3=1
14	X_1=0,X_3=2
15	X_1=1,X_3=2
16	X_1=0,X_3=3
17	X_1=1,X_3=3
18	X_2=0,X_3=1
19	X_2=1,X_3=1
20	X_2=0,X_3=2
21	X_2=1,X_3=2
22	X_2=0,X_3=3
23	X_2=1,X_3=3
24	X_1=0,X_2=0,X_3=1
25	X_1=1,X_2=0,X_3=1
26	X_1=0,X_2=1,X_3=1
27	X_1=1,X_2=1,X_3=1
28	X_1=0,X_2=0,X_3=2
29	X_1=1,X_2=0,X_3=2
30	X_1=0,X_2=1,X_3=2
31	X_1=1,X_2=1,X_3=2
32	X_1=0,X_2=0,X_3=3
33	X_1=1,X_2=0,X_3=3
34	X_1=0,X_2=1,X_3=3
35	X_1=1,X_2=1,X_3=3

Note that subgroups 1 and 2, for example, are subpopulations defined by X_1 only, whereas subgroups 24 to 35, for example, are defined by a combination of X_1, X_2, and X_3. Performing a subgroup search over the entire K possible subgroups may be expensive, and might inflate the error rate of incorrectly reporting an unimportant subgroup due to the number of comparisons made. Note further that the subgroups can be classified based on which variables are important in defining the subgroup. We can think of each distinct "classification" as a distinct model. Thus, the subgroups having the same subgroup-defining variables can be thought to come from the same model. For example, subgroups 1 and 2 are from the model $g(\mu(X,T)) = \alpha_0 + \beta_T T + \boldsymbol{\beta}_X X_1 + \boldsymbol{\gamma} X_1 T$. The possible models are shown below in Table 5.2.

Table 5.2: Categories of the subgroups based on the variables defining the subgroup. Each category correspond to a model.

Model	Subgroup-defining Variable	Subgroups
0	None	0
1	X_1	1,2
2	X_2	3,4
3	X_1,X_2	5,6,7,8
4	X_3	9,10,11
5	X_1,X_3	12,13,14,15,16,17
6	X_2,X_3	18,19,20,21,22,23
7	X_1,X_2,X_3	24,25,26,27,28,29,30,31,32,33,34,35

Note that in general, for p variables, there are $M = 2^p$ possible models to consider, and we denote them as model 0,1,...,$M-1$, with model 0 denoting the model with just the overall population.

We view the prospective subgroup selection as a two-stage Bayesian decision problem. The first stage involves the decision of whether there exists enough evidence of effect to warrant further subgroup search or not (in which case we proceed with testing the effect in the overall population). If the decision in the first stage was to search for subgroup(s), the second stage involves deciding which subgroup(s) to report. We discuss the proposed method for these two stages further in the next two sections.

5.2.2 Model Selection Approach

We view the decision of reporting either the overall population or certain subgroup(s) first as a Bayesian model selection problem. This is analogous to testing between the model (M_0) which assess the effect of treatment, adjusting for all predictor variables $\mathbf{X}$, that is,

$$g(E[y]) = \alpha_0 + \beta_T T + \boldsymbol{\beta}_X \mathbf{X}$$

vs. a particular model with subgroup $(M_j, j = 1, 2, ..., M-1)$:

$$g(E[y]) = \alpha_0 + \beta_T T + \boldsymbol{\beta}_X \mathbf{X} + \boldsymbol{\gamma}_j \mathbf{A}_j T.$$

where $\mathbf{A}_j$ is the covariate matrix (or vector) based on the model j. In our example in Section 5.2.1, under model 1 (M_1), $\mathbf{A}_1 = X_1$, whereas under model 3 (M_3), $\mathbf{A}_3 = [X_1\ X_2]$.

Note that we can recast the above problem as testing between the models $\tilde{M}_0$ and $\tilde{M}_j$, where $\tilde{M}_0$ is defined by

$$\begin{aligned} g(E[y|\tilde{M}_0]) &= \alpha_0 + \beta_T T + \boldsymbol{\beta}_X \mathbf{X} + \boldsymbol{\gamma} \mathbf{X} T, \\ \pi(\boldsymbol{\theta}|\sigma^2, \tilde{M}_0) &\sim \mathcal{N}(\boldsymbol{\theta}_0, \sigma^2 \mathbf{M}_0), \\ \pi(\sigma^2|\tilde{M}_0) &\sim \mathcal{IG}(a, b) \end{aligned}$$

where the rows of $\boldsymbol{\theta}_0$ corresponding to the parameters $\boldsymbol{\gamma}$ are 0, and the submatrix of $\mathbf{M}_0$ corresponding to the parameters $\boldsymbol{\gamma}$ is a diagonal matrix with diagonal values δ, where δ is a small number. Likewise, the model $\tilde{M}_j$, for $j = 1, ..., M-1$, is defined by

$$\begin{aligned} g(E[y|\tilde{M}_1]) &= \alpha_0 + \beta_T T + \boldsymbol{\beta}_X \mathbf{X} + \boldsymbol{\gamma} \mathbf{X} T, \\ \pi(\boldsymbol{\theta}|\sigma^2, \tilde{M}_1) &\sim \mathcal{N}(\boldsymbol{\theta}_j, \sigma^2 \mathbf{M}_j), \\ \pi(\sigma^2|\tilde{M}_1) &\sim \mathcal{IG}(a, b) \end{aligned}$$

where the rows of $\boldsymbol{\theta}_j$ corresponding to the interaction parameters that are not of interest under model M_j are 0, and the submatrix of $\mathbf{M}_j$ corresponding to the interaction parameters that are not of interest under model M_j is a diagonal matrix with diagonal values δ, where δ is a small number. In other words, we assume a common probabilistic model for the outcome y, but with a prior that has a large mass at 0 for the the interaction term(s) that are not of interest under model M_j (j=0, ..., M-1). The prior is less informative for the parameters of interest under model M_j.

The comparison between any model $\tilde{M}_j$ and model $\tilde{M}_0$ is performed by computing the posterior probabilities of the models,

$$\begin{aligned} \frac{p(\tilde{M}_j|y)}{p(\tilde{M}_0|y)} &= \frac{p(y|\tilde{M}_j)}{p(y|\tilde{M}_0)} \times \frac{p(\tilde{M}_j)}{p(\tilde{M}_0)} \\ \text{posterior ratio} &= \text{Bayes Factor} \times \text{prior ratio} \end{aligned}$$

The Bayes Factor is given by the ratio of marginal probabilities for each model $\tilde{M}_j; j = 0, 1, .., M-1$ each calculated as

$$p(y|\tilde{M}_j) = \int\int p(y|\boldsymbol{\theta}, \sigma^2, \tilde{M}_j)\pi(\boldsymbol{\theta}|\sigma^2, \tilde{M}_j)\pi(\sigma^2|\tilde{M}_j).$$

5.2.2.1 Estimation of model-specific marginal probabilities

The model-specific marginal probabilities $p(y|\tilde{M}_j)$ can be estimated via several methods [37]. One method is via Monte Carlo integration by simulating from the prior. Here, we obtain $\sigma^{2(i)}$ samples from the prior distribution $\pi(\sigma^2|\tilde{M}_j)$, then obtaining $\boldsymbol{\theta}^{(i)}$ samples based on the sampled $\sigma^{2(i)}$ from the prior distribution $\pi(\boldsymbol{\theta}|\sigma^2, \tilde{M}_j)$. Let N denote the number of samples from the distribution. Then, the marginal probability is estimated by the Monte Carlo average

$$\hat{p}(y|\tilde{M}_j) = \frac{1}{N}\sum_{i}^{N} p(y|\boldsymbol{\theta}^{(i)}, \sigma^{2(i)}, \tilde{M}_j); j = 0, 1, ..., M-1$$

with the Bayes factors estimated by $\frac{\hat{p}(y|\tilde{M}_j)}{\hat{p}(y|\tilde{M}_0)}$ for all $j = 1, ..., M-1$. One limitation to this method is that the prior distribution may be quite different from the likelihood, and hence samples from the prior distribution may lead to very small estimated marginal likelihoods. There are several computational alternatives to address this issue. Kass and Raftery (1995) discuss several methods for estimating the Bayes factor. In particular, there is an approximation via the Bayes Information Criterion (BIC) which we use in our work.

Note that $\tilde{M}_j$'s and $\tilde{M}_0$ contain the same parameters and predictors, and without the constraint that some parameters are tightly concentrated at zero, the BICs computed under the models are exactly the same. Hence, when using the BIC, we compare models M_0 and M_j, instead, since they constrain the parameters directly (instead of via prior distributions). Define the quantity S_j as the negative half of the difference of the BIC's for the two models, i.e.

$$S_j \;=\; -\frac{1}{2}(BIC_j - BIC_0).$$

S_j has been shown to approach the Bayes Factor (BF) asymptotically, i.e.

$$\frac{S_j - \log(\mathrm{BF}_j)}{\log(\mathrm{BF}_j)} \;\rightarrow\; 0.$$

5.2.2.2 Reporting based on the Bayes Factor

We propose using a $(0-1)$ utility function for reporting the overall population vs. searching for certain subgroup(s). That is, the utility is 1 for choosing model j ($a_M = j$) when M_j is true, and 0 otherwise (see Table 5.3).

Table 5.3: 0-1 utility function.

		Truth			
		M_0	M_1	...	M_{M-1}
	0	1	0		0
a_M	1	0	1		0
	⋮			⋱	
	$M-1$	0	0		1

where $a_M = j$ denote selecting model M_j, $j = 0, 1, ..., M-1$. Thus, the posterior expected utility is given by

$$\begin{aligned} E[U|y] &= p(M_0|y)1[a_M = 0] + \sum_{j=1}^{M-1} p(M_j|y)1[a_M = j] \\ &= p(M_0|y)\left(1[a_M = 0] + \sum_{j=1}^{M-1} \mathrm{BF}_j o_j 1[a_M = j]\right) \end{aligned}$$

where $o_j = p(M_j)/p(M_0)$. The expected utility is maximized by $a_M = 0$ when all of the $\mathrm{BF}_j o_j$'s are less than 1. Otherwise, if for some $\tilde{j}$, $\mathrm{BF}_{\tilde{j}} o_{\tilde{j}} > 1$, then the action $a_M = \underset{j \in \{1,...,M-1\}}{\mathrm{argmax}} \mathrm{BF}_j o_j$.

5.2.2.3 Comparison of Methods for Bayes Factor Estimation

Computationally, the Bayes Factor computed using the BIC approximation is the easiest and least intensive to compute, and hence it is the estimate of Bayes Factor that we use in our method. To justify this choice, we performed a small simulation study to validate Bayes Factors computed via the BIC approximation relative to other methods. Specifically, we compared the estimates of BF obtained via an analytical expression, prior sampling or BIC approximation under the linear model with a single binary variable X.

Let $\mathbf{D} = [\mathbf{1}\ T\ X\ XT]$ denote the design matrix. We assume that the priors for the regression parameters are specified with $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_1 = \mathbf{0}$, $\mathbf{M}_1 = 100\ \mathbf{I}$, where $\mathbf{I}$ denotes the

identity matrix, and

$$\mathbf{M}_0 = \begin{pmatrix} 100 & 0 & 0 & 0 \\ 0 & 100 & 0 & 0 \\ 0 & 0 & 0.00000001 & 0 \\ 0 & 0 & 0 & 0.00000001 \end{pmatrix}$$

and that the prior for the variance component σ^2 is specified with $a = b = 1$. Under the above priors, the marginal probability of y under each model can be computed in closed form, i.e.

$$p(y|\tilde{M}_j) \quad = \quad \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{|\mathbf{\Sigma}_j|^{1/2}}{|\mathbf{M}_j|^{1/2}} \frac{\Gamma(\tilde{a})}{\tilde{b}^{\tilde{a}}}$$

where n denotes the sample size, $\tilde{a} = 1 + \frac{n}{2}$, $\tilde{b} = 1 + \frac{1}{2}y^T y - \frac{1}{2}\hat{\theta}_j^T \mathbf{\Sigma}_j^{-1} \hat{\theta}_j$, $\mathbf{\Sigma}_j = \left(\mathbf{D}^T\mathbf{D} + \mathbf{M}_j^{-1}\right)^{-1}$, and $\hat{\theta}_j = \mathbf{\Sigma}_j \mathbf{D}^T y$. We call the Bayes Factor computed using the ratio of the closed form marginal probabilities as the "Analytical" Bayes Factor.

In our simulation we set $\sigma^2 = 1$, $n = 126$ and considered the null scenario (i.e. $\boldsymbol{\theta} = \mathbf{0}$). For a simulation with 1,000 replications, the summary of the differences between the Bayes Factor estimates (in relation to that obtained via the BIC approximation) are given in Table 5.4 below.

Table 5.4: Summary of the difference between the Bayes Factor estimates obtained from the BIC approximation to the estimates obtained from alternative estimation methods, namely, Prior Sampling, and Analytical.

Percentile	BIC - Prior	BIC - Analytical
5	-0.04	-1.41
25	0.01	-0.18
50	0.01	-0.05
75	0.03	-0.01
95	0.17	0.06

From the above Table 5.4, we observe that the difference between the BF estimate from BIC and those under the other methods are mostly close to 0. We conducted additional simulations with increasing subgroup effect size γ. The proportion of replications selecting M_1 instead of M_0 under the different methods are shown in Figure 5.1.

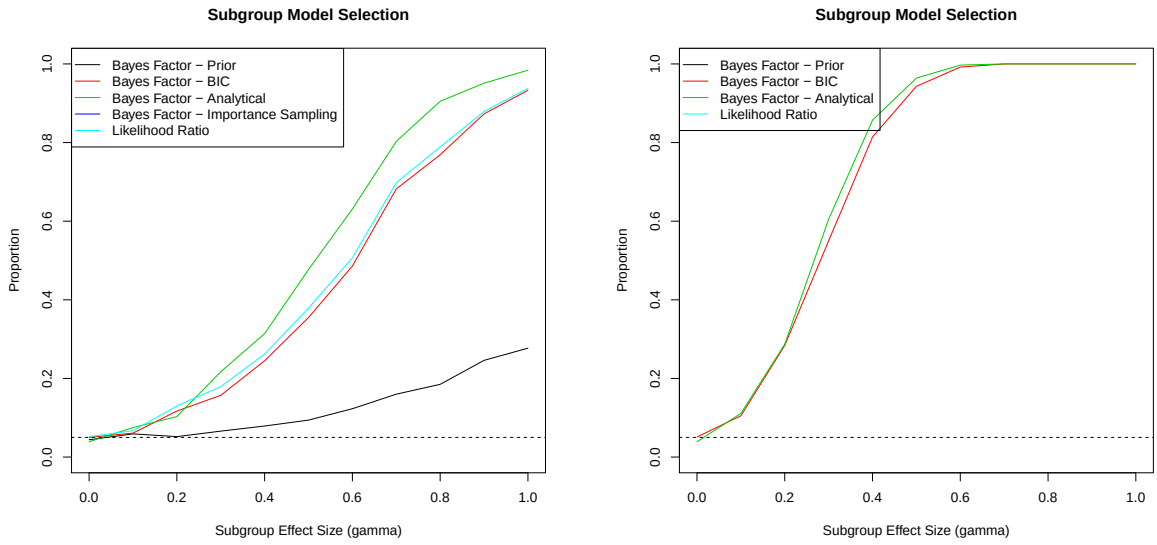


Figure 5.1: The proportion of replications selecting M_1 instead of M_0 under each estimation method for increasing effect size γ. Left panel: $n = 126$. Right panel: $n = 500$.

From Figure 5.1 (left panel), we note that the proportion of simulations selecting M_1 using BIC appears to be close to the proportions of selecting M_1 from the analytical Bayes Factor calculation and from the likelihood ratio test. The proportion is much lower when using prior samples for computing the Bayes Factor, due to the probabilities being close to 0 for many prior samples.

As discussed previously, the Bayes Factor estimate via BIC is consistent. We performed another set of simulation with the same setup using a larger sample size ($n = 500$). The results under the different methods are shown in Figure 5.1 (right panel). We observe that the proportion of simulations selecting M_1 using BIC appears to be very close to the proportions

of selecting M_1 from the analytical Bayes Factor calculation. Thus, the Bayes Factor estimate via BIC provides a close approximation to the true Bayes Factor.

The above simulations validate the use of the BIC as an approximation to the BF. This approximation is particularly useful in non-linear models for which the Bayes Factor is not available in closed form solution.

5.2.3 Bayesian Decision Approach to Reporting Subgroups

Once the model has been selected via Bayes factor, we proceed with the second step involving reporting of subgroup(s), if any. We view the prospective subgroup reporting as a Bayesian decision problem. Given the selected model M_j, let $\mathcal{K}_j$ denote the set of all possible subgroups under model M_j. In our illustration in Section 5.2.1, for example, under model M_6, $\mathcal{K}_6 = \{18, 19, 20, 21, 22, 23\}$. Thus, we consider all subgroups $k \in \mathcal{K}_j$. We note that the union of all the subgroups in $\mathcal{K}_j$ necessarily equals the overall population.

Let η_k denote the true treatment effect comparing treatment vs. control in subgroup k. Furthermore, let e_k denote the estimate of the treatment effect in subgroup k. Let $a_k \in \{0, 1\}$ denote whether to report subgroup k (a_k=1) or not (a_k=0). Let f_k denote the size of subgroup k relative to the overall population. We define the "benefit" for reporting subgroup k (v_k) as given below.

$$v_k \quad = \quad -(e_k - \eta_k)^2 + (\eta_k - 0)^2 \, (f_k)^{K_2}$$

where K_2 is a tuning parameter. Thus, v_k combines the inference loss by estimating η_k with e_k (which we will call the "uncertainty" part of the utility) as well as a payoff for reporting subgroup k if η_k is different from 0, penalized by the relative size and complexity of subgroup k (which we will call the "effect" part of the utility). Furthermore, let E_k denote the event whether the subgroup k is truly of interest (be it to report as beneficial or to flag as harmful), i.e. $E_k = 1$ if subgroup k is of interest and $E_k = 0$ otherwise. The event E_k is defined based on the parameter value η_k. For example, given a minimum clinically important difference

(MCID) denoted d, one possible event of interest is $E_k = 1$ if $|\eta_k| \geq d$ and $E_k = 0$ if $|\eta_k| < d$. The utility for the action of reporting subgroup k ($a_k = 1$) or not ($a_k = 0$) is then defined as follows

$$u_k = v_k 1[E_k = 1] a_k + K_1 1[E_k = 0](1 - a_k).$$

The posterior expected utility is then given by

$$\begin{aligned} U_k &= E[u_k] \\ &= a_k \int v_k 1[E_k = 1] d\pi(\boldsymbol{\theta}|y) + K_1(1 - a_k) P(E_k = 0|y) \end{aligned}$$

and this is maximized by action $a_k = 1$ when $\int v_k 1[E_k = 1] d\pi(\boldsymbol{\theta}|y) > K_1 P(E_k = 0|y)$. This is equivalent to reporting the subgroup k when $Q_k = \frac{\int v_k 1[E_k=1] d\pi(\boldsymbol{\theta}|y)}{P(E_k=0|y)} > K_1$. The quantity Q_k can be calculated via Monte Carlo integration. Note that in this case, the action a_k involves only flagging the subgroup k and does not specify whether the particular subgroup is benefited or harmed by the treatment. The assessment of benefit or harm can be done by assessing the direction of the estimate of the treatment effect e_k.

Note that the above procedure reports any subgroup $k \in \mathcal{K}_j$ where $Q_k > K_1$. Conceptually, K_1 represents a "trade-off" between falsely reporting a non-important subgroup and falsely not reporting an important subgroup. On the other hand, K_2 is associated with the penalty for subgroup size. The values of K_1 and K_2 can be either pre-defined by the investigators, or calibrated via simulations to achieve desired operating characteristics.

5.2.4 Calibration of Tuning Parameters

The tuning parameters in the utility function can be calibrated via simulations to achieve some desirable operating characteristics, such as meeting the desired Frequentist type-I error. The calibration of K_1 and K_2 is done a-priori to maximize the proportion of correctly reporting a subgroup under the condition that the average proportion of falsely reporting

any subgroup under the null is less than some number α. For calibration, pre-specification of E_k, MCID, and $f_k^{(0)}$, which is the minimum f_k of interest under η_k =MCID, is needed. The $f_k^{(0)}$ depicts the minimum relative size that is of interest, for the subgroup in which the treatment just barely demonstrates MCID.

In performing calibration, we assume that the number of baseline variables to be investigated is known. The calibration is done by generating datasets under two scenarios, each scenario having B (a large number) datasets. Under the null scenario we assume $\boldsymbol{\theta} = \mathbf{0}$. Under the alternative scenario, we simulate datasets assuming that there is a subgroup of size $f_k^{(0)}$ with an effect of size MCID, and that there is no effect in all of the other subgroups. Without loss of generality, we take the pre-determined subgroup for calibration to be a subgroup defined by one variable. We are also assuming equal variance among subgroups.

For a given K_2, the K_1 is selected such that the proportion of replications not selecting any subgroup under the null scenario is $(1-\alpha)$, where α is the desired type-I error.

Note that the type-I error α can be decomposed as

$$\begin{aligned}\alpha &= \sum_{j=0}^{M-1} P(\text{Select a subgroup}|\text{sel } M_j)P(\text{sel } M_j) \\ &= P(\text{Select a subgroup}|\text{sel } M_0)P(\text{sel } M_0) + P(\text{Select a subgroup}|\text{sel not } M_0)P(\text{sel not } M_0).\end{aligned}$$

One approach for calibration of K_1 given a value of K_2 is to obtain all the largest Q_k's (see Section 5.2.3 for the definition of Q_k) of the selected model under each replication under the null scenario and taking the $100(1-\alpha)$th quantile of the Q_k quantities. However, this approach may not be desirable because: (1). the quantity Q_k for the overall population does not depend on the value of K_2, and (2). the replications for which the overall model M_0 is not selected under the null tends to have a random high bias for the interaction effects, thus selecting one of the models with interaction. Thus, pooling them together and calibrating both K_1 and K_2 together may not be completely desirable as it leads to a more conservative selection with a lower type-I error and lower power.

Alternatively, we recommend (and we utilize) a separate threshold based on the selected

model in each replication. Specifically, we "split" the type-I error based on the selected model: $\alpha/2$ for calibrating $K_1^{(0)}$ when selecting M_0, and $\alpha/2$ for calibrating K_1 and K_2 when models other than M_0 are selected. In other words, we perform the calibration such that $P(\text{Select a subgroup}|M_0)P(M_0) = P(\text{Select a subgroup}|\text{not}M_0)P(\text{not}M_0) = \alpha/2$. We note that this is just one way of "splitting" the type-I error. The splitting of the type-I error into halves is intuitive and is an approach "naive" investigators and statisticians may take unless there are specific preferences based on specific interests. With more effort and thought, a better approach would likely consider the likelihood ratio (Bayes factor) that considered the "split" of power relative to the "split" of the type-I error, but that is outside the scope of this dissertation. Thus, under the null-scenario, $\alpha B/2$ replications are (incorrectly) selecting a subgroup under replications selecting model M_0 and $\alpha B/2$ replications are (incorrectly) selecting a subgroup under the replications not selecting M_0.

Under the null scenario, let $p_0^{(M)}$ denote the proportion of replications for which model M_0 is selected. Thus, under the null scenario, we select $K_1^{(0)}$ as the $100\left(1 - \frac{(\alpha/2)}{p_0^{(M)}}\right)$th quantile of the Q_0's, the quantities under the replications that select M_0.

Similarly, for a given value of K_2, the value of K_1 is selected as the $100\left(1 - \frac{(\alpha/2)}{1-p_0^{(M)}}\right)$th quantile of the quantities Q_k's for those replications not selecting model M_0. The values of K_1 and K_2 can be calibrated to have the proportion of replications selecting the subgroup of interest under the alternative scenario maximized. This is done using the forward slice procedure described in Chapter 3.

5.2.5 Inference Based on Utility

The inference following the analysis using the utility approach is based on the actions a_k's. Note that the actions a_k's represent only the decision that the subgroup(s) is to be of notice. However, we are not only interested in whether the subgroup(s) is important to notice, but more importantly whether the subgroup is important to notice due to benefit or harm from the treatment. We call this variable for reporting the benefit or harm (or no subgroup importance), the "report" r_k. Simply, r_k is guided by the action a_k and the direction of the

estimated effect in the subgroup k, i.e. $r_k = a_k \text{sign}(e_k)$.

Under this scheme, given that a particular model $M_j, j \neq 0$ was selected at the previous stage, it is possible to flag one or more of the subgroups, all of the subgroups, or none of the subgroups in $\mathcal{K}_j$. The subgroups that are flagged and reported for benefit are taken for the subsequent trials and investigations. On the other hand, the negatively reported subgroups are noted down for possible counter-indications of the treatment. For example, using our illustration in Section 5.2.1, suppose in the analysis of post–phase II trial, we selected model M_6, and obtained $(r_{18}, r_{19}, r_{20}, r_{21}, r_{22}, r_{23}) = (-1, -1, 0, 1, 0, 0)$. Thus, our recommended action would be to enroll subgroup 21 ($X_2 = 1, X_3 = 2$) for the subsequent phase III trial and to exclude subgroups 18 and 19 from indication (or put in for counter-indication). We note that although the estimate of the effects e_k is also obtained, we recommend to not use this estimate to power the subsequent trial. Following the results from Sanchez (2014) [58], it is inefficient to power the subsequent phase III trial based on the estimates obtained in the phase II trial. This is because the estimates from the accepted smaller phase II trial tends to be biased away from the null, and thus powering the subsequent trial based on this estimate would reduce the power of the subsequent phase III trial. Thus, the inference based on the analysis would consist only on reporting subgroup(s) (if any) that would be further investigated in a future study or trial.

We note the existence of other possible scenarios in which all of the subgroups in the model M_j are flagged and have the same directions (that is, same r_k). An example would be reporting $(r_{18}, r_{19}, r_{20}, r_{21}, r_{22}, r_{23}) = (-1, -1, -1, -1, -1, -1)$ or $(r_{18}, r_{19}, r_{20}, r_{21}, r_{22}, r_{23}) = (1, 1, 1, 1, 1, 1)$ under model M_6. Note, however, in such cases we are essentially reporting the overall population for harm or benefit, depending on the sign of r_k. Similarly, when all of the subgroups are not flagged under model M_j, we are essentially not reporting the overall population. The interpretation for reporting the overall population (or not reporting the overall population) under $M_j, j \neq 0$ and under M_0 is subtly different. The overall population reported under M_0 assumes that the effect is homogeneous across all subpopulations, whereas the overall population reported under $M_j, j \neq 0$ reports that there is some heterogeneity of

effects across subpopulations, but the effect in each subpopulation is significant enough to be of notice. Note, however, that in both cases, the action taken is the same, which is to enroll the overall population for the subsequent investigation.

In addition to the reported decision, the components of the utility may be of interest to the investigators. Specifically, the utility function incorporates the information regarding the uncertainty of the subgroup effect estimate, the effect size, the relative size of the subgroup, and the probability that the effect in the subgroup is deemed "important". Thus, two or more subgroups with differing values of the different components may have the same expected utility. For example, a smaller subgroup with a larger effect size may have the same utility as a larger subgroup with moderate effect size. An investigator may be interested in further examining the individual components of the utility (i.e. the effect size, the uncertainty of the effect size estimate, and the subgroup size) that lead to the reported decision. These components are available as by-products for the final decision, and can be reported alongside the final decision.

Given that the estimate of the effect size can be reported alongside the reported subgroup, it might seem "natural" to use these estimates in designing the subsequent studies. However, based on the investigations of Sanchez (2014) [58], careful considerations need to be exercised in using these estimates for the subsequent trial. Specifically, Sanchez (2014) [58] pointed out that the Phase-II trials that are "positive" tend to have effects that are biased away from the null. Thus, designing the subsequent trials based on these estimates lead to smaller trials that have reduced power. This problem is further exacerbated in conducting subgroup analysis, since the estimates are obtained from a smaller sample size, due to it being a subset of the sample. This might cause the estimate to have a higher uncertainty. The failure in incorporating the uncertainty of the estimates leads to loss of power in the subsequent studies. The methods for incorporating the uncertainty in designing a study is available, but outside of the scope of this dissertation.

5.3 Continuous Outcome – Linear Model

5.3.1 Overview - Bayesian Linear Model

We denote the subgroup-defining variable by X, and without loss of generality, let X be a binary variable such that the subgroups are defined by the 'negative subgroup' ($X = 0$) and the 'positive subgroup' ($X = 1$). We also denote the treatment variable by T, where $T = 0$ denotes the control group and $T = 1$ denotes the experimental group. Furthermore, we assume that the outcome y is normally distributed. For any observation i,

$$y_i = \alpha_0 + \beta_X X_i + \beta_T T_i + \gamma X_i T_i + \epsilon \tag{5.2}$$

where ϵ is the error term having a Normal(0,σ^2) distribution, and σ^2 is the variance of the error term. Thus, note that we work under the assumption of equal variances across subgroups. Extensions allowing for unequal variances can be made analogously.

As before, let the priors on the parameters $\boldsymbol{\theta}$ be a multivariate normal distribution with mean 0 and covariance matrix $\sigma^2\mathbf{M}_j$ depending on the model M_j, and the prior for σ^2 be an inverse-gamma(a, b) distribution. These choices are due to the conjugacy properties of these distributions leading to closed-form posterior distributions. We note that these assumptions may have large impact on (the comparability of) the distribution of unadjusted response across homogeneous vs. heterogeneous scenarios. Other prior distributions reflecting the a-priori belief of the investigators may be utilized instead. Estimation and inference could then be carried out utilizing Markov chain Monte Carlo methods.

Let $\mathbf{D}$ denote the design matrix, i.e. $\mathbf{D} = [\mathbf{1}\ T\ X\ XT]$. It can be shown that the posterior distribution of the regression parameters are given by

$$(\boldsymbol{\theta}|y, \sigma^2) \sim \text{Normal}(\hat{\boldsymbol{\theta}}, \sigma^2(\mathbf{D}^T\mathbf{D} + \mathbf{M}_j^{-1})^{-1}),$$

where $\hat{\boldsymbol{\theta}} = (\mathbf{D}^T\mathbf{D} + \mathbf{M}_j^{-1})^{-1}\mathbf{D}^T Y$, and that the posterior distribution of σ^2 has an inverse-

gamma($\tilde{a}, \tilde{b}$) distribution, where $\tilde{a} = a+n/2$ and $\tilde{b} = b+\frac{1}{2}Y^TY-\frac{1}{2}\hat{\boldsymbol{\theta}}^T(\mathbf{D}^T\mathbf{D}+\mathbf{M}_j^{-1})^{-1}\hat{\boldsymbol{\theta}}$. Based on the regression model, the treatment effect for subgroup $X = 0$ is given by β_T, whereas the treatment effect for subgroup $X = 1$ is given by $\beta_T + \gamma$. The posterior distributions of the treatment effects for both subgroups are also available in closed-form solutions as each is a linear combination of the model parameters.

5.3.2 Simulation Studies

We apply the method to the analysis of continuous outcomes using linear models. For our simulations, we assume that the minimum clinically important difference (MCID) is $d = 0.5$ units.

In our simulations, we also compared our utility approach with a few Frequentist approaches for subgroup selection, some of which are discussed in Sanchez (2014) [58]:

- Interaction-stratified analysis: This approach involves testing the full covariate-by-treatment interactions and determining which variable-interactions are significant. If no interaction is significant, the overall effect is tested for significance. Otherwise, stratified analyses are conducted for each stratum of the previously determined significant variables combination. This method is in light of the comments made by Pocock et al [54] [2] regarding interaction tests.

- Prefer Overall: This approach first looks for the significance of treatment effect in the overall population. If no significant treatment effect is observed in the overall population, we search for the subpopulations that might benefit.

- Smallest p-value: This approach looks for the subgroup for which the treatment effect appears to be the most "significant". Hence, we are conducting the tests on all of the subgroups (including overall population) and selecting the subgroup with the smallest p-value.

- Choice of Significant Subgroup: This approach first tests for the effect in the overall population. Regardless of its significance, it tests for the effects in the different subgroups. The approach reports the most significant subgroup if the effect in the complement subgroup is not significant [58]. For the dissertation, we will call this the "Choice" approach or the "Significant Subgroup" approach.

For each of the approaches considered above, we conduct correction for multiple testing such that the type-I error is preserved. We consider the case where there are three binary variables being considered, with X_2 prognostic of the outcome. Here, we take the effect of the prognostic variable to be $\beta_{X,2} = 0.7$. Without loss of generality, we assume that for each variable, the subgroups are of the same size on average.

The sample size for each "trial" in each replication is determined using a standard sample size calculation for the study. For the calculation of the sample size, the overall null hypothesis is that the overall treatment effect is 0. The alternative is that there is an overall effect of size d, and the variance in both groups is taken to be 1 unit. Under randomization, assuming equal probability of treatment assignment and assuming type I error of 0.05 and power of 0.80 to detect the alternative, the sample size for each treatment group is 63. Finally, we assume that the size of both subgroups are the same.

Under the setting of three binary variables as potential subgroup-definers, there are 27 different subgroups considered, shown in Table 5.5, under 8 different models, shown in Table 5.6.

We conducted simulation studies to assess the performance of the method with 1000 replications under six scenarios: (1) there is an overall treatment effect and no differential effect in the subgroups, (2) there is only effect in subgroup $X_3 = 1$ and zero treatment effect in the complement subgroup (subgroup 9), (3) there is no overall effect, due to the effect in subgroup $X_3 = 0$ and subgroup $X_3 = 1$ cancelling each other, with the effect in subgroup $X_3 = 1$ being positive, (4) there is a positive effect in subgroups with $X_2 = 1$ or $X_3 = 1$, and (5) there is a positive effect in subgroups with $X_1 = 1$ or $X_3 = 1$. Implicit and embedded in

each scenario is scenario (0), where there is no treatment effect in any of the subgroups.

Also note that there is a distinction in what the value of σ^2 represents in each scenario. In scenarios (0) and (1), σ^2 corresponds to both the variance of the response variable in the overall population and in each of the subgroups. On the other hand, in scenarios (2) and (3), σ^2 corresponds to the variance of the response variable in subgroup $X_3 = 0$ and in subgroup $X_3 = 1$. In scenario (4), σ^2 corresponds to the variance of the response variable in subgroups defined by combinations of values of X_2 and X_3, and similar is true for scenario (5). In other words, the response variable y is generated from a $N(\mathbf{X}\boldsymbol{\beta}, \sigma^2)$ distribution under the "correct" model in each scenario. In the latter scenarios, the variance in the overall population will be greater than σ^2.

Furthermore, we define $E_k = 0$ when $-d<\eta_k<d$ and thus $E_k = 1$ when $\eta_k \geq d$ or $\eta_k \leq -d$. The calibration of the tuning parameters is done to attain the type I error rate $\alpha = 0.05$, using $B = 1,000$ simulated replications under the null as described in Section 5.2.4. The calibrated tuning parameters are $(K_1^{(0)}, K_1, K_2) = (0.159, 0.320, 1.960)$.

Table 5.5: Enumeration of subgroups based on three variables.

Subgroup	(Sub)Population
0	Overall
1	X_1=0
2	X_1=1
3	X_2=0
4	X_2=1
5	X_1=0,X_2=0
6	X_1=1,X_2=0
7	X_1=0,X_2=1
8	X_1=1,X_2=1
9	X_3=0
10	X_3=1
11	X_1=0,X_3=0
12	X_1=1,X_3=0
13	X_1=0,X_3=1
14	X_1=1,X_3=1
15	X_2=0,X_3=0
16	X_2=1,X_3=0
17	X_2=0,X_3=1
18	X_2=1,X_3=1
19	X_1=0,X_2=0,X_3=0
20	X_1=1,X_2=0,X_3=0
21	X_1=0,X_2=1,X_3=0
22	X_1=1,X_2=1,X_3=0
23	X_1=0,X_2=0,X_3=1
24	X_1=1,X_2=0,X_3=1
25	X_1=0,X_2=1,X_3=1
26	X_1=1,X_2=1,X_3=1

Table 5.6: Categories of the subgroups based on the variables defining the subgroup. Each category correspond to a model.

Model	Subgroup-defining Variable	Subgroups
0	None	0
1	X_1	1,2
2	X_2	3,4
3	X_1,X_2	5,6,7,8
4	X_3	9,10
5	X_1,X_3	11,12,13,14
6	X_2,X_3	15,16,17,18
7	X_1,X_2,X_3	19,20,21,22,23,24,25,26

For this simulation study, we restrict the search for subgroups under the "Prefer Overall" and "Smallest p-value" approaches to be among subgroups defined by one variable and not the higher-order combinations between the variables (i.e. subgroups 0, 1, 2, 3, 4, 9, and 10 only), which reduces the number of tests conducted under these approaches.

The simulation results are displayed in Figure 5.2 by scenarios. Furthermore, Table 5.7 shows the proportion of replications selecting the overall population under the different approaches when the overall effect is MCID for scenarios 1 through 3. Table 5.8 shows the proportion of replications selecting the "correct" decision under the different approaches when the effect in the subgroups is at MCID in scenarios 4-5. We next describe the main results by scenario.

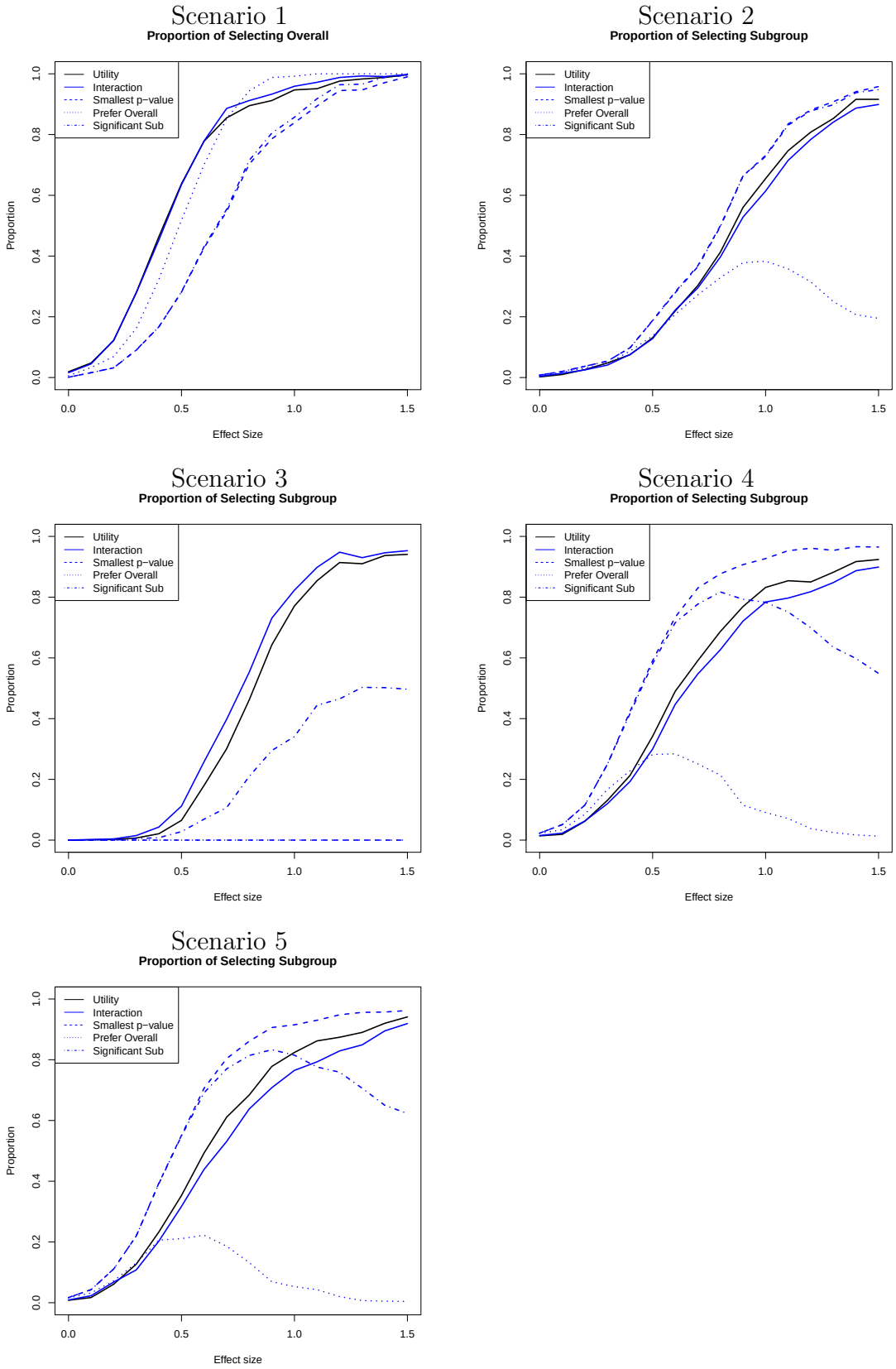


Figure 5.2: Proportion of replications selecting the overall population for benefit under different approaches and simulation scenarios.

Table 5.7: Type-1 error and the power at MCID under different approaches and simulation scenarios 1-3.

Approach	Type-I Error	Power at MCID
Scenario 1		
Utility	0.046	0.638
Interaction	0.052	0.635
Smallest p	0.050	0.281
Prefer Overall	0.051	0.518
Choice	0.051	0.282
Scenario 2		
Utility	0.049	0.131
Interaction	0.052	0.128
Smallest p	0.050	0.195
Prefer Overall	0.051	0.137
Choice	0.051	0.194
Scenario 3		
Utility	0.046	0.065
Interaction	0.052	0.112
Smallest p	0.050	0
Prefer Overall	0.051	0
Choice	0.055	0.028

5.3.2.1 Scenario 1: Homogeneous overall effect

In this scenario we performed simulations with 1000 replications under the case of no subgroup effect ($\gamma_3 = 0$), and varying the overall effect size (β_T) from 0 to 1.5. This corresponds to a relative effect size of 0 to 3 compared to the MCID (d).

The simulation results displayed in Figure 5.2 and Table 5.7 show that the "Utility", "Interaction", and "Prefer Overall" approaches appear to have higher power for detecting an overall effect at MCID. Furthermore, the power for reporting the overall population increases for all of the methods as the effect size grows. The power is low for the "Smallest p-value" and the "Choice" approaches. This is not unexpected for the "Smallest p-value" approach, as the approach looks for the subgroup with the smallest p-value. In this case, when the overall treatment effect increases, the treatment effect in each of the subgroups

Table 5.8: Type-1 error, of making a "correct" decision, and the proportion of selecting overall population at MCID under different approaches and simulation scenarios 4-5

Approach	Type-I Error	Power at MCID	Proportion Selecting Overall
Scenario 4			
Utility	0.038	0.342	0.360
Interaction	0.052	0.300	0.436
Smallest p	0.052	0.591	0.049
Prefer Overall	0.055	0.282	0.408
Choice	0.055	0.581	0.029
Scenario 5			
Utility	0.036	0.353	0.353
Interaction	0.052	0.317	0.430
Smallest p	0.050	0.551	0.005
Prefer Overall	0.051	0.211	0.439
Choice	0.051	0.549	0.021

also increases. Thus, by chance alone, one or more of the subgroups may exhibit "higher significance" than the overall population. The power for detecting the overall population grows to 1 under the "Choice" approach as the effect size grows. However, the power is still low at MCID.

5.3.2.2 Scenario 2: Effect only in subgroup $X_3 = 1$

In this scenario we performed simulations with 1000 replications under the case where there is a positive effect in subgroup $X_3 = 1$ and no effect in subgroup $X_3 = 0$. In this case, we vary the treatment effect in subgroup $X_3 = 1$ from 0 to 1.5.

The simulation results displayed in Figure 5.2 and Table 5.7 show that the "Utility" approach has a slightly lower type-I error rate. The power at MCID for detecting a subgroup effect under the "Utility", "Interaction", and "Prefer Overall" methods are lower compared to the power under the "Smallest p-value" and the "Choice" approaches. Furthermore, the power for reporting the overall population increases for all of the methods as the effect size grows, with the exception of the "Prefer Overall" approach. This is due to that as the

treatment effect in the subgroup increases, the treatment effect in the overall population, which is half of the treatment effect of the subgroup, also increases. For large effect sizes, the treatment effect in the overall population becomes significant enough that the "Prefer Overall" approach does not conduct any subgroup analysis afterwards.

5.3.2.3 Scenario 3: Effect in subgroup $X_3 = 1$, negative effect in subgroup $X_3 = 0$, zero overall

We performed simulations with 1000 replications under the case where there is a positive effect in subgroup $X_3 = 1$ and a negative effect in subgroup $X_3 = 0$, averaging to a zero effect in the overall population. In this case, we vary the treatment effect in subgroup $X_3 = 1$ from 0 to 1.5, and the treatment effect in subgroup $X_3 = 0$ is negative of equal magnitude.

The simulation results showing the proportion of replications making the "correct" decision (defined by selecting $X_3 = 1$ for benefit and $X_3 = 0$ for harm) under each approach are displayed in Figure 5.2 and Table 5.7. We note that the "Utility" approach has slightly lower Type-I error rate. Moreover, we observe that both the "Utility" and the "Interaction" approaches have an increasing power in reporting the "correct" decision. Both the "Smallest p-value" and "Prefer Overall" methods have zero power of making the "correct" decision as they only report one subgroup at the end. The "Choice" approach has a lower power of detecting subgroup $X_3 = 1$ for benefit and $X_3 = 0$ for harm compared to the utility and the interaction approaches.

5.3.2.4 Scenario 4: Effect in subgroups defined by $X_2 = 1$ or $X_3 = 1$

We performed simulations under the case where there is a positive effect in subgroups defined by $X_2 = 1$ or $X_3 = 1$, and 0 otherwise. In this case, we vary the treatment effect from 0 to 1.5.

The simulation results showing the proportion of replications making the "correct" decision, that is, selecting a beneficial subgroup and not selecting any subgroup with null effect,

under each approach are given in Figure 5.2. The proportion of replications with the "correct" decision under the different approaches when the effect in the subgroups defined by $X_2 = 1$ or $X_3 = 1$ is at MCID is given in Table 5.8.

We observe that the "Smallest p-value" and the "Choice" approaches make a "correct" decision more often than under the other methods. This is due to the fact that three-quarters of the population has an effect, whereas only one-quarter of the population has no effect. This causes the "Utility", "Interaction", and "Prefer Overall" effect to have a select the overall population more often than a subpopulation. The "Smallest p-value" approach, on the other hand, selects a subgroup more often because the treatment effect is expected to be "more significant" in a subgroup rather than in the overall population where the treatment effect is "diluted".

5.3.2.5 Scenario 5: Effect in subgroups defined by $X_1 = 1$ or $X_3 = 1$

We performed simulations under the case where there is a positive effect in subgroups defined by $X_1 = 1$ or $X_3 = 1$, and 0 otherwise. In this case, we vary the treatment effect from 0 to 1.5.

The simulation results showing the proportion of replications making the "correct" decision, that is, selecting a beneficial subgroup and not selecting any subgroup with null effect, under each approach are given in Figure 5.2. The proportion of replications with the "correct" decision under the different approaches when the effect in the subgroups defined by $X_1 = 1$ or $X_3 = 1$ is MCID is given in Table 5.8.

As expected, we observe the phenomenon similar to that under scenario 4 in that the "Smallest p-value" approach makes a "correct" decision more often than the other methods. This is again due to the fact that three-quarters of the population has an effect, whereas only one-quarter of the population has no effect. This causes the "Utility", "Interaction", and "Prefer Overall" approaches to select more often the overall population rather than a subpopulation. And, as before, the "Smallest p-value" approach, on the other hand, selects a subgroup more often because the treatment effect is expected to be "more significant" in

a subgroup rather than in the overall population where the treatment effect is "diluted".

5.3.2.6 Incorporation of Prior Information

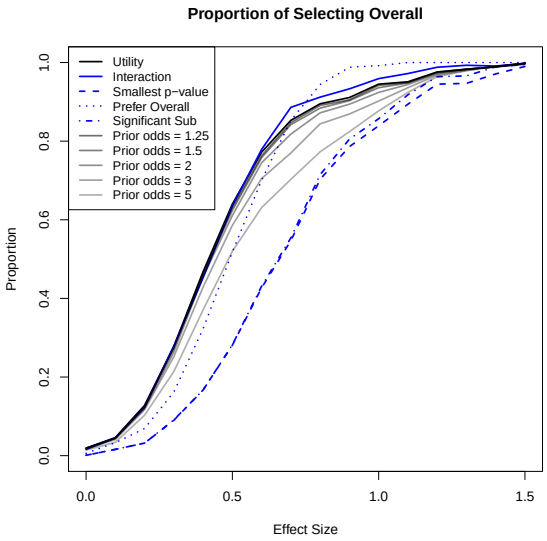
In the utility approach priors can be utilized to guide model selection, and, thus, subgroup reporting. The prior information may come into the selection via the values of o_j's (introduced in Section 5.2.2.2 as prior odds) in the model selection process as well as in terms of the priors for the specific parameters within a given model.

We conducted some simulation studies to assess the performance of the utility approach under different values of the prior odds o_j's and priors on the model parameters. Our simulation setup is the same as in scenarios (1) and (2) before, where we consider X_2 as a prognostic variable and X_3 as an important predictive (subgroup-defining) variable.

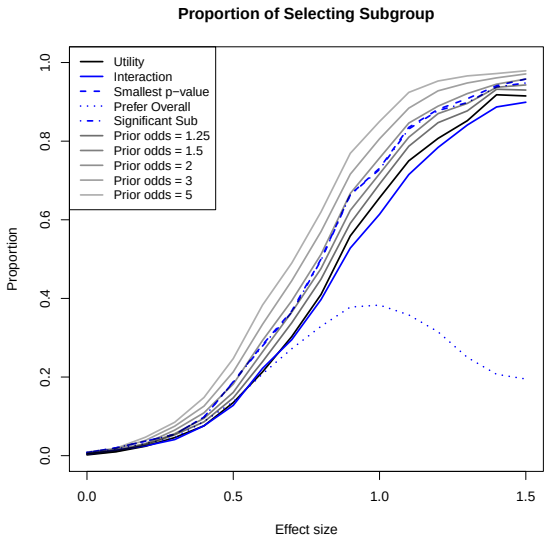
We conducted the simulations under two different conditions: 1. We assign a prior on the correct model M_4 with γ_3 under M_4 distributed as $N(0.5, 100)$ and 2. we assign a prior on the incorrect model M_2, where the prior for γ_2 under M_4 is distributed as $N(0.5, 100)$. Basically, we changed the prior for the treatment effect interaction such that the prior is centered at the MCID, but still is quite non-informative. We set that the priors for the other models (other than those under investigation) is the same as model M_0. Further, we consider the prior odds values of 1, 1.25, 1.5, 2, 3, and 5.

The simulation results are shown in Figure 5.3 and Tables 5.9 and 5.10. We discuss next some of the results.

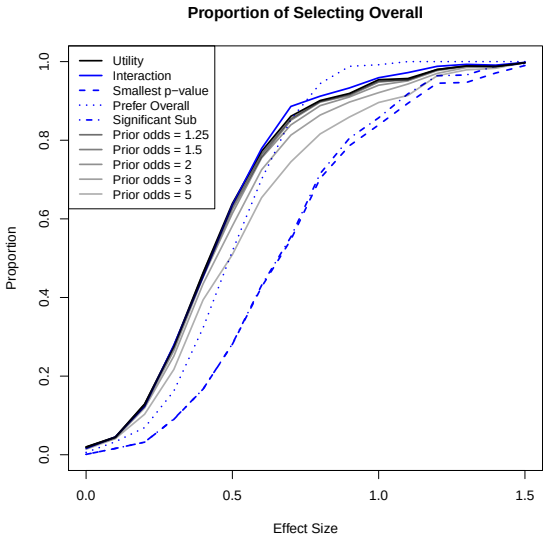
Scenario 1, prior on correct model



Scenario 2, prior on correct model



Scenario 1, prior on wrong model



Scenario 2, prior on wrong model

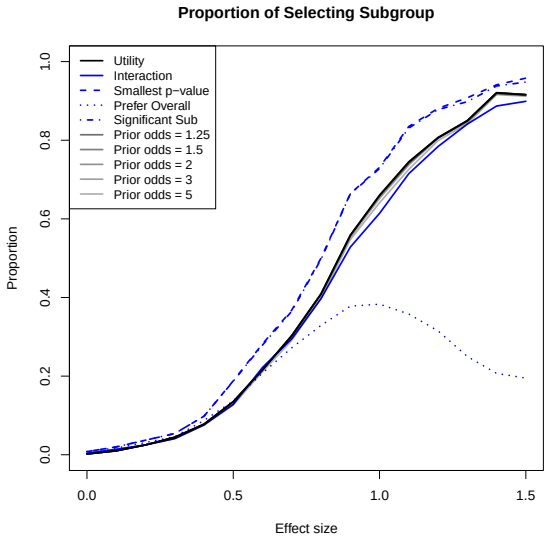


Figure 5.3: Proportion of replications selecting the overall population under the different approaches.

Prior on correct model.

The simulation results shown in Figure 5.3 (top panels) and Table 5.9 indicate that as

Table 5.9: Type-1 error and the proportion of replications selecting the overall population at MCID under different approaches.

Approach	Type-I Error	Power at MCID
Scenario (1), prior on correct model:		
Utility, $o_4 = 1$	0.049	0.640
Utility, $o_4 = 1.25$	0.053	0.631
Utility, $o_4 = 1.5$	0.053	0.624
Utility, $o_4 = 2$	0.052	0.611
Utility, $o_4 = 3$	0.052	0.582
Utility, $o_4 = 5$	0.052	0.520
Interaction	0.052	0.635
Smallest p	0.050	0.281
Prefer Overall	0.051	0.518
Choice	0.051	0.282
Scenario (2), prior on correct model		
Utility, $o_4 = 1$	0.049	0.135
Utility, $o_4 = 1.25$	0.050	0.147
Utility, $o_4 = 1.5$	0.051	0.161
Utility, $o_4 = 2$	0.050	0.182
Utility, $o_4 = 3$	0.051	0.213
Utility, $o_4 = 5$	0.052	0.247
Interaction	0.052	0.128
Smallest p	0.050	0.195
Prefer Overall	0.051	0.137
Choice	0.051	0.194

we place a more informative prior on the correct model M_4, the proportion of replications selecting the overall population for benefit under scenario (1) decreases, while the power for reporting the subgroup under scenario (2) increases. Further, we note that the type-1 error also increases as we place more informative priors on M_4. Recall that the K_1 and K_2 have been calibrated to achieve the desired pre-specified operating characteristics in a setting where the priors were less informative. The "increase" is due to replications for which the Bayes Factor is comparatively larger, and thus, they are more likely to have imbalance in the treatment effects across the two subgroups. Hence, there is a higher probability of selecting a subgroup, resulting in an increase in the type-1 error. However, note that the desired type-I

Table 5.10: Type-1 error and proportion of replications selecting the overall population at MCID under different approaches.

Approach	Type-I Error	Power at MCID
Scenario (1), prior on incorrect model:		
Utility, $o_2 = 1$	0.041	0.639
Utility, $o_2 = 1.25$	0.042	0.633
Utility, $o_2 = 1.5$	0.042	0.625
Utility, $o_2 = 2$	0.045	0.614
Utility, $o_2 = 3$	0.048	0.582
Utility, $o_2 = 5$	0.056	0.509
Interaction	0.052	0.635
Smallest p	0.050	0.281
Prefer Overall	0.051	0.518
Choice	0.051	0.282
Scenario (2), prior on incorrect model:		
Utility, $o_2 = 1$	0.043	0.135
Utility, $o_2 = 1.25$	0.044	0.135
Utility, $o_2 = 1.5$	0.044	0.135
Utility, $o_2 = 2$	0.047	0.135
Utility, $o_2 = 3$	0.049	0.133
Utility, $o_2 = 5$	0.056	0.127
Interaction	0.052	0.128
Smallest p	0.050	0.195
Prefer Overall	0.051	0.137
Choice	0.051	0.194

error is attained for larger prior ratio o_4 values.

Prior on incorrect model

Similarly, we conducted simulations under the case where we place a more informative prior on the incorrect model. Figure 5.3 (bottom panels) shows the proportion of replications selecting the overall population for benefit under scenario (1) and the proportion selecting the subgroup $X_3 = 1$ for benefit under scenario (2). Furthermore, the type-1 error and the power at MCID for the different approaches under both scenarios are shown in Table 5.10.

As we place a more informative prior on model M_2, the proportion of replications selecting the overall population for benefit under scenario (1) decreases, while the power for reporting

the subgroup under scenario (2) decreases by a small amount. We note that the type-1 error also increases. Further, the reduction in power for detecting the subgroup under scenario (2) for the utility approach appears to be small, even when the prior odds o_2 is much larger.

5.3.2.7 Prior Distributions on Subgroups

In our method, priors are specified for the model and for the parameters under each model. This implies the prior distributions on the overall population and the subgroups. In our simulation studies, the model parameters are normally distributed a priori. Thus, given a model, the prior distribution for the treatment effect under each subgroup is also normally distributed. Hence, the prior distribution of the treatment effect under each subgroup is a mixture of Normal distributions, with the prior model probabilities as the mixture weights.

For illustration, consider our simulation setting where we placed a higher prior probability on model M_4 with $o_4 = 3$ and placing a prior distribution centered at d instead of at 0 for the subgroup $X_3 = 1$. Since the proportion of 1's for each variable is 0.5, the priors on the overall treatment effect and the subgroup $X_3 = 1$ under each model are then normal distributions with means given in Table 5.11.

Table 5.11: Prior means of the treatment effects in the overall population and in subgroup $X_3 = 1$ under each model

Model	Overall Population	Subgroup $X_3 = 1$
0	0.00	0.00
1	0.00	0.00
2	0.00	0.00
3	0.00	0.00
4	0.25	0.50
5	0.25	0.50
6	0.25	0.50
7	0.25	0.50

Thus, the prior distributions of the treatment effect in the overall population (π_0) and in the subgroup $X_3 = 1$ (π_{10}) are mixtures of normal distributions with means:

$$
\begin{aligned}
E[\pi_0] &= \frac{1}{10}0 + \frac{1}{10}0 + \frac{1}{10}0 + \frac{1}{10}0 + \frac{3}{10}0.25 + \frac{1}{10}0.25 + \frac{1}{10}0.25 + \frac{1}{10}0.25 \\
&= 0.15. \\
E[\pi_{10}] &= \frac{1}{10}0 + \frac{1}{10}0 + \frac{1}{10}0 + \frac{1}{10}0 + \frac{3}{10}0.5 + \frac{1}{10}0.5 + \frac{1}{10}0.5 + \frac{1}{10}0.5 \\
&= 0.30.
\end{aligned}
$$

5.3.2.8 Sensitivity of Results

We conducted simulations to assess the sensitivity of the reported decisions to various factors: relative subgroup size, σ, and normality of errors.

Sensitivity to σ

We conducted simulations to assess the sensitivity of the reported decisions to the value of σ. In these simulations, the effect size is fixed at d and we consider the σ values ranging from 1 to 4, and we consider the case under scenario (2) as before. The sensitivity of the reported decision is shown in Figure 5.4.

We observe that the proportion of replications reporting the subgroup decreases as the value of σ increases. This is expected, as the value of σ affects the "inference loss" component of the utility. Furthermore, as the value of σ increases, the negative "inference loss" component also increases, leading to lower expected utility, thus leading to lower proportion of reporting.

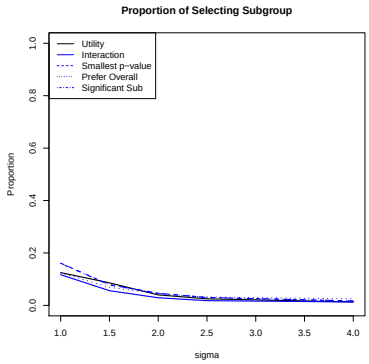


Figure 5.4: Proportion of replications selecting the subgroup for various values of σ.

Sensitivity to relative subgroup size

We conducted simulations to assess the sensitivity of the reported decisions to the relative subgroup size. In these simulations, the effect size is fixed at d and σ fixed at 1, and we consider the relative size of the effective subgroup ranging from 0.2 to 0.8, and we consider the case under scenario (2) as before. The sensitivity of the reported decision is shown in Figure 5.5.

Selecting subgroup $X_3 = 1$ Selecting overall population

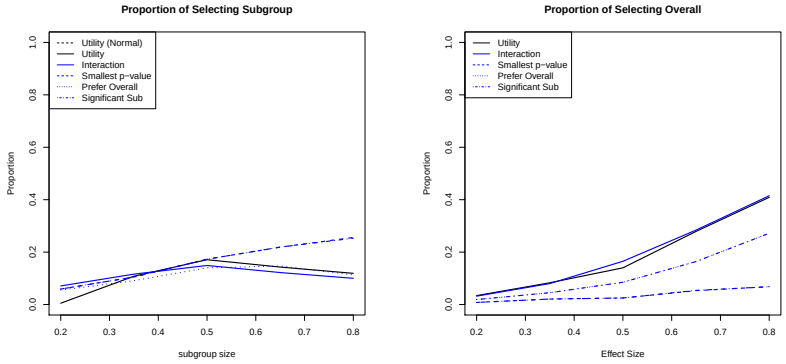


Figure 5.5: Proportion of replications selecting the subgroup and the overall population for various relative subgroup sizes.

We observe that the proportion of replications reporting the subgroup decreases as the relative subgroup size decreases. We also observe that the proportion of selecting the subgroup also is lower for larger relative size close to 0.8. This is due to a higher proportion of selecting the overall population as the effective subgroup makes up larger proportion of the overall population. In addition, we also observe that for all the methods, the proportion of replications selecting the overall population increases as the relative subgroup size increases. We conducted further simulations in the same setup but placing a higher prior probability for the subgroup, specifically with $o_4 = 5$. The results are given in 5.6.

Selecting subgroup $X_3 = 1$ Selecting overall population

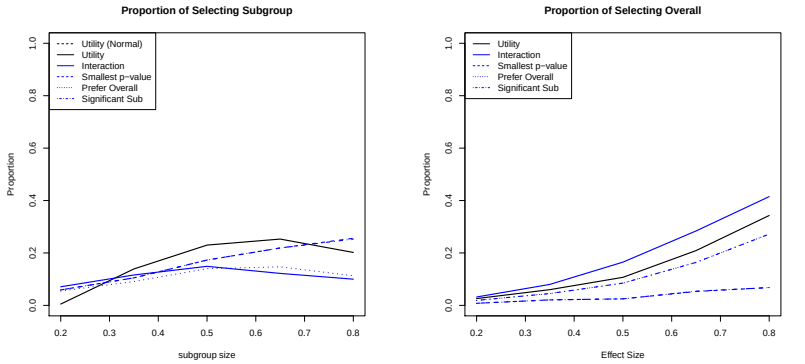


Figure 5.6: Proportion of replications selecting the subgroup and the overall population for various relative subgroup sizes for $o_4 = 5$.

We observe that the proportion of replications reporting the subgroup there is more of a trend that the proportion of selecting the subgroup increases as the relative subgroup size increases. This is due to a higher prior probability placed on selecting the subgroup.

Sensitivity to normality of errors

We conducted simulations to assess the sensitivity of the reported decisions to the distribution of errors. In these simulations, the effect size is fixed at d, while the errors are distributed $U(-\sqrt{3}, \sqrt{3})$, such that the errors have mean 0 and variance 1. The sensitivity

of the reported decision is shown in Figure 5.7.

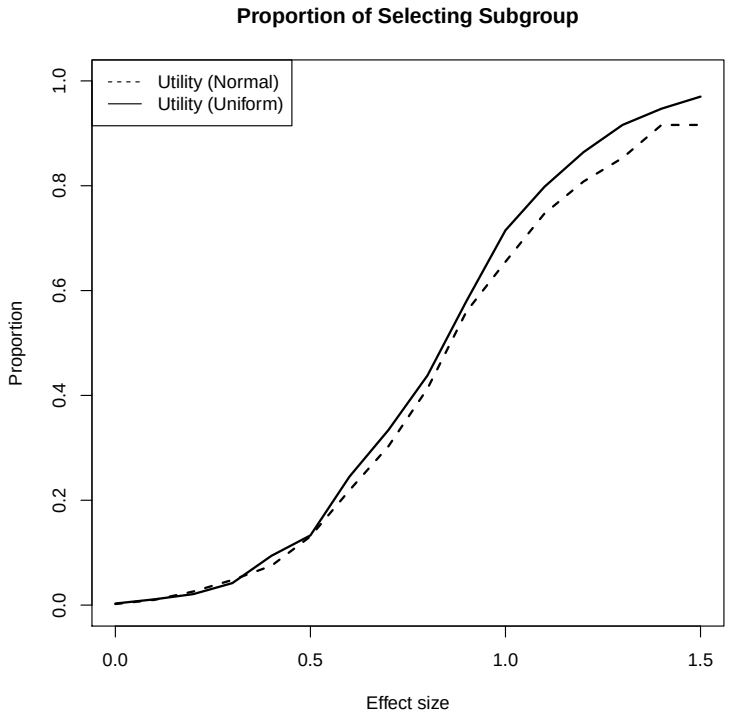


Figure 5.7: Proportion of replications selecting the subgroup for uniformly distributed errors.

We observe that the reported decisions using the utility approach do not appear to differ dramatically under uniformly distributed errors. Thus, we observe that the decisions appear to be quite robust to the distributional assumption of the errors.

5.3.2.9 Utility Components and Decision

The decision as to report or not a given subgroup is a based on whether the expected utility is above the calibrated threshold. The utility function in our case contains three components: the inference loss, the effect size, and the relative subgroup size. We conducted an inspection of the reported decisions in all of our simulated results to assess whether one or more of the

components appear to have a higher contribution in driving the decision to report or not report a given subgroup. Specifically, we focus our attention to subgroup $X_3 = 1$. Here, we pooled all of our simulations from all of the scenarios.

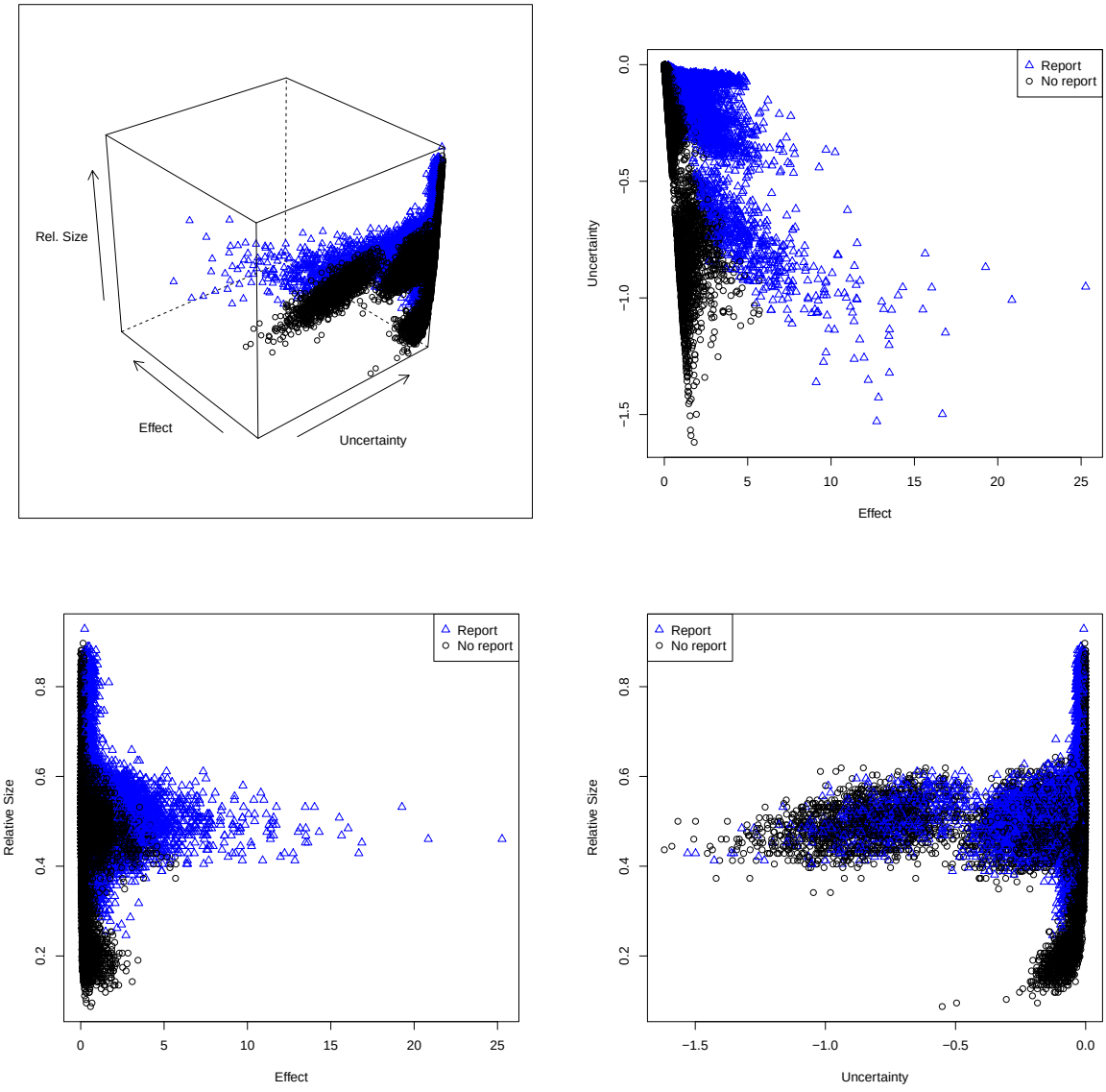


Figure 5.8: The reported decision and the utility components for all of the scenarios.

From Figure 5.8, we observe that the decision appears to be driven more by the effect

size and the inference loss than the relative size.

5.4 Binary Outcome

5.4.1 Overview

Denote the subgroup-defining variable by X, and without loss of generality, let X be a binary variable such that the subgroups are defined by the 'negative subgroup' ($X = 0$) and the 'positive subgroup' ($X = 1$). Denote the treatment variable by T, where $T = 0$ denotes the control group and $T = 1$ denotes the experimental group. Furthermore, the outcome $y \sim$ Bernoulli(p), where p denotes the probability of a positive (beneficial) outcome. For any observation i,

$$g(p_i) \quad = \quad \alpha_0 + \beta_X X_i + \beta_T T_i + \gamma X_i T_i \tag{5.3}$$

where g is the link function. Some common link functions (g) are the 'logit' link and the 'probit' link. The likelihood is given by

$$l(y|\boldsymbol{\theta}) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \tag{5.4}$$

where $p_i = g^{-1}(\mathbf{D}_i\boldsymbol{\theta})$ and $\mathbf{D}_i$ is the i^{th} row of the design matrix $\mathbf{D}$, as defined before in Section 5.3. The posterior distribution of $\boldsymbol{\theta}$ is proportional to the prior distribution $\pi(\boldsymbol{\theta})$ times the likelihood $l(y|\boldsymbol{\theta})$.

$$p(\boldsymbol{\theta}|y) \propto l(y|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{5.5}$$

Similar to Section 5.3, a measure of the within-subgroup treatment effects for subgroups defined by $X = 0$ and $X = 1$ are given by β_T and $\beta_T + \gamma$ respectively. However, the posterior distribution is not available in closed form and, thus, we utilize Markov Chain Monte Carlo (MCMC) methods to draw samples from the posterior distribution of the model parameters

and corresponding functions of interest such as those defining subgroup treatment effects.

5.4.2 Simulation Studies

We consider a simulation setup similar to that presented in Section 5.3.2 in the continuous outcome setting, considering three baseline variables. For the binary outcome, we assume that a "success" ($y = 1$) is considered desirable, and the event of "failure" is considered undesirable. We assume a logistic model, for which under the null, the probability of success on both the control and the treatment group is 0.5. We consider the MCID to be an improvement to a probability of success of 0.7. Thus, on the model parameter scale, the MCID corresponds to an effect in the scale of the difference in log odds of $\text{logit}(0.7) = 0.847$. For a study with level 0.05 and 80% power to detect the MCID, the sample size required is calculated to be 180.

Similar to the investigations in Section 5.3.2, we consider two scenarios: (1) there is a homogeneous overall effect, and (2) there is an effect in the subgroup defined by $X_3 = 1$, but zero in the other subgroup. We compare the utility approach with alternative approaches for subgroup analysis under the logistic model. A simulation with 1000 replications were conducted. The calibrated tuning parameters are $(K_1^{(0)}, K_1, K_2) = (0.463, 2.787, 0.931)$. The simulation results are shown below in Figure 5.9 and Table 5.12.

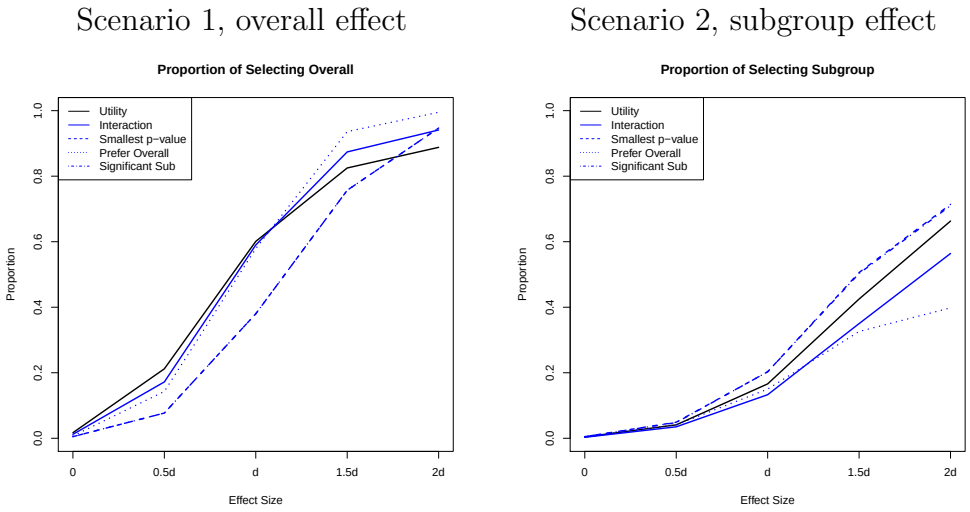


Figure 5.9: Proportion of replications selecting the overall population under the different approaches.

Table 5.12: Type-I error and power of detecting MCID under the two scenarios for the various approaches

		Scenario (1)	Scenario (2)
	Type-I Error	Overall Power at MCID	Subgroup Power at MCID
Utility	0.049	0.601	0.166
Interaction	0.045	0.590	0.133
Smallest p	0.043	0.380	0.203
Prefer Overall	0.043	0.579	0.150
Choice	0.043	0.380	0.203

We observe in this case that the "Utility" and "Prefer overall" approaches have lower power for detecting the subgroup in scenario (2) compared to the "Smallest p-value" and the "Choice of significant subgroup" approaches. The "Interaction" approach appears to have the lowest power of detecting the subgroup effect under scenario (2), but the highest power in detecting the overall population under scenario (1). Further, we note the same trade-off between power for detecting the overall population under scenario (1) and the power for

detecting the subgroup under scenario (2) as observed in the linear model setting also applies to the logistic model setting.

5.5 Censored Outcome

5.5.1 Overview

In this section we consider the application of the utility method to a right-censored survival outcomes. As in previous sections we denote the subgroup-defining variable by X which, without loss of generality, is binary and such that the 'negative subgroup' has $X = 0$ while the 'positive subgroup' has $X = 1$. We also denote the treatment variable by T, where $T = 0$ denotes the control group whereas $T = 1$ denotes the experimental group. Furthermore, assume the outcome y denote the time-to-event variable, under (non-informative) administrative censoring. We assume a parametric distribution with $y \sim Weibull(k, \lambda)$, where k denotes the shape parameter and λ is the scale parameter. Thus, the baseline hazard (i.e. when $\boldsymbol{\theta} = \mathbf{0}$) is

$$h_0(y|\boldsymbol{\theta} = \mathbf{0}) \quad = \quad \frac{k}{\lambda}\left(\frac{y}{\lambda}\right)^{k-1} \tag{5.6}$$

and so

$$h(y|\alpha_0, \beta_X, \beta_T, \gamma) \quad = \quad h_0(y)\exp\left(\alpha_0 + \beta_X X + \beta_T T + \gamma XT\right) \tag{5.7}$$

Similar to Section 5.3, the treatment effect quantified as the log hazard ratio in the negative subgroup is β_T while in the positive subgroup is given by $\beta_T+\gamma$. As with the binary outcomes, there is no closed form solutions of the posterior distribution of the model parameters. Thus, we utilize MCMC methods to draw samples from the posterior distribution of the model parameters and corresponding functions of interest such as those defining subgroup treatment effects.

5.5.2 Simulation Studies

We considered a simulation setup similar to that presented in Section 5.3.2 in the continuous outcome setting, considering three baseline variables. For the censored outcome, we assumed that an event is considered undesirable (e.g. death). We assume that the baseline hazard function is from the Weibull(scale=1,shape=1.3) distribution. We consider the MCID to be a hazard ratio of 2/3. For a study with level 0.05 and 80% power to detect the MCID, the number of events required is 190. We consider enrolling 200 subjects and take the administrative censoring time to be such that 190 events are observed.

Similar to the investigations performed in Section 5.3.2, we considered two scenarios: (1) there is a homogeneous overall effect, and (2) there is an effect in the subgroup defined by $X_3 = 1$, but zero in the other subgroup. We compared the utility approach with alternative approaches for subgroup analysis using the Cox model. A simulation with 1,000 replications were conducted. The calibrated tuning parameters are $(K_1^{(0)}, K_1, K_2) = (0.146, 0.161, 2.381)$. The simulation results are shown below in Figure 5.10 and Table 5.13

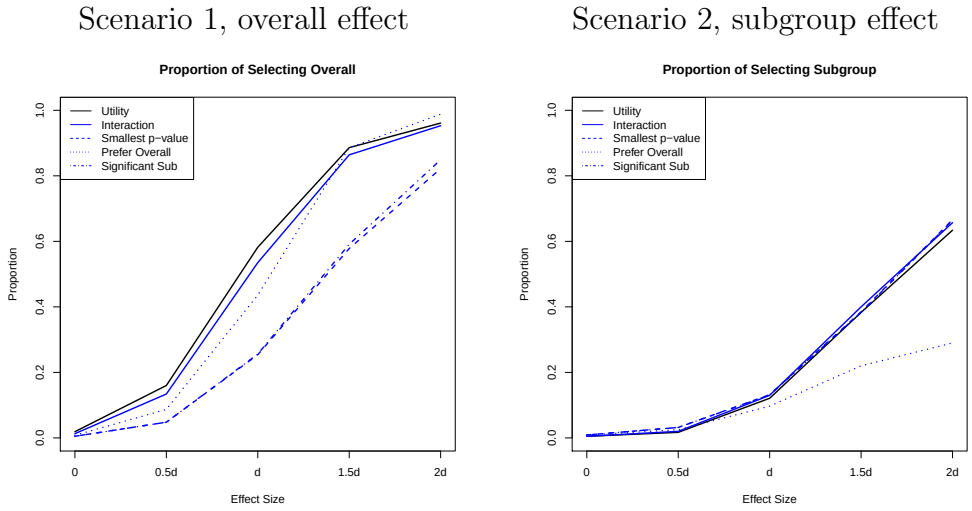


Figure 5.10: Proportion of replications selecting the overall population under the different approaches.

Table 5.13: Type-I error and power of detecting MCID under the two scenarios for the various approaches

		Scenario (1)	Scenario (2)
	Type-I Error	Overall Power at MCID	Subgroup Power at MCID
Utility	0.048	0.582	0.121
Interaction	0.045	0.535	0.130
Smallest p	0.049	0.254	0.132
Prefer Overall	0.050	0.436	0.097
Choice	0.050	0.256	0.131

We observe in this case that the "Utility", "Interaction", and "Prefer overall" approaches have lower power for detecting the subgroup in scenario (2) compared to the "Smallest p-value" and the "Choice of significant subgroup" approaches. On the other hand, the "Utility", "Interaction", and "Prefer overall" approaches have higher power to detect the overall population in scenario (1) compared to the "Smallest p-value" and the "Choice of significant subgroup" approaches. We note the same trade-off between power for detecting the overall population under scenario (1) and the power for detecting the subgroup under scenario (2) as observed in the linear model setting also applies to the censored data setting.

5.6 Contiguity Considerations

In some cases, it might be of interest or importance to the investigator that the reported subgroup(s), if any, are contiguous with one another. This concern of contiguity may arise when some of the reported subgroup(s) may lead to contradicting scientific inferences. For example, a reporting of subgroups of younger male and older female may be contradictory for a given treatment that has certain mechanisms. Under these concerns, we propose an approach to post-processing the results from the method to address contiguity issues.

First, we define a "distance" metric D between any two reported subgroups k_1 and k_2. Let $X^O = \{X_1^O, ..., X_o^O\}$ and $X^N = \{X_1^N, ..., X_n^N\}$ denote the sets of ordinal and nominal subgroup-defining categorical variables for k_1 and k_2. Furthermore, let $X_i(k)$ denote the

value of variable X_i for subgroup k, Thus, the metric D is defined as follows:

$$D(k_1, k_2) = \sum_{X_i \in X^O} |X_i(k_1) - X_i(k_2)| + \sum_{X_i \in X^N} 1[X_i(k_1) \neq X_i(k_2)]$$

Note that $D \geq 0$ and $D = 0$ iff $k_1 = k_2$. Furthermore, since the variables $\mathbf{X}$ are categorical, $D \in \mathbb{N}$. Thus, we call two subgroups k_1 and k_2 contiguous if $D(k_1, k_2) \leq 1$ (i.e. by definition a subgroup is contiguous to itself). A more general case can be to consider subgroups that are "within a neighborhood" of one another. Two subgroups k_1 and k_2 are said to be "within a neighborhood d_0" of one another if $D(k_1, k_2) \leq d_0$, where d_0 is a pre-specified number.

Under model $M_j, j \neq 0$, we have the posterior expected utility U_k for all $k \in \mathcal{K}_j$. Let $U_{(1)} \leq U_{(2)} \leq ... \leq U_{(k_j)}$ be the ordered expected utilities for the subgroups in $\mathcal{K}_j$, and we assume that some are larger than the threshold (i.e. some subgroups are reported). One approach to preserving the contiguity is by selecting the subgroup with the highest utility (i.e. subgroup corresponding to $U_{(k_j)}$), and dropping other selected subgroups that are not within d_0 distance. We note, however, that the selection of the subgroups for the contiguity concerns should be motivated scientifically.

5.6.1 The Utility Approach and Contiguity

We investigated if the utility approach reported discontiguous subgroups with subgroup treatment effects sharing the same "sign" (that is, either all "harm" or all "benefit") considering the simulation results under scenarios (2) and (5).

We found that in scenario (2), only one out of 16,000 decisions reported a discontiguous subgroup, whereas under scenario (5), no discontiguous subgroup was reported.

This can be attributed to the estimates of the effects in the utility function being based on a linear combination of treatment and interaction parameters from the interaction model

– and that we only consider lower-order interactions between treatment and other predictors. Hence, the estimates lying on a $p-$dimensional hyperplane. Thus, using the utility approach, we are unlikely to report two discontiguous subgroups. It is, however, possible to have a discontiguous pair of subgroups being both flagged, but with opposite directions of the report.

5.7 Impact of Subgroup Analysis

While it is important to consider the type-1 error and the power of the approaches under the different conditions, it is of great importance as well to understand the impact of conducting subgroup analyses. Particularly, it is important to consider the impact of the analysis on subsequent trials or studies, and more importantly, the impact on the public. In assessing the public health impact, it is important to consider the predictive values, which we discuss in the following section, and we will conduct simulation studies to assess the performance of these different approaches.

5.7.1 Error Rates and Predictive Values

As we conduct subgroup analyses, additional error rates may need to be considered beyond the traditional type-I and type-II error rates. Consider the following error rates in conducting a subgroup analysis, given in Table 5.14.

Table 5.14: The error rates following a subgroup analysis

		Truth		
		Null	Overall Alternative	Subgroup k
Action	Null	$1-\alpha_0-\sum_k \alpha_k$	β_N	γ_{Nk}
	Alt	α_0	$1-\beta_N-\sum_k \beta_k$	γ_{0k}
	Sub k	α_k	β_k	$1-\gamma_{Nk}-\gamma_{0k}-\Sigma_{j\neq k}\gamma_{jk}$

Let us denote π_N, π_0,and π_k as the prevalences on the null, overall alternative, and subgroup k, respectively. The negative, positive, and subgroup k predictive values are then

given by Bayes rule as follows:

$$\begin{aligned}
NPV &= \text{P(no effect} \mid \text{selected null)} \\
&= \frac{\pi_N(1-\alpha_0-\sum_k \alpha_k)}{\pi_N(1-\alpha_0-\sum_k \alpha_k)+\pi_0\alpha_0+\sum_k \pi_k\alpha_k} \\
PPV &= \text{P(overall effect} \mid \text{selected overall)} \\
&= \frac{\pi_0(1-\beta_N-\sum_k \beta_k)}{\pi_N\beta_N+\pi_0(1-\beta_N-\sum_k \beta_k)+\sum_k \pi_k\beta_k} \\
SPV_k &= \text{P(effect in sub } k \mid \text{selected sub } k) \\
&= \frac{\pi_k(1-\gamma_{Nk}-\gamma_{0k}-\Sigma_{j\neq k}\gamma_{jk})}{\pi_N\gamma_{Nk}+\pi_0\gamma_{0k}+\pi_k(1-\gamma_{Nk}-\gamma_{0k}-\Sigma_{j\neq k}\gamma_{jk})+\Sigma_{j\neq k}\gamma_{jk}}
\end{aligned}$$

5.7.2 Simulation

For further investigation of the "Utility", "Interaction" and "Smallest p-value" approaches, we performed additional simulation studies to investigate the resulting analysis following a phase II trial with subsequent planning of a phase III trial based on the reported subgroup(s). We conducted the simulations as before, considering 3 predictive binary variables, where X_2 is prognostic. Again, the subgroup proportions of $X_1 = 0$,vs $X_1 = 1$,$X_2 = 0$ vs $X_2 = 1$, and $X_3 = 0$, vs $X_3 = 1$ are assumed to be 0.5.

We assume that the phase II trial was planned to have 80% power of detecting the MCID=0.5 with $\sigma = 1$ at 0.05 two-sided level. Thus, the sample size for the phase II trial is determined to be 126. Further, we assume that the subsequent phase III trial is planned for the same MCID, σ, and level, but with 95% power of detecting the MCID. The sample size for the phase-III trial is determined to be 240.

In this setting, we call a treatment "Null" (abbreviated N) if the treatment effect is zero in the overall population, "Overall" (abbreviated O) if the treatment effect is MCID in the overall population, and "Sub" (abbreviated S) if the treatment effect is of size MCID in subgroup $X_3 = 1$ and zero everywhere else.

We performed the simulations under two scenarios: (1) 80% of the treatments are "Null"

and 20% of the treatments are "Sub", and (2) 70% of the treatments are "Null", 10% of the treatments are "Overall", and 20% of the treatments are "Sub". For both cases, we assume that the number of treatments tested at the beginning of phase II is 10,000. We assess the performance of the "Utility", "Interaction, and "Smallest p-value" approaches.

5.7.2.1 Scenario 1

The results of the simulated trials under scenario 1 is given in Table 5.15. Under scenario 1, the negative and the subgroup predictive values are comparable between the three approaches. Furthermore, the power for detecting the subgroup is higher under the "smallest p-value" approach compared to the other two approaches.

Table 5.15: Simulation results under Scenario 1

			Utility	Interaction	Smallest p
Overall	Accepted Treatments	Indication			
	"Null"	"Null"	7989	7990	7985
		"Overall"	8	5	3
		"Sub"	0	2	3
	"Sub"	"Null"	1603	1628	1547
		"Overall"	119	126	36
		"Sub"	234	221	332
	Type-I error		0.00138	0.00125	0.00188
	Subgroup power		0.117	0.111	0.166
	NPV		83.3%	83.1%	83.8%
	SPV_k		100%	99.1%	99.1%

5.7.2.2 Scenario 2

The results of the simulated trials under scenario 2 is given in Table 5.16.

Table 5.16: Simulation results under Scenario 1

			Utility	Interaction	Smallest p
Overall	Accepted Treatments	Indication			
	"Null"	"Null"	6989	6993	6990
		"Overall"	2	5	1
		"Sub"	0	0	3
	"Overall"	"Null"	330	348	424
		"Overall"	602	607	269
		"Sub"	14	8	48
	"Sub"	"Null"	1556	1661	1578
		"Overall"	137	105	16
		"Sub"	261	209	339
	Type-I error		0.00157	0.0010	0.00143
	Overall power		0.602	0.607	0.269
	Subgroup power		0.1305	0.1045	0.1695
	NPV		78.7%	77.7%	77.7%
	PPV		81.2%	84.7%	94.1%
	SPV_k		94.9%	96.3%	86.9%

Under scenario 2, the negative predictive value is comparable between the three approaches. On the other hand, the positive predictive value is larger for the "smallest p-value" approach compared to the other two approaches. This is due to the small proportion of replications selecting the overall population under all three effects types (N, O, and S), and especially small for N and S. On the other hand, we also note that the power for detecting the overall population is much smaller for the "smallest p-value" approach. Furthermore, the subgroup predictive value is the lowest for the "smallest p-value approach".

5.8 Application

We illustrate the method using an example data application. We consider a clinical trial examining the efficacy of BCNU polymer for the treatment of recurring gliomas, following a "positive" Phase-I study of BCNU [8]. The trial is a multi-center study, enrolling 222 patients randomized to receive either a BCNU polymer vs. placebo polymer following a resection surgery, with survival being the event of interest. The data for the trial is provided in

Piantadosi (1997) [53]. The variables that are presented in Piantadosi (1997) [53] are: treatment indicator, sex, event indicator (1=dead,0=alive), previous nitrosoureas (1=yes,0=no), time, resection (1=above 75%,0=below 75%), pathology (1=glioblastoma, 0=other), Karnofsky performance score (1=70 or above, 0=below 70), race (1=white, 0=other), and grade (1=active, 0=quiescent). The descriptive statistics is given in Table 5.17 and the Kaplan-Meier estimates by treatment group is given in Figure 5.11.

Table 5.17: Descriptive statistics of mean (sd) or number (percentages) of the variables for each treatment groups in the BCNU data

		Placebo	BCNU
N		112	110
Age		47.6 (13.6)	48.1 (12.3)
Sex	Male	69 (61.6%)	74 (67.3%)
	Female	43 (38.4%)	36 (32.7%)
Resection	>75%	82 (73.2%)	82 (74.5%)
	<75%	30 (26.8%)	28 (25.5%)
Nitrosoureas	yes	49 (43.8%)	54 (49.1%)
	no	63 (56.2%)	56 (50.9%)
Pathology	Glioblastoma	73 (65.2%)	76 (69.1%)
	Other	39 (34.8%)	34 (30.9%)
Karnofsky score	≥ 70	56 (50.0%)	61 (55.5%)
	<70	56 (50.0%)	49 (44.5%)
Grade	Active	100 (89.3%)	103 (93.6%)
	Quiescent	12 (10.7%)	7 (6.4%)
Race	White	103 (92.0%)	100 (90.9%)
	Other	9 (8.0%)	10 (9.1%)

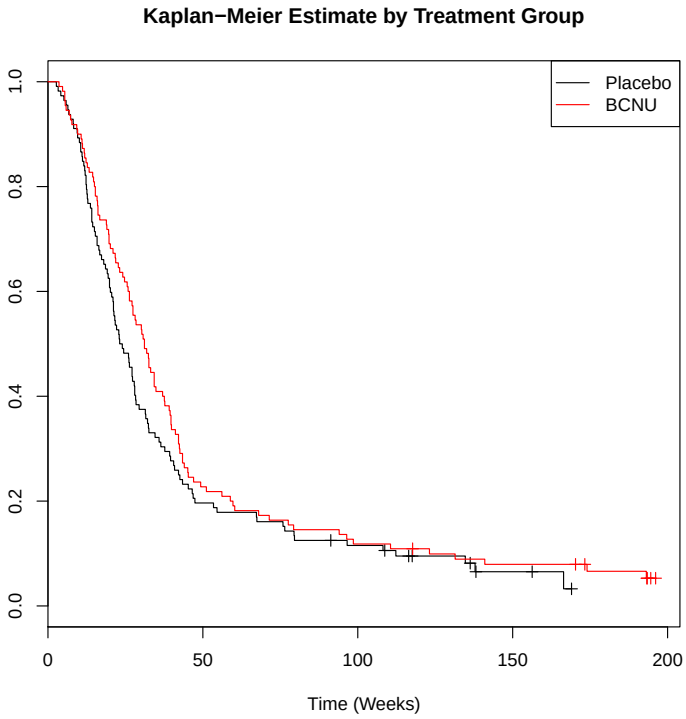


Figure 5.11: Kaplan-Meier estimates by treatment group for the BCNU data

A priori, the Karnofsky performance score, previous use of nitrosoureas, and pathological type are considered important variables associated with the event [8]. The number of deaths in the dataset is 215. Based on the number of deaths, the "assumed" MCID in designing the trial assuming 5% level and 80% power is back-calculated to be a hazard ratio of 0.682.

5.8.1 Inferential Analysis - Frequentist

The treatment effects for the overall population and the subgroups defined by one variable using the Cox regression are given in Table 5.18

From the frequentist analysis using Cox model, we observe that the null hypothesis of no effect would not be rejected nor any subgroups would be reported using any of the approaches that we considered (interaction, smallest p-value, choice of significant subgroup).

Table 5.18: Frequentist inference of the BCNU data using Cox Regression

Subgroup	log(HR)	Std. Err.	p
Overall	-0.219	0.142	0.123
Previous nitrosoureas	-0.156	0.203	0.440
No previous nitrosoureas	-0.219	0.194	0.259
Glioblastoma	-0.206	0.167	0.217
Other	-0.280	0.264	0.289
Kanofsky score ≥ 70	-0.254	0.197	0.199
Kanofsky score <70	-0.122	0.203	0.547

5.8.2 Inferential Analysis - Proposed Method

Using the back-calculated MCID, we performed the calibration with $B = 1000$ and obtained the calibrated parameters $(K_1^{(0)}, K_1, K_2) = (0.0852, 0.0996, 2.381)$. Suppose X_1, X_2, and X_3 denote the indicator for previous use of nitrosoureas, pathological type, and Karnofsky performance score respectively. We used a Weibull model for the death times and put non-informative priors on both the model probabilities and the model parameters. The Bayes factors for the models are given in Table 5.19.

Table 5.19: Bayes factors of the models in the analysis. X_1 = previous nitrosoureas use, X_2 = pathological type (glioblastoma), X_3 = Karnofsky performance score

Model	Subgroup-defining Variable	Bayes Factor
M_1	X_1	0.0908
M_2	X_2	0.0919
M_3	X_1,X_2	0.0067
M_4	X_3	0.1247
M_5	X_1,X_3	0.0090
M_6	X_2,X_3	0.0094
M_7	X_1,X_2,X_3	0.0007

Thus, the model selected is the model M_0 denoting testing only in the overall population. The expectation of the components of the utility for the overall population is given in Table 5.20. The quantities presented are for the log hazard ratio scale.

Table 5.20: Utility components and inference of the BCNU data. The expected components of the utility function as well as the optimal action for the overall population are given.

	Value
Expected "Uncertainty" Component	-0.0002
Expected "Effect" Component	0.013
Probability of subgroup 0 being important, i.e. $P(E_0 = 0\|y)$	0.933
Computed "quantity" for subgroup 0, Q_0	0.014
Threshold $K_1^{(0)}$	0.085
Optimal Action (a_0)	0 (no subgroup)
Posterior Mean	-0.176
Posterior SD	0.132

Thus, based on our method for the BCNU data, we decide not to report any subgroup nor the overall population. We obtained similar results from conducting the analysis using frequentist methods.

These conclusions are consistent with that by Brem et al [8] and Piantadosi [53], reporting an estimated hazard ratio of 0.82 (95% CI = (0.45,1.03), p=0.19) comparing BCNU vs. placebo in the overall population.

5.9 Discussion

In this chapter, we have presented an approach for a post-hoc subgroup analysis under the Bayesian setting, combining both Bayesian model selection and a utility-based approach for reporting any subgroup(s) of interest.

In the selection of the subgroups, we favored reporting subgroups that: exhibit larger treatment effect, are larger, and are simpler (less complex). The contribution of the treatment effect size and the subgroup size are formalized in the reporting payoff, whereas the contribution of subgroup complexity is in the model selection component of the method. The decision rule for the model selection is based on the Bayes Factors and the ratio of prior probabilities of the models. We computed an approximation of the Bayes Factor via BIC, which contains a penalty term for model complexity, quantified by the number of predictors in the model.

We compared our approach to a few different approaches for subgroup selection: interaction, smallest p-value, and prefer overall approaches. We noticed that the approaches that tend to have a larger power for detecting a subpopulation (e.g. the smallest p-value approach) may perform poorly when the effect is in fact homogeneous in the overall population. On the other hand, the approaches that tend to favor the inference in the overall population may have lower power to detect a subpopulation, especially when the effect in the subpopulation is large enough such that the "average effect" in the overall population starts to be noticeable.

Further, we noticed that both the utility and the interaction approaches provide a "middle ground" for detecting a subpopulation when there is truly one, while conserving the power for reporting the overall population when there is homogeneous effect in the overall population. Furthermore, the power for reporting the overall population when the effect is homogeneous in the overall population using the utility and interaction approaches are comparable to that of the "Prefer Overall" approach. We also note that in the simulation studies, the subgroup investigations for the "prefer overall", "smallest p-value", and "choice" approaches

are limited to subgroups defined by only one variable. This reduces the number of tests performed, thus alleviating the correction on the critical p-value, leading to higher power. On the other hand, the utility and interaction methods allow testing for all 27 potential subgroups.

In our investigations, we assumed that the subgroups have equal variance. In general, we might not expect this to be true in practice. In practice, we might expect to have various subgroups, each with different uncertainties. Furthermore, the approach currently does not handle the case where the difference in uncertainties is of interest. We might imagine the case where there are two subgroups having the same average effect, but different variances, and the variance in the other subgroup might be large enough such that it is desirable to report one, but not the other subgroup. As of now, the utility takes into account the uncertainty (or the inference loss) of the posterior subgroup effects. However, in the current parameterization of the linear model, the uncertainty is a function of the common parameter σ^2 and the sample size. If the two subgroup sizes are the same, then we expect to have the same report for both subgroups, which may not be what is of interest.

5.9.1 Frequentist Subgroup Selection Approaches

The Frequentist approaches considered in the simulation studies (i.e. interaction-stratified, prefer overall, smallest p-value, and choice of significant subgroup) are decision rules that may result from the investigator's preference (i.e. the investigator's utility). The interaction-stratified approach correspond quite closely to our formulation of utility function. The prefer overall approach may be utilized by a sponsor whose interest is to have the treatment marketed to a larger population whenever possible. In this case, the sponsor places a higher utility for the overall population compared to that in the subgroups, favoring reports on the overall population.

The smallest p-value approach, on the other hand, may reflect a sponsor whose interest is to obtain the subpopulation for which the treatment effect appears to be most "significant", perhaps so that the treatment has a higher chance for "passing" the subsequent trials. In

this case, conceptually, the investigator places equal utility across all subpopulations, thus not favoring specifically a particular subgroup.

The choice of significant subgroup approach may reflect a regulatory agency's perspective, where the treatment is approved in a particular subgroup if the effect is significant in the subgroup and non-significant in the complement subgroup. In other words, the regulator prefers to report only a subpopulation (rather than the overall population), thus limiting the indication of the treatment.

Note that these approaches may correspond to a "non-linear" utility function, which in turns produces "non-linear" decision rules, but nevertheless can be used.

Chapter 6

OVERALL CONCLUSIONS AND FUTURE WORK

6.1 Forward Slice in Optimization and Optimal Design

In this research, we proposed a method for stochastic optimization called "forward slice" and applied the method to some design problems. We showed that the forward slice selects the global maximum with higher rate than some of the existing optimization methods readily implemented in R, such as Nelder-Mead, conjugate gradient, BFGS, and simulated annealing. Further, in optimal design problems, we found the forward slice is able to select (locally) D-optimal design points.

One drawback to the forward slice method is the computing time in higher dimensions. At each iteration, the forward slice method obtains a proposal candidate that is drawn from the design space. While this allows for selection of the global optimum (and not being stuck on a local mode), it is computationally more expensive as the number of iterations grows. This is due to the slice portion S getting smaller at every iteration, and hence there is a higher rate of rejected proposals for samples from the design space. Thus, there is a natural trade-off with the use of the forward slice procedure in that while it enables attaining the optimal solution, this may come at the expense of additional computing time. Future work should consider ways to reduce the computational time, for example, via parallel processing.

6.2 Post-Hoc Subgroup Selection

In deciding whether to perform (or not) a subgroup analysis following the end of a trial, some considerations need to be taken into account. When no subgroup analysis is (pre-planned and) conducted, a significant effect in the overall population does not necessarily mean that the treatment is effective for the whole study population. In fact, the treatment

may be effective in one or more subpopulations, but it may not be effective and may be even harmful, in others. Thus, it may be important to identify which subpopulations are harmed by the treatment, even though the "average" effect over the population shows significant benefit. On the other hand, if a subgroup analysis is conducted, it is also possible that only one or more subpopulations are reported, when in fact the treatment is effective in the overall population. In this case, we are excluding some subpopulations to a potentially beneficial treatment. Thus questions of whether/how to conduct subgroup analysis are not trivial. The underlying statistical errors that can be made by performing (or not) subgroup analysis is also an important consideration.

In this research, we proposed and investigated a method for a post-hoc subgroup selection via a utility formulation. Specifically, we evaluated our method in the context of randomized clinical trial. We evaluated the performance of the utility-based subgroup analysis for continuous, binary, and censored outcomes. The utility-based method allows for incorporation of a-priori knowledge or hypothesis via the priors in the subgroup analysis. Under a relatively non-informative prior distribution on the models and non-informative conjugate priors on the parameters, the performance of the utility-based method is comparable to that of using a combination of interaction test and stratified analyses. By Bayes, the reported actions are stage-wise optimal with respect to the utility-prior pair. The utility-based approach provides an approach for detecting a subpopulation that is benefited (or harmed) by a particular treatment, without sacrificing much power in detecting an overall population effect.

Among the approaches that we considered in our investigations (and that may be commonly applied in practice), there are trade-offs between the power to detect a subpopulation and the power to detect an overall effect. The "smallest p-value" approach, although has a higher power for detecting a subpopulation, has an abysmal power for detecting an overall effect. On the other hand, the "prefer overall" approach might conserve power for detecting an overall effect, but at a cost of decreased power for detecting a true subpopulation effect, especially when the subpopulation effect is large. Both the utility and the interaction ap-

proaches allow for conservation of power for detecting the overall effect, while still able to detect subpopulations effect. The power of utility and interaction approaches is lower than that under the "smallest p-value" approach. In our study, we did not find an approach for subgroup analysis that is "best in all settings", but both the utility and the interaction approaches seem to provide overall desirable operating characteristics. We also note that in the simulation studies, the subgroup investigations for the "prefer overall", "smallest p-value", and "choice" approaches are limited to subgroups defined by only one variable. This reduces the number of tests performed, thus alleviating the correction on the critical p-value, leading to higher power. On the other hand, the utility and interaction methods allow testing for all 27 potential subgroups.

One advantage of the utility approach over the interaction approach is that it allows for incorporation of prior information into the analysis. This can be done in two ways: via the prior probabilities of the models or via the prior distribution of the model parameters. Under a more informative prior on some subgroups, the power for detecting the overall population is decreased. Furthermore, placing a more informative prior on an incorrect model results in power loss for detecting an overall effect, while not gaining power to detect a subgroup. Thus, it is important for the investigators to carefully consider the priors that are used, and have the use of priors be motivated by a scientific basis or based on previous (trustworthy) investigations.

We also note that the utility approach has been embedded within a framework where utilities are set to attain certain Frequentist operating characteristics. The utilities may be set based on the particular trade-offs the decision-maker/analyst is willing to accept. For example, Table 5.3 assigns the same utility for selecting any model M_j when it is the true model. Such utilities do not need to be the same. Indeed, for a more conservative approach for subgroup selection, one may place a higher utility for model M_0 relative to the other models.

One possible direction for future research is extending the approach to handle continuous predictors. In many cases, many of the baseline variables that are of interest are measured on

a continuous scale. Although using the continuous variable (compared to using a categorized version of the continuous variable) in an interaction model may increase power to detect a differential effect, the question arises as to where in the spectrum of the continuous variable do we report a significant effect? In other words, even when a continuous variable is predictive linearly, categories based on the continuous variable provide a more meaningful and practical interpretation. It would be of interest to extend the utility-based approach to investigate the "cutoffs" of the continuous variables – such a choice may also be viewed as a decision problem. Another possible future investigation is to extend the utility-based approach to handle the cases where the uncertainties among the different subgroups are unequal.

The utility formulation provides a general framework to decision-making in subgroup selection that reflects the investigator's interest. Conceptually, the Frequentist approaches for subgroup selection are decision rules. Thus, another possible direction for future work is to derive the specific utility functions that give rise to these (Frequentist) decision rules.

BIBLIOGRAPHY

[1] Khidir Mohamed Abdelbasit and RL Plackett. Experimental design for binary data. *Journal of the American Statistical Association*, 78(381):90–98, 1983.

[2] Susan F Assmann, Stuart J Pocock, Laura E Enos, and Linda E Kasten. Subgroup analysis and other (mis) uses of baseline data in clinical trials. *The Lancet*, 355(9209):1064–1069, 2000.

[3] James Babb, Andre Rogatko, and Shelemyahu Zacks. Cancer phase i clinical trials: efficient dose escalation with overdose control. *Statistics in medicine*, 17(10):1103–1120, 1998.

[4] José M Bernardo. Expected information as expected utility. *The Annals of Statistics*, pages 686–690, 1979.

[5] Scott M Berry, Bradley P Carlin, J Jack Lee, and Peter Muller. *Bayesian adaptive methods for clinical trials*. CRC press, 2010.

[6] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[7] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[8] Henry Brem, S Piantadosi, PC Burger, M Walker, R Selker, NA Vick, K Black, M Sisti, S Brem, G Mohr, et al. Placebo-controlled trial of safety and efficacy of intraoperative controlled delivery by biodegradable polymers of chemotherapy for recurrent gliomas. *The Lancet*, 345(8956):1008–1012, 1995.

[9] Sara T Brookes, Elise Whitely, Matthias Egger, George Davey Smith, Paul A Mulheran, and Tim J Peters. Subgroup analyses in randomized trials: risks of subgroup-specific analyses;: power and sample size for the interaction test. *Journal of clinical epidemiology*, 57(3):229–236, 2004.

[10] Sarah T Brookes, Elise Whitley, Tim J Peters, Paul A Mulheran, Matthias Egger, and G Davey Smith. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. *Health Technology Assessment*, 5(33):1–56, 2001.

[11] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[12] Kathryn Chaloner and Kinley Larntz. Optimal bayesian design applied to logistic regression experiments. *Journal of Statistical Planning and Inference*, 21(2):191–208, 1989.

[13] Kathryn Chaloner, Isabella Verdinelli, et al. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

[14] Probal Chaudhuri and Per A Mykland. Nonlinear experiments: Optimal design and inference based on likelihood. *Journal of the American Statistical Association*, 88(422):538–546, 1993.

[15] Herman Chernoff. Locally optimal designs for estimating parameters. *The Annals of Mathematical Statistics*, pages 586–602, 1953.

[16] Ying Kuen Cheung. Stochastic approximation and modern model-based designs for dose-finding clinical trials. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(2):191, 2010.

[17] P Damien, Jonathan Wakefield, and Stephen Walker. Gibbs sampling for bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):331–344, 1999.

[18] Morris H DeGroot. Concepts of information based on utility. In *Recent Developments in the Foundations of Utility and Risk Theory*, pages 265–275. Springer, 1986.

[19] Morris H DeGroot et al. Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, 33(2):404–419, 1962.

[20] Dennis O Dixon and Richard Simon. Bayesian subset analysis. *Biometrics*, pages 871–881, 1991.

[21] Hovav A Dror and David M Steinberg. Sequential experimental designs for generalized linear models. *Journal of the American Statistical Association*, 103(481):288–298, 2008.

[22] Gustav Elfving et al. Optimum allocation in linear regression theory. *The Annals of Mathematical Statistics*, 23(2):255–262, 1952.

[23] Valerii Vadimovich Fedorov. *Theory of optimal experiments*. Elsevier, 1972.

[24] Ronald Aylmer Fisher. Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 22, pages 700–725. Cambridge Univ Press, 1925.

[25] Reeves Fletcher and Colin M Reeves. Function minimization by conjugate gradients. *The computer journal*, 7(2):149–154, 1964.

[26] Reeves Fletcher and Colin M Reeves. Function minimization by conjugate gradients. *The computer journal*, 7(2):149–154, 1964.

[27] Jared C Foster, Jeremy M G Taylor, and Stephen J Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 2011.

[28] Dimitris Fouskakis and David Draper. Stochastic optimization: a review. *International Statistical Review*, 70(3):315–349, 2002.

[29] Fred Glover. Tabu search: A tutorial. *Interfaces*, 20(4):74–94, 1990.

[30] David Edward Goldberg et al. *Genetic algorithms in search, optimization, and machine learning*, volume 412. Addison-wesley Reading Menlo Park, 1989.

[31] M Hamada, HF Martz, CS Reese, and AG Wilson. Finding near-optimal bayesian experimental designs via genetic algorithms. *The American Statistician*, 55(3):175–181, 2001.

[32] Darrall Henderson, Sheldon H Jacobson, and Alan W Johnson. The theory and practice of simulated annealing. In *Handbook of metaheuristics*, pages 287–319. Springer, 2003.

[33] Wenyu Jiang, Boris Freidlin, and Richard Simon. Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *Journal of the National Cancer Institute*, 99(13):1036–1043, 2007.

[34] Holland John. *Adaptation in natural and artificial systems*. MIT Press, Cambridge, MA, 1992.

[35] Hayley E Jones, David I Ohlssen, Beat Neuenschwander, Amy Racine, and Michael Branson. Bayesian models for subgroup analysis in clinical trials. *Clinical Trials*, 2011.

[36] Hayley E Jones, David I Ohlssen, Beat Neuenschwander, Amy Racine, and Michael Branson. Bayesian models for subgroup analysis in clinical trials. *Clinical Trials*, 8(2):129–143, 2011.

[37] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[38] André I Khuri, Bhramar Mukherjee, Bikas K Sinha, Malay Ghosh, et al. Design issues for generalized linear models: A review. *Statistical Science*, 21(3):376–399, 2006.

[39] Jack Kiefer. Optimum experimental designs. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 272–319, 1959.

[40] Edward S Kim, Roy S Herbst, Ignacio I Wistuba, J Jack Lee, George R Blumenschein, Anne Tsao, David J Stewart, Marshall E Hicks, Jeremy Erasmus, Sanjay Gupta, et al. The battle trial: personalizing therapy for lung cancer. *Cancer discovery*, 1(1):44–53, 2011.

[41] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, pages 79–86, 1951.

[42] Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pages 986–1005, 1956.

[43] Dennis Victor Lindley. *Bayesian statistics: A review*. SIAM, 1972.

[44] Ilya Lipkovich and Alex Dmitrienko. Biomarker identification in clinical trials. 2013.

[45] Ilya Lipkovich, Alex Dmitrienko, Jonathan Denne, and Gregory Enas. Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in medicine*, 30(21):2601–2621, 2011.

[46] Peter Muller. Simulation-based optimal design. *Bayesian statistics*, 6:459–474, 1999.

[47] Peter Muller, Giovanni Parmigiani, and Kenneth Rice. Fdr and bayesian multiple comparisons rules. 2006.

[48] Radford M Neal. Markov chain monte carlo methods based onslicing'the density function. *Technical Report*, 1997.

[49] Radford M Neal. Slice sampling. *Annals of statistics*, pages 705–741, 2003.

[50] John A Nelder and Roger Mead. A simplex method for function minimization. *Computer journal*, 7(4):308–313, 1965.

[51] Barry T Neyer. A d-optimality-based sensitivity test. *Technometrics*, 36(1):61–70, 1994.

[52] John O'Quigley, Margaret Pepe, and Lloyd Fisher. Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics*, pages 33–48, 1990.

[53] Steven Piantadosi. *Clinical trials: a methodologic perspective*. John Wiley & Sons, 1997.

[54] Stuart J Pocock, Susan E Assmann, Laura E Enos, and Linda E Kasten. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. *Statistics in medicine*, 21(19):2917–2930, 2002.

[55] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[56] William F Rosenberger and Linda M Haines. Competing designs for phase i clinical trials: a review. *Statistics in Medicine*, 21(18):2757–2770, 2002.

[57] Michael Rosenblum and Mark J van der Laan. Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment. *Biometrika*, 98(4):845–860, 2011.

[58] Brittany Sanchez. Evaluation of strategies for the phase ii to phase iii progression in treatment discovery. Master's thesis, University of Washington, 2014.

[59] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 1948.

[60] Noah Simon and Richard Simon. Adaptive enrichment designs for clinical trials. *Biostatistics*, 14(4):613–625, 2013.

[61] Richard Simon. Bayesian subset analysis: application to studying treatment-by-gender interactions. *Statistics in Medicine*, 2002.

[62] Siva Sivaganesan, Purushottam W Laud, and Peter Müller. A bayesian subgroup analysis with a zero-enriched polya urn scheme. *Statistics in medicine*, 30(4):312–323, 2011.

[63] M Stone. Application of a measure of information to the design and comparison of regression experiments. *The Annals of Mathematical Statistics*, pages 55–70, 1959.

[64] Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M Nickerson, and Bogong Li. Subgroup analysis via recursive partitioning. *The Journal of Machine Learning Research*, 10:141–158, 2009.

[65] Peter F Thall, Richard Simon, and Susan S Ellenberg. A two-stage design for choosing among several experimental treatments and a control in clinical trials. *Biometrics*, pages 537–547, 1989.

[66] Madeleine B Thompson and Radford M Neal. Slice sampling with adaptive multivariate steps: The shrinking-rank method. *arXiv preprint arXiv:1011.4722*, 2010.

[67] Nicole White, Helen Johnson, Peter Silburn, George Mellick, Nadeeka Dissanayaka, and Kerrie Mengersen. Probabilistic subgroup identification using bayesian finite mixture modelling: A case study in parkinson's disease phenotype identification. *Statistical Methods in Medical Research*, 2010.

[68] John Whitehead and Hazel Brunier. Bayesian decision procedures for dose determining experiments. *Statistics in medicine*, 14(9):885–893, 1995.

[69] SJ Wright and J Nocedal. *Numerical optimization*, volume 2. Springer New York, 1999.

[70] CF Jeff Wu. Efficient sequential designs with binary data. *Journal of the American Statistical Association*, 80(392):974–984, 1985.

[71] Xian Zhou, Suyu Liu, Edward S Kim, Roy S Herbst, and J Jack Lee. Bayesian adaptive design for targeted therapy development in lung cancer—a step toward personalized medicine. *Clinical Trials*, 5(3):181–193, 2008.

Appendix A

ALGORITHMS

A.1 Algorithms for Selected Optimization Methods

A.1.1 Conjugate Gradient

The algorithm for the general conjugate gradient method, as given by Fletcher and Reeves (1964) [26], is as follows [69]:

1. Start with an initial point x_0. Compute $f_0 = f(x_0)$, $\nabla f_0 = \nabla f(x_0)$, and $p_0 = -\nabla f_0$. Set $k = 0$ and the tolerance for convergence ϵ.

2. Compute α_k, which is the minimizer of f along the direction of p_k, estimated using line search. Fletcher and Reeves described the line search method as having three stages: the first stage estimates the order of magnitude of α_k, the second stage establishes bounds on the estimates, and the third stage interpolates the value. Compute $x_{k+1} = x_k + \alpha_k p_k$.

3. Compute ∇f_{k+1}.

   - If $|\nabla f_{k+1}| \leq \epsilon$, stop the algorithm and report x_{k+1} as the value of x that optimizes $f(x)$.
   - Otherwise, compute $\beta_{k+1} = (\nabla f_{k+1}^T \nabla f_{k+1})/(\nabla f_k^T \nabla f_k)$ and $p_{k+1} = -\nabla f_{k+1} + \beta_{k+1} p_k$. Update $k \leftarrow k + 1$.

4. Repeat steps 2 and 3 until convergence.

A.1.2 BFGS Method

The algorithm for the BFGS method is:

1. Set a starting point x_0, tolerance for convergence ϵ and an inverse Hessian approximation $H_0 = B_0^{-1}$.

2. Compute the search direction $p_k = -H_k \nabla f_k$, where $H_k = B_k^{-1}$.

3. Compute α_k, the minimizer along the direction of p_k, by line search procedure to satisfy the Wolfe condition. Set $x_{k+1} = x_k + \alpha_k p_k$.

4. Define $s_k = x_{k+1} - x_k$ and $y_k = \nabla f_{k+1} - \nabla f_k$. Compute $H_{k+1} = (1 - \rho_k s_k y_k^T) H_k (1 - \rho_k s_k y_k^T) + \rho_k s_k s_k^T$, where $\rho_k = \frac{1}{y_k^T s_k}$.

5. Calculate $\|\nabla f_k\|$.

   - If $\|\nabla f_k\| < \epsilon$, convergence is achieved.
   - Otherwise, update $k \leftarrow k + 1$ and repeat steps 2-5.

A.1.3 Nelder-Mead

Suppose that $x_1, x_2, ..., x_{n+1}$ are $(n+1)$ points in the n-dimensional space defining the current "simplex". By convention, the vertices $\{x_1, ..., x_{n+1}\}$ are ordered such that $f(x_1) \leq ... \leq f(x_{n+1})$. Initially, $x_1, x_2, ..., x_{n+1}$ are randomly selected from the n-dimensional space. The centroid of the best n points at that iteration is given by $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$. The points along the line joining $\bar{x}$ and the worst vertex x_{n+1} is given by the function $\tilde{x}(t) = \bar{x} + t(x_{n+1} - \bar{x})$. The procedure for one iteration of the Nelder-Mead method [69] [50] is:

1. Compute $\tilde{x}(-1)$ and $f_{-1} = f(\tilde{x}(-1))$.

2. Update the $n + 1$ vertices as follows:

- If $f(x_1) \leq f_{-1} < f(x_n)$, replace x_{n+1} with $\tilde{x}(-1)$; proceed to the next iteration.
- If $f_{-1} < f(x_1)$, compute $\tilde{x}(-2)$ and $f_{-2} = f(\tilde{x}(-2))$
  - If $f_{-2} < f_{-1}$, replace x_{n+1} with $\tilde{x}(-2)$; proceed to the next iteration.
  - Otherwise, replace x_{n+1} with $\tilde{x}(-1)$; proceed to the next iteration.
- If $f_{-1} \geq f(x_n)$
  - If $f(x_n) \leq f_{-1} < f(x_{n+1})$, calculate $\tilde{x}(-1/2)$ and $f_{-1/2}$.
    - If $f_{-1/2} < f_{-1}$, replace x_{n+1} by $\tilde{x}(-1/2)$; proceed to the next iteration.
    - Otherwise, replace x_i by $(1/2)(x_1 + x_i)$ for each i=2,3,...,$n+1$; proceed to the next iteration.
  - Otherwise, calculate $\tilde{x}(1/2)$ and $f_{1/2}$.
    - If $f_{1/2} < f_{n+1}$, replace x_{n+1} by $\tilde{x}(1/2)$; proceed to the next iteration.
    - Otherwise, replace x_i by $(1/2)(x_1 + x_i)$ for each i=2,3,...,$n+1$; proceed to the next iteration.

A.1.4 Simulated Annealing

The algorithm for simulated annealing [32] is:

1. Select an initial solution ω, initial temperature $t_0 \geq 0$, the temperature "schedule" t_k, $k = 1, 2, 3, ...$, and the number, M_k, of iterations at each temperature t_k, and tolerance ϵ. Start with the initial temperature t_0.

2. Perform the following:

   (i) Set m=0.

   (ii) Generate a candidate next point $\omega' \in N(\omega)$, where $N(\omega)$ denotes the neighborhood of ω.

   (iii) Calculate $\Delta_{\omega,\omega'} = f(\omega') - f(\omega)$:

- If $\Delta_{\omega,\omega'} \leq 0$, take $\omega = \omega'$,
- Otherwise, take $\omega = \omega'$ with probability $\exp(-\Delta_{\omega,\omega'}/t_k)$.

(iv) Check for convergence or maximum runs at the temperature:

- If $|\Delta_{\omega,\omega'}| \leq \epsilon$, stop the procedure. Report ω as the optimum point,
- Otherwise, if $m = M_k$, move to step 3.
- Otherwise, update $m \leftarrow m + 1$, and repeat steps (ii) to (iv).

3. Set the current temperature to be $t_k = t_{k+1}$. Repeat step 2.

A.2 Algorithms for Selected Optimal Designs

A.2.1 Method by Chaudhuri and Mykland

The algorithm for the method proposed by Chaudhuri and Mykland [14] is:

1. Obtain an initial design consisting of n_1 points, $X_1, X_2, ..., X_{n_1}$.

2. Obtain parameter estimates using MLE based on these first n_1 points.

3. Select the next design point as the one that maximizes the D-optimality criterion given the design points selected so far and the current parameter estimate.

4. Update the MLE parameter estimate based on the observations thus far.

5. Repeat until the maximum sample size N is reached.

A.2.2 Method by Dror and Steinberg

The proposed algorithm by Dror and Steinberg [21] for obtaining the design points for a fully sequential experiment, in which the design points are obtained one at a time, starting with a null design (no observations) is as follows:

1. Determine the augmentation horizon m at the start of the experiment, given the prior median and the assumed model.

2. Given the current design, find locally D-optimal m observations that when added to the current design, would maximize ϕ_2. Call this set of m observations C_m.

3. Generate a candidate set, which contains the m points obtained from the previous step (C_m) and their coordinatewise median. Hence, the candidate set will contain $m+1$ observations.

4. If the design so far provide a nonsingular information matrix, the next design point is chosen from the candidate set as the design point that when added to the current design, would give the best ϕ_1. This selection of the design point from the candidate set is done using Fedorov's exchange algorithm [23] (see Appendix A1 for details).

5. If the current design so far does not provide a nonsingular information matrix, then the next design point is chosen among the candidates that when added to the design so far and the m-run augmentation (i.e. added to $[X_0\ C_m]^T$, where X_0 denote the current design) would give the best ϕ_1.

6. Repeat until the maximum sample size N is reached.

A.3 Algorithms of Dose-Finding Studies

A.3.1 Continual Reassessment Method (CRM)

Denote the dose–response function by $\psi(x, a)$ where x denotes the dose and $a \in \mathcal{A}$ is the parameter describing the function. The CRM procedure [52] is given as follows:

1. Set the doses x_d $(d = 1, 2, ..., k)$ for experimentation and the target probability, θ, of a dose-limiting toxicity (DLT). Set $\Omega_1 = \emptyset$. Assume a prior $f(a, \Omega_1)$ for a and obtain the prior mean $\mu_a(1)$. Set $i = 1$.

2. Treat the next patient i at dose x^*, where $x^* = \arg\min_{x\in\{x_1,\ldots,x_k\}}\{\mid \psi(x,a) - \theta \mid\}$, that is, the dose that is closest to reaching the DLT.

3. Observe the outcome, y_i, for the i^{th} patient. Set $\Omega_{i+1} = \Omega_i \cup \{y_i\}$.

4. Update the information about a by obtaining $f(a, \Omega_{j+1})$ from $f(a, \Omega_i)$ via Bayes Theorem. Update $i \leftarrow i + 1$.

5. Repeat steps 1-4 until the estimate of a is sufficiently precise, or until the maximum size is reached.

A.3.2 Escalation with Overdose Control (EWOC)

The EWOC procedure [3] is given as follows:

1. Assume a function, F, for the dose-response model, a prior on the parameters, and a proportion of maximum overdose, α. Set $\Omega_0 = \emptyset$, $x_1 = X_{min}$ and $i = 1$.

2. Treat patient i at dose x_i and observe the outcome y_i.

3. Update $\Omega_i = \Omega_{i-1} \cup \{y_i\}$. Given the data, Ω_i, obtain the posterior CDF of the MTD, $\pi_i(z)$. Obtain the optimal dose selected for the next patient as $x^* = \pi^{-1}(\alpha)$.

4. Update $i \leftarrow i + 1$ and $x_i = x^*$.

5. Repeat steps 2-4 until the estimate of the MTD is sufficiently precise, or until the maximum sample size.

B. Fedorov Exchange Algorithm

Fedorov (1972) developed an algorithm that considers exchanges between the design point and the points in a candidate list which is a discretized version of the design space. At each iteration, the exchange is made such that it maximizes the pre–defined optimality criterion.

Let $\mathbf{I}$ denote the information matrix and $\mathcal{G} = \{x_1, ..., x_n\}$ be the set consisting of n candidate design points. Under the D-optimality criterion, we select design points to maximize the determinant of the information matrix, that is, $|\mathbf{I}|$. If we append the current design matrix $\mathbf{X}$ with a new design point x to obtain the newly augmented design matrix $\tilde{\mathbf{X}}$, Fedorov (1972) showed that the following result holds:

$$|\tilde{\mathbf{I}}| = |\mathbf{I}|(1 + x^T \mathbf{I}(\mathbf{X})x),$$

where $\tilde{\mathbf{I}}$ is the information matrix given the augmented design matrix $\tilde{\mathbf{X}}$. The above equation alleviates the algorithm from the additional calculation of the information matrix for every selected design point. Thus, the Fedorov exchange algorithm using D-optimality criterion for appending a design matrix is given as follows.

Fedorov Exchange Algorithm:

1. Begin with a non-singular design matrix $\mathbf{X}$ and initialize the candidate set $\mathcal{G}$.

2. For each $j \in \{1, ..., n\}$, calculate $x_j^T \mathbf{I}(\mathbf{X}) x_j$, where x_j is the jth element of the candidate set $\mathcal{G}$.

3. Choose $j^* = \text{argmax}_{j \in \{1,...,n\}}(x_j^T \mathbf{I}(\mathbf{X}) x_j)$ and select x_{j^*} as the next design point to be appended.

Appendix B

PROOF OF CONVERGENCE OF THE FORWARD SLICE ALGORITHM

Under the conditions stated in Section 3.4, we show that the iterates of the forward slice procedure converge to a global maximum.

Lemma 2.1. For any given point $x_0 \in \mathcal{X}$, for which x_0 is not a global maximum, one iteration of the forward slice procedure gives a point $x_1 \in \mathcal{X}$ such that $f(x_1)>f(x_0)$ a.s.

Proof. Given the initial point x_0, we obtain the slice level $y = f(x_0)$ which defines the horizontal 'slice' $S_0 = \{x : f(x) \geq f(x_0)\}$. Based on our slice procedure, the next point x_1 is sampled from $S_0^X = S_0 \cap \mathcal{X} \subseteq S_0$. Note that $S_0^X \neq \emptyset$, since $x_0 \in S_0^X$. Since $\forall x \in S_0$, $f(x) \geq f(x_0)$ by definition of S_0, we have $f(x_1) \geq f(x_0)$. Let $A_0^X = \{x : f(x) = f(x_0)\} \cap \mathcal{X}$ and $B_0^X = \{x : f(x)>f(x_0)\} \cap \mathcal{X}$. Thus, clearly $S_0^X = A_0^X \cup B_0^X$, and $A_0^X \cap B_0^X = \emptyset$. Since x_1 is sampled from S_0^X, so either $x_1 \in A_0^X$ or $x_1 \in B_0^X$. However, given the continuity and absence of local plateau, A_0^X is a set of finitely many points, i.e. $l(A_0^X) = 0$, where l denotes the Lebesgue measure. Since x_1 is sampled uniformly from S_0^X, then x_1 is from A_0^X with probability $l(A_0^X)/l(S_0^X)$ or x_1 is from B_0^X with probability $l(B_0^X)/l(S_0^X)$. Since $l(A_0^X) = 0$, then $x_1 \in B_0^X$ a.s., which means $f(x_1)>f(x_0)$ a.s.

□

Recall the notation that for the k^{th} iteration of the forward slice procedure, x_k denotes the current iterate, x_{k+1} denotes the next iterate, and S_k denotes the 'slice' at the current iterate, i.e. $S_k = \{x : f(x_k) \geq f(x)\}$. Let $X^* = \{x_1^*, x_2^*, ..., x_N^*\}$ denote the points that give the global maxima, i.e. $f(x_i^*) \geq f(x)$, $\forall x \in \mathcal{X}$, $\forall i = 1, 2, ..., N$. Given condition (2) of no local plateau in the design space $\mathcal{X}$, X^* contains finitely many points.

Lemma 2.2. For a given point $x_0 \in \mathcal{X}$, for which x_0 is a global maximum, i.e. $x_0 \in X^*$, an iteration of the forward slice procedure gives a point $x_1 \in X^*$ a.s.

Proof. Given the initial point $x_0 \in X^*$, the slice region is given by $S^X = X^*$ since x_0 is already a global maximum. Hence, the next design point x_1 obtained from an iteration of the slice procedure would result in uniform sampling of the points from X^* and hence $x_1 \in X^*$.

□

For an arbitrary k^{th} iteration, define $S_k^X = S_k \cap \mathcal{X}$. Hence, $S_k^X \subseteq S_{k-1}^X$ because $f(x_k) \geq f(x_{k-1})$. So, given the initial input x_0, $S_0^X \supseteq S_1^X \supseteq \ldots \supseteq S_k^X \supseteq \ldots$. Then, we have that $\cap_{n=1}^k S_n^X = S_k^X$. Furthermore, S_k^X is compact for any k^{th} iteration. By Cantor's interesection theorem, $\cap_{n=1}^\infty S_n^X \neq \emptyset$. First, we prove that the limit of the slice regions exist, and is non-empty.

Lemma 2.3. The limit of S_k^X exists and is non-empty

Proof.

$$\liminf_{k\to\infty} S_k^X \quad = \quad \bigcup_{k\in\mathbb{N}} \bigcap_{i\geq k} S_k^X \quad = \quad \bigcup_{k\in\mathbb{N}} (S_k^X \cap S_{k+1}^X \cap \ldots)$$

Since $S_k^X \subseteq S_{k-1}^X$ for any k, then $(S_k^X \cap S_{k+1}^X \cap \ldots) \supseteq (S_{k+1}^X \cap S_{k+2}^X \cap \ldots) \supseteq \ldots$. Thus, we have

$$\liminf_{k\to\infty} S_k^X \quad = \quad \bigcup_{k\in\mathbb{N}} (S_k^X \cap S_{k+1}^X \cap \ldots) \quad = \quad (S_1^X \cap S_2^X \cap \ldots) \quad = \quad \bigcap_{k\in\mathbb{N}} S_k^X$$

and

$$\limsup_{k\to\infty} S_k^X \quad = \quad \bigcap_{k\in\mathbb{N}} \bigcup_{i\geq k} S_k^X \quad = \quad \bigcap_{k\in\mathbb{N}} S_k^X.$$

Hence, $\lim_{k\to\infty} S_k^X = \limsup_{k\to\infty} S_k^X = \liminf_{k\to\infty} S_k^X = \bigcap_{k\in\mathbb{N}} S_k^X$. By Cantor's intersection theorem, $\bigcap_{k\in\mathbb{N}} S_k^X \neq \emptyset$. □

Moreover, $\forall x \in X^*, x \in S_k^X$ for any $k = 1, 2, ...$ and the smallest possible slice within the design space is $S^X = X^*$. We then prove that the limit of the slice regions S_k^X is X^*.

Theorem 2.4. $\lim_{k\to\infty} S_k^X = X^*$.

Proof. First, note that $\forall x^* \in X^*$, $x^* \in S_k^X$ for any $k \in \mathbb{N}$, since x^* is a global maximum. Hence, $X^* \subseteq \bigcap_{k\in\mathbb{N}} S_k^X$. What is left to be proven is that the limit of S_k^X is not a set strictly greater than X^*.

Suppose that $\tilde{S}$ is any set strictly greater than X^*, that is, $\tilde{S} \cap (X^*)^C \neq \emptyset$. Denote $S_C = \tilde{S} \cap (X^*)^C$. For any $k, S_k^X \subseteq S_{k-1}^X$, to have $\lim_{k\to\infty} S_k^X = \tilde{S}$, we need that $x_c \in S_k^X, \forall x_c \in S_c$ $\forall k \in \mathbb{N}$. By definition of S_C, $f(x_c) < f(x^*)$ for any $x_c \in S_C, x^* \in X^*$.

Given any iteration k, for $\tilde{S}$ to be the limit of S_k^X, we need $\tilde{S} \subseteq S_k^X$. By the continuity of the function f and Lemma 2.1, then for any given $x_c \in S_C$, $\exists$ an interval I_c around x^* for any $x^* \in X^*$ such that $f(x_I) > f(x_c), \forall x_I \in I_c$. Let I_C denote the largest of such intervals, i.e. $I_C = \bigcup_c I_c$. Then, the given iteration k, there is a non-zero probability that the next iterate from the forward procedure selects the next iterate from I_C. Specifically, the probability of obtaining the next iterate from I_C is given by $P_k^C = l(I_C)/l(S_k^X)$, where $l()$ denotes the Lebesgue measure. Then:

- If the next iterate $x_{k+1} \in I_C$, then the next slice region S_{k+1}^X would exclude x_c entirely, since $f(x_c) < f(x_{k+1})$.

- If the next iterate $x_{k+1} \notin I_C$, then the next slice region S_{k+1}^X would still include x_c. However, note that since $S_k^X \supset S_{k+1}^X$, $l(S_k^X) > l(S_{k+1}^X)$, and hence $P_k^C < P_{k+1}^C$. Thus, for

a sequence of iterates $x_k, x_{k+1}, ... \notin I_C$, $P_k^C < P_{k+1}^C < ...$, and hence $P_k^C \to 1$. Then, $\exists$ a number n such that $x_{k+n} \in I_C$ a.s.

Thus, $\exists$ a slice level S_{k+n}^X such that $x_c \notin S_{k+n}^X$ a.s., and so $\forall m \geq n, x_c \notin S_{k+m}^X$. Hence, $x_c \notin \bigcap_{k\in\mathbb{N}} S_k^X$. Since x_c is an arbitrary point from S_C, this is true for all points in S_C, and hence $\forall x_c \in S_C, x_c \notin \bigcap_{k\in\mathbb{N}} S_k^X$. So, we have $S_C = \emptyset$, which is a contradiction to the definition of S_C. This is true for an arbitrary $\tilde{S}$, and hence true for any $\tilde{S}$. Hence, there does not exist a set $\tilde{S}$ strictly larger than X^* such that $\lim_{k\to\infty} S_k^X = \tilde{S}$. Therefore, $\lim_{k\to\infty} S_k^X = X^*$. $\square$

It follows from the above Theorem 2.4 that the iterates converge to a global maximum, i.e. $x_k \to x^*, x^* \in X^*$.

VITA

Bob Salim was born in Jakarta, Indonesia. He moved to Seattle, Washington in 2006 to pursue Bachelor degrees in Materials Science and Engineering and in Applied and Computational Mathematical Sciences. Upon his graduation in 2010, he enrolled in the Biostatistics department at the University of Washington to pursue his doctoral degree.