Optimal Designs for Maximum Likelihood Estimation and Factorial Structure Design

by

Monsur Ahmed Chowdhury

A Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfilment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

Department of Statistics
University of Manitoba
Winnipeg, Canada.

Abstract

This thesis develops methodologies for the construction of various types of optimal designs with applications in maximum likelihood estimation and factorial structure design. The methodologies are applied to some real data sets throughout the thesis.

We start with a broad review of optimal design theory including various types of optimal designs along with some fundamental concepts. We then consider a class of optimization problems and determine the optimality conditions. An important tool is the directional derivative of a criterion function. We study extensively the properties of the directional derivatives. In order to determine the optimal designs, we consider a class of multiplicative algorithms indexed by a function, which satisfies certain conditions. The most important and popular design criterion in applications is D-optimality. We construct such designs for various regression models and develop some useful strategies for better convergence of the algorithms.

The remaining thesis is devoted to some important applications of optimal design theory. We first consider the problem of determining maximum likelihood estimates of the cell probabilities under the hypothesis of marginal homogeneity in a square contingency table. We formulate the Lagrangian function and remove the Lagrange parameters by substitution. We then transform the problem to one of maximizing some functions of the cell probabilities simultaneously. We apply this problem to some real data sets, namely, a US Migration data, and a data on grading of unaided distance vision. We solve another estimation problem to determine the maximum

likelihood estimation of the parameters of the latent variable models such as Bradley-Terry model where the data come from a paired comparisons experiment. We approach this problem by considering the observed frequency having a binomial distribution and then replacing the binomial parameters in terms of optimal design weights. We apply this problem to a data set from American League Baseball Teams.

Finally, we construct some optimal structure designs for comparing test treatments with a control. We introduce different structure designs and establish their properties using the incidence and characteristic matrices. We also develop methods of obtaining optimal R-type structure designs and show how such designs are trace, A- and MV-optimal.

Acknowledgments

First of all, I would like to express my deepest gratitude to my Ph.D. supervisor Dr. Saumen Mandal for his suggestion of the topics, his enthusiastic guidance, his huge support, patience and encouragement during the course of my research and making this work a success. Without his support, this thesis would not have been successfully completed.

Secondly, I sincerely thank my Ph.D. advisory committee members, Dr. Saman Muthukumarana, Dr. Po Yang of the Department of Statistics and Dr. S. S. Appadoo of the Department of Supply Chain Management, University of Manitoba for their valuable contributions.

I am grateful to the various members of the Department of Statistics and many friends for their generous support and assistance for the last couple of years.

I gratefully acknowledge the financial support from the Faculty of Graduate Studies, the Faculty of Science, the Department of Statistics and from Dr. Mandal's NSERC research grant.

Finally, I take this opportunity to express my profound gratitude to my wife Farmuda. Without her constant support, encouragement and love, this day would have never happened. Above all my gratefulness to the Almighty God without his mercy nothing is possible.

Dedication Page

This dissertation is dedicated to my Late Parents, Wife, my Son and Daughter

Contents

List of Tables

List of Figures

Chapter 1

Introduction

There are many problems in statistics which demand the calculation of one or more optimizing probability distributions or measures (see, e.g., Atkinson et al. (2007), Silvey (1980), Pukelsheim (1993), Berger and Wong (2009), Cook and Fedorov (1995), Torsney (1977), Mandal and Torsney (2006)). Construction of optimizing probability distributions plays an important role in many areas of statistical research. Examples arise in optimal design, optimal response-adaptive design, parameter estimation, stratified sampling, image processing and optimal structure design. Sometimes optimizing distributions are constructed subject to some specific constraints of interest. Constructing optimal designs under constraints is an important topic because the constrained optimization technique can be applied to any circumstances where the restrictions are needed. As our work is based on optimal design theory, we start with some general description of optimal design theory with some fundamental concepts and definitions.

Experiments are performed in various fields of study to answer certain questions of interest. An experiment is a well defined act or an investigation conducted to discover the underlying facts about a phenomenon, which are utilized to test some hypotheses of interest, to verify the results of previous investigations or to study

the effects of new treatments. We can define experiment as a test or series of tests in which some purposeful changes are made in the input variables of a system or process to observe and identify the reason for changes in the response output. Usually, statistical experiments are conducted in situations in which researchers can manipulate the conditions of the experiment and can control the factors that are irrelevant to the research objectives. Design of experiment is the process of planning the experiment so that the appropriate data can be collected and analyzed by some statistical methods to meet some specific objectives. Planning an experiment is important in order to ensure that the right type of data, sufficient sample size and power are available to answer the research questions of interest as clearly and efficiently as possible.

1.1 Stages of a Statistically Designed Experiment

In order to use the statistical approach in designing and analyzing an experiment, it is important and necessary for the researchers to have a clear idea of what is to be studied, how the data are to be collected and an understanding of how these data are to be analyzed. These are the guidelines for designing an experiment.

(i) **Formulation of the research problem.** In scientific research, many problems are formulated as a relationship between a set of explanatory variables X and the outcome variable Y. Therefore, it is important to first identify the set of explanatory variables X and the outcome variable Y of interest.

(ii) **Choice of the research design.** In selecting the design, it is necessary to develop the structure of the research and have the clear ideas about the objectives of the experiment. In addition to the selection of the number of independent variables,

choice of research design also involves the distinction between qualitative versus quantitative variables, random versus fixed variables and crossed or nested relationship and the selection of the number of measurements, time points and subjects within groups. Also it is important to decide on the sources and control the amount of unwanted variation in the design.

(iii) **Choice of statistical model.** In design of experiments, it is also important to choose a statistical model that describes the relationship between the response variable and the explanatory variables. A model is usually specified by the mathematical equation that describes the outcome variable Y as a function of the explanatory variables X and the error term. A statistical model describes how the change of explanatory variables will affect the response variable.

(iv) **Data collection and performing the experiment.** In this step, the data are collected based on the design chosen in Step (ii). When conducting the experiment, it is crucial to observe the process carefully to ensure that the experiment is being done according to the proposed plan.

(v) **Statistical analysis of the data.** In this stage, based on the chosen statistical model in Step (iii), statistical methods should be used to analyze the data. Regression diagnostics are useful tools to check model assumptions and model adequacy.

(vi) **Conclusions.** After the data have been properly analyzed, the researcher must carefully draw practical conclusions about the results and recommend the future course of action.

It is important to note that optimal design plays an important role in Step (ii). In a

statistical model, the focus is in good estimation of the parameters of the model. There are a variety of criteria defining good estimation. We choose an appropriate design to optimize a chosen criterion. The way of doing this is called optimal design. The general theory of optimal design was originally developed for linear models. We will discuss some basic concepts of optimal design theory such as the definition of a design measure, variance function, information matrix, various criterion functions and their properties.

Some advantages of optimal designs are:

(1) Using optimal design we can obtain the best selection of the input (factors) for which the optimal value of each response occurs.

(2) Using optimal design theory statistical models can be estimated with fewer experimental runs which reduce the costs of experimentation. If we select the design points in an efficient way, we can obtain better precision or good estimation of the parameters by taking a small or moderate sample size.

(3) Optimal designs alllow multiple types of factors, such as process, mixture, and discrete factors.

(4) Designs can be optimized when the design space is constrained, for example, when the mathematical process involves factor settings that are practically infeasible due to safety concerns.

(5) Different algorithms are available for the construction of optimal designs. Later on we will discuss construction of a flexible design, called an approximate design.

We now consider the problem of selecting an experimental design and assume a simple probability model of the type

$$y \sim g(y \mid \underline{x}, \underline{\theta}, \sigma), \tag{1.1}$$

where y is the response variable, $\underline{x}$ is a vector of the design variables such that $\underline{x} \in \mathcal{X} \subseteq \mathbb{R}^m$, $\mathcal{X}$ is the design space. Typically the design space is continuous but can be discrete. The vector $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)^T$ are unknown parameters. The true values of $\underline{\theta}$ belong to a set $\Theta \in \mathbb{R}^k$. The parameter σ is fixed and unknown, and is considered as a nuisance parameter. $g(.)$ is a probability model.

The experimental conditions from the given experimental domain $\mathcal{X}$ can be chosen freely by the experimenter. The experimental domain $\mathcal{X}$ is considered to be compact in most of the applications. That is, the design space is closed and bounded. For each $x \in \mathcal{X}$, we observe the response variable $y = y(\underline{x})$ which is a random variable with $var(y(\underline{x})) = \sigma^2$. We generally assume that σ is independent on the experimental condition $\underline{x}$.

In linear regression design the conditional mean of $y(\underline{x})$ should be linear in the unknown parameters $\underline{\theta}$ and $y(\underline{x})$ has an expected value of the explicit form:

$$E(y|\underline{x}, \underline{\theta}, \sigma) = \underline{f}^T(\underline{x})\underline{\theta}, \tag{1.2}$$

where $\underline{f}(\underline{x}) = (f_1(\underline{x}), f_2(\underline{x}), \ldots, f_k(\underline{x}))^T$ is a vector of k real valued functions defined on $\mathcal{X}$. The functions $f_1, f_2, \ldots, f_k$ are known to the experimenter.

A value for $\underline{x}$ must be chosen at first from $\mathcal{X}$ in order to get an observation on y. It is supposed that $\underline{x}$ can be set to any chosen value in the design space $\mathcal{X}$. A natural

question is at what values of the design variables $\underline{x}$, observations, say n, should be taken on y in order to obtain the best inference or as reliable inference as possible for all or some of the parameters $\underline{\theta}$. Such a 'best' selection of the values of the design variables or allocation of the n observations to the elements of the design space is termed an optimal regression design or simply an optimal design. The methods of optimal design were originally developed for the choice of those values of the explanatory variables $\underline{x}$ in a regression model at which observations should be taken in order to obtain good estimation of the parameters.

As an example, in a chemical experiment, there may be several factors, such as time of reaction, temperature, pressure and catalyst concentration, that affect the response which is a smooth function of these variables. The question is at what combinations of these variables should measurements be taken in order to obtain good estimates of the dependence of responses on these variables.

1.2 Examples

We consider two simple examples, namely simple linear regression and quadratic regression in the case of a single quantitative variable (Atkinson et al. (2007)).

The simple linear regression model is

$$E(y|x) = \theta_0 + \theta_1 x. \tag{1.3}$$

It is important to specify the design matrix X in order to design the experiment. For a design having $n = 3$, the design matrix can be figured out by the following matrix

$$\tilde{X} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

and by the model. Suppose the design space is $\mathcal{X} = [-1, 1]$. The design problem can be formulated to choose n points in $\mathcal{X}$ so that the linear relationship between y and x given by (1.2) can be estimated as precisely as possible. One possible choice for this purpose consists of trials at three equally spaced values of x, that is,

$$\tilde{X} = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}.$$

Another possibility is that the design may consist of two trials at one end of the design space and one at the other. Then matrix $\tilde{X}$ looks like as follows

$$\tilde{X} = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}.$$

Now consider another model, say, a quadratic regression model, which is given by

$$E(y|x) = \theta_0 + \theta_1 x + \theta_2 x^2. \tag{1.4}$$

Suppose the design space here also is $\mathcal{X} = [-1, 1]$. This model allows curvature in the regression of y on x and trials are required for at least three different values of x in order to estimate the three parameters. One possible choice is that the design consists of trials at four equally spaced values of x. So the matrix $\tilde{X}$ becomes

$$\tilde{X} = \begin{pmatrix} -1 \\ -1/3 \\ 1/3 \\ 1 \end{pmatrix}.$$

This would allow detection of departures from the quadratic model. Then the full design matrix for the quadratic model is

$$\begin{pmatrix} 1 & -1 & 1 \\ 1 & -1/3 & 1/9 \\ 1 & 1/3 & 1/9 \\ 1 & 1 & 1 \end{pmatrix},$$

where the final column gives the values of x^2.

Now, the question naturally arises how we select the values of the design variable. We discuss this in the following. First, we have to decide the mode of inference. Suppose that it is on point estimation. However the solution proposed for this case will also hold good to other modes of inference too.

It is now desired to choose n values $(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n)$ to yield 'best' point estimates $\hat{\underline{\theta}}$ of some or all of the parameters $\underline{\theta}$. Suppose the estimator $\hat{\underline{\theta}}$ of $\underline{\theta}$ is obtained by some method of point estimation and $\hat{\underline{\theta}}$ is unbiased estimator for $\underline{\theta}$. Usually the components $\hat{\theta}_j$ will be correlated. Possibly then the dispersion matrix (variance covariance matrix) of order $k \times k$ of $\hat{\underline{\theta}}$ about $\underline{\theta}$ is $D(\hat{\theta})$= $E([\hat{\underline{\theta}} - \underline{\theta}][\hat{\underline{\theta}} - \underline{\theta}]^T)$. It holds information about the accuracy of $\hat{\underline{\theta}}$ not only in its diagonal elements, which of course measure the mean squared deviation of $\hat{\theta}_j$ from θ_j, $j = 1, 2, \ldots, k$, but also in its off-diagonal elements. In general, the "smaller" is $D(\hat{\underline{\theta}})$ the better is the accuracy of $\hat{\underline{\theta}}$.

Now suppose the model (1.2) is true. Let y_i denote the observations obtained at $\underline{x}_i$ so that

$$E(y_i) = \underline{f}^T(x_i)\,\underline{\theta}, i = 1, 2, \ldots, n. \tag{1.5}$$

Suppose that y_i's are independent random variables with equal variance σ^2. Note that there may be several equalities between the $\underline{x}_i$'s, where more than one observation can be taken at the same $\underline{x}$ value. Then the standard linear model is given by

$$E(\underline{Y}) = X\underline{\theta}, D(\underline{Y}) = \sigma^2 I_n, \tag{1.6}$$

where $\underline{Y} = (y_1, y_2, \ldots, y_n)$, X is the $n \times k$ matrix whose $(i, j)th$ element is $f_j(\underline{x}_i)$, I_n is the $n \times n$ identity matrix and $D(\underline{Y})$ denotes the variance covariance matrix of $\underline{Y}$.

The best linear unbiased estimators (BLUE) of the $\underline{\hat{\theta}}$ of $\underline{\theta}$ is the solution of the equations

$$(X^T X)\underline{\hat{\theta}} = X^T \underline{Y}. \tag{1.7}$$

The matrix $(X^T X)$ is the information matrix for $\underline{\theta}$ of order $k \times k$. The larger the matrix $(X^T X)$, the greater is the information in the experiment. If all the parameters $\underline{\theta}$ are to be estimated, then $\underline{x}$ must at least be chosen to ensure that the matrix $(X^T X)$ is non-singular. Then the unique solution for (1.7) is given by:

$$\underline{\hat{\theta}} = (X^T X)^{-1} X^T \underline{Y} \tag{1.8}$$

with

$$\mathrm{E}(\underline{\hat{\theta}}) = \underline{\theta},$$
$$\mathrm{D}(\underline{\hat{\theta}}) = \sigma^2 (X^T X)^{-1}.$$

The predicted value of the response at $\underline{x}$ is

$$\begin{aligned}\hat{Y}(\underline{x}) &= f_1(\underline{x})\hat{\theta}_1 + f_2(\underline{x})\hat{\theta}_2 + \ldots + f_k(\underline{x})\hat{\theta}_k \\ &= \underline{f}^T(\underline{x})\underline{\hat{\theta}},\end{aligned}$$

where $\underline{f}(\underline{x}) = (f_1(\underline{x}), f_2(\underline{x}), \ldots, f_k(\underline{x}))^T$.

From the above, it is clearly noticed that the variance covariance matrix of $\underline{\hat{\theta}}$ does not depend on $\underline{\theta}$. But it is proportional to the parameter σ^2. In order to obtain a better inference for $\underline{\hat{\theta}}$, we need to select $\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n\}$ to make the matrix $D(\underline{\hat{\theta}})$ as small as possible, namely a $\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n\}$ which makes the matrix $(X^T X)$ large in some sense.

1.3 Discretizing the Design Space

The linear model in (1.2) can be written as:

$$E(y|\underline{v}, \underline{\theta}, \sigma) = \underline{v}^T \underline{\theta}, \tag{1.9}$$

where

$$\begin{aligned}\underline{v} &= (f_1(\underline{x}), f_2(\underline{x}), \ldots, f_k(\underline{x}))^T, \underline{v} \in \mathcal{V}, \\ \mathcal{V} &= \{\underline{v} \in \mathbb{R}^k : \underline{v} = (f_1(\underline{x}), f_2(\underline{x}), \ldots, f_k(\underline{x}))^T, \underline{x} \in \mathcal{X}\}.\end{aligned}$$

From the above, it is obvious that choosing a vector $\underline{x}$ in the design space $\mathcal{X}$ is equivalent to choosing a k-vector $\underline{v}$ in the closed bounded k-dimensional space $\mathcal{V} = \underline{f}(\mathcal{X})$, where $\underline{f}$ is the vector valued function $(f_1, f_2, \ldots, f_k)^T$. $\mathcal{V}$ is the image under f of $\mathcal{X}$. So, we call $\mathcal{V}$ as an induced design space. Usually this design space

is continuous but we can assume that $\mathcal{V}$ is discrete. We will justify this later using Caratheodory's theorem.

Suppose, $\mathcal{V}$, the discrete design space, consists of J distinct vectors $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_J$. In order to obtain an observation on y, a value for $\underline{v}$ must first be chosen from the J elements of $\mathcal{V}$ to be the point at which the observation is taken.

After discretization, we can express the design problem more concisely. At what points $\underline{v}_j$ should observations be taken and, if n observations in total are allowed, how many observations should be taken at these points in order to obtain 'best' least squares estimators of $\underline{\theta}$? Given that we have n observations in total, we decide how many of these, say n_j to take at $\underline{v}_j$ subject to $\sum_{j=1}^{J} n_j = n$. Thus, the matrix $(X^T X)$ can be expressed in the form:

$$X^T X = M(\underline{n}), \ \underline{n} = (n_1, n_2, \ldots, n_J)^T, \tag{1.10}$$

where

$$M(\underline{n}) = \sum_{j=1}^{J} n_j \underline{v}_j \underline{v}_j^T = VNV^T, \quad V = [\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_J], \quad N = diag(n_1, n_2, \ldots, n_J).$$

1.3.1 Exact vs. Approximate Design

In this section, we focus on the distinction between two types of designs: exact and approximate designs. Suppose our goal is now to choose the n_j's to make the matrix $M(\underline{n})$ big in some sense. Given the condition that the n_j's must be integers, the problem becomes an integer programming problem and in the optimal design context this is described as an 'exact design' problem.

The integer programming problems are usually hard to solve even without additional constraints, mainly because the theory of calculus cannot be used to identify the optimal solutions. Note also that, a solution would have to be worked out separately for different values of n. There is no formula for an optimal exact design that would express it as a function of n.

Approximate design is another way to assign a given number of n subjects. Unlike the optimal exact designs, the optimal approximate design is independent of the value of n. We can formulate an approximate design in the following way. First note that, $M(\underline{n})$ can be written as:

$$M(\underline{n}) = nM(p), \tag{1.11}$$

where

$$M(p) = \sum_{j=1}^{J} p_j \underline{v}_j \underline{v}_j^T \tag{1.12}$$

$$= VPV^T \tag{1.13}$$

and $P = diag(p_1, p_2, \ldots, p_J)$. Here $p_j = n_j/n$, the proportion of n observations taken at $\underline{v}_j$ such that $p_j \geq 0$, $\sum_{j=1}^{J} p_j = 1$. Thus, $p = (p_1, p_2, \ldots, p_J)$ can be thought of as a resultant probability distribution on the induced design space $\mathcal{V}$.

So our problem is now to choose p to make $M(p)$ large subject to $p_j = n_j/n$. Thus, we have the basic constraints on design weights, as given by, $p_j \geq 0$ and $\sum_{j=1}^{J} p_j = 1$. This generates an 'approximate design' problem.

As we see, an approximate design problem is a simpler or more flexible problem to solve and is not much visibly different from the original exact design problem.

Note that the information matrix $M(p)$ is actually the expectation of $\underline{v}\underline{v}^T$ by considering p as a probability distribution on $\mathcal{V}$. That is,

$$M(p) = E_p[\underline{v}\underline{v}^T]. \tag{1.14}$$

We will see that this information matrix plays an important role in the construction of optimal designs throughout the thesis.

1.3.2 Design Measure

Earlier we denoted p as the vector of weights $(p_1, p_2, \ldots, p_J)$. We also considered p as a probability distribution on $\mathcal{V}$. Since $\mathcal{V}$ is an induced design space, we can also think of p as a probability distribution or measure on the original design space $\mathcal{X}$. Thus, we can write the design measure as given by

$$p = \left\{ \begin{array}{cc} \underline{x}_1 & \underline{x}_2 \cdots \underline{x}_J \\ p_1 & p_2 \ldots p_J \end{array} \right\}, \tag{1.15}$$

where $\underline{x}_j$'s are the design points and p_j's are the associated design weights, $0 \leq p_j \leq 1$, $j = 1, 2, \ldots, J$, $\sum_{j=1}^{J} p_j = 1$.

1.3.3 Support of a Design Measure

The support of the design measure p is defined to be those vertices (or design points) which have nonzero weights under p. In terms of the induced design space, it is given by

$$Supp\,(p) = \{\underline{v}_j \in \mathcal{V} : p_j > 0, j = 1, 2, \ldots, J\}.$$

Consider now the case of $\mathcal{V}$ continuous. The question arises that what will be an optimal solution in this case? Note that an exact design would allow a discrete probability distribution on $\mathcal{V}$. It would allocate weights to a finite set of points in $\mathcal{V}$ and zero weight to all other points.

However it will be no less difficult to discover an optimal exact design in the case of $\mathcal{V}$. An approximating problem arises in this case. It is called the continuous optimal design, and the problem can be stated as:
Find a probability measure p^* on the continuous space $\mathcal{V}$ which maximizes a function of the information matrix $M(p)$ where

$$M(p) = E_p(vv^T) = \int_{\mathcal{V}} vv^T dp(v).$$

An optimizing p^* would be a continuous probability measure on $\mathcal{V}$ and the above problem would be more difficult to solve. However, we will see that Caratheodory's theorem (Silvey (1980)) guarantees that there is at least one solution to the above problem which is a discrete solution. We now state the Caratheodory's Theorem.

Let $\mathcal{M} = \{M(p) : p \text{ is any probability measure on } \mathcal{V}\}$. Each point in M of the convex hull $\mathcal{M}$ of any subset $\mathcal{U}$ of n-dimensional space can be represented as a convex combination of at most $n + 1$ points of $\mathcal{U}$ given by

$$M = \sum_{j=1}^{n+1} \alpha_j u_j, \alpha_j \geq 0, \sum_{j=1}^{n+1} \alpha_j = 1, u_j \in \mathcal{U}.$$

If M is a boundary point of $\mathcal{M}$, then $\alpha_{(n+1)}$ can be set to zero. Such representations are not unique (Fedorov (1972)).

The set $\mathcal{M}$ is the convex hull of the set $\{\underline{v}\,\underline{v}^T : \underline{v} \in \mathcal{V}\}$. Now applying the above theroem in our case, it follows that each $M \in \mathcal{M}$ has at least one representation of the form

$$M = \sum_{j=1}^{J} p_j \underline{v}_j \underline{v}_j^T,$$

where $\underline{v}_j \in \mathcal{V}$, $j = 1, 2, \ldots, J$ and $J \leq \{[k(k+1)/2] + 1\}$. Also by the same theorem if M is a boundary point of $\mathcal{M}$, then the inequality becomes $J \leq k(k+1)/2$.

Typically J will be smaller than the above limits. If $J < k$ then M will be singular, since then the rank of M can be at most J.

Thus we can justify that any continuous optimal design measure can be replaced by at least one finite discrete probability distribution. So we have provided a justification for having initially assumed induced design space discrete.

A standard discretization would be some form of uniform grid on a continuous $\mathcal{V}$, but usually this is hard to determine when $\mathcal{V}$ is an image under some $\underline{f}$ of $\mathcal{X}$. In reality, the discretization that is used in obtaining an optimal solution is the image under $\underline{f}$ of a uniform grid on $\mathcal{X}$.

1.3.4 Standardized Variance of the Predicted Response

The standardized variance of the predicted response on y at $\underline{x}$ for the design (1.15) is given by

$$d(\underline{x}, p) = \underline{f}^T(\underline{x}) M^{-1}(p) \underline{f}(\underline{x}) \tag{1.16}$$

where $M(p)$ is the information matrix.

Proof. We know the predicted value of the response at $\underline{x}$ is

$$\hat{Y}(\underline{x}) \quad = \quad \underline{f}^T(\underline{x})\hat{\underline{\theta}}.$$

So the variance of the predicted value of the response is

$$Var(\hat{Y}(\underline{x})) = \underline{f}^T(\underline{x})Var(\hat{\underline{\theta}})\underline{f}(\underline{x}) = \sigma^2\underline{f}^T(\underline{x})(X^TX)^{-1}\underline{f}(\underline{x}).$$

If the design has n trials, then following (1.11) the standardized variance of the predicted response on y at $\underline{x}$ is

$$d(\underline{x}, p) = n\frac{Var(\hat{Y}(\underline{x}))}{\sigma^2} = n\frac{\sigma^2\underline{f}^T(\underline{x})(X^TX)^{-1}\underline{f}(\underline{x})}{\sigma^2} = \underline{f}^T(\underline{x})M^{-1}(p)\underline{f}(\underline{x}).$$

1.3.5 Properties of the Information Matrix

The definition of the information matrix (1.12) is given here again

$$M(p) \quad = \quad \sum_{j=1}^{J} p_j\underline{v}_j\underline{v}_j^T \quad = \quad VPV^T.$$

First note that, the information matrix is symmetric and nonnegative definite. The symmetry of $M(p)$ follows from the above expression. The nonnegativeness of the appropriate quadratic form can be verified by the following:

$$\begin{aligned}
\underline{x}^TM(p)\underline{x} \quad &= \quad \underline{x}^TE_p[\underline{v}\underline{v}^T]\underline{x} \\
&= \quad E_p[\underline{x}^T\underline{v}\underline{v}^T\underline{x}] \\
&= \quad E_p[(\underline{x}^T\underline{v})^2] \geq 0.
\end{aligned}$$

The information matrix is widely used in optimal experimental design. The inverse of the variance-covariance matrix (dispersion matrix) of $\underline{\hat{\theta}}$ is actually the information matrix. In the following section, we will see that many of the design criteria are functions of the above information matrix $M(p)$.

1.4 Criteria of Optimality and Their Properties

We may have different designs in terms of the values of the determinant of the information matrix $M(p)$ or its inverse, average or total variance of the least squares estimators of the parameters, the standardized variance $d(\underline{x}, p)$ over the design space $\mathcal{X}$, or the largest eigenvalue of the inverse of the information matrix.

As we have seen earlier that the information matrix $M(p)$ plays an important role in optimal design theory and in obtaining a best inference for all or some of the unknown parameters $\underline{\theta} \in \Theta$ by making the matrix $M(p)$ large in some sense. Different experiments may have different ways of making $M(p)$ large. So we maximize some real valued function $\phi(p) = \psi\{M(p)\}$. Such a function ϕ is called the criterion function, and in turn, the criterion defined by the function ϕ is usually called ϕ-optimality. A design maximizing $\phi(p)$ is called a ϕ-optimal design.

There is an extensive literature available for different types of criteria in the literature. For example, see Kiefer (1959), Fedorov (1972), Silvey (1980), Atkinson et al. (2007), Pukelsheim (1993), John and Draper (1975), Berger and Wong (2009), Shah and Sinha (1989), and Wynn (1972).

Now we consider some of the design criteria of interest and their properties. In general, we can divide the set of criteria into two cases. We first consider the case

in which interest is in inference about all of the parameters $\underline{\theta}$ of the linear model (1.9). We assume that $J > k$ so that the information matrix $M(p)$ is non-singular and hence positive definite. Possible criteria in this case include D-optimality, A-optimality, E-optimality and G-optimality. The second case is about the criteria when the experimenter is only interested in some of the unknown parameters or some linear functions of the parameters of the linear model (1.9). Such criteria include D_A-optimality, D_S-optimality, linear optimality, c-optimality and E_A-optimality. Note that, to keep uniformity, we express the criterion function of each of these criteria in terms of a maximization problem.

1.4.1 *D*-optimality

The D-optimality criterion is defined by the criterion function:

$$\phi_D(p) = \psi_D\{M(p)\} = logdet\{M(p)\} = -logdet\{M^{-1}(p)\}.$$

This is the most important and popular design criterion which seeks to maximize the determinant of the information matrix of the design. In other words, a D-optimal design minimizes the generalized variance of the parameter estimates.

Various motivations exists for D-optimality. A D-optimal design minimizes the product of the squared lengths of the axes of the ellipsoid. This design thus minimizes the volume of the confidence ellipsoid. This property provides a natural interpretation of this criterion in terms of confidence intervals for the parameters in a regression model. Other motivations of the D-optimality lie in hypothesis testing problems under the assumption of a normal linear model. An advantage of D-optimality is

that the optimum designs for quantitative factors do not depend upon the scale of the independent variable. That is, if we change the design space, we can easily obtain the D-optimal design directly from the one constructed over the original design space. This property is not in general hold for other optimality criteria. This is why D-optimality is the most extensively studied of all design criteria.

The D-optimality criterion $[\phi_D(p)]$ has several other useful properties. We will discuss all these properties in detail and construct D-optimal designs for some regression models in Chapter 3.

1.4.2 A-optimality

A-optimality is defined by the criterion function:

$$\phi_A(p) = \psi_A\{M(p)\} = -Trace\{M^{-1}(p)\}.$$

As we see the criterion function, an A-optimum design seeks to minimize the trace of the inverse of the information matrix. So this criterion results in minimizing the average or total variance of the parameter estimates, but does not take correlations between these estimates into account. An important reference of this criterion is Elfving (1952). Note that the greatest lower bound for the trace of the covariance matrix of the least squares estimator can be obtained by using the idea of majorization and Schur convexity. See Chan and Li (1989) and Chan (1987) for further details.

Note that this criterion $\phi_A(p)$ is simple to evaluate since it only requires the computation of the k diagonal elements of the matrix $M^{-1}(p)$. Note that, unlike D-optimality, the A-optimality may not be invariant under linear transformations of the

scale of the independent variables. This means that each scale may lead to another optimal design.

Properties of A-optimality

(i) ϕ_A is an increasing function over the set of positive definite symmetric matrices.

(ii) ϕ_A is concave on $\mathbb{M}$, where $\mathbb{M}$ is the set of all positive definite symmetric matrices.

(iii) ϕ_A is differentiable whenever it is finite, and the first derivative is given by

$$\frac{\partial \phi_A}{\partial p_j} = \underline{v}_j^T M^{-2}(p) \underline{v}_j.$$

1.4.3 *G*-optimality

G-optimality is defined by the criterion function:

$$\phi_G(p) = \psi_G\{M(p)\} = -\max_{\underline{v} \in \mathcal{V}} \underline{v}^T M^{-1}(p) \underline{v}.$$

As we see, in this optimality, we minimize the maximum value of $\underline{v}^T M^{-1}(p) \underline{v}$ which is proportional to the variance of $\underline{v}^T \underline{\hat{\theta}}$.

Properties of G-optimality

(i) $\phi_G(p)$ is an increasing function over the set of positive definite symmetric matrices.

(ii) $\phi_G(p)$ is concave on the set of positive definite symmetric matrices.

(iii) $\phi_G(p)$ is invariant under a non-singular linear transformation on the induced design space $\mathcal{V}$. This can be seen by the following proof. Suppose $\mathcal{V} = [\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_J]$ is transformed to $\mathcal{W} = [\underline{w}_1, \underline{w}_2, \ldots, \underline{w}_J]$ under the linear transformation $\underline{w}_j = A\underline{v}_j$, where A is a $k \times k$ non-singular matrix. Then a design assigning weight p_j to $\underline{w}_j$ has the criterion function:

$$\begin{aligned}
\phi_G(p) &= \psi_G\{M_\omega(p)\} \\
&= -\max_{\underline{\omega} \in \mathcal{W}} \underline{\omega}^T M_\omega^{-1}(p) \underline{\omega} \\
&= -\max_{A\underline{v} \in \mathcal{W}} (A\underline{v})^T (AVPV^TA^T)^{-1} (A\underline{v}) \\
&= -\max_{\underline{v} \in A^{-1}\mathcal{W}} \underline{v}^T A^T (A^T)^{-1} (VPV^T)^{-1} A^{-1} A\underline{v} \\
&= -\max_{\underline{v} \in \mathcal{V}} \underline{v}^T M^{-1}(p) \underline{v} \\
&= \psi_G\{M(p)\} \\
&= \phi_G(p).
\end{aligned}$$

(iv) Suppose that uniquely $\underline{v}_j^T M^{-1}(p) \underline{v}_j = \max_t \underline{v}_t^T M^{-1}(p) \underline{v}_t$. Then the G-optimality crietrion has unique partial derivatives corresponding to positive weights, and are given by

$$\frac{\partial \phi_G}{\partial p_j} = [\underline{v}_j^T M^{-1}(p) \underline{v}_j]^2.$$

Using this criterion we obtain an accurate and efficient prediction of the outcome variable of interest. Kiefer and Wolfowitz (1960) proved the equivalence of this

criterion and the D-optimal criterion. It is well known that the standardized variance of the predicted response for a G-optimal design (say p^*) is always less than or equal to the number of parameters (k) in the model. That is, $d(\underline{x}, p^*) \leq k$ with equality at the support points. Because of the equivalency of D and G optimality this inequality can be used to check whether a design is D-optimal or not.

1.4.4 *E*-optimality

In this optimality, the squared length of the largest axis of the confidence ellipsoid is minimized. Thus we see that the name E of this criterion stands for the extreme axis of the confidence ellipsoid. This optimality criterion is defined by the criterion function:

$$\phi_E(p) = \psi_E\{M(p)\} = -\lambda_{max}[M^{-1}(p)]$$

where $\lambda_{max}[M^{-1}(p)]$ denotes the largest eigenvalue of $M^{-1}(p)$ [Kiefer (1974)].

Properties of E-optimality

(i) $\phi_E(p)$ is an increasing fumction over the set of positive definite symmetric matrices.

(ii) $\phi_E(p)$ is a concave function on the set of positive definite symmetric matrices.

(iii) If λ_{max} is unique then this criterion function has unique partial derivatives corresponding to positive weights.

1.4.5 D_A-optimality

Sometimes our interest is not in all of the parameters of the linear model. We may be interested only in some of the unknown parameters or some linear functions of the parameters. Note that we have a total of k parameters in the model. Suppose that we are interested in s linear combinations of $\underline{\theta}$ which are the elements of $A\underline{\theta}$, where A is a $s \times k$ matrix of rank $s \leq k$. The variance-covariance matrix of the least squares estimator of $A\underline{\theta}$ is proportional to $AM^{-1}(p)A^T$. So, a modification of the D-optimality criterion can be adjusted based on the matrix $AM^{-1}(p)A^T$ for this purpose.

The criterion function for this optimality is then defined by

$$\phi_{D_A}(p) = \psi_{D_A}\{M(p)\} = -log|AM^{-1}(p)A^T|.$$

Sibson (1974) called this as D_A-optimality criterion to indicate that the design criterion is dependent on the matrix of coefficients A.

Properties of D_A-optimality:

(i) $\phi_{D_A}(p)$ is an increasing function over the set of positive definite symmetric matrices.

(i) $\phi_{D_A}(p)$ is a concave function over the set of positive definite symmetric matrices.

(iii) $\phi_{D_A}(p)$ has unique partial derivatives corresponding to positive weights. These are given by

$$\frac{\partial \phi_{D_A}}{\partial p_j} = \underline{v}_j^T M^{-1}(p)\, A^T [AM^{-1}(p)A^T]^{-1} AM^{-1}(p)\underline{v}_j.$$

1.4.6 D_S-optimality

D_S-optimality is a special case of D_A-optimality criterion. This criterion is appropriate when we are interested in estimating a subset of s parameters in the linear model.

Let $A = [I_s : O]$, where I_s is the $s \times s$ identity matrix and O is the $s \times (k - s)$ zero matrix. Then we can partition the information matrix $M(p)$ as follows:

$$M(p) = \begin{bmatrix} M_{11}^{s\times s} & M_{12}^{s\times(k-s)} \\ M_{12}^{T} & M_{22}^{(k-s)\times(k-s)} \end{bmatrix}.$$

The matrix $(AM^{-1}(p)A^T)^{-1}$ can be written as $M_{11} - M_{12}M_{22}^{-1}M_{12}^T$ [Rohde (1965), Torsney (1981)]. So in this criterion we maximize the determinant of this matrix.

So maximizing the D_A-optimality criterion $\phi_{D_A}(p)$ in this case is equivalent to maximizing:

$$\phi_{D_S}(p) = logdet\{M_{11} - M_{12}M_{22}^{-1}M_{12}^T\}.$$

This criterion is known as the D_S-optimal criterion [see Karlin and Studden (1966), Atwood (1969), Silvey and Titterington (1973) and Silvey (1980)].

As D_S-optimality is a special case of D_A-optimality, this criterion shares similar properties to the D or D_A-optimality.

1.4.7 E_A-optimality

This is a special case of E-optimality. E_A-optimality is defined by the following criterion function:

$$\phi_{E_A}(p) = \psi_{E_A}\{AM^{-1}(p)A^T\} = max[-\lambda_{max}(AM^{-1}(p)A^T)],$$

where $\lambda_{max}(AM^{-1}(p)A^T)$ denotes the largest eigenvalue of the matrix $AM^{-1}(p)A^T$.

The properties of E_A-optimality are similar to those of E-optimality.

1.4.8 Linear Optimality

Linear optimality criterion is a very flexible optimality criterion. It is a class of criteria because it generates several other optimality criteria as special cases. Let L be a $k \times k$ matrix of coefficients. The criterion function for L-optimality is defined as

$$\phi_L(p) = \psi_L\{M(p)\} = -tr\{M^{-1}(p)L\}.$$

As we can see, this criterion function is linear in the elements of the covariance matrix $M^{-1}(p)$. If the rank of the matrix L is s with $s \leq k$, the form of the matrix L can be expressed as $L = A^TA$ where A is a $s \times k$ matrix of rank s.

Then we can wrtie the criterion function as given by

$$\phi_L(p) = -tr\{M^{-1}(p)L\} = -tr\{M^{-1}(p)A^TA\} = -tr\{AM^{-1}(p)A^T\}.$$

Note that, the above form of the criterion function indicates a relationship with the D_A-optimum design criterion function $\phi_{D_A}(p)$, where the determinant of $-AM^{-1}(p)A^T$ is maximized.

This criterion is discussed in detail by Fedorov (1972), Tsay (1976) and Tsay (1977). If the coefficient matrix L is an identity matrix I, then the criterion becomes the A-optimality criterion.

Properties of Linear Optimality:

(i) $\phi_L(p)$ is an increasing function over the set of positive definite symmetric matrices.

(ii) The criterion function $\phi_L(p)$ is a concave function over the set of positive definite symmetric matrices.

(iii) The partial derivatives of $\phi_L(p)$ corresponding to positive weights are given by

$$\frac{\partial \phi_L}{\partial p_j} = \underline{v}_j^T M^{-1}(p) A^T A M^{-1}(p) \underline{v}_j.$$

1.4.9 *c*-optimality

This is an important special case of linear optimality.

Note that in c-optimality we treat A as $A = \underline{c}^T$, where $\underline{c}$ is a $k \times 1$ vector. An important reference of this criterion in the literature is Elfving (1952). This optimality seeks to maximize the criterion function:

$$\phi_c(p) = -\underline{c}^T M^{-1}(p) \underline{c}.$$

In other words, this criterion seeks to minimize $\underline{c}^T M^{-1}(p) \underline{c}$. Thus, as we can see, in c-optimality, our interest is in estimating the linear parametric function $\underline{c}^T \underline{\theta}$ with minimum variance.

Similar to the linear optimality criterion, the partial derivatives of ϕ_c are given by

$$\frac{\partial \phi_c}{\partial p_j} = [\underline{c}^T M^{-1}(p) \underline{v}_j]^2.$$

As a special case of linear optimality, c-optimality possesses similar properties as linear optimality.

The above criteria are the popular ones in optimal design literature. There are several classes of optimization problems or optimal designs in which all the traditional optimality criteria can be considered. Instead of using the optimality criteria separately, one may consider the universal optimality that includes almost all alphabetical optimality such as A, D and E-optimality. In this case, one may study the robustness property of the optimality criteria. According to the theory of universal optimality of Kiefer (1974), the robustness with respect to the changes in an optimality criterion is greater than the robustness with respect to changes in the model under consideration.

The rest of the thesis is as follows: In Chapter 2, we study a class of optimization problems and determine the optimality conditions. An important tool is the directional derivative of a criterion function. We define two types of such directional derivatives and study their properties extensively. We then discuss a class of multiplicative algorithms along with its properties for constructing an optimal design. In Chapter 3, we focus on constructing some D-optimal designs and attempt to improve the convergence of the algorithm by exploring the properties of the directional derivatives. The remaining chapters are devoted to some important applications of optimal design theory. In Chapter 4, we solve an estimation problem for determining the maximum likelihood estimates of the cell probabilities under the hypothesis of marginal homogeneity for a square contingency table. In Chapter 5, we consider another estimation problem of determining maximum likelihood estimates of the parameters of the Bradley-Terry model where the data comes from a paired comparisons experiment. In Chapter 6, we construct optimal structure design for comparing test treatments with a control and

establish properties of the structure (k_1), structure (k_2) and structure (k_1k_2) designs using the characteristic and incidence matrix of the design. In Chapter 7, we conclude the thesis by summarizing the main findings and highlighting some potential future research work.

Chapter 2

Optimality Conditions and a Class of Algorithms

2.1 Introduction

In this chapter we determine the optimality conditions in order to find the optimal designs for our optimization problems. We define the optimality conditions in terms of the vertex directional derivatives of a criterion function $\phi(.)$. The directional derivative $F_\phi\{p, q\}$ of a criterion function $\phi(.)$ at p in the direction of q is an important tool. We discuss the properties of the directional derivatives. We also discuss further properties of the directional derivatives when the the criterion function is differentiable. We also consider some important optimality theorems relevant to our optimization problems in this thesis.

In order to find an optimizing distribution, we first consider a class of optimization problems. Optimal regression designs are particular examples of such optimization problems. Other examples of such optimization problems are maximum likelihood estimation, stratified sampling, image processing, optimal response-adaptive design

and optimal structure design.

Our problem is to maximize a criterion function $\phi(p)$ subject to the basic constraints of the design weights. That is,

$$Maximize \;\; \phi(p) \;\; over \;\; \mathcal{P} \equiv \left\{ p = (p_1, p_2, \ldots, p_J) : p_j \geq 0, \sum_{j=1}^{J} p_j = 1 \right\}. \qquad (2.1)$$

We call this optimization problem as our general problem.

Thus, our goal is to find an optimal design based on a chosen criterion function. We choose the proportion p_j of observations, taken at the design point x_j to obtain good estimators of the parameters $\underline{\theta}$. Recall that we can think of p as a probability distribution on the induced design space $\mathcal{V}$, where $\mathcal{V} = (\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_J)$. Since $\mathcal{V}$ is an induced design space, we can also consider p as a probability distribution on the original design space $\mathcal{X}$.

Now we consider the above general problem in order to find an optimizing distribution (say, p^*). In order to find an optimal design or an optimizing distribution, we need to determine conditions for optimality. Before we determine the optimality conditions, we need to consider a class of optimization problems.

2.2 A Class of Optimization Problems

We start with the above general problem.

Problem (P1)

Maximize a criterion $\phi(p)$ over $\mathcal{P} \equiv \left\{ p = (p_1, p_2, \ldots, p_J) : p_j \geq 0, \sum_{j=1}^{J} p_j = 1 \right\}$.

This is a constraint optimization problem in terms of the design weights p_j's. The full constraint region is a closed bounded convex set.

Problem (P2)

Maximize $\psi(x)$ over the convex hull (of the points $G(\underline{v}_1), G(\underline{v}_2), \ldots, G(\underline{v}_J)$)

$$\mathcal{CH}\{\mathcal{G}(\mathcal{V})\} = \{x = x(p) = \sum_{j=1}^{J} p_j G(\underline{v}_j) : p = (p_1, p_2, \ldots, p_J) \in \mathcal{P}\},$$

where $G(.)$ is a given one to one function and $\mathcal{V} = \{\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_J\}$ is a known set of vertices of fixed dimension. From the above expression of $x(p)$, we could write $x(p) = E_p[G(\underline{v})]$.

So we can solve Problem (P1) for $\phi(p) = \psi\{E_p[G(\underline{v})]\}$, $x = E_p[G(\underline{v})] = \sum_{j=1}^{J} p_j G(\underline{v}_j)$.

Problems (P1) and (P2) share many properties. We will discuss them in detail later in this chapter.

Problem (P3)

Maximize $\phi(p)$ over $\mathcal{P}' = \{p = (p_1, p_2, \ldots, p_J) : p_j \geq 0, Cp = \underline{a}\}$

where C is an s × J matrix of rank s, $\underline{a}$ is a $s \times 1$ vector, and the system of equations $Cp = \underline{a}$ is consistent.

It is clear that Problem (P3) is a generalization of Problem (P1). Note that Problem (P3) can generate an example of Problem (P2) and hence of Problem (P1). One example of Problem (P3) arises when we test a linear hypothesis about the parameters in multinomial models. These parameters are the probabilities. Mandal and Torsney (2000) considers an example of Problem (P3) by using the vertices of the feasible

region in determining the maximum likelihood estimates of the cell probabilities under the hypothesis of marginal homogeneity for data in a square $n \times n$ contingency table. They solved this problem by transforming it to an example of Problem (P2). We consider Problem (P3) to determine the maximum likelihood estimate of the cell probabilities using a Lagrangian approach and simultaneous optimization techniques in Chapter 4.

There are two approaches for solving the above optimization problems. We could directly find an optimizing p^* or first find an x^* maximizing $\psi(x)$ over $\mathcal{CH}\{\mathcal{G}(\mathcal{V})\}$ and then find an optimizing p^* such that $x(p^*) = x^*$. We focus on the former approach, which requires optimality conditions explicitly for an optimizing p^*. We determine such optimality conditions in terms of point to point directional derivatives of a criterion function.

2.3 Directional Derivatives

2.3.1 Definition 1.

This definition refers to the directional derivative of Whittle (1973). The directional derivative $F_\phi\{p, q\}$ of a criterion function $\phi(.)$ at p in the direction of q is defined as

$$F_\phi\{p, q\} = \lim_{\epsilon \downarrow 0} \frac{\phi\{(1-\epsilon)p + \epsilon q\} - \phi(p)}{\epsilon}. \tag{2.2}$$

Note that we define the above directional derivative in terms of a criterion function $\phi(.)$, but this could be any function with no constraints on the design weights p. This directional derivative exists even if $\phi(.)$ is not differentiable.

In the context of influence curves, the term $F_\phi\{p, q\}$ has been referred to by Andrews et al. (1972, p.20), as a Von Mises derivative (Von Mises (1947)). For further details we also refer to Hampel (1968), Hampel (1971) and Eplett (1980).

2.3.2 Definition 2.

Another directional derivative of a criterion function $\phi(.)$ is defined as

$$G_\phi\{p, m\} = \lim_{\epsilon \downarrow 0} \frac{\phi\{p + \epsilon m\} - \phi(p)}{\epsilon}. \tag{2.3}$$

$G_\phi\{p, m\}$ is called Gâteaux derivative of $\phi(.)$ at p in the direction of m.

Looking at the above two kinds of directional derivatives, it is clear that

$$F_\phi\{p, q\} = G_\phi\{p, m\}, \tag{2.4}$$

where $m = q - p$, while $G_\phi\{p, m\} = F_\phi\{p, p + m\}$.

It is interesting to note that $G_\phi\{p, e_j\} = \frac{\partial^+ \phi}{\partial p_j}$, the right hand partial derivative of $\phi(.)$ with respect to the j^{th} component of p, e_j being the j^{th} unit vector.

Whittle (1971) used this alternative but equivalent definition of $F_\phi\{p, q\}$. Rockafellar (1970) also used this derivative. Kiefer (1974) also used this concept of Gâteaux derivative in his design theory. However, he did not call it a directional derivative. The above representation of $F_\phi\{p, q\}$ in terms of $G\{p, m\}$ is useful for studying their properties. In particular, $G\{p, m\}$ is useful for deriving the partial derivatives of a criterion function. However, we will see that the directional derivative $F_\phi\{p,\ q\}$ will

serve better for determining the optimality conditions. Mandal (2000) studied the properties of these two directional derivatives including the cases when the criterion function is differentiable and not differentiable. In the following section, we will see that differentiability of the criterion function plays an important simplifying role in our optimization problems.

2.4 Properties of the Directional Derivatives

(PR1) Let $p, q \in S$, where S is a convex set. Then $\{(1-\epsilon)p + \epsilon q\}$ also belongs to S. This is an advantage if one wishes $F_\phi\{p, q\}$ only for $p, q \in S$.

(PR2) Let $\phi(.)$ be a concave function. Then, $F_\phi\{p, q\} \geq \phi(q) - \phi(p)$.

Proof:

$$\begin{aligned} F_\phi\{p, q\} &= \lim_{\epsilon \downarrow 0}[\phi\{(1-\epsilon)p + \epsilon q\} - \phi(p)]/\epsilon \\ &\geq \lim_{\epsilon \downarrow 0}[(1-\epsilon)\phi(p) + \epsilon\phi(q) - \phi(p)]/\epsilon \\ &= \phi(q) - \phi(p). \end{aligned}$$

(PR3) $F_\phi\{p, p\} = 0$. This is clearly a desirable property since no change is effected in $\phi(.)$ if we do not move from p. However, $G_\phi\{p, p\} = F_\phi\{p, 2p\} \neq 0$.

(PR4) The directional derivative $F_\phi\{p, q\}$ measures the rate of change in $\phi(.)$ at p in the direction of q. However, it depends on the units of measurements which also depend on the distance between p and q. Thus $F_\phi\{p, q\}$ depends on the distance between p and q.

(PR5) A converse concept would be the directional derivatives of $\phi(.)$ at p as p is approached from the direction of q. This could be defined as

$$\bar{F}_\phi\{p, q\} = \lim_{\delta \uparrow 0}[\phi\{(1+\delta)p - \delta q\} - \phi(p)]/\delta. \tag{2.5}$$

Based on the above definition, the following relationship between $\bar{F}_\phi\{p, q\}$ and $F_\phi\{p, q\}$ can be established.

$$\begin{aligned}
\bar{F}_\phi\{p, q\} &= \lim_{\delta \uparrow 0}[\phi\{(1+\delta)p - \delta q\} - \phi(p)]/\delta \\
&= -\lim_{\epsilon \downarrow 0}[\phi\{p + \epsilon(q-p)\} - \phi(p)]/\epsilon, \epsilon = -\delta \\
&= -F_\phi\{p, q\}.
\end{aligned}$$

(PR6) Based on the definition of $F_\phi\{p, q\}$, the higher order directional derivatives of $\phi(.)$ at p in the direction of q can be defined as follows

$$F_\phi^{(n)}\{p, q\} = \left.\frac{d^n f(\epsilon)}{d\epsilon^n}\right|_{\epsilon=0^+}, \quad f(\epsilon) = \phi\{(1-\epsilon)p + \epsilon q\}.$$

So far we did not make any assumption about the differentiability of the criterion function $\phi(p)$. As we mentioned earlier, a function does not need to be differentiable at a point p in order that it should have well defined directional derivatives in all possible directions. However, the differentiability of the criterion function $\phi(p)$ plays an important simplifying role in the optimization of $\phi(p)$.

Now we attempt to redefine the concept in terms of $F_\phi(p, q)$. Note that the criterion $\phi(.)$ should change smoothly in all directions. The $\phi(.)$-surface should just touch or

cross in parallel a unique linear hyper-plane, the tangent plane to $\phi(.)$ at p. Thus this plane will provide a linear approximation to $\phi(.)$ at p in any direction. Consider then the form of the directional derivative of the linear function $L(p) = a^T p + b$. Then the directional derivative of L at p in the direction of q can be derived as

$$\begin{aligned} F_L\{p, q\} &= \lim_{\epsilon \downarrow 0}[L\{p + \epsilon(q - p)\} - L(p)]/\epsilon \\ &= \lim_{\epsilon \downarrow 0}[a^T[p + \epsilon(q - p)] - a^T p]/\epsilon \\ &= a^T(q - p) \\ &= L(q) - L(p). \end{aligned}$$

Similarly we can simplify

$$\begin{aligned} G_L\{p, q\} &= a^T q \\ &= L(q) - b. \end{aligned}$$

The vector of partial derivatives of L is $\frac{\partial L}{\partial p} = a$.

If $\phi(.)$ is differentiable, then

$$\begin{aligned} F_\phi(p, q) &= (q - p)^T d \text{ for all } q \\ &= \sum_{j=1}^{J}(q_j - p_j)d_j, \end{aligned}$$

where $d_j = \frac{\partial \phi}{\partial p_j}$, $j = 1, 2, \ldots, J$, and $d = \frac{\partial \phi}{\partial p}$.

The Gâteaux derivative can also be written as

$$G_\phi\{p, q\} = q^T \frac{\partial \phi}{\partial p} = q^T d \text{ for all } q.$$

It is worth mentioning that the condition on $G_\phi\{p, q\}$ is a familiar definition of differentiability. We explore this idea for finding the partial derivatives of our criterion functions.

In particular, when $p \in \mathcal{P}$ of Problem (P1),

$$F_\phi(p, e_j) = \frac{\partial \phi}{\partial p_j} - \sum_{i=1}^{J} p_i \frac{\partial \phi}{\partial p_i}, \tag{2.6}$$

where e_j is the jth unit vector in $\mathcal{R}^J$. We call $F_\phi(p, e_j)$ as F_j, the vertex directional derivative of $\phi(.)$ at p.

In many instances in this thesis, we will see that the criterion function $\phi(p)$ is concave. So we further study the properties based on this. A concave function $\phi(p)$ is differentiable at p if

$$F_\phi\left\{p, \sum_r c_r q_r\right\} = \sum c_r F_\phi\{p, q_r\} + \left(\sum_r c_r - 1\right) F_\phi\{p, 2p\} \tag{2.7}$$

or

$$G_\phi\left\{p, \sum_r c_r q_r\right\} = \sum_r c_r G_\phi\{p, q_r\}. \tag{2.8}$$

The above two conditions are equivalent as we will see below.

Now we discuss some properties which follow from this definition. All of them assume the differentiability at the point $p \in S$, where S is a convex set.

(PR7) $G_\phi\{p, q\} = q^T d$ since $q = (q_1, q_2, \ldots, q_J)^T = \sum_{i=1}^{J} q_i e_i$ and $d_i = \frac{\partial \phi}{\partial p_i} = G_\phi\{p, e_i\}$.

Then

$$G_\phi\left\{p, \sum_r c_r q_r\right\} = \left[\sum_r c_r q_r\right]^T d = \sum_r c_r q_r^T d = \sum_r c_r G_\phi\{p, q_r\}.$$

From this we can say that condition (2.8) is equivalent to requiring that $G_\phi\{p, q\} = q^T d$.

(PR8) $F_\phi\{p, q\} = G_\phi\{p, q - p\} = G_\phi\{p, q\} - G_\phi\{p, p\} = (q - p)^T d$.

Theorem 2.1

$G_\phi\left\{p, \sum_r c_r q_r\right\} = \sum_r c_r G_\phi\{p, q_r\}$ implies $F_\phi\left\{p, \sum_r c_r q_r\right\} = \sum c_r F_\phi\{p, q_r\} + \left(\sum_r c_r - 1\right) F_\phi\{p, 2p\}$.

Proof:

$$\begin{aligned}
F_\phi\left\{p, \sum_r c_r q_r\right\} &= G_\phi\left\{p, \sum_r c_r q_r\right\} - G_\phi\{p, p\} \\
&= \sum_r c_r G_\phi\{p, q_r\} - G_\phi\{p, p\} \\
&= \sum_r c_r \left[G_\phi\{p, q_r\} - G_\phi\{p, p\}\right] + \left(\sum_r c_r - 1\right) G_\phi\{p, p\} \\
&= \sum_r c_r F_\phi\{p, q_r\} + \left(\sum_r c_r - 1\right) F_\phi\{p, 2p\} \text{ (by property PR3).}
\end{aligned}$$

Theorem 2.2

$F_\phi\left\{p, \sum_r c_r q_r\right\} = \sum c_r F_\phi\{p, q_r\} + \left(\sum_r c_r - 1\right) F_\phi\{p, 2p\}$ implies $G_\phi\left\{p, \sum_r c_r q_r\right\} = \sum_r c_r G_\phi\{p, q_r\}$.

Proof:

$$
\begin{aligned}
G_\phi\left\{p, \sum_r c_r q_r\right\} &= F_\phi\left\{p, p + \sum_r c_r q_r\right\} \\
&= F_\phi\{p,p\} + \sum_r c_r F_\phi\{p, q_r\} + \left[1 + \sum_r c_r - 1\right] F_\phi\{p, 2p\} \quad (by\ (2.7)) \\
&= \sum_r c_r F_\phi\{p, q_r\} + \left[\sum_r c_r\right] F_\phi\{p, 2p\} \quad (by\ PR3) \\
&= \sum_r c_r \left[F_\phi\{p, q_r\} + F_\phi\{p, 2p\}\right] \\
&= \sum_r c_r \left[F_\phi\{p,p\} + F_\phi\{p, q_r\} + F_\phi\{p, 2p\}\right] \\
&= \sum_r c_r \{F_\phi\{p, p + q_r\} = \sum_r c_r G_\phi\{p, q_r\} \ \text{(by (2.7))}.
\end{aligned}
$$

(PR9) $F_\phi\{p, 2p - q\} = 2F_\phi\{p,p\} - F_\phi\{p,q\} + [(2-1)-1]F_\phi\{p, 2p\}$

i.e., $F_\phi\{p, 2p - q\} = -F_\phi\{p,q\}$ (by PR3)

or $\bar{F}_\phi\{p, 2p - q\} = F_\phi\{p,q\}$.

As we pass through p in the direction of q, the rate of change in $\phi(.)$ should be the same on the approach to and the departure from p.

In the case of a function $g(x)$ of a variable x, a consequence is that there is no need to distinguish between right and left hand derivatives. That is,

$$
\lim_{\epsilon \uparrow 0}[\{g(x+\epsilon) - g(x)\}/\epsilon] = \lim_{\epsilon \downarrow 0}[\{g(x+\epsilon) - g(x)\}/\epsilon].
$$

(PR10) If $\sum_r c_r = 1$ then it is obvious that $F_\phi\{p, \sum_r c_r q_r\} = \sum_r c_r F_\phi\{p, q_r\}$. This property is a very useful result, when the criterion function $\phi(.)$ or $\psi(.)$ is defined on a convex set S.

For example, consider the case of Problem (P1) for $S = \mathcal{P}$ and the case of Problem (P2) for $S = \mathcal{CH}\{\mathcal{G}(\mathcal{V})\}$. For instance, if $y \in \mathcal{CH}\{\mathcal{G}(\mathcal{V})\}$ then $y = x(q) = \sum_j q_j G(\underline{v}_j)$, $\sum_j q_j = 1$, $q_j \geq 0$.

Then we have

$$F_\psi\{x, y\} = \sum_j q_j F_\psi \left\{x, G(\underline{v}_j)\right\}. \tag{2.9}$$

In Problem (P1), the criterion $\phi(.)$ is a function of $p \in \mathcal{P}$.

Then for $q \in S$,

$$F_\phi\{p, q\} = \sum_j q_j F_\phi \left\{p, e_j\right\}. \tag{2.10}$$

If we have $\phi(p) = \psi\{x(p)\}$, $x(p) \in \mathcal{CH}\{\mathcal{G}(\mathcal{V})\}$ then

$$\begin{aligned} \frac{\partial \phi}{\partial p_j} &= G_\phi\{p, e_j\} \\ &= G_\psi\{x(p), G(\underline{v}_j)\}. \end{aligned}$$

Considering $S = \mathcal{M}$ and ψ is defined on $\mathcal{M}$, where $\mathcal{M} = \{M : M = M(p), p \in \mathcal{P}\}$, we conclude that

$$F_\psi\{M(p), M(q)\} = \sum q_j F_\psi\{M(p), \underline{v}_j \underline{v}_j^T\}. \tag{2.11}$$

(PR11) For $S = \mathcal{CH}\{\mathcal{G}(\mathcal{V})\}$, $\sum_j p_j F_\psi\{x(p), G(v_j)\} = 0$.

This is true because $\sum_j p_j F_\psi\{x(p), G(\underline{v}_j)\} = F_\psi\{x(p), x(p)\}$.

In particular, note that, $\sum_j p_j F_\psi\{M(p), \underline{v}_j \underline{v}_j^T\} = 0$ when $G(\underline{v}_j) = \underline{v}_j \underline{v}_j^T$ in Problem (P2).

Theorem 2.3

Consider Problem (P1). If $S = \mathcal{P}$ then

$$\begin{aligned} \underset{q \in S}{max} F_\phi\{p, q\} &= \underset{1 \le j \le J}{max} F_\phi\{p, e_j\} \\ \underset{q \in S}{min} F_\phi\{p, q\} &= \underset{1 \le j \le J}{min} F_\phi\{p, e_j\}. \end{aligned}$$

Proof:

Since $q \in \mathcal{P}$ we have

$$F_\phi\{p, q\} = \sum_j q_j F_\phi\{p, e_j\}$$

and

$$\left[\sum_j q_j\right] \underset{1 \le t \le J}{min} F_\phi\{p, e_t\} \le F_\phi\{p, q\} \le \left[\sum_j q_j\right] \underset{1 \le s \le J}{max} F_\phi\{p, e_s\}.$$

This is true since $\sum_j q_j = 1$.

(PR12) Consider Problem (P2). For $S = \mathcal{CH}\{\mathcal{G}(\mathcal{V})\}$, $\underset{y \in S}{max} F_\psi\{x, y\} \ge 0$, $\underset{y \in S}{min} F_\psi\{x, y\} \le 0$.

(PR13) Consider Problem (P1). For $S = \mathcal{P}$, $\underset{q \in S}{max} F_\phi\{p, q\} \geq 0$, $\underset{q \in S}{min} F_\phi\{p, q\} \leq 0$.

These results follow from (PR11) and Theorem 2.3.

2.5 Vertex Direction Optimality Theorem

Theorem 2.4

In the context of Problem (P2), consider $S = \mathcal{CH}\{\mathcal{G}(\mathcal{V})\}$ and assume that $\psi(x)$ is concave on S and $x(p^*)$ is a differentiable point of $\psi(.)$.

Then $x(p^*)$ maximizes $\psi(.)$ on S iff

$$\begin{array}{ll} F_\psi\{x(p^*), G(\underline{v}_j)\} = 0 & \text{when} \quad p_j^* > 0 \\ F_\psi\{x(p^*), G(\underline{v}_j)\} \leq 0 & \text{when} \quad p_j^* = 0. \end{array} \tag{2.12}$$

This is the key theorem in optimal design theory, and is known as General Equivalence Theorem.

Whittle (1973) derived the theorem in general optimal design problem. Kiefer (1974) considered this using Gâteaux derivative (2.3). Some authors too have derived the theorem by using Lagrangian theory. See, for example, Sibson (1974) and Silvey and Titterington (1974). Wu (1976) derived it by considering the Kuhn-Tucker theroem in a more general setting.

The general equivalence theorem plays an important role for the construction of optimal designs, specifying a finite set of optimality conditions (for example, see Mandal and Torsney (2006)). Differentiability is an essential requirement.

Corollary (i)

Consider Problem (P2). If $S = \mathcal{M}$, and $M(p^*)$ is a differentiable point of $\psi(.)$, then $M(p^*)$ maximizes $\psi(.)$ on $\mathcal{M}$ iff

$$\begin{aligned} F_\psi\{M(p^*), \underline{v}_j \underline{v}_j^T\} &= 0 \quad \text{when} \quad p_j^* > 0 \\ F_\psi\{M(p^*), \underline{v}_j \underline{v}_j^T\} &\leq 0 \quad \text{when} \quad p_j^* = 0. \end{aligned} \tag{2.13}$$

Corollary (ii)

Consider Problem (P1). If $S = \mathcal{P}$, and p^* is a differentiable point of $\phi(.)$ on $\mathcal{P}$, then p^* maximizes $\phi(.)$ on $\mathcal{P}$ iff

$$\begin{aligned} F_j^* = \frac{\partial \phi}{\partial p_j^*} - \sum_{i=1}^{J} p_i^* \frac{\partial \phi}{\partial p_i^*} &= 0 \quad \text{when} \quad p_j^* > 0 \\ F_j^* = \frac{\partial \phi}{\partial p_j^*} - \sum_{i=1}^{J} p_i^* \frac{\partial \phi}{\partial p_i^*} &\leq 0 \quad \text{when} \quad p_j^* = 0. \end{aligned} \tag{2.14}$$

This is a simplified version of the general equivalence theorem in the context of Problem (P1). We refer this as the first order condition of optimality. We check whether or not these conditions are satisfied by a postulated solution obtained by numerical techniques.

Once we determine the optimality conditions (based on the general equivalence theorem) for an optimization problem, we often need to find an appropriate algorithm for finding the optimal design. In the following section we consider a class of multiplicative algorithms for finding the optimal design.

2.6 A Class of Algorithms

In order to construct an optimizing distribution, it is usually not possible to evaluate an optimal solution p^* explicitly. Numerical techniques such as some algorithms must be needed to find an optimal solution. The following algorithms have been developed particularly for the design problem which requires the calculation of an optimizing probability distribution.

First note that, Problems (P1) and (P2) have a unique set of constraints, namely

(a) $p_1, p_2, \ldots, p_J$ must be nonnegative, i.e. $p_j \geq 0$, and

(b) p_j's sum to unity, i.e. $\sum_{j=1}^{J} p_j = 1$.

Let δ be a free positive parameter, d be the partial derivatives of $\phi(p)$ with respect to p, and $f(d, \delta)$ be a function that satisfies the following conditions:

(i) $f(d, \delta) > 0$.

(ii) $\dfrac{\partial f(d, \delta)}{\partial d} > 0$, i.e. $f(.)$ is a strictly increasing function.

(iii) $f(d, 0)$ is constant.

Then a class of algorithms which neatly satisfies the above basic constraints of the optimal weights would take the form

$$p_j^{(r+1)} \propto p_j^{(r)} f(d_j^{(r)}, \delta) \tag{2.15}$$

or, the full form being

$$p_j^{(r+1)} = \frac{p_j^{(r)} f(d_j^{(r)}, \delta)}{\sum_{i=1}^{J} p_i^{(r)} f(d_i^{(r)}, \delta)}, \tag{2.16}$$

where

$$d_j^{(r)} = \left.\frac{\partial \phi}{\partial p_j}\right|_{p=p^{(r)}}$$

are the partial derivatives of the criterion function $\phi(p)$ at rth iterate $p = p^{(r)}$. For simplicity we sometime denote the function $f(d, \delta)$ by $f(d)$ in this thesis.

Torsney (1977) first proposed this kind of algorithm by considering $f(d) = d^{\delta}$, where $\delta > 0$. The partial derivatives need to be positive for this choice of the function. Silvey et al. (1978) is an empirical study of the choice of δ when $f(d) = d^{\delta}$. Torsney (1988) considered the choice of $f(d) = e^{\delta d}$ in a variety of problems, including some estimation and image processing problems, for which one criterion $\phi(.)$ could have negative derivatives. Torsney and Alahmadi (1992) continued these investigations by exploring other choices of $f(.)$. Mandal and Torsney (2000) considered some systematic choices of $f(.)$.

Mandal and Torsney (2006) used the algorithm in a clustering approach for more than one optimizing distributions. Mandal et al. (2005) used the algorithm for constructing designs subject to additional constraints. Mandal et al. (2005) and Mandal and Torsney (2006) considered objective choices of $f(.)$ for constructing optimal designs.

Titterington (1976) proved monotonicity of $f(d) = d$ with $\delta = 1$ for D-optimality. Torsney (1983) explored monotonicity of particular values of δ for some criterion function $\phi(p)$. Torsney (1983) also established a sufficient condition for monotonicity for $f(d) = d^{\delta}$, where $\delta = 1/(t + 1)$. He also considered the criterion $\phi(p)$ as a homogeneous function of degree $-t$, $t > 0$ with positive derivatives and proved this

condition to hold in the case of linear design criteria such as c-optimality and A-optimality when $t = 1$. The value of $\delta = 1$ can be shown to yield an EM algorithm which is known to be monotonic and convergent. See Dempster et al. (1977). However, the EM algorithm is known to have slow convergence rate.

There are several other algorithms exist in the literature. Fedorov (1972) and Wynn (1972) considered vertex direction algorithms which perturb one p_j and change the others proportionately. These are useful when many of the p_j are zero at the optimum as happens in design problems (Mandal and Torsney (2006)). When all optimal weights are positive or when it has already been established which weights are positive, constrained steepest ascent or Newton type iterations may be appropriate. For further details, see Wu (1978), Atwood (1976) and Atwood (1980). Yu (2010) established monotonic convergence of a class of multiplicative algorithm for computing optimal designs. Molchanov and Zuyev (2000) considered steepest descent algorithms based on some gradient functions.

2.7 Properties of the Algorithms

As we mentioned before, Problems (P1) and (P2) have a distinctive set of constraints on the design weights p_j's. Iteration (2.16) neatly satisfy these constraints and possess the following nice properties. It is important to note that the function $f(d)$ is positive and strictly increasing.

(i) $p^{(r)}$ is always feasible.

(ii) $F_\phi\left\{p^{(r)}, p^{(r+1)}\right\} \geq 0$ with equality when the d_j's corresponding to nonzero p_j's have

a common value, d, i.e. $d_j = \sum p_i d_i = d$ so that

$$\begin{aligned} p_j^{(r+1)} &= \frac{p_j^{(r)} f(d_j)}{\sum_{i=1}^{J} p_i^{(r)} f(d_i)} \\ &= \frac{p_j^{(r)} f(d)}{f(d) \sum_{i=1}^{J} p_i^{(r)}} \\ &= p_j^{(r)}. \end{aligned}$$

(iii) If $\delta = 0$ there is no change in $p^{(r)}$, given $f(d, 0)$ is constant.

(iv) An iterate $p^{(r)}$ is a fixed point of the iteration if the partial derivatives $\partial\phi/\partial p_j^{(r)}$ corresponding to nonzero $p_j^{(r)}$ are all equal.

(v) Let $supp(p) = \left\{\underline{v}_j \in \mathcal{V} : p_j > 0\right\}$ denote the support of the design measure p in the induced design space $\mathcal{V}$. Under the above iteration, $supp(p^{(r+1)}) \subseteq supp(p^{(r)})$, but some weights can converge to zero at an optimizing p^*.

The importance of this algorithm is that it neatly preserves the basic constraints on the design weights; i.e., iterations are always feasible. This algorithm can be used even if the partial derivatives of a criterion function are both positive and negative (Torsney and Mandal (2006)). In property (ii), the directional derivative $F_\phi\left\{p^{(r)}, p^{(r+1)}\right\}$ can be expressed as $Cov(D, f(D))/E(f(D))$, where D is a random variable taking the value d_j with probability p_j. As $f(D)$ is an increasing function, the covariance between D and $f(D)$ must be nonnegative. Property (iii) implies no changes in the design weights and hence in the criterion value if $\delta = 0$. Property (iv) also holds if the vertex directional

derivatives $F_j^{(r)}$ (corresponding to the partial derivatives) are zero. Thus in view of the conditions for optimality, a solution is a fixed point of the iteration at the optimum. Property (v) implies that the set of support points of the design becomes more compact as we move on with the algorithm.

Chapter 3

Construction of D-optimal Designs

3.1 Introduction

As we discussed in Chapter 1, we can construct an optimizing probability distribution in two ways, namely, exact design and approximate design. In an exact design, we need to find the 'exact' integer values of the number of trials or observations (n_j's) at different design points of the original design space $\mathcal{X}$ or at the vertices of the induced design space $\mathcal{V}$. In an approximate design, we find the 'proportion' of observations (p_j's) corresponding to the design points so that $p_j \geq 0$ and $\sum p_j = 1$. The weights $p = (p_1, p_2, \ldots, p_J)$ represent the probability distribution or measure on the induced design space $\mathcal{V}$. In practice, we follow the latter approach. This is simple and more flexible problem to solve and is not much visibly different from the exact design problem.

In this chapter we construct the approximate optimal designs and study the performance of the algorithms discussed in the previous chapter. In particular, we construct the D-optimal designs and study some important properties. The most popular design

criterion in applications is D-optimality. In D-optimality criterion, the generalized variance of the parameter estimates, or its logarithm is minimized. Because of the reciprocity property of the covariance matrix and the information matrix, minimizing the determinant of the covariance matrix is equivalent to maximizing the determinant of the information matrix. Thus, in terms of a maximization problem, the criterion function is given by

$$\phi_D(p) = \psi_D\{M(p)\} = logdet\{M(p)\} = -logdet\{M^{-1}(p)\}. \tag{3.1}$$

There are various motivations for D-optimality, especially for many practical problems in statistical inference. This goes beyond the idea of point and interval estimation. One interesting statistical interpretation of D-optimal design is that if we assume normality of the errors in the linear model (1.9), then the general form of the joint confidence region for the vector of the unknown parameters $\underline{\theta} \in \Theta$ is described by an ellipsoid:

$$\{\underline{\theta} : (\underline{\theta} - \underline{\hat{\theta}})^T M(p)(\underline{\theta} - \underline{\hat{\theta}}) \leq c\}, \tag{3.2}$$

for some critical value c, where $\underline{\hat{\theta}}$ is the least squares estimate or the maximum likelihood estimate of $\underline{\theta}$. The D-optimal criterion chooses the design in such a way that the volume of the above ellipsoid is as small as possible. The reason is that the volume of the ellipsoid is proportional to $[det\{M(p)\}]^{-\frac{1}{2}}$. The value of the D-optimality criterion is finite if and only if the information matrix $M(p)$ is non-singular, i.e., when all the unknown parameters are estimable. The above ideas can be expressed more formally in terms of the eigenvalues of the information matrix $M(p)$. Let the eigenvalues of the information matrix $M(p)$ be $\lambda_1, \lambda_2, \ldots, \lambda_k$. The eigenvalues of $M^{-1}(p)$ are then $1/\lambda_1, 1/\lambda_2 \ldots, 1/\lambda_k$ and are proportional to the squares of the lengths

of the axes of the confidence ellipsoid. In terms of these eigenvalues, the D-optimality criteria is as follows: Minimize the product of the eigenvalues of $M^{-1}(p)$, that is, find

$$min \prod_{i=1}^{k} \frac{1}{\lambda_i}.$$

Other motivations for D-optimality lie in hypothesis testing problems under normal linear models.

This is the most extensively studied criterion in optimal design literature. The references include Kiefer (1959), Fedorov (1972), Silvey (1980), Berger and Wong (2009), Atkinson et al. (2007), Shah and Sinha (1989), Pukelsheim (1993), Titterington (1975), Dette et al. (2008), Mandal et al. (2005), Torsney (1983), and Torsney (1988). Mandal (2000) studies the properties of D-optimality in a variety of applications. Mandal and Torsney (2006) constructs the D-optimal designs using a clustering approach.

3.2 Properties of D-optimality

(i) The D-optimal design maximizes the determinant of the information matrix or equivalently, it minimizes the determinant of the inverse of the information matrix. Sometimes it is more convenient to consider the problem of maximizing $log|M(p)|$ or minimizing -$log|M(p)|$.

(ii) In some cases, it is desired to compare a certain design to the D-optimal design. Consider that we have a design p for a given model with k parameters and the D-optimal design is p^*. The D-efficiency of the design p is defined as

$$D_{eff} = \left\{ \frac{|M(p)|}{|M(p^*)|} \right\}^{1/k}.$$

Taking the ratio of the determinants to the k^{th} root results in an efficiency measure which is proportional to design size, irrespective of the dimension of the model. So two replicates of a design p for which $D_{eff} = 0.5$ would be as efficient as one replicate of the D-optimal design.

(iii) If p_1^* and p_2^* are two D-optimal designs, the design (a convex combination of p_1^* and p_2^*)

$$p^* = \alpha p_1^* + (1 - \alpha) p_2^* \ \ (0 \leq \alpha \leq 1)$$

is also D-optimal.

(iv) Suppose that the number of support points of the design is n. Then there exists a D-optimal design p^* with $k \leq n \leq k(k+1)/2$, where k is the number of parameters in the model.

(v) ψ_D is an increasing function over the set of positive definite symmetric matrices. That is for $M_1, M_2 \in \mathbb{M}$,

$$\psi_D(M_1 + M_2) \geq \phi_D(M_1),$$

where $\mathbb{M}$ is the set of all positive definite symmetric matrices.

(vi) ψ_D is a concave function of the positive definite symmetric matrices.

(vii) ϕ_D is differentiable when the criterion function is finite. The first partial derivatives are given by

$$\frac{\partial \phi_D}{\partial p_j} = \underline{v}_j^T M^{-1}(p) \underline{v}_j. \tag{3.3}$$

(viii) The D-optimality criterion is model dependent. However ϕ_D is invariant under a non-singular linear transformation of $\mathcal{V}$, where $\mathcal{V}$ is the induced design space.

This property can be easily seen to follow from the expression (1.13) for $M(p)$. Suppose the induced design space $\mathcal{V} = [\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_J]$ is transformed to $\mathcal{W} = [\underline{w}_1, \underline{w}_2, \ldots, \underline{w}_J]$ under the linear transformation $\underline{w}_j = A\underline{v}_j$, where A is a $k \times k$ matrix. Then a design assigning weight p_j to $\underline{w}_j$ has information matrix:

$$\begin{aligned} M_w(p) &= \mathcal{W}P\mathcal{W}^T \\ &= AVPV^TA^T. \end{aligned}$$

Then

$$\begin{aligned} \phi_D\{M_w(p)\} &= logdet\{M_w(p)\} \\ &= logdet\{AVPV^TA^T\} \\ &= log[det\{VPV^T\} \times det\{A\}^2] \\ &= logdet\{M(p)\} + logdet\{A\}^2 \\ &= \phi_D\{M(p)\} + constant. \end{aligned}$$

A D-optimal design is linked with the standardized variance of the predicted response. We consider this in the following theorem.

Theorem 3.1.

Suppose that, for a given model, we have a design variable x. The weighted sum of the standardized variances of the predicted response $d(x, p)$, taken over all points of the design p, is equal to the number of parameters. That is,

$$\sum_{j=1}^{J} p_j d(x_j, p) = k. \tag{3.4}$$

Proof.

From the definition of the standardized variances of the predicted response we can write $d(x_j, p)$ as

$$d(x_j, p) = \underline{f}^T(x_j)M^{-1}(p)\underline{f}(x_j). \tag{3.5}$$

So we can write

$$\begin{aligned}
\sum_{j=1}^{J} p_j d(x_j, p) &= \sum_{j=1}^{J} p_j \underline{f}^T(x_j)M^{-1}(p)\underline{f}(x_j) \\
&= tr\left\{M^{-1}(p)\sum_{j=1}^{J} p_j[\underline{f}(x_j)\underline{f}^T(x_j)]\right\} \\
&= tr\left\{M^{-1}(p)M(p)\right\} \\
&= tr\{I_k\} \\
&= k.
\end{aligned}$$

Hence the theorem.

Theorem 3.2

Consider Problem (P2) of Section 2.2. The optimal design p^* solves $\underset{p\in\mathcal{P}}{min}\ \underset{y\in\mathcal{CH}\{\mathcal{G}(\mathcal{V})\}}{max}[F_\psi\{x(p), y\}]$.

Proof: From Theorem 2.3, we can write $\underset{y\in\mathcal{CH}\{\mathcal{G}(\mathcal{V})\}}{max}[F_\psi\{x, y\}] = \underset{1\le j\le J}{max}[F_\psi\{x, G(\underline{v}_j)\}]$.

It is obvious that $\underset{1\le j\le J}{max}[F_\psi\{x(p^*), G(\underline{v}_j)\}] = 0$.

Then we have $\underset{y\in\mathcal{CH}\{\mathcal{G}(\mathcal{V})\}}{max}[F_\psi\{x(p^*), y\}] = 0$.

From (PR12) of Section 2.4, $\underset{y\in\mathcal{CH}\{\mathcal{G}(\mathcal{V})\}}{max}[F_\psi\{x(p), y\}] \ge 0$ for all p.

Thus p^* attains what is a lower bound for other p. Hence the theorem.

In the D-optimal version of the criterion $\psi(.)$, the above theorem establishes the equivalence of D-optimality and G-optimality. This follows from the fact that

$$F_{\psi}\left\{M(p), \underline{v}_j \underline{v}_j^T\right\} = \underline{v}_j^T M^{-1}(p) \underline{v}_j - k, \, [\text{rank}(M(p)) = k].$$

Hence the above theorem implies that p^* solves $\underset{p \in \mathcal{P}}{min} \underset{1 \leq j \leq J}{max} \{\underline{v}_j^T M^{-1}(p) \underline{v}_j\}$, which is the G-optimal criterion defined in Chapter 1. Kiefer and Wolfowitz (1960) used this result and proved the equivalence between D-optimality and G-optimality.

Although the above results provide the methods for the construction of optimum designs, but it says nothing about the number of support points (n) of the design. A bound on this number can be obtained from the nature of the information matrix which is a $k \times k$ symmetric matrix.

The D-optimality criterion also has some drawbacks.

(i) It is not easy to calculate the value of D-optimality criterion using a pocket calculator when the number of parameters is large.

(ii) The minimization of the determinant of a variance-covariance matrix of the parameter estimates may lead to elongation in the direction of one axis of the confidence ellipsoid. This occurs when the length of the confidence interval of only one of the parameters is short and all the others are long. This leads to the situation that only one of the parameters is estimated efficiently while the others are not.

(iii) The D-optimal design may be very inefficient for estimating certain linear combinations of the parameters in the model because of the correlations among the parameter estimates.

In the D-optimality criterion, we assumed that the information matrix $M(p)$ is non-singular. If this matrix is singular, then one may deal with a generalized inverse. A generalized inverse of the matrix M is defined as any matrix M^{-} satisfying the condition $MM^{-}M = M$. Note that this generalized inverse exists for each matrix M, however, it is not unique. In this case, one may consider $M^{-} = M^{+}$, where M^{+} is the Moore-Penrose inverse. This Moore-Penrose inverse not only satisfies $MM^{+}M = M$, but also $M^{+}MM^{+} = M^{+}$. This generalized inverse is applicable when we consider the D_A-optimality criterion, in which case, one can consider the elements of $A\underline{\theta}$ which are estimable in the linear model under consideration. For further details, we refer to Silvey (1978).

In the following section we construct D-optimal designs using the multiplicative algorithm discussed in the previous chapter. An analytic solution of the problem of constructing D-optimal designs is possible only in simple cases.

3.3 Construction of D-optimal Designs: Analytic Approach

We consider some polynomial regression models on constructing D-optimal designs for which explicit solutions can be obtained. Here we consider a standardized continuous design space. We can construct the D-optimal design using the Legendre polynomial

of Fedorov (1972). The discrete D-optimal design is unique, and has a minimal support of k points which are the k roots of the polynomials

$$(1 - x^2)P'_{k-1}(x),$$

where $P_k(x)$ is the k^{th} Legendre polynomial

$$P_k(x) = \sum_{n=0}^{N} \left[\frac{(-1)^n (2k - 2n)! x^{k-2n}}{2^k n!(k - n)!(k - 2n)!} \right], \tag{3.6}$$

where

$$N = \begin{cases} k/2 & \text{if k is even} \\ (k - 1)/2 & \text{if k is odd.} \end{cases}$$

Note here that, in a minimal support design, since $Supp(p^*)$ contains k points, the D-optimal design on it assigns weight $(1/k)$ to each of these.

We know the polynomial regression model in one variable of order $k - 1$ is given by

$$E(y|\underline{v}_x) = \underline{v}_x^T \underline{\theta} \tag{3.7}$$

where $\underline{v}_x = (1, x, x^2, \ldots, x^{k-1})^T, x \in [-1, 1]$ and $\underline{\theta} = (\theta_0, \theta_1, \ldots, \theta_{k-1})^T$.

$\underline{v}_x \in \mathcal{V} = \{\underline{v}_x : \underline{v}_x = (1, x, x^2, \ldots, x^{k-1})^T, -1 \leq x \leq 1\}$, the induced design space.

For simple linear regression, we take $k = 2$ in (3.7). So the model is

$$E(y|x) = \theta_0 + \theta_1 x.$$

So using (3.6) we obtain

$$(1 - x^2)P'_1(x) = (1 - x^2).$$

So the support points of the D-optimal design p^* are given by

$$x = \pm 1.$$

So we obtain the D-optimal design for the simple linear regression model as

$$p^* = \left\{ \begin{array}{cc} -1 & 1 \\ 0.5 & 0.5 \end{array} \right\}.$$

For the quadratic regression model, we take $k = 3$ in (3.7). The model is

$$E(y|x) = \theta_0 + \theta_1 x + \theta_2 x^2.$$

Therefore, using (3.6) we obtain

$$(1 - x^2)P_2^{'}(x) = 3x(1 - x^2).$$

Hence, the support points of p^* are given by

$$x = \pm 1, 0.$$

Thus, we obtain the D-optimal design for the quadratic regression model as

$$p^* = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ 0.33333 & 0.33333 & 0.33333 \end{array} \right\}.$$

For the cubic regression model, we take $k = 4$ in (3.7). So the model is

$$E(y|x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3.$$

So (3.6) simplifies to

$$(1 - x^2)P_3^{'}(x) = \frac{(15x^2 - 3)(1 - x^2)}{2}.$$

Hence, the support points of p^* are given by

$$x = \pm 1, \pm \frac{1}{\sqrt{5}} \approx \pm 0.45.$$

Thus, we obtain the D-optimal design for the cubic regression model as

$$p^* = \left\{ \begin{array}{cccc} -1 & -0.45 & 0.45 & 1 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{array} \right\}.$$

For the quartic model, we take $k = 5$ in (3.7). So the model is given by

$$E(y|x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4.$$

Simplifying (3.6) we obtain

$$(1 - x^2)P_4'(x) = \frac{(35x^3 - 15x)(1 - x^2)}{2}.$$

Hence, the support points of p^* are given by

$$x = \pm 1, 0, \pm \sqrt{3/7} \approx \pm 0.655.$$

Thus, we obtain the D-optimal design for the quartic regression model as

$$p^* = \left\{ \begin{array}{ccccc} -1 & -0.655 & 0 & 0.655 & 1 \\ 0.20 & 0.20 & 0.20 & 0.20 & 0.20 \end{array} \right\}.$$

It will be of interest to compare numerically constructed designs with the above analytic solutions. Note that we can do this comparison after we consider some algorithms for constructing optimal designs in the following section.

3.4 Construction of D-optimal Designs: Algorithmic Approach

In the previous section, we constructed D-optimal design for polynomial regression problems by some analytic approach. In this section, we will construct the D-optimal design for those problems using the class of multiplicative algorithms (2.16).

We now develop some important useful strategies for constructing the D-optimal designs and for better convergence of the algorithms (2.16). Convergence of the algorithm (2.16) can be slow if we do not objectively choose the function $f(.)$ and its arguments. Some choices of $f(.)$ may not be good because D-optimal derivatives are positive and centred at k, the number of parameters. For example, the choice of $f(d) = \Phi(\delta d)$ (the normal c.d.f.) may not be good because the d_j' s are positive and centred at k. Also, $\Phi(\delta d)$ may change slowly at k whereas it changes more quickly at zero. However, $f(d) = d^{\delta}$ proved to be a natural choice for particular values of δ. Note that, in particular, $\delta = 1$ for D-optimality and $\delta = 1/2$ for c-optimality yield monotonic iterations.

We now attempt to improve convergence by considering some choices of $f(.)$ for which we replace the partial derivatives d_j by the corresponding directional derivatives F_j. Note that a criterion has both positive and negative vertex directional derivatives. So the function $f(.)$ needs to be defined for positive and negative F_j's. From Chapter 2, we have that $F_j = d_j - \sum p_j d_j$. Thus, $\sum p_j F_j = 0$. Also, recall that first order conditions for a local maximum p^* are

$$F_j \begin{cases} = 0 & \text{for} \quad p_j^* > 0 \\ \leq 0 & \text{for} \quad p_j^* = 0. \end{cases}$$

The above suggests that we should choose a function which is centred at zero and changes reasonably quickly about $F = 0$. We should also treat positive and negative directional derivatives symmetrically, at least when all the weights are positive.

Two choices of the function $f(.)$ with the potential to satisfy the above requirements are $f(F) = \Phi(\delta F)$ (the normal c.d.f.), and $f(F) = exp(\delta F)/1 + exp(\delta F)$ (the logistic c.d.f.). These functions change quickly at zero.

Note that the choice of $f(.)$ depends on a free parameter δ which should be positive. Clearly the value of δ is crucial for the convergence of the algorithm. Keeping this in mind, we explore various choices of the function $f(.)$ and the parameter δ for improving the convergence rate of the algorithm.

In order to satisfy the first order optimality conditions (2.14), we run the algorithm (2.16) and record, for $n = 1, 2, 3, 4, 5$, the number of iterations needed to achieve

$$\max_{1 \le j \le J} \{F_j\} \le 10^{-n}, \tag{3.8}$$

where F_j are the directional derivatives. We also make sure that F_j's are less than or equal to zero for the zero weights.

We now construct the optimal designs for the following regression models.

3.4.1 Simple Linear Regression

Our simple linear regression model is

$$E(y|x) = \theta_0 + \theta_1 x\,.$$

We consider the design space $-1 \leq x \leq 1$. As we discussed in Chapter 1, we discretize the design space approximated by a grid of 21 points equally spaced at intervals of 0.1 between -1 and 1. So basically we are considering a design of the form

$$p = \left\{ \begin{array}{cc} x_1 & x_2 \ldots x_{21} \\ p_1 & p_2 \ldots p_{21} \end{array} \right\}.$$

To evaluate how efficiently algorithm (2.16) performs in calculating D-optimal design we first take the argument x as partial derivative (d) of the D-optimal criterion function and choose 4 choices of $f(.)$, namely, $f(d) = d^{\delta}$, $f(d) = exp(\delta d)$, $f(d) = exp(\delta d)/[1 + exp(\delta d)]$, the logistic c.d.f. and $f(d) = \Phi(\delta d)$, the normal c.d.f. with equal initial weights, i.e $p_j^{(0)} = 1/J$, ($J = 21$). We consider various choices of δ in each case and run the algorithm (2.16). We record the number of iterations needed to achieve the condition (3.8) for $n = 1, 2, 3, 4, 5$. The results are given in Table 3.1. Under each choice of $f(.)$, the best choices of δ are given in bold font. The best choices of δ corresponds to least number of iterations.

From Table 3.1, we see that the two choices of $f(d)$ such as $f(d) = d^{\delta}$ and $f(d) = exp(\delta d)$ give us better convergence than the others. As for instance, with $f(d) = exp(\delta d)$ and $\delta = 1.0$, the number of iterations needed to achieve the condition (3.8) for $n = 5$ is 50. Now consider the suitable choices of δ. For example, $f(d) = exp(\delta d)/1 + exp(\delta d)$ with $\delta = 2.0$, the number of iterations needed to achieve the condition for $n = 5$ is 1238, whereas for $\delta = 0.7$, the number of iterations needed is 351. This also happens with other choices of $f(d)$.

As we mentioned earlier, we now attempt to explore the performance of the algorithm by objectively choosing the function $f(.)$. We replace the partial derivatives

Table 3.1: Simple linear regression with $x = d$
Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$.

$f(d) = d^{\delta}$					
δ	n=1	n=2	n=3	n=4	n=5
0.8	13	36	65	93	122
0.9	12	33	58	83	109
1.5	8	20	35	50	66
1.7	7	18	31	45	58
2.0	**6**	**15**	**27**	**38**	**50**
$f(d) = exp(\delta d)$					
δ	n=1	n=2	n=3	n=4	n=5
0.3	16	47	85	126	166
0.4	12	35	64	94	125
0.5	10	28	52	76	100
0.8	7	20	37	54	72
0.9	6	16	29	42	56
1.0	**5**	**15**	**26**	**38**	**50**
$f(d) = exp(\delta d)/[1 + exp(\delta d)]$					
δ	n=1	n=2	n=3	n=4	n=5
0.4	39	114	207	301	397
0.6	36	104	185	268	352
0.7	**36**	**105**	**186**	**268**	**351**
0.8	38	108	191	276	360
1.2	52	147	255	364	474
1.5	73	206	354	502	650
1.8	107	304	518	731	945
2.0	142	402	681	960	1238
$f(d) = \Phi(\delta d)$					
δ	n=1	n=2	n=3	n=4	n=5
0.3	36	103	186	271	357
0.4	**34**	**98**	**175**	**254**	**333**
0.5	36	101	179	258	338
0.7	46	130	225	321	416
0.8	57	158	271	384	498

d_j by the directional derivatives F_j for some suitable choices of $f(.)$. Any criterion has both positive and negative vertex directional derivatives. Also note that $\sum p_j F_j = 0$. This is one of the reasons for replacing d_j by F_j. So when we replace d_j by F_j, we need to choose the function $f(x)$ in such a way that the function is centred at zero and changes reasonably quickly about $F = 0$. With this set-up, the algorithm (2.16) becomes

$$p_j^{(r+1)} = \frac{p_j^{(r)} f(F_j^{(r)}, \delta)}{\sum_{i=1}^{J} p_i^{(r)} f(F_i^{(r)}, \delta)}. \tag{3.9}$$

We consider two choices of the function $f(.)$ with $f(F) = exp(\delta F)/1 + exp(\delta F)$ and $f(F) = \Phi(\delta F)$. The results are given in Table 3.2. Comparing Table 3.1 and Table 3.2, we observe that the convergence is improved. For example, using the partial derivatives with $f(d) = [exp(\delta d)]/[1 + exp(\delta d)]$, $\delta = 0.7$ and $n = 5$, the number of iterations needed is 351 (see Table 3.1), whereas using the directional derivatives, for $\delta = 2.0$ and $n = 5$, the number of iterations needed is only 50 (see Table 3.2). We even get better convergence by choosing $f(F) = \Phi(F)$, with $\delta = 2.0$ and $n = 5$. For this choice, the number of iterations needed is only 31.

The solution converged to the optimal design:

$$p^* = \begin{Bmatrix} -1 & 1 \\ 0.5 & 0.5 \end{Bmatrix}. \tag{3.10}$$

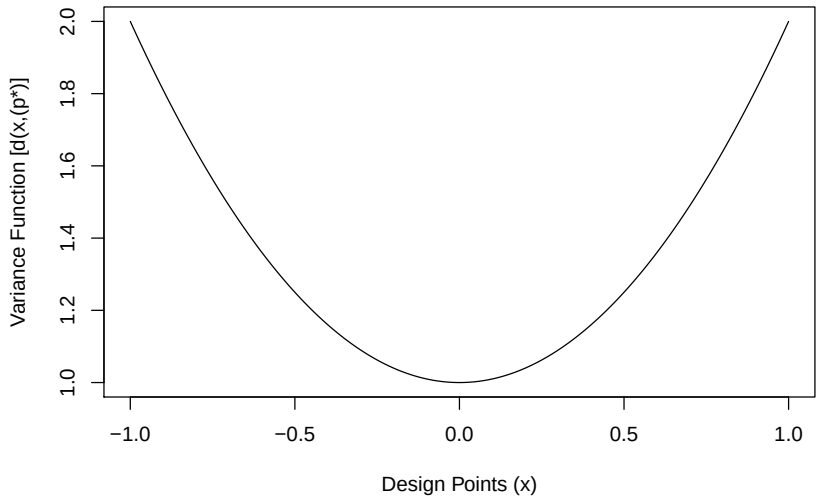
which is exactly the same design we obtained using the analytic approach in the previous section.

Now we investigate the D-optimal design by plotting the standardized variance of the predicted response $d(x, p^*)$ versus the design variable x. This is given in Figure

Table 3.2: Simple linear regression with $x = F$
Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$.

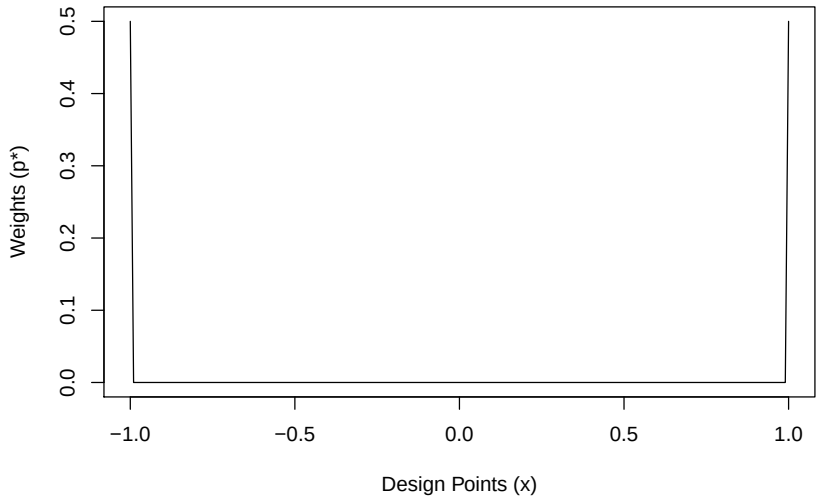
$f(F) = exp(\delta F)/1 + exp(\delta F)$					
δ	n=1	n=2	n=3	n=4	n=5
0.1	93	276	508	748	989
0.3	32	94	171	250	330
0.8	13	37	65	94	124
0.9	12	33	58	84	110
1.0	11	30	53	76	99
1.2	10	26	44	63	83
1.8	8	18	30	43	55
1.9	7	17	29	41	52
2.0	**7**	**17**	**28**	**39**	**50**
$f(F) = \Phi(\delta F)$					
δ	n=1	n=2	n=3	n=4	n=5
0.1	59	174	319	469	620
0.3	21	60	108	157	206
0.8	9	24	42	60	77
0.9	9	22	37	53	69
1.0	8	20	34	48	62
1.2	7	17	29	40	52
1.8	6	13	20	27	35
1.9	6	12	19	26	33
2.0	**6**	**12**	**18**	**25**	**31**

Figure 3.1: Variance Function vs Design Points for Simple Linear Regression Model



3.1. We see that the maximum value of the variance function is 2, the number of the parameters in simple linear regression. This maximum occurs at the design points. The optimal weights are plotted in Figure 3.2.

Figure 3.2: Weights vs Design Points for Simple Linear Regression Model



3.4.2 Quadratic Regression

Now consider the quadratic regression model to construct the D-optimal design. The model is given by

$$E(y|x) = \theta_0 + \theta_1 x + \theta_2 x^2.$$

Now we discretize the design interval by a grid of 21 points equally spaced at intervals of 0.1 between -1 and 1. As we did before, we first take the argument of $f(.)$ as the partial derivative of the criterion function and consider the four choices of $f(.)$. We consider the suitable choices of δ in each case and run algorithm (2.16) until the first order conditions are satisfied.

We start with the equal initial weights $p_j^{(0)} = 1/J$ and record, for $n = 1, 2, 3, 4, 5$, the number of iterations needed to achieve the condition (3.8). The results are given in Table 3.3.

From Table 3.3, we observe that the choices $f(d) = d^\delta$ and $f(d) = exp(\delta d)$ give us faster convergence among the four choices. For example with $f(d) = d^\delta$, $\delta = 1.9$, the number of iteration needed to achieve the condition (3.8) for $n = 5$ is 320. Now we note here the iteration results for suitable choices of δ. As for instance, with $f(d) = \Phi(\delta d)$, $\delta = 1.0$, the number of iterations needed to achieve the condition for $n = 5$ is 43428, whereas for $\delta = 0.3$ the number of iterations needed is 2063. We can see similar results in other choices of $f(d)$.

In the quadratic regression model, we also replace the partial derivative d_j by the directional derivative F_j to obtain the better convergence of the algorithm. We again focus on the two choices $f(F) = \Phi(\delta F)$ and $f(F) = exp(\delta F)/1 + exp(\delta F)$. The results

Table 3.3: Quadratic regression with $x = d$
Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$.

$f(d) = d^{\delta}$					
δ	n=1	n=2	n=3	n=4	n=5
0.2	73	643	1482	2258	3028
0.5	30	258	593	904	1212
0.8	19	161	371	565	758
0.9	17	144	330	502	674
1.7	9	76	175	266	357
1.8	7	72	165	252	337
1.9	**9**	**69**	**157**	**239**	**320**
$f(d) = exp(\delta d)$					
δ	n=1	n=2	n=3	n=4	n=5
0.2	26	213	494	755	1013
0.3	18	142	330	503	676
0.4	13	107	247	378	507
0.5	**11**	**85**	**198**	**302**	**405**
$f(d) = exp(\delta d)/[1 + exp(\delta d)]$					
δ	n=1	n=2	n=3	n=4	n=5
0.1	119	1001	2320	8 3543	4757
0.2	71	602	1394	2127	2855
0.3	58	493	1139	1738	2331
0.4	**54**	**463**	**1067**	**1626**	**2181**
0.5	54	471	1083	1649	2211
1.0	88	908	2076	3152	4220
$f(d) = \Phi(\delta d)$					
δ	n=1	n=2	n=3	n=4	n=5
0.1	82	691	1600	2443	3279
0.2	55	466	1076	1641	2202
0.3	**50**	**439**	**1010**	**1538**	**2063**
0.5	64	621	1421	2158	2890
0.8	150	2369	5385	8150	10893
0.9	273	4516	10254	15503	20708
1.0	537	9488	21537	32529	43428

are reported in Table 3.4. We now compare Table 3.3 and Table 3.4 and see that we improve the convergence rates. For example, with $f(d) = [exp(\delta d)]/[1 + exp(\delta d)]$, $\delta = 0.4$ and $n = 5$, the number of iterations needed is 2181 (see Table 3.3), whereas using the directional derivatives, for $\delta = 1.3$ and $n = 5$, the number of iterations needed is 311 (see Table 3.4). Similar things happen in the results for $f(F) = \Phi(\delta F)$.

The solutions converge to the design:

$$p^* = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ 0.3333 & 0.3333 & 0.3333 \end{array} \right\}. \tag{3.11}$$

This is the same solution that we obtained using the analytic approach earlier.

Table 3.4: Quadratic regression with $x = F$
Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$.

$f(F) = exp(\delta F)/1 + exp(\delta F)$					
δ	n=1	n=2	n=3	n=4	n=5
0.1	101	852	1975	3016	4049
0.3	34	285	659	1005	1349
0.5	21	172	396	603	809
0.8	13	108	248	377	505
0.9	12	96	220	335	449
1.0	11	87	199	302	404
1.1	10	79	181	274	367
1.2	9	73	166	251	337
1.3	**9**	**67**	**153**	**232**	**311**
$f(F) = \Phi(\delta F)$					
δ	n=1	n=2	n=3	n=4	n=5
0.1	64	534	1238	1890	2537
0.3	22	179	413	630	845
0.5	13	108	248	378	506
0.6	11	91	207	315	422
0.7	10	78	178	270	361
0.8	**9**	**69**	**156**	**236**	**316**

Now we also investigate the D-optimal design by plotting the standardized variance of the predicted response $d(x, p^*)$ versus the design variable x. This is given in Figure 3.3. From the figure we see that the maximum value of the variance function is 3, the number of the parameters in the quadratic regression model. This maximum value of 3 occurs at the design points. We plot optimal weights in Figure 3.4.

Figure 3.3: Variance Function vs Design Points for Quadratic Regression Model

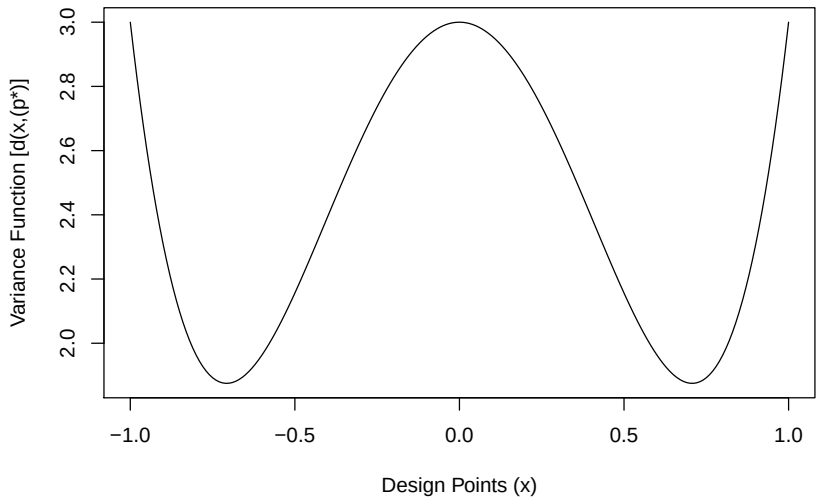
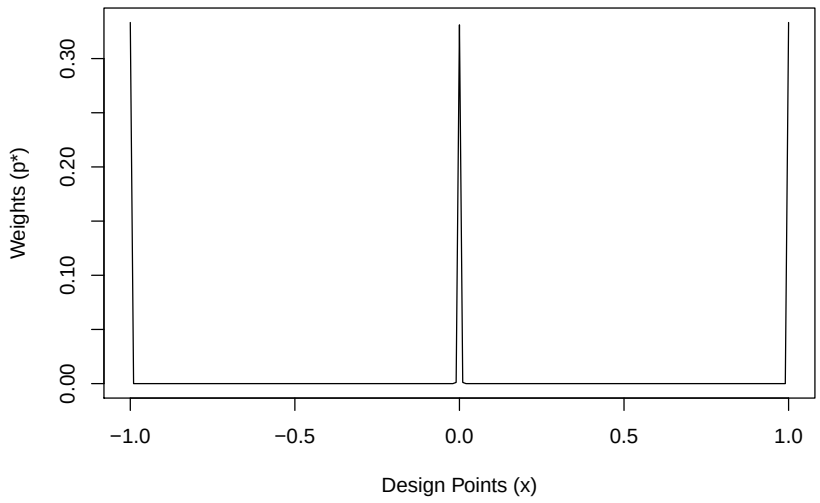


Figure 3.4: Weights vs Design Points for Quadratic Regression Model



3.4.3 Cubic Regression

Now consider the cubic regression model to construct the D-optimal design. The model is given by

$$E(y|x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3.$$

We consider the design interval to be $-1 \leq x \leq 1$ and discretize the design interval by a grid of 221 points equally spaced at intervals of 0.01 between -1 and 1. For this model, we change the increment of x from 0.1 to 0.01 for more accuracy. We first take the argument of $f(.)$ as the partial derivatives of the criterion function and consider the four choices of the function as considered in the previous models. We consider the appropriate choices of δ in each case and run the algorithm (2.16) until the first-order conditions are satisfied.

We start with the initial equal weight $p_j^{(0)} = 1/J$ and record for $n = 1, 2, 3, 4, 5$ the number of iterations needed to achieve the condition (3.8). The results are given in Table 3.5. In the cubic regression model, we also replace the partial derivatives d_j by the directional derivatives F_j to get better convergence results of the algorithm. In Table 3.6 we report the corresponding iteration results.

From Table 3.5, we see that the function $f(d) = d^\delta$ and $f(d) = exp(\delta d)$ gives faster convergence result. For example, with $f(d) = d^\delta$ and $\delta = 1.99$, the number of needed to achieve the first order condition for $n = 5$ is 13971 and for $f(d) = exp(\delta d)$ with $\delta = 0.37$, the number of iteration needed to achieve for $n = 5$ is 18781.

Now if we compare the results of Table 3.5 and Table 3.6, it is obvious that the convergence is improved a great deal by replacing the partial derivatives with the

Table 3.5: Cubic regression with $x = d$
Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$.

$f(d) = d^{\delta}$					
δ	n=1	n=2	n=3	n=4	n=5
1.5	14	131	1217	6521	18534
1.8	15	109	1015	5434	15445
1.9	19	104	961	5148	14632
1.93	25	101	946	5068	14405
1.95	31	69	937	5017	14257
1.97	47	113	927	4966	14113
1.99	**115**	**317**	**533**	**4916**	**13971**
$f(d) = exp(\delta d)$					
δ	n=1	n=2	n=3	n=4	n=5
0.1	51	486	4557	24444	69495
0.2	26	243	2279	12222	34747
0.3	17	162	1519	8148	23165
0.35	14	138	1302	6983	19855
0.37	**13**	**130**	**1230**	**6605**	**18781**
$f(d) = exp(\delta d)/[1 + exp(\delta d)]$					
δ	n=1	n=2	n=3	n=4	n=5
0.1	124	1211	11356	60912	173173
0.2	80	785	7353	39428	112085
0.3	**73**	**702**	**6568**	**35208**	**100084**
0.4	79	726	6790	36391	103439
$f(d) = \Phi(\delta d)$					
δ	n=1	n=2	n=3	n=4	n=5
0.1	89	866	8112	43507	123687
0.2	**70**	**663**	**6204**	**33259**	**94543**
0.3	86	741	6932	37146	105581

directional derivatives. For example, with $f(d) = exp(\delta d)/1 + exp(\delta d)$, for $\delta = 0.3$ and $n = 5$, the number of iterations needed is 100084, whereas using the directional

Table 3.6: Cubic regression with $x = F$
Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$.

$f(F) = exp(\delta F)/1 + exp(\delta F)$					
δ	n=1	n=2	n=3	n=4	n=5
0.4	26	244	2281	12225	34750
0.6	19	163	1521	8151	23168
0.8	16	123	1142	6114	17377
0.9	14	109	1015	5435	15446
0.95	**14**	**104**	**962**	**5149**	**14633**
$f(F) = \Phi(\delta F)$					
δ	n=1	n=2	n=3	n=4	n=5
0.5	15	123	1145	6130	17423
0.6	14	103	954	5109	14519
0.61	14	101	939	5026	14281
0.62	**14**	**95**	**924**	**4945**	**14051**

derivatives, for $\delta = 0.95$ and $n = 5$, the number of iterations needed is only 14633. Similar things happen in the results for other choices as well.

Here we see that the support points can be viewed as consisting of clusters of points. For this cubic regression model, there are two clusters centered on the two middle points, namely around -0.44, -0.45 and 0.44, 0.45. There are two peaks at the ends -1 and 1. This is given by the following design.

$$p^* = \left\{ \begin{array}{cccccc} -1 & -0.45 & -0.44 & 0.44 & 0.45 & 1 \\ 0.25 & 0.223244 & 0.02677738 & 0.02677738 & 0.2232244 & 0.25 \end{array} \right\}. \quad (3.12)$$

This suggests that the solution for the continuous space is a 4-point design, with the 4 support points contained 'within' the clusters and each point having the total design weight of its cluster.

We then take the convex combination of the relevant cluster members (convex weights being proportional to design weights). For example, taking the convex combination of the first cluster, we get the support point

$$\frac{(-0.45)(0.223244) + (-0.44)(0.02677738)}{0.223244 + 0.02677738} \approx -0.45$$

with the corresponding weight 0.25. Similarly we obtain the other support point 0.45 with the weight 0.25.

Thus, we obtain the optimal design

$$p^* = \left\{ \begin{array}{cccc} -1 & -0.45 & 0.45 & 1 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{array} \right\}. \tag{3.13}$$

This is the same solution that we obtained using the analytic approach in the previous section.

Now we also investigate the D-optimal design by plotting the standardized variance of the predicted response $d(x, p^*)$ versus the design variable x. This is given in Figure 3.5. From the figure we see that the maximum value of the variance function is 4, the number of the parameters in the cubic regression model. This maximum value occurs at the four design points. The optimal weights are plotted in Figure 3.6.

Figure 3.5: Variance Function vs Design Points for Cubic Regression Model

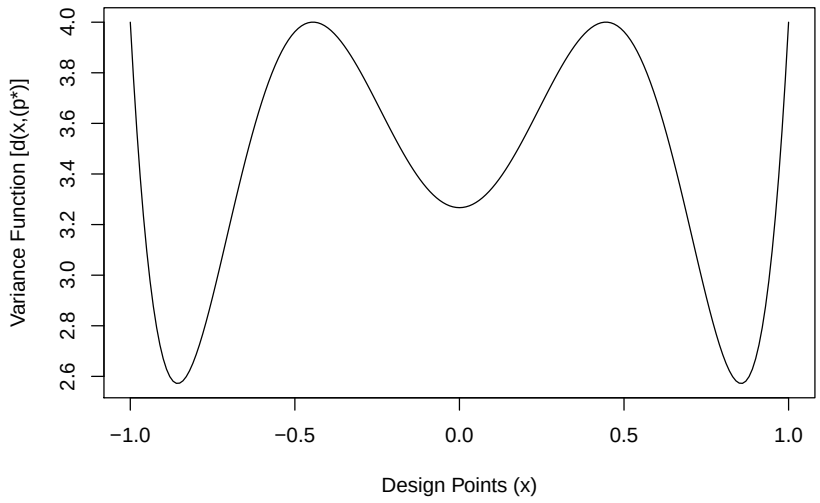
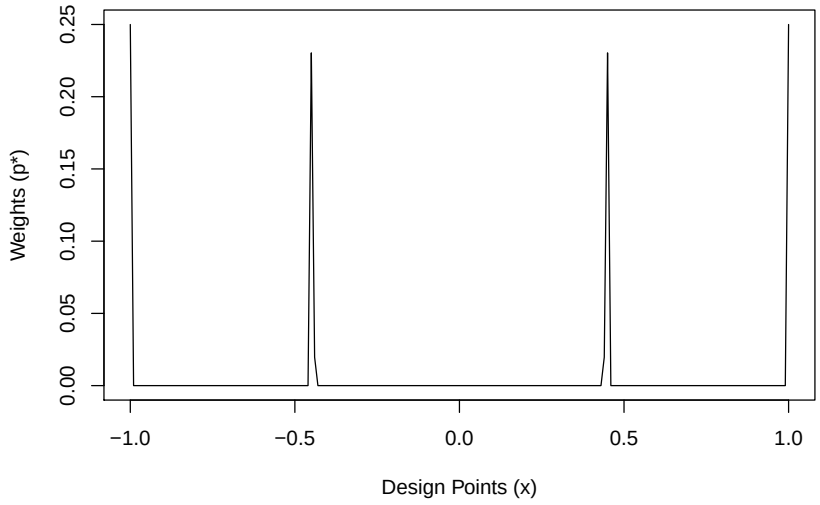


Figure 3.6: Weights vs Design Points for Cubic Regression Model



3.4.4 Quartic Regression

Finally we consider the quartic regression model. The model is

$$E(y|x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4.$$

We consider the design interval to be $-1 \leq x \leq 1$ and discretize the design interval by a grid of 221 points equally spaced at intervals of 0.01 between -1 and 1.

We first take the argument of the function $f(.)$ as the partial derivatives of the criterion function and consider the four choices of the function as considered before. We consider the appropriate choices of δ in each case and run the algorithm (2.16) until the first-order conditions are satisfied. We again start with the initial equal weight $p_j^{(0)} = 1/J$ and record, for $n = 1, 2, 3, 4, 5$, the number of iterations needed to achieve (3.8). The results are given in Table 3.7. In the quartic regression model, we also replace the partial derivative d_j by the directional derivative F_j to get the better convergence of the algorithm. In Table 3.8 we report the corresponding iteration results.

Now if we compare the results of Table 3.7 and Table 3.8, it is obvious that the convergence is improved a lot by replacing the partial derivatives with the directional derivatives. For example, with $f(d) = exp(\delta d)/1 + exp(\delta d)$, for $\delta = 0.3$ and $n = 5$, the number of iterations needed is 73053, whereas using the directional derivatives, for $\delta = 0.7$ and $n = 5$, the number of iterations needed is only 11424. Also with $f(d) = \Phi(\delta d)$, for $\delta = 0.2$ and $n = 5$, the number of iterations needed is 69506, whereas using the directional derivatives, for $\delta = 0.5$ and $n = 5$, the number of iterations needed is only 10022.

Table 3.7: Quartic regression with $x = d$
Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$.

$f(d) = d^{\delta}$					
δ	n=1	n=2	n=3	n=4	n=5
0.5	50	490	4985	23178	39981
0.7	36	350	3561	16556	28558
0.8	32	306	3116	14487	24988
0.9	28	272	2770	12877	22212
1.0	26	245	2493	11590	19991
1.2	22	205	2078	9658	16659
1.5	18	164	1662	7727	13327
1.7	16	145	1467	6818	11760
1.9	**23**	**130**	**1313**	**6100**	**10522**
$f(d) = exp(\delta d)$					
δ	n=1	n=2	n=3	n=4	n=5
0.07	73	698	7117	33106	57112
0.08	64	611	6227	28968	49973
0.09	57	543	5535	25749	44421
0.1	51	489	4982	23174	39979
0.2	**26**	**244**	**2491**	**11587**	**19989**
$f(d) = exp(\delta d)/[1 + exp(\delta d)]$					
δ	n=1	n=2	n=3	n=4	n=5
0.1	133	1294	13197	61384	105893
0.2	93	909	9265	43089	74327
0.3	**90**	**893**	**9108**	**42352**	**73053**
0.4	101	1024	10455	48614	83849
$f(d) = \Phi(\delta d)$					
δ	n=1	n=2	n=3	n=4	n=5
0.1	99	960	9786	45518	78520
0.2	**86**	**849**	**8666**	**40296**	**69506**
0.3	112	1171	11974	55675	96021

Table 3.8: Quartic regression with $x = F$
Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$.

$f(F) = exp(\delta F)/1 + exp(\delta F)$					
δ	n=1	n=2	n=3	n=4	n=5
0.1	101	977	9965	46350	79958
0.2	51	489	4983	23176	39979
0.3	34	326	3323	15451	26653
0.4	26	245	2493	11589	19990
0.5	21	196	1994	9272	15992
0.6	18	164	1662	7727	13327
0.7	**15**	**141**	**1425**	**6624**	**11424**
0.8	13	75	12074	31148	50104
$f(F) = \Phi(\delta F)$					
δ	n=1	n=2	n=3	n=4	n=5
0.2	32	307	3123	14524	25054
0.3	22	205	2083	9684	16703
0.4	17	154	1563	7263	12528
0.5	**13**	**75**	**1243**	**5811**	**10022**

Here we see that the support points can be viewed as consisting of clusters of points. For this quartic regression model, there are two clusters centred on the two middle points, namely around -0.66, -0.65 and 0.65, 0.66. There are three peaks at the point -1, 0 and 1. This is given by the following design.

$$p^* = \left\{ \begin{array}{ccccccc} -1 & -0.66 & -0.65 & 0.00 & 0.65 & 0.66 & 1 \\ 0.20 & 0.08475309 & 0.1152696 & 0.20 & 0.1152696 & 0.08475309 & 0.20 \end{array} \right\} \quad (3.14)$$

This suggests that the solution for the continuous space is a 5-point design, with the 5 support points contained within the clusters and each point having the total design weight of its cluster.

We then take the convex combination of the relevant cluster members (convex weights being proportional to design weights). For example, taking the convex combination of the first cluster, we get the support point

$$\frac{(-0.66)(0.08475309) + (-0.65)(0.1152696)}{0.08475309 + 0.1152696} \approx -0.65$$

with the corresponding weight 0.20. Similarly we obtain the other support point 0.65 with the weight 0.20.

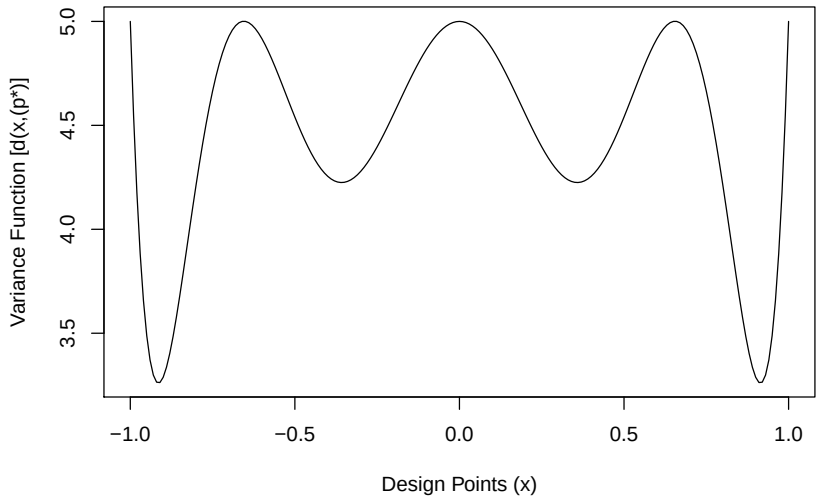
Thus, we obtain the optimal design

$$p^* = \left\{ \begin{array}{ccccc} -1 & -0.65 & 0 & 0.65 & 1 \\ 0.20 & 0.20 & 0.20 & 0.20 & 0.20 \end{array} \right\}. \quad (3.15)$$

This is the same solution that we obtained using the analytic approach earlier.

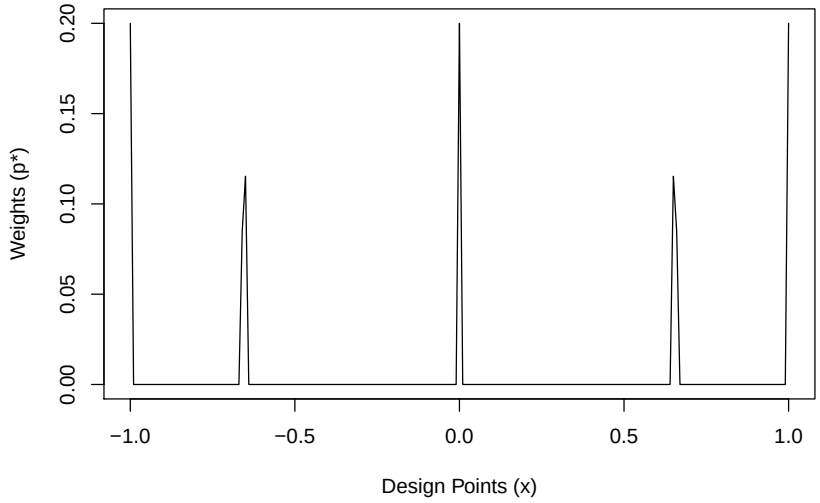
Now we also investigate the D-optimal design by plotting the standardized variance of the predicted response $d(x, p^*)$ versus the design variable x. This is given in Figure

Figure 3.7: Variance Function vs Design Points for Quartic Regression Model



3.7. We see that the maximum value of the variance function is 5, the number of the parameters in this model. This maximum value occurs at the five design points. The optimal weights are plotted in Figure 3.8.

Figure 3.8: Weights vs Design Points for Quartic Regression Model



Chapter 4

Maximum Likelihood Estimation of the Cell Probabilities under the Hypothesis of Marginal Homogeneity

4.1 Introduction

In this chapter, we present a quite flexible methodology to solve a maximum likelihood estimation problem using optimal design theory and simultaneous optimization techniques. We consider the problem of determining the maximum likelihood estimates of the cell probabilities under the hypothesis of marginal homogeneity in a square contingency table. This is an optimization problem with respect to variables that satisfy several constraints based on the marginal homogeneity conditions. We first formulate the Lagrangian function with the constraints and then transform the problem to that of maximizing some functions of the cell probabilities simultaneously. The functions have a common maximum of zero which is simultaneously attained at the optimal design weights. We apply the methodologies in some data sets for which the hypothesis of marginal homogeneity is of interest.

4.2 Formulation of the Constrained Optimization Problem

We first add some structure to the general problem which we discussed in Chapter 2. Our more general problem is to maximize $\phi(\underline{\theta})$ over $\Theta \equiv (\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)^T : \theta_j \geq 0,$ $C\underline{\theta} = \underline{a})$ where C is an $s \times k$ matrix of rank s and $\underline{a}$ is in the range space of C. This problem arises in testing linear hypothesis about the parameters in multinomial models for categorical data. These parameters are probabilities so that $C\underline{\theta} = \underline{a}$ must include a component that $\underline{1}^T\theta = 1$, where $\underline{1}$ is a vector of 1's. As an application of this optimization problem, we wish to consider a problem of determining the maximum likelihood estimates of the cell probabilities under the hypothesis of marginal homogeneity in a $n \times n$ contingency table.

There has been some work in the literature for testing marginal homogeneity in this context. For example, Ireland et al. (1969) used the principle of minimum discrimination information estimation; Wedderburn (1974) used a generalized linear models approach; and Mandal and Torsney (2000) used the vertices of the feasible region to estimate the cell frequencies. However, as the dimension of the contingency table increases, the estimation becomes intractable. Also, some of these works assume symmetry or quasi-symmetry of the cell probabilities. Here we propose a quite flexible method using a Lagrangian approach and simultaneous optimization techniques. We refer this work to our manuscript Chowdhury and Mandal (2016) which is in preparation.

Given the observed frequencies O_{ij}, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, n$ and assuming a

single multinomial distribution conditional on $N = \sum_{i=1}^{n} \sum_{j=1}^{n} O_{ij}$, with cell probabilities θ_{ij}, we wish to maximize the log likelihood function $\phi(\underline{\theta}) = \sum_{i=1}^{n} \sum_{j=1}^{n} O_{ij} ln(\theta_{ij})$ subject to $\theta_{ij} \geq 0$, $\sum_{i=1}^{n} \sum_{j=1}^{n} \theta_{ij} = 1$ and the marginal homogeneity constraint $\sum_{j=1}^{n} \theta_{rj} = \sum_{j=1}^{n} \theta_{jr}$ for $r = 1, 2, \ldots, n$. Some simplification of the problem is possible in view of the fact that at the solution $\theta_{ii} = O_{ii}/N, i = 1, 2, \ldots, n$, and also that one of the marginal homogeneity constraints, for example, that corresponding to $r = n$, can be removed since they are linearly dependent. Thus we need to consider only $n - 1$ constraints based on the marginal homogeneity conditions.

Now we can formulate the problem of maximizing the log likelihood function $\phi(\underline{\theta})$ subject to the marginal homogeneity constraints $h_i(\underline{\theta}) = c_i$, $i = 1, 2, \ldots, n - 1$ with $\theta_{ij} \geq 0$, $\sum_{i=1}^{n} \sum_{j=1}^{n} \theta_{ij} = 1$ where c_i's are some constants. Later on we set the constants c_i's to zero as these constants will be zero in the marginal homogeneity conditions $h_i(\underline{\theta}) = \sum_{j=1}^{n} \theta_{rj} - \sum_{j=1}^{n} \theta_{jr} = 0$, $i = 1, 2, \ldots, n - 1$. Note that the cell probabilities θ_{ij}'s have two suffixes i and j. Suppose that we have J cell probabilities under consideration. For simplicity we rewrite these probabilities in one suffix and denote the vector $\underline{\theta}$ as $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_J)^T$.

We formulate the Lagrangian function as given by

$$L(\phi, \underline{\theta}, \underline{\lambda}, \mu) = \phi(\underline{\theta}) + \sum_{i=1}^{n-1} \lambda_i (h_i(\underline{\theta}) - c_i) + \mu(\sum_{j=1}^{J} \theta_j - 1).$$

The Lagrange multiplier λ_i is the rate of change of the likelihood function being

maximized as a function of the constraint parameter.

We first consider the partial derivatives of the Lagrangian function with respect to θ_j, $j = 1, 2, \ldots, J$.

These can be written as

$$\begin{aligned} d_j^L &= \frac{\partial L}{\partial \theta_j} \\ &= \frac{\partial \phi}{\partial \theta_j} + \lambda_1 \frac{\partial h_1}{\partial \theta_j} + \lambda_2 \frac{\partial h_2}{\partial \theta_j} + \ldots + \lambda_{n-1} \frac{\partial h_{n-1}}{\partial \theta_j} + \mu \\ &= d_j^{\phi} + \lambda_1 d_j^{h_1} + \lambda_2 d_j^{h_2} + \ldots + \lambda_{n-1} d_j^{h_{n-1}} + \mu \\ &= d_j^{\phi} + \sum_{i=1}^{n-1} \lambda_i \frac{\partial d_j^{h_i}}{\partial \theta_j} + \mu, \end{aligned} \tag{4.1}$$

where $d_j^{\phi} = \frac{\partial \phi}{\partial \theta_j}$, $d_j^{h_1} = \frac{\partial h_1}{\partial \theta_j}$, $d_j^{h_2} = \frac{\partial h_2}{\partial \theta_j}$, $\ldots$, $d_j^{h_{n-1}} = \frac{\partial h_{n-1}}{\partial \theta_j}$.

As for $\theta_j > 0$, $j = 1, 2, \ldots, J$, we should have $d_j^L = 0 \; \forall j$. It implies that $\sum_{j=1}^{J} \theta_j d_j^L = 0$ so that $\mu = -\sum_{j=1}^{J} \theta_j \left(d_j^{\phi} + \lambda_1 d_j^{h_1} + \lambda_2 d_j^{h_2} + \ldots + \lambda_{n-1} d_j^{h_{n-1}} \right)$.

Hence, we can write

$$d_j^{\phi} + \lambda_1 d_j^{h_1} + \lambda_2 d_j^{h_2} + \ldots + \lambda_{n-1} d_j^{h_{n-1}} = -\mu = \sum_{j=1}^{J} \theta_j \left(d_j^{\phi} + \lambda_1 d_j^{h_1} + \lambda_2 d_j^{h_2} + \ldots + \lambda_{n-1} d_j^{h_{n-1}} \right).$$

Hence, from the above, the vertex directional derivatives of L are as follows

$$\begin{aligned} F_j^L &= d_j^L - \sum_{j=1}^{J} \theta_j d_j^L \\ &= \left(d_j^\phi + \lambda_1 d_j^{h_1} + \lambda_2 d_j^{h_2} + \ldots + \lambda_{n-1} d_j^{h_{n-1}}\right) - \sum_{j=1}^{J} \theta_j \left(d_j^\phi + \lambda_1 d_j^{h_1} + \lambda_2 d_j^{h_2} + \ldots + \lambda_{n-1} d_j^{h_{n-1}}\right) \\ &= F_j^\phi + \sum_{i=1}^{n-1} \lambda_i F_j^{h_i} \\ &\equiv 0, \end{aligned} \tag{4.2}$$

where

$$F_j^\phi = d_j^\phi - \sum_{j=1}^{J} \theta_j d_j^\phi$$

and

$$F_j^{h_i} = d_j^{h_i} - \sum_{j=1}^{J} \theta_j d_j^{h_i}$$

are the directional derivatives of ϕ and h_i respectively, $i = 1, 2, \ldots, n-1$.

Now from (4.2), we can write $\underline{F}^L = \underline{F}^\phi + \sum_{i=1}^{n-1} \lambda_i \underline{F}^{h_i} = \underline{0}$. Hence,

$$F^h \underline{\lambda} = -\underline{F}^\phi, \tag{4.3}$$

where $F^h = \left[\underline{F}^{h_1}, \underline{F}^{h_2}, \ldots, \underline{F}^{h_{n-1}}\right]$ and $\underline{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_{n-1})^T$.

The set of solutions to the above system of equations (4.3), if solutions exists, is given by $\underline{\lambda} = (F^h)^-(-\underline{F}^\phi) + (I - (F^h)^-(F^h))\underline{z}$ for any $\underline{z}$, where $(F^h)^-$ is any generalized inverse of F^h.

If $(F^h)^T(F^h)$ is nonsingular, then one choice is the Moore-Penrose inverse of F^h, and is given by

$$(F^h)^- = \left[(F^h)^T(F^h)\right]^{-1}(F^h)^T.$$

Then,

$$\underline{\hat{\lambda}} = (F^h)^-(-\underline{F}^\phi) = -\left[(F^h)^T(F^h)\right]^{-1}(F^h)^T\underline{F}^\phi. \tag{4.4}$$

Also the optimum $\underline{\theta}^*$ must satisfy

$$F^h\underline{\hat{\lambda}} = -\underline{F}^\phi. \tag{4.5}$$

Now substituting the value of $\hat{\lambda}$ from (4.4) in the above equation (4.5), we obtain

$$F^h\left[(F^h)^T(F^h)\right]^{-1}(F^h)^T\underline{F}^\phi = \underline{F}^\phi \tag{4.6}$$

That is, the optimum $\underline{\theta}^*$ must satisfy

$$\underline{k} = (k_1, k_2, \ldots, k_J)^T \equiv \underline{0} \tag{4.7}$$

where $\underline{k} = F^h\left[(F^h)^T(F^h)\right]^{-1}(F^h)^T\underline{F}^\phi - \underline{F}^\phi$. The elements of $\underline{k}$ are functions of the directional derivatives of ϕ and h_i, $i = 1, 2, \ldots, n-1$.

Since, in general, $\underline{k}^T\underline{k} \geq 0$, θ^* should minimize $\underline{k}^T\underline{k}$ or maximize $K(\underline{\theta}) = \left[-\underline{k}^T\underline{k}\right]$ with a maximum value of zero.

Now let us consider the possibility of zero weights. By the equivalence theorem, we have

$$F_j^L \begin{cases} = & 0 \quad \text{if } \theta_j^* > 0 \\ \leq & 0 \quad \text{if } \theta_j^* = 0. \end{cases}$$

One way to approach this problem is to consider

$$g_i = \theta_i k_i = 0 \quad \text{for } i = 1, 2, \ldots, J$$

$$\text{i.e.} \quad \underline{g} = \Omega \underline{k} = \underline{0}$$

where $\Omega = \text{diag}(\theta_1, \theta_2, \ldots, \theta_J)$. So we replace $\underline{k}$ in the argument above by $\underline{g} = \Omega \underline{k}$ and conclude that θ^* should maximize $G = -\underline{k}^T \Omega^2 \underline{k}$. This was considered by Torsney and Mandal (2001). However, we do not need to consider this case as we do not have any zero weights in our optimization problem.

In respect of $\underline{\theta} = (\theta_1, \ldots, \theta_J)$ we have transformed our problem to that of an optimization problem, namely maximization of $K(\underline{\theta})$ (or any increasing function of $K(\underline{\theta})$) with respect to $\underline{\theta}$. At this point we also need to ensure that the constraints satisfy the first order conditions with respect to the Lagrange multipliers. These can be done by simultaneously maximizing $H_i(\underline{\theta}) = -[h_i(\underline{\theta}) - c_i]^2$.

Thus, we have transformed the above problem to that of maximizing n functions $(K(\underline{\theta}), H_i(\underline{\theta})), i = 1, 2, \ldots, n-1)$ simultaneously.

We can solve this problem in two ways:

(i) As each of these functions is negative and has a common maximum of zero, we can maximize their sum, $K(\underline{\theta}) + \sum_{i=1}^{n-1} H_i(\underline{\theta})$, which will have a maximum of zero at the

common optimizing $\underline{\theta}$.

(ii) Alternatively, we can solve a *Maximin* problem in which we maximize $min\{K(\underline{\theta}), H_i(\underline{\theta}), i = 1, \ldots, n-1\}$, which will be zero at the optimum.

4.3 Constrained Optimization Problem for the 4 × 4 Case

In this section, we consider a particular case for $n = 4$, i.e., a 4×4 contingency table, and derive the expressions of the previous section explicitly. For convenience, we use the following notations for the observed and expected frequencies. Let

$$\begin{aligned} y_1, y_2, y_3, y_4, y_5, y_6 &= O_{12}, O_{13}, O_{14}, O_{21}, O_{23}, O_{24} \\ y_7, y_8, y_9, y_{10}, y_{11}, y_{12} &= O_{31}, O_{32}, O_{34}, O_{41}, O_{42}, O_{43} \end{aligned}$$

and

$$\begin{aligned} x_1, x_2, x_3, x_4, x_5, x_6 &= E_{12}, E_{13}, E_{14}, E_{21}, E_{23}, E_{24} \\ x_7, x_8, x_9, x_{10}, x_{11}, x_{12} &= E_{31}, E_{32}, E_{34}, E_{41}, E_{42}, E_{43} \end{aligned}$$

where $E_{ij} = N\theta_{ij}$, $i = 1, 2, 3, 4$, $j = 1, 2, 3, 4$ and hence are expected frequencies.

As we mentioned earlier, some simplification of the problem is possible in view of the fact that at the solution $\theta_{ii} = O_{ii}/N, i = 1, 2, 3, 4$. So we can write the maximum likelihood problem subject to the constraints of marginal homogeneity as given in the following:

Maximize $\phi(x) = \sum_{t=1}^{12} y_t ln(x_t)$ subject to

$$x_t \geq 0, t = 1, 2, \ldots, 12$$

$$\sum_{t=1}^{12} x_t = b = (N - \sum_{i=1}^{4} O_{ii})$$

$$x_1 + x_2 + x_3 - x_4 - x_7 - x_{10} = 0$$

$$-x_1 + x_4 + x_5 + x_6 - x_8 - x_{11} = 0$$

$$-x_2 - x_5 + x_7 + x_8 + x_9 - x_{12} = 0.$$

The last three equations are simplified according to the marginal homogeneity constrains. As we discussed in the previous section, one of the marginal homogeneity constraints, for example, that corresponding to $r = 4$, can be removed since they are linearly dependent.

Let us define $p_t = x_t/b$. So we can now write the above maximum likelihood function in this form

$$\phi = \phi(\underline{p}) = \sum_t y_t ln(p_t) + \sum_t y_t ln(b).$$

Note that $\sum_t p_t = 1$ as $\sum x_t = b$. So our problem now is to maximize $\phi\left(\underline{p}\right) = \sum_t y_t ln(p_t)$

subject to $p_t \geq 0, t = 1, 2, \ldots, 12$, $\sum_{t=1}^{12} p_t = 1$, and

$$h_1\left(\underline{p}\right) = p_1 + p_2 + p_3 - p_4 - p_7 + p_{10} = 0$$

$$h_2\left(\underline{p}\right) = -p_1 + p_4 + p_5 + p_6 - p_8 - p_{11} = 0$$

$$h_3\left(\underline{p}\right) = -p_2 - p_5 + p_7 + p_8 + p_9 - p_{12} = 0$$

i.e., $C\underline{p} = \underline{a}$, where

$$C = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 & 1 & 1 & 0 & -1 & 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 & -1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}, \ \underline{a} = (1,0,0,0)^T, \ \underline{p} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \\ p_6 \\ p_7 \\ p_8 \\ p_9 \\ p_{10} \\ p_{11} \\ p_{12} \end{pmatrix}.$$

Now we can formulate the problem of maximizing the likelihood function $\phi(\underline{p})$ subject to the marginal homogeneity constraints using Lagrangian approach. The Lagrangian function is

$$L(\phi, \underline{p}, \underline{\lambda}, \mu) = \phi\left(\underline{p}\right) + \sum_{i=1}^{3} \lambda_i \left(h_i(\underline{p}) - c_i\right) + \mu(\sum_{t=1}^{12} p_t - 1),$$

where $h_1(\underline{p}) = c_1$, $h_2(\underline{p}) = c_2$ and $h_3(\underline{p}) = c_3$ with $p_t \geq 0$ and $\sum p_t = 1$, for constants c_1, c_2 and c_3 equal to zero. The partial derivatives of the Lagrangian function with respect to p_t are

$$d_t^L = \tfrac{\partial L}{\partial p_t} = d_t^{\phi} + \lambda_1 d_t^{h_1} + \lambda_2 d_t^{h_2} + \lambda_3 d^{h_3} + \mu,$$

where $d_t^{\phi} = \tfrac{\partial \phi}{\partial p_t}$, $d_t^{h_1} = \tfrac{\partial h_1}{\partial p_t}$, $d_t^{h_2} = \tfrac{\partial h_2}{\partial p_t}$ and $d_t^{h_3} = \tfrac{\partial h_3}{\partial p_t}$. We must have $d_t^L = 0 \ \forall t$, which means $\sum_{t=1}^{12} p_t d_t^L = 0$ so that $\mu = -\sum_{t=1}^{12} p_t \left(d_t^{\phi} + \lambda_1 d_t^{h_1} + \lambda_2 d_t^{h_2} + \lambda_3 d_t^{h_3}\right)$.

Now considering the vertex directional derivatives of L and substituting the values of μ we can obtain the vertex directional derivatives of L in this form

$$F_t^L = F_t^\phi + \lambda_1 F_t^{h_1} + \lambda_2 F_t^{h_2} + \lambda_3 F_t^{h_3},$$

where $F_t^\phi = d_t^\phi - \sum p_t d_t^\phi$, $F_t^{h_1} = d_t^{h_1} - \sum p_t d_t^{h_1}$, $F_t^{h_2} = d_t^{h_2} - \sum p_t d_t^{h_2}$ and $F_t^{h_3} = d_t^{h_3} - \sum p_t d_t^{h_3}$ are the directional derivatives of ϕ, h_1, h_2 and h_3 respectively.

The directional derivatives of L must be zero at the optimum for $p_t > 0$. That is, in the vector notation, we have

$$\begin{aligned} \underline{F}^L &= \underline{0} \\ i.e.,\ F^h \underline{\lambda} &= -\underline{F}^\phi, \end{aligned}$$

where $F^h = \left[\underline{F}^{h_1}, \underline{F}^{h_2}, \underline{F}^{h_3}\right]$ and $\underline{\lambda} = (\lambda_1, \lambda_2, \lambda_3)^T$. If the above system of equation is consistent, the set of solutions to the system is

$$\underline{\lambda} = -(F^h)^-(\underline{F}^\phi) + (I - (F^h)^-(F^h))\underline{z}$$

for any $\underline{z}$ and $(F^h)^-$ is any g-inverse of F^h. If $(F^h)^T(F^h)$ is nonsingular, one choice of $(F^h)^-$ is

$$\begin{aligned} (F^h)^- &= \left[(F^h)^T(F^h)\right]^{-1}(F^h)^T \\ &= \frac{1}{b}\begin{bmatrix} (\underline{F}^{h_2})^T(\underline{F}^{h_2})(\underline{F}^{h_3})^T(\underline{F}^{h_3})(\underline{F}^{h_1})^T - (\underline{F}^{h_2})^T(\underline{F}^{h_3})(\underline{F}^{h_3})^T(\underline{F}^{h_2})(\underline{F}^{h_1})^T \\ -(\underline{F}^{h_1})^T(\underline{F}^{h_2})(\underline{F}^{h_3})^T(\underline{F}^{h_3})(\underline{F}^{h_2})^T + (\underline{F}^{h_1})^T(\underline{F}^{h_3})(\underline{F}^{h_3})^T(\underline{F}^{h_2})(\underline{F}^{h_2})^T \\ +(\underline{F}^{h_1})^T(\underline{F}^{h_2})(\underline{F}^{h_2})^T(\underline{F}^{h_3})(\underline{F}^{h_3})^T - (\underline{F}^{h_1})^T(\underline{F}^{h_3})(\underline{F}^{h_2})^T(\underline{F}^{h_2})(\underline{F}^{h_3})^T \\ \\ (\underline{F}^{h_3})^T(\underline{F}^{h_1})(\underline{F}^{h_2})^T(\underline{F}^{h_3})(\underline{F}^{h_1})^T - (\underline{F}^{h_2})^T(\underline{F}^{h_1})(\underline{F}^{h_3})^T(\underline{F}^{h_3})(\underline{F}^{h_1})^T \\ -(\underline{F}^{h_3})^T(\underline{F}^{h_1})(\underline{F}^{h_1})^T(\underline{F}^{h_3})(\underline{F}^{h_2})^T + (\underline{F}^{h_1})^T(\underline{F}^{h_1})(\underline{F}^{h_3})^T(\underline{F}^{h_3})(\underline{F}^{h_2})^T \\ +(\underline{F}^{h_1})^T(\underline{F}^{h_3})(\underline{F}^{h_2})^T(\underline{F}^{h_1})(\underline{F}^{h_3})^T - (\underline{F}^{h_1})^T(\underline{F}^{h_1})(\underline{F}^{h_2})^T(\underline{F}^{h_3})(\underline{F}^{h_3})^T \\ \\ -(\underline{F}^{h_3})^T(\underline{F}^{h_1})(\underline{F}^{h_2})^T(\underline{F}^{h_2})(\underline{F}^{h_1})^T + (\underline{F}^{h_2})^T(\underline{F}^{h_1})(\underline{F}^{h_3})^T(\underline{F}^{h_2})(\underline{F}^{h_1})^T \\ +(\underline{F}^{h_3})^T(\underline{F}^{h_1})(\underline{F}^{h_1})^T(\underline{F}^{h_2})(\underline{F}^{h_2})^T - (\underline{F}^{h_1})^T(\underline{F}^{h_1})(\underline{F}^{h_3})^T(\underline{F}^{h_2})(\underline{F}^{h_2})^T \\ -(\underline{F}^{h_2})^T(\underline{F}^{h_1})(\underline{F}^{h_1})^T(\underline{F}^{h_2})(\underline{F}^{h_3})^T + (\underline{F}^{h_1})^T(\underline{F}^{h_1})(\underline{F}^{h_2})^T(\underline{F}^{h_2})(\underline{F}^{h_3})^T \end{bmatrix} \end{aligned}$$

where

$$b = (\underline{F}^{h_1})^T(\underline{F}^{h_1})(\underline{F}^{h_2})^T(\underline{F}^{h_2})(\underline{F}^{h_3})^T(\underline{F}^{h_3}) - (\underline{F}^{h_1})^T(\underline{F}^{h_1})(\underline{F}^{h_2})^T(\underline{F}^{h_3})(\underline{F}^{h_3})^T(\underline{F}^{h_2})$$

$$- (\underline{F}^{h_1})^T(\underline{F}^{h_2})(\underline{F}^{h_2})^T(\underline{F}^{h_1})(\underline{F}^{h_3})^T(\underline{F}^{h_3}) + (\underline{F}^{h_1})^T(\underline{F}^{h_2})(\underline{F}^{h_2})^T(\underline{F}^{h_3})(\underline{F}^{h_3})^T(\underline{F}^{h_1})$$

$$+ (\underline{F}^{h_1})^T(\underline{F}^{h_3})(\underline{F}^{h_2})^T(\underline{F}^{h_1})(\underline{F}^{h_3})^T(\underline{F}^{h_2}) - (\underline{F}^{h_1})^T(\underline{F}^{h_3})(\underline{F}^{h_2})^T(\underline{F}^{h_2})(\underline{F}^{h_3})^T(\underline{F}^{h_1}).$$

Then,

$$\hat{\underline{\lambda}} = -(F^h)^-(\underline{F}^{\phi})$$

$$= -\frac{1}{b}\begin{bmatrix}
((\underline{F}^{h_2})^T(\underline{F}^{h_2}))((\underline{F}^{h_3})^T(\underline{F}^{h_3}))((\underline{F}^{h_1})^T(F^{\phi})) - ((\underline{F}^{h_2})^T(\underline{F}^{h_3}))((\underline{F}^{h_3})^T(\underline{F}^{h_2}))((\underline{F}^{h_1})^T(F^{\phi})) \\
-((\underline{F}^{h_1})^T(\underline{F}^{h_2}))((\underline{F}^{h_3})^T(\underline{F}^{h_3}))((\underline{F}^{h_2})^T(F^{\phi})) + ((\underline{F}^{h_1})^T(\underline{F}^{h_3}))((\underline{F}^{h_3})^T(\underline{F}^{h_2}))((\underline{F}^{h_2})^T(F^{\phi})) \\
+((\underline{F}^{h_1})^T(\underline{F}^{h_2}))((\underline{F}^{h_2})^T(\underline{F}^{h_3}))((\underline{F}^{h_3})^T(F^{\phi})) - ((\underline{F}^{h_1})^T(\underline{F}^{h_3}))((\underline{F}^{h_2})^T(\underline{F}^{h_2}))((\underline{F}^{h_3})^T(F^{\phi})) \\
\\
((\underline{F}^{h_3})^T(\underline{F}^{h_1}))((\underline{F}^{h_2})^T(\underline{F}^{h_3}))((\underline{F}^{h_1})^T(F^{\phi})) - ((\underline{F}^{h_2})^T(\underline{F}^{h_1}))((\underline{F}^{h_3})^T(\underline{F}^{h_3}))((\underline{F}^{h_1})^T(F^{\phi})) \\
-((\underline{F}^{h_3})^T(\underline{F}^{h_1}))((\underline{F}^{h_1})^T(\underline{F}^{h_3}))((\underline{F}^{h_2})^T(F^{\phi})) + ((\underline{F}^{h_1})^T(\underline{F}^{h_1}))((\underline{F}^{h_3})^T(\underline{F}^{h_3}))((\underline{F}^{h_2})^T(F^{\phi})) \\
+((\underline{F}^{h_1})^T(\underline{F}^{h_3}))((\underline{F}^{h_2})^T(\underline{F}^{h_1}))((\underline{F}^{h_3})^T(F^{\phi})) - ((\underline{F}^{h_1})^T(\underline{F}^{h_1}))((\underline{F}^{h_2})^T(\underline{F}^{h_3}))((\underline{F}^{h_3})^T(F^{\phi})) \\
\\
-((\underline{F}^{h_3})^T(\underline{F}^{h_1}))((\underline{F}^{h_2})^T(\underline{F}^{h_2}))((\underline{F}^{h_1})^T(F^{\phi})) + ((\underline{F}^{h_2})^T(\underline{F}^{h_1}))((\underline{F}^{h_3})^T(\underline{F}^{h_2}))((\underline{F}^{h_1})^T(F^{\phi})) \\
+((\underline{F}^{h_3})^T(\underline{F}^{h_1}))((\underline{F}^{h_1})^T(\underline{F}^{h_2}))((\underline{F}^{h_2})^T(F^{\phi})) - ((\underline{F}^{h_1})^T(\underline{F}^{h_1}))((\underline{F}^{h_2})^T(\underline{F}^{h_3}))((\underline{F}^{h_3})^T(F^{\phi})) \\
-((\underline{F}^{h_2})^T(\underline{F}^{h_1}))((\underline{F}^{h_1})^T(\underline{F}^{h_2}))((\underline{F}^{h_3})^T(F^{\phi})) + ((\underline{F}^{h_1})^T(\underline{F}^{h_1}))((\underline{F}^{h_2})^T(\underline{F}^{h_2}))((\underline{F}^{h_3})^T(F^{\phi}))
\end{bmatrix}$$

and p^* should be such that it must satisfy the following

$$F^h\hat{\underline{\lambda}} = -\underline{F}^{\phi}.$$

Now substituting the value of $\hat{\lambda}$ in the above equation, we obtain

$$\underline{k} = \begin{bmatrix}
((\underline{F}^{h_2})^T(\underline{F}^{h_2}))((\underline{F}^{h_3})^T(\underline{F}^{h_3}))((\underline{F}^{h_1})^T(F^{\phi})) - ((\underline{F}^{h_2})^T(\underline{F}^{h_3}))((\underline{F}^{h_3})^T(\underline{F}^{h_2}))((\underline{F}^{h_1})^T(F^{\phi})) \\
-((\underline{F}^{h_1})^T(\underline{F}^{h_2}))((\underline{F}^{h_3})^T(\underline{F}^{h_3}))((\underline{F}^{h_2})^T(F^{\phi})) + ((\underline{F}^{h_1})^T(\underline{F}^{h_3}))((\underline{F}^{h_3})^T(\underline{F}^{h_2}))((\underline{F}^{h_2})^T(F^{\phi})) \\
+((\underline{F}^{h_1})^T(\underline{F}^{h_2}))((\underline{F}^{h_2})^T(\underline{F}^{h_3}))((\underline{F}^{h_3})^T(F^{\phi})) - ((\underline{F}^{h_1})^T(\underline{F}^{h_3}))((\underline{F}^{h_2})^T(\underline{F}^{h_2}))((\underline{F}^{h_3})^T(F^{\phi}))
\end{bmatrix}\underline{F}^{h_1}$$

$$+\begin{bmatrix}
((\underline{F}^{h_3})^T(\underline{F}^{h_1}))((\underline{F}^{h_2})^T(\underline{F}^{h_3}))((\underline{F}^{h_1})^T(F^{\phi})) - ((\underline{F}^{h_2})^T(\underline{F}^{h_1}))((\underline{F}^{h_3})^T(\underline{F}^{h_3}))((\underline{F}^{h_1})^T(F^{\phi})) \\
-((\underline{F}^{h_3})^T(\underline{F}^{h_1}))((\underline{F}^{h_1})^T(\underline{F}^{h_3}))((\underline{F}^{h_2})^T(F^{\phi})) + ((\underline{F}^{h_1})^T(\underline{F}^{h_1}))((\underline{F}^{h_3})^T(\underline{F}^{h_3}))((\underline{F}^{h_2})^T(F^{\phi})) \\
+((\underline{F}^{h_1})^T(\underline{F}^{h_3}))((\underline{F}^{h_2})^T(\underline{F}^{h_1}))((\underline{F}^{h_3})^T(F^{\phi})) - ((\underline{F}^{h_1})^T(\underline{F}^{h_1}))((\underline{F}^{h_2})^T(\underline{F}^{h_3}))((\underline{F}^{h_3})^T(F^{\phi}))
\end{bmatrix}\underline{F}^{h_2}$$

$$+\begin{bmatrix} -((\underline{F}^{h_3})^T(\underline{F}^{h_1}))((\underline{F}^{h_2})^T(\underline{F}^{h_2}))((\underline{F}^{h_1})^T(F^{\phi})) + ((\underline{F}^{h_2})^T(\underline{F}^{h_1}))((\underline{F}^{h_3})^T(\underline{F}^{h_2}))((\underline{F}^{h_1})^T(F^{\phi})) \\ +((\underline{F}^{h_3})^T(\underline{F}^{h_1}))((\underline{F}^{h_1})^T(\underline{F}^{h_2}))((\underline{F}^{h_2})^T(F^{\phi})) - ((\underline{F}^{h_1})^T(\underline{F}^{h_1}))((\underline{F}^{h_3})^T(\underline{F}^{h_2}))((\underline{F}^{h_2})^T(F^{\phi})) \\ -((\underline{F}^{h_2})^T(\underline{F}^{h_1}))((\underline{F}^{h_1})^T(\underline{F}^{h_2}))((\underline{F}^{h_3})^T(F^{\phi})) + ((\underline{F}^{h_1})^T(\underline{F}^{h_1}))((\underline{F}^{h_2})^T(\underline{F}^{h_2}))((\underline{F}^{h_3})^T(F^{\phi})) \end{bmatrix} \underline{F}^{h_3}$$

$$-\begin{bmatrix} ((\underline{F}^{h_1})^T(\underline{F}^{h_1}))((\underline{F}^{h_2})^T(\underline{F}^{h_2}))((\underline{F}^{h_3})^T(\underline{F}^{h_3})) - ((\underline{F}^{h_1})^T(\underline{F}^{h_1}))((\underline{F}^{h_2})^T(\underline{F}^{h_3}))((\underline{F}^{h_3})^T(\underline{F}^{h_2})) \\ -((\underline{F}^{h_1})^T(\underline{F}^{h_2}))((\underline{F}^{h_2})^T(\underline{F}^{h_1}))((\underline{F}^{h_3})^T(\underline{F}^{h_3})) + ((\underline{F}^{h_1})^T(\underline{F}^{h_2}))((\underline{F}^{h_2})^T(\underline{F}^{h_3}))((\underline{F}^{h_3})^T(\underline{F}^{h_1})) \\ +((\underline{F}^{h_1})^T(\underline{F}^{h_3}))((\underline{F}^{h_2})^T(\underline{F}^{h_1}))((\underline{F}^{h_3})^T(\underline{F}^{h_2})) - ((\underline{F}^{h_1})^T(\underline{F}^{h_3}))((\underline{F}^{h_2})^T(\underline{F}^{h_2}))((\underline{F}^{h_3})^T(\underline{F}^{h_1})) \end{bmatrix} \underline{F}^{\phi}$$

where

$$\begin{aligned} k_i = &(\alpha_2\alpha_3\alpha_4 - \alpha_8^2\alpha_4 - \alpha_9\alpha_3\alpha_5 + \alpha_7\alpha_8\alpha_5 + \alpha_9\alpha_8\alpha_6 - \alpha_7\alpha_2\alpha_6)F_i^{h_1} \\ &+(\alpha_7\alpha_8\alpha_4 - \alpha_9\alpha_3\alpha_4 - \alpha_7^2\alpha_5 + \alpha_1\alpha_3\alpha_5 + \alpha_7\alpha_9\alpha_6 - \alpha_1\alpha_8\alpha_6)F_i^{h_2} \\ &+(-\alpha_7\alpha_2\alpha_4 + \alpha_9\alpha_8\alpha_4 + \alpha_7\alpha_9\alpha_5 - \alpha_1\alpha_8\alpha_5 - \alpha_9^2\alpha_6 + \alpha_1\alpha_2\alpha_6)F_i^{h_3} \\ &-(\alpha_1\alpha_2\alpha_3 - \alpha_1\alpha_8^2 - \alpha_9^2\alpha_3 + 2\alpha_9\alpha_8\alpha_7 - \alpha_7^2\alpha_2)F_i^{\phi} \end{aligned}$$

$$\begin{aligned} \alpha_1 &= (F^{h_1})^T(F^{h_1}),\ \alpha_2 = (F^{h_2})^T(F^{h_2}),\ \alpha_3 = (F^{h_3})^T(F^{h_3}), \\ \alpha_4 &= (F^{h_1})^T(F^{\phi}),\ \ \alpha_5 = (F^{h_2})^T(F^{\phi}),\ \ \alpha_6 = (F^{h_3})^T(F^{\phi}), \\ \alpha_7 &= (F^{h_1})^T(F^{h_3}),\ \ \alpha_8 = (F^{h_2})^T(F^{h_3}),\ \ \alpha_9 = (F^{h_1})^T(F^{h_2}), \\ \alpha_{10} &= (F^{\phi})^T(F^{\phi}). \end{aligned}$$

Now, we can write

$$\begin{aligned} \underline{k}^T\underline{k} &= \left[\beta_1F^{h_1} + \beta_2F^{h_2} + \beta_3F^{h_3} - \beta_4F^{\phi}\right]^T\left[\beta_1F^{h_1} + \beta_2F^{h_2} + \beta_3F^{h_3} - \beta_4F^{\phi}\right] \\ &= \beta_1^2\alpha_1 + \beta_2^2\alpha_2 + \beta_3^2\alpha_3 + \beta_4^2\alpha_{10} + 2\beta_1\beta_2\alpha_9 + 2\beta_1\beta_3\alpha_7 - 2\beta_1\beta_4\alpha_4 + 2\beta_2\beta_3\alpha_8 \\ &\quad -2\beta_2\beta_4\alpha_5 - 2\beta_3\beta_4\alpha_6, \end{aligned}$$

where

$$\begin{aligned} \beta_1 &= \alpha_2\alpha_3\alpha_4 - \alpha_8^2\alpha_4 - \alpha_9\alpha_3\alpha_5 + \alpha_7\alpha_8\alpha_5 + \alpha_9\alpha_8\alpha_6 - \alpha_7\alpha_2\alpha_6 \\ \beta_2 &= \alpha_7\alpha_8\alpha_4 - \alpha_9\alpha_3\alpha_4 - \alpha_7^2\alpha_5 + \alpha_1\alpha_3\alpha_5 + \alpha_7\alpha_9\alpha_6 - \alpha_1\alpha_8\alpha_6 \\ \beta_3 &= -\alpha_7\alpha_2\alpha_4 + \alpha_9\alpha_8\alpha_4 + \alpha_7\alpha_9\alpha_5 - \alpha_1\alpha_8\alpha_5 - \alpha_9^2\alpha_6 + \alpha_1\alpha_2\alpha_6 \\ \beta_4 &= \alpha_1\alpha_2\alpha_3 - \alpha_1\alpha_8^2 - \alpha_9^2\alpha_3 + 2\alpha_9\alpha_8\alpha_7 - \alpha_7^2\alpha_2. \end{aligned}$$

Here $\underline{k}^T\underline{k}$ must be zero at the optimum. Since, in general, $\underline{k}^T\underline{k} \geq 0$, p^* should minimize $\underline{k}^T\underline{k}$ or maximize $K(\underline{p}) = \left[-\underline{k}^T\underline{k}\right]$ with a maximum value of zero. Hence we have transformed our problem to that of an optimization problem in respect of $\underline{p}$, namely maximization of $K(\underline{p})$ (or any increasing function of $K(\underline{p})$).

At this point we also need to ensure that the constraints satisfy the first order conditions with respect to the Lagrange multipliers. These can be done by simultaneously maximizing $H_1(\underline{p}) = -[h_1(\underline{p}) - c_1]^2$, $H_2(\underline{p}) = -[h_2(\underline{p}) - c_2]^2$ and $H_3(\underline{p}) = -[h_3(\underline{p}) - c_3]^2$. In our case, we have $h_1(\underline{p}) = 0$, $h_2(\underline{p}) = 0$ and $h_3(\underline{p}) = 0$. So the corresponding $H_1(\underline{p})$, $H_2(\underline{p})$ and $H_3(\underline{p})$ for this problem will be

$$H_1(\underline{p}) = -[h_1(\underline{p})]^2$$

$$H_2(\underline{p}) = -[h_2(\underline{p})]^2$$

$$H_3(\underline{p}) = -[h_3(\underline{p})]^2.$$

Therefore, the maximum value of the $H_1(\underline{p})$, $H_2(\underline{p})$ and $H_3(\underline{p})$ are zero at the optimum.

Thus the optimal p^* should simultaneously maximize the four functions K, H_1, H_2 and H_3, with a common maximum of zero. Since it is usually not possible to evaluate an explicit solution directly, we need an algorithm to find out the optimal solution. In the following section, we will discuss an appropriate algorithm for finding the optimal solution.

4.4 Proposed Algorithms

In the previous section, we have transformed the constrained optimization problem to that of maximization problems, i.e., maximizing the functions K, H_1, H_2 and H_3 of the cell probabilities simultaneously. As each of these functions is negative and has a common maximum of zero, we can maximize the sum of the functions $K+H_1+H_2+H_3$ that has a common maximum of zero at the optimum p^*.

On the other hand, we can consider a *maximin* problem in which we can maximize the minimum of K, H_1, H_2 and H_3 i.e., min(K, H_1, H_2, H_3) which will be zero at the optimum p^*. So, for this problem, we would suggest an appropriate extension of the algorithm outlined in Chapter 2. The full form of the algorithm at the $(r+1)^{th}$ step is

$$p_j^{(r+1)} = \frac{p_j^{(r)} f(x_j^{(r)}, \delta)}{\sum_{j=1}^{J} p_j^{(r)} f(x_j^{(r)}, \delta)}, \tag{4.8}$$

where $f(x, \delta)$ is a positive and strictly increasing function in x and may depend on a positive parameter δ. The choices of the function $f(x, \delta)$ and its first argument x depend on what type of optimization problem is used. For the above two optimization problems, the choices of x are

$$x_j = \begin{cases} F_j^K + F_j^{H_1} + F_j^{H_2} + F_j^{H_3} & \text{if we maximize } K + H_1 + H_2 + H_3 \\ F_j^{min(K,H_1,H_2,H_3)} & \text{if we maximize } min\{K, H_1, H_2, H_3\} \end{cases}$$

where F_j^K, $F_j^{H_1}$, $F_j^{H_2}$, $F_J^{H_3}$ are the directional derivatives of K, H_1, H_2, H_3 respectively, which are given by

$$F_j^K = d_j^K - \sum_{j=1}^{J} p_j d_j^K, F_j^{H_1} = d_j^{H_1} - \sum_{j=1}^{J} p_j d_j^{H_1}, \quad F_j^{H_2} = d_j^{H_2} - \sum_{j=1}^{J} p_j d_j^{H_2}, \quad F_j^{H_3} = d_j^{H_3} - \sum_{j=1}^{J} p_j d_j^{H_3},$$

where d_j^K, $d_j^{H_1}$, $d_j^{H_2}$, $d_j^{H_3}$ are first partial derivatives for the functions K, H_1, H_2 and H_3 respectively.

The partial derivatives of the function K depend on the choice of K. Depending on the magnitude of the partial and directional derivatives we may need to consider some transformations of the function K so that the maximum of the transformed function is still zero. For the original expression of K for the 4×4 case ($K = -\underline{k}^T \underline{k}$), d_j^K is

$$\begin{aligned} d_j^K = {} & \beta_1^2 \frac{\partial \alpha_1}{\partial p_j} + \beta_2^2 \frac{\partial \alpha_2}{\partial p_j} + \beta_3^2 \frac{\partial \alpha_3}{\partial p_j} + \beta_4^2 \frac{\alpha_{10}}{\partial p_j} + 2\beta_1\beta_2 \frac{\partial \alpha_9}{\partial p_j} + 2\beta_1\beta_3 \frac{\partial \alpha_7}{\partial p_j} - 2\beta_1\beta_4 \frac{\partial \alpha_4}{\partial p_j} \\ & + 2\beta_2\beta_3 \frac{\partial \alpha_8}{\partial p_j} - 2\beta_2\beta_4 \frac{\partial \alpha_5}{\partial p_j} - 2\beta_3\beta_4 \frac{\partial \alpha_6}{\partial p_j} + 2(\alpha_1\beta_1 + \alpha_9\beta_2 + \alpha_7\beta_3 - \alpha_4\beta_4) \frac{\partial \beta_1}{\partial p_j} \\ & + 2(\alpha_2\beta_2 + \alpha_9\beta_1 + \alpha_8\beta_3 - \alpha_5\beta_4) \frac{\partial \beta_2}{\partial p_j} + 2(\alpha_3\beta_3 + \alpha_7\beta_1 + \alpha_8\beta_2 - \alpha_6\beta_4) \frac{\partial \beta_3}{\partial p_j} \\ & + 2(\alpha_{10}\beta_4 - \alpha_6\beta_3 - \alpha_4\beta_1 - \alpha_5\beta_2) \frac{\partial \beta_4}{\partial p_j}, \end{aligned}$$

where the expressions $\frac{\partial \alpha_i}{\partial p_j}$ and $\frac{\partial \beta_i}{\partial p_j}$ contain $\frac{\partial F_i^{\phi}}{\partial p_j}$, $\frac{\partial F_i^{h_1}}{\partial p_j}$, $\frac{\partial F_i^{h_2}}{\partial p_j}$, $\frac{\partial F_i^{h_3}}{\partial p_j}$ and the 1st and 2nd-order partial derivatives of ϕ, h_1, h_2 and h_3.

We approach this problem by considering the choice of x as $x_j = F_j^K + F_j^{H_1} + F_j^{H_2} + F_j^{H_3}$. Convergence rates of the algorithm depends on the choice of the function $f(.)$ and

the positive parameter δ. We have endeavoured to further improve the convergence rate of the algorithm by considering the suitable transformation of K. A natural choice of $f(x,\delta)$ with the potential to satisfy the requirements is $f(x,\delta) = exp(\delta x)/(1+exp(\delta x))$, i.e. the logistic c.d.f. evaluated at δx. Sometime the magnitude of the partial and directional derivatives of the function K could be very large and it is difficult to attain the first order conditions. In such a situation, Torsney and Mandal (2001) considered a suitable transformation for an optimal design problem. We use this transformation for the function K. This is given by

$$K_1 = -K = \begin{cases} (-K)^t & \text{for } -K < 1 \\ (-K)^{1/t} & \text{for } -K > 1 \end{cases} \tag{4.9}$$

for some t.

4.5 Applications

We now apply the above methodologies to the following problems.

4.5.1 Unaided Distance Vision Data

Here we consider a real data set given in Placket (1981) for a square 4×4 contingency table for which the hypothesis of marginal homogeneity is of interest. The data set is given for a grading of the unaided distance vision of each eye of 3242 men employees in Royal Ordnance factories (Table 7.3, page 86). Here we have two responses that are defined by vision in the right and left eyes, and the responses are ordered in four categories, namely, highest, second highest, third highest

and the lowest. These same categories were used for each eye. A hypothesis of possible interest is that the margins for the categories are homogeneous. Based on our formulation of Section 4.3, we have the following observed frequencies $(O_{12}, O_{13}, O_{14}, O_{21}, O_{23}, O_{24}, O_{31}, O_{32}, O_{34}, O_{41}, O_{42}, O_{43}) = (112, 85, 35, 116, 145, 27, 72, 151, 87, 43, 34, 106)$ with $b = 1013$. We start with the initial design using the observed frequencies, and consider the choice of $f(x, \delta)$ as the logistic c.d.f. with x as the sum of the directional derivatives and use the transformation K_1 with $t = 4$ as given in (4.9) and $\delta = 0.001$. After 98006 iterations we obtain the following results

$$(H_1, H_2, H_3, K_1) = (-1.6877 \times 10^{-13}, -1.6578 \times 10^{-13}, -1.1759 \times 10^{-13}, -3.4145 \times 10^{-05})$$

$$-8.216422 \times 10^{-07} \leq F_j^{H_1} \leq 8.216429 \times 10^{-07}$$

$$-8.143123 \times 10^{-07} \leq F_j^{H_2} \leq 8.14313 \times 10^{-07}$$

$$-6.858219 \times 10^{-07} \leq F_j^{H_3} \leq 6.858224 \times 10^{-07}$$

$$-3.048315 \times 10^{-08} \leq F_j^{K_1} \leq 7.006246 \times 10^{-08}.$$

From the above results we see that the directional derivatives closely satisfy the first order conditions. We obtain the optimal solution p^* of the cell probabilities as given by (0.10860845, 0.08203665, 0.03779397, 0.11660967, 0.14245095, 0.02974073, 0.07273632, 0.14978607, 0.09635066, 0.03909350, 0.03040724, 0.09438579).

We also explore another data set given in Placket (1981) for women employees. The data set is given for a grading of the unaided distance vision of each eye of 7477 women employees in Royal Ordnance factories. Here we have the following observed frequencies $(O_{12}, O_{13}, O_{14}, O_{21}, O_{23}, O_{24}, O_{31}, O_{32}, O_{34}, O_{41}, O_{42}, O_{43})$ = (266,124, 66, 234, 432,78, 117, 362, 205, 36, 82, 179) and $b = 2181$. We start with the initial design

using the observed frequencies, and consider the choice of $f(x, \delta)$ as the logistic c.d.f. with x as the sum of the directional derivatives and use the transformation K_1 with $t = 4$ as given in (4.9) and $\delta = 0.0001$. After 250000 iterations we obtain the following results

$$(H_1, H_2, H_3, K_1) = (-1.3068 \times 10^{-23}, -1.0251 \times 10^{-23}, -8.9064 \times 10^{-24}, -2.7544 \times 10^{-06})$$

$$-7.229953 \times 10^{-12} \leq F_j^{H_1} \leq 7.229953 \times 10^{-12}$$

$$-6.403322 \times 10^{-12} \leq F_j^{H_2} \leq 6.403322 \times 10^{-12}$$

$$-5.968726 \times 10^{-12} \leq F_j^{H_3} \leq 5.968726 \times 10^{-12}$$

$$-3.146292 \times 10^{-11} \leq F_j^{K_1} \leq 1.487525 \times 10^{-10}.$$

Here also we see that the directional derivatives closely satisfy the first order conditions. The design converged to the optimal solution p^* of the cell probabilities is (0.11576438, 0.05128057, 0.02611916, 0.11335953, 0.18772011, 0.03236367, 0.06018734, 0.17566839, 0.08952702, 0.01961724, 0.04201054, 0.08638206).

4.5.2 Migration Data

Agresti (2002) provides a migration data of U.S. residents from 1980 to 1985 in four regions, namely, Northeast, Midwest, South and West for a total 55,981 residents (Table 10.6, page 423). According to the data, relatively few people changed region, that is, 95% of the observations falling on the main diagonals. A hypothesis of possible interest is that the margins for the four regions (categories) are homogeneous. Based on our formulation of Section 4.3, we have the following observed frequencies $(O_{12}, O_{13}, O_{14}, O_{21}, O_{23}, O_{24}, O_{31}, O_{32}, O_{34}, O_{41}, O_{42}, O_{43})$ = (100, 366, 124, 87, 515, 302, 172, 255, 270, 63, 176, 286) and $b = 2716$. We apply our methodologies to this

data set and run the multiplicative algorithm (4.8). We start with the initial design using the observed frequencies, and consider the choice of $f(x, \delta)$ as the logistic c.d.f. with x as the sum of the directional derivatives and $\delta = 0.000001$ and the transformation K_1 with $t = 4$ as given in (4.9). With this set up, after 1540000 iterations, we obtain the following results

$$(H_1, H_2, H_3, K_1) = (-2.1856 \times 10^{-24}, -4.7312 \times 10^{-25}, -1.1990 \times 10^{-24}, -1.6766 \times 10^{-06})$$

$$-2.956732 \times 10^{-12} \leq F_j^{H_1} \leq 2.956732 \times 10^{-12}$$

$$-1.375677 \times 10^{-12} \leq F_j^{H_2} \leq 1.375677 \times 10^{-12}$$

$$-2.189998 \times 10^{-12} \leq F_j^{H_3} \leq 2.189998 \times 10^{-12}$$

$$-1.181109 \times 10^{-11} \leq F_j^{K_1} \leq 1.368265 \times 10^{-11}.$$

We see that all the directional derivatives satisfy the first order conditions of optimality. The solution converged to the optimal p^* of the cell probabilities, which is given by (0.03521801, 0.09848969, 0.03445615, 0.03355777, 0.14334748, 0.08689845, 0.10024039, 0.13863741, 0.10390003, 0.03436569, 0.08994827, 0.10094067).

4.6 Constrained Optimization Problem for the 3×3 Case

We now consider the case for $n = 3$, i.e., a 3×3 contingency table. For convenience, we use the following notations for the observed and expected frequencies. Let

$$(y_1, y_2, y_3, y_4, y_5, y_6) = (O_{12}, O_{13}, O_{21}, O_{23}, O_{31}, O_{32}) \tag{4.10}$$

and

$$(x_1, x_2, x_3, x_4, x_5, x_6) = (E_{12}, E_{13}, E_{21}, E_{23}, E_{31}, E_{32}) \tag{4.11}$$

where $E_{ij} = N\theta_{ij}$, $i = 1, 2, 3$, $j = 1, 2, 3$. Similar to the 4×4 case, we write the maximum likelihood problem subject to the constraints of marginal homogeneity as given in the following:

Maximize $\phi(x) = \sum_{t=1}^{6} y_t ln(x_t)$ subject to

$$x_t \geq 0, t = 1, 2, \ldots, 6$$

$$\sum_{t=1}^{6} x_t = b = (N - \sum_{i=1}^{3} O_{ii})$$

$$x_1 - x_2 - x_4 + x_5 = 0$$

$$-x_1 + x_3 + x_4 - x_6 = 0.$$

As we mentioned before, one of the marginal homogeneity constraints, for example, that corresponding to $r = 3$, can be removed since they are linearly dependent.

Let us define $p_t = x_t/b$. So we can now write the above maximum likelihood problem in this form

$$\phi = \phi(\underline{p}) = \sum_t y_t ln(p_t) + \sum_t y_t ln(b).$$

Note that $\sum_t p_t = 1$ as $\sum x_t = b$.

So our problem now is to maximize $\phi\left(\underline{p}\right) = \sum_t y_t ln(p_t)$ subject to $p_t \geq 0, t = 1, 2, \ldots, 6, \sum_{t=1}^{6} p_t = 1$, and

$$h_1\left(\underline{p}\right) = p_1 - p_2 - p_4 + p_5 = 0$$

$$h_2\left(\underline{p}\right) = -p_1 + p_3 + p_4 - p_6 = 0$$

i.e., $C\underline{p} = \underline{a}$, where

$$C = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 0 & -1 & 1 & 0 \\ -1 & 0 & 1 & 1 & 0 & -1 \end{pmatrix}, \quad \underline{a} = (1, 0, 0)^T, \quad \underline{p} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \\ p_6 \end{pmatrix}.$$

The Lagrangian function is given by

$$L(\phi, \underline{p}, \underline{\lambda}, \mu) = \phi\left(\underline{p}\right) + \sum_{i=1}^{2} \lambda_i \left(h_i(\underline{p}) - c_i\right) + \mu(\sum_{t=1}^{6} p_t - 1),$$

where $h_1(\underline{p}) = c_1$, and $h_2(\underline{p}) = c_2$ with $p_t \geq 0$ and $\sum p_t = 1$, for constants c_1 and c_2 equal to zero.

Now considering the vertex directional derivatives of L and substituting the values of μ we can obtain the vertex directional derivatives of L in this form

$$F_t^L = F_t^{\phi} + \lambda_1 F_t^{h_1} + \lambda_2 F_t^{h_2},$$

where $F_t^{\phi} = d_t^{\phi} - \sum p_t d_t^{\phi}$, $F_t^{h_1} = d_t^{h_1} - \sum p_t d_t^{h_1}$ and $F_t^{h_2} = d_t^{h_2} - \sum p_t d_t^{h_2}$ are the directional derivatives of ϕ, h_1 and h_2 respectively.

The directional derivatives of L must be zero at the optimum for $p_t > 0$. That is, in the vector notation, we have $\underline{F}^L = \underline{0}$, that is, $F^h \underline{\lambda} = -\underline{F}^\phi$, where $F^h = \left[\underline{F}^{h_1}, \underline{F}^{h_2}\right]$ and $\underline{\lambda} = (\lambda_1, \lambda_2)$. If the above system of equations is consistent, the set of solutions to the system is $\underline{\lambda} = (F^h)^-(-\underline{F}^\phi) + (I - (F^h)^-(F^h))\underline{z}$ for any $\underline{z}$ and $(F^h)^-$ is any g-inverse of F^h. If $(F^h)^T(F^h)$ is nonsingular, one choice of $(F^h)^-$ is

$$(F^h)^- \quad = \quad \left[(F^h)^T(F^h)\right]^{-1}(F^h)^T.$$

Then

$$\underline{\hat{\lambda}} \quad = \quad (F^h)^-(-\underline{F}^\phi)$$

and p^* should be such that it must satisfy the following

$$F^h \underline{\hat{\lambda}} = -\underline{F}^\phi.$$

Now substituting the value of $\hat{\lambda}$ in the above equation, we get

$$\begin{aligned}
\underline{k} = \quad & \left[\; ((\underline{F}^{h_2})^T(\underline{F}^{h_2}))((\underline{F}^{h_1})^T(F^\phi)) - ((\underline{F}^{h_1})^T(\underline{F}^{h_2}))((\underline{F}^{h_2})^T(F^\phi)) \; \right]\underline{F}^{h_1} \\
& + \left[\; ((\underline{F}^{h_1})^T(\underline{F}^{h_1}))((\underline{F}^{h_2})^T(F^\phi)) - ((\underline{F}^{h_1})^T(\underline{F}^{h_2}))((\underline{F}^{h_1})^T(F^\phi)) \; \right]\underline{F}^{h_2} \\
& - \left[\; ((\underline{F}^{h_1})^T(\underline{F}^{h_1}))((\underline{F}^{h_2})^T(\underline{F}^{h_2})) - ((\underline{F}^{h_1})^T(\underline{F}^{h_2}))^2 \; \right]\underline{F}^\phi,
\end{aligned}$$

where

$$k_i = \; (\alpha_2\alpha_4 - \alpha_3\alpha_5)F_i^{h_1} + (\alpha_1\alpha_5 - \alpha_3\alpha_4)F_i^{h_2} - (\alpha_1\alpha_2 - \alpha_3^2)F_i^\phi,$$

$$\begin{aligned}\alpha_1 &= (F^{h_1})^T(F^{h_1}),\ \alpha_2 = (F^{h_2})^T(F^{h_2}),\ \alpha_3 = (F^{h_1})^T(F^{h_2}),\\ \alpha_4 &= (F^{h_1})^T(F^{\phi}),\ \ \alpha_5 = (F^{h_2})^T(F^{\phi}),\ \ \alpha_6 = (F^{\phi})^T(F^{\phi}).\end{aligned}$$

Now, we can write

$$\begin{aligned}\underline{k}^T\underline{k} &= \left[\beta_1 F^{h_1} + \beta_2 F^{h_2} - \beta_3 F^{\phi}\right]^T \left[\beta_1 F^{h_1} + \beta_2 F^{h_2} - \beta_3 F^{\phi}\right]\\ &= \beta_1^2\alpha_1 + \beta_2^2\alpha_2 + \beta_3^2\alpha_6 + 2\beta_1\beta_2\alpha_3 - 2\beta_1\beta_3\alpha_4 - 2\beta_2\beta_3\alpha_5,\end{aligned}$$

where

$$\beta_1 = \alpha_2\alpha_4 - \alpha_3\alpha_5$$

$$\beta_2 = \alpha_1\alpha_5 - \alpha_3\alpha_4$$

$$\beta_3 = \alpha_1\alpha_2 - \alpha_3^2.$$

Here $\underline{k}^T\underline{k}$ must be zero at the optimum. Since, in general, $\underline{k}^T\underline{k} \geq 0$, p^* should minimize $\underline{k}^T\underline{k}$ or maximize $K(\underline{p}) = \left[-\underline{k}^T\underline{k}\right]$ with a maximum value of zero.

Hence we have transformed our problem to that of an optimization problem in respect of $\underline{p}$, namely maximization of $K(\underline{p})$ (or any increasing function of $K(\underline{p})$) .

We also need to ensure that the constraints satisfy the first order conditions with respect to the Lagrange multipliers. These can be done by simultaneously maximizing $H_1(\underline{p}) = -[h_1(\underline{p}) - c_1]^2$ and $H_2(\underline{p}) = -[h_2(\underline{p}) - c_2]^2$. In our case, we have $h_1(\underline{p}) = 0$ and $h_2(\underline{p}) = 0$. So the corresponding $H_1(\underline{p})$ and $H_2(\underline{p})$ for this problem will be

$$H_1(\underline{p}) = -[h_1(\underline{p})]^2$$

$$H_2(\underline{p}) = -[h_2(\underline{p})]^2.$$

Therefore, the maximum value of the $H_1(\underline{p})$ and $H_2(\underline{p})$ are zero at the optimum.

Thus the optimal p^* should simultaneously maximize the three functions K, H_1 and H_2, with a common maximum of zero. In the following section, we will discuss an appropriate algorithm for finding the optimal solution.

4.7 Proposed Algorithms

We have now transformed the constrained optimization problem to that of maximization problems, i.e., maximizing the functions K, H_1 and H_2 of the cell probabilities simultaneously. As each of these functions is negative and has a common maximum of zero, we can maximize the sum of the functions $K + H_1 + H_2$ that has a common maximum of zero at the optimum p^*.

Alternatively, we can consider a *maximin* problem in which we can maximize the minimum of K, H_1 and H_2 i.e., min(K,H_1,H_2) which will be zero at the optimum p^*. So, for this problem, we would suggest an appropriate extension of the algorithm given in (4.8). For the above two optimization problems, the choices of x are

$$x_j = \begin{cases} F_j^K + F_j^{H_1} + F_j^{H_2} & \text{if we maximize } K + H_1 + H_2 \\ F_j^{min(K,H_1,H_2)} & \text{if we maximize } min\{K, H_1, H_2\} \end{cases}$$

where F_j^K, $F_j^{H_1}$, $F_j^{H_2}$ are the directional derivaties of K, H_1, H_2 respectively, which are given by

$$F_j^K = d_j^K - \sum_{j=1}^{J} p_j d_j^K, F_j^{H_1} = d_j^{H_1} - \sum_{j=1}^{J} p_j d_j^{H_1}, \quad F_j^{H_2} = d_j^{H_2} - \sum_{j=1}^{J} p_j d_j^{H_2},$$

where d_j^K, $d_j^{H_1}$, $d_j^{H_2}$ are first partial derivatives for the functions K, H_1 and H_2 respectively.

Depending on the magnitude of the partial and directional derivatives of the function K we may need to consider some transformations of the function K so that the maximum of the transformed function is still zero. For the original expression of K for the 3×3 case ($K = -\underline{k}^T \underline{k}$), d_j^K is

$$\begin{aligned} d_j^K = {} & \beta_1^2 \frac{\partial \alpha_1}{\partial p_j} + \beta_2^2 \frac{\partial \alpha_2}{\partial p_j} + \beta_3^2 \frac{\partial \alpha_6}{\partial p_j} + 2\beta_1\beta_2 \frac{\partial \alpha_3}{\partial p_j} - 2\beta_1\beta_3 \frac{\partial \alpha_4}{\partial p_j} - 2\beta_2\beta_3 \frac{\partial \alpha_5}{\partial p_j} \\ & + 2(\alpha_1\beta_1 + \alpha_3\beta_2 - \alpha_4\beta_3) \frac{\partial \beta_1}{\partial p_j} + 2(\alpha_2\beta_2 + \alpha_3\beta_1 - \alpha_5\beta_3) \frac{\partial \beta_2}{\partial p_j} \\ & + 2(\alpha_6\beta_3 - \alpha_4\beta_1 - \alpha_5\beta_2) \frac{\partial \beta_3}{\partial p_j}, \end{aligned}$$

where the expressions $\frac{\partial \alpha_i}{\partial p_j}$ and $\frac{\partial \beta_i}{\partial p_j}$ contain $\frac{\partial F_i^{\phi}}{\partial p_j}$, $\frac{\partial F_i^{h_1}}{\partial p_j}$, $\frac{\partial F_i^{h_2}}{\partial p_j}$, and the 1st and 2nd-order partial derivatives of ϕ, h_1 and h_2.

As we mentioned in the previous case, we approach this problem by considering the choice of x as $x_j = F_j^K + F_j^{H_1} + F_j^{H_2}$. Convergence rates of the multiplicative algorithm depends on the choice of the functions $f(.)$ and the positive parameter δ. We attempt to further improve the convergence rate of the algorithm by considering the

suitable transformation of K. A natural choice of $f(x, \delta)$ with the potential to satisfy the requirements is $f(x, \delta) = exp(\delta x)/(1 + exp(\delta x))$, i.e. the logistic c.d.f. evaluated at δx. Sometime the magnitude of the partial and directional derivatives of the function K could be very large and it is difficult to attain the first order conditions. So we use the same transformation that is given in (4.9).

4.8 Applications

4.8.1 Unaided Distance Vision Data for 3×3 Case

Here we consider the same data set given in Placket (1981) but we merge the data into a 3×3 contingency table. Based on our formulation of Section 4.6, we have the following observed frequencies $(O_{12}, O_{13}, O_{21}, O_{23}, O_{31}, O_{32}) = (230, 62, 223, 87, 77, 106)$ with $b = 785$. We start with the initial design using the observed frequencies, and consider the choice of $f(x, \delta)$ as the logistic c.d.f. with x as the sum of the directional derivatives and the transformation K_1 given in (4.9). We also explore by choosing different values of t and initial δ values.

First we choose $t = 2$ and $\delta = 0.0005$ for the transformation K_1. We obtain the following results after 90000 iterations

$$(H_1, H_2, K_1) = (-3.188867 \times 10^{-12}, -2.829583 \times 10^{-12}, -4.265083 \times 10^{-10})$$

$$-3.571473 \times 10^{-06} \leq F_j^{H_1} \leq 3.571486 \times 10^{-06}$$

$$-3.364267 \times 10^{-06} \leq F_j^{H_2} \leq 3.364278 \times 10^{-06}$$

$$-1.633616 \times 10^{-07} \leq F_j^{K_1} \leq 2.514841 \times 10^{-07}.$$

From the above results we see that the directional derivatives closely satisfy the first order conditions.

Now we change the value of δ to 0.0008. Starting with the same setting as in the previous case, we obtain the following results after 50059 iterations

$$(H_1, H_2, K_1) = (-2.529459 \times 10^{-11}, -2.244391 \times 10^{-11}, -9.427831 \times 10^{-10})$$

$$-1.00587 \times 10^{-05} \leq F_j^{H_1} \leq 1.00588 \times 10^{-05}$$

$$-9.474956 \times 10^{-06} \leq F_j^{H_2} \leq 9.475045 \times 10^{-06}$$

$$-4.401903 \times 10^{-07} \leq F_j^{K_1} \leq 6.762648 \times 10^{-07}.$$

Here also we see that the directional derivatives closely satisfy the first order conditions.

Now changing the value of δ to 0.001 and starting with the same setting as above, after 50000 iterations we get the following results

$$(H_1, H_2, K_1) = (-3.939987 \times 10^{-13}, -3.496329 \times 10^{-13}, -1.986063 \times 10^{-10})$$

$$-1.255386 \times 10^{-06} \leq F_j^{H_1} \leq 1.255387 \times 10^{-06}$$

$$-1.182595 \times 10^{-06} \leq F_j^{H_2} \leq 1.182596 \times 10^{-06}$$

$$-6.285332 \times 10^{-08} \leq F_j^{K_1} \leq 9.71094 \times 10^{-08}.$$

We see that the directional derivatives closely satisfy the first order conditions.

We further change the value of δ to 0.002 and obtain the following values after 20036

iterations

$$(H_1, H_2, K_1) = (-2.499348 \times 10^{-11}, -2.217673 \times 10^{-11}, -9.384416 \times 10^{-10})$$

$$-9.998645 \times 10^{-06} \leq F_j^{H_1} \leq 9.998745 \times 10^{-06}$$

$$-9.418391 \times 10^{-06} \leq F_j^{H_2} \leq 9.41848 \times 10^{-06}$$

$$-4.376546 \times 10^{-07} \leq F_j^{K_1} \leq 6.723749 \times 10^{-07}.$$

As we can see the value of δ plays an important role in the convergence rate of the algorithm. We see that all the directional derivatives satisfy the first order conditions of optimality.

In all of the above cases, we obtain the optimal solution p^* of the cell probabilities as given by (0.28962010, 0.08732993, 0.28742507, 0.12414264, 0.08952995, 0.12195232).

4.8.2 Migration Data for 3×3 Case

We again consider the same data set given in Agresti (2002) but we merge the data into a 3×3 contingency table. We merged the two categories Midwest and West into West. That is, we considered the migration data set in three different regions namely, Northeast, West and South of 56011 peoples resulted in the following observed frequencies $(O_{12}, O_{13}, O_{21}, O_{23}, O_{31}, O_{32}) = (224, 366, 150, 801, 172, 525)$ and $b = 2238$.

We apply the methodologies discussed in the earlier section and run the multiplicative algorithm (4.8). We start with the initial design using the observed frequencies, and

consider the choice of $f(x, \delta)$ as the logistic c.d.f. with x as the sum of the directional derivatives and $\delta = 0.0001$ and the transformation K_1 with $t = 4$ as given in (4.9).

We obtain the following results after 140000 iterations

$$(H_1, H_2, K_1) = (-2.184725 \times 10^{-24}, -3.774823 \times 10^{-30}, -4.316032 \times 10^{-06})$$

$$-2.956163 \times 10^{-12} \leq F_j^{H_1} \leq 2.956163 \times 10^{-12}$$

$$-3.885781 \times 10^{-15} \leq F_j^{H_2} \leq 3.885781 \times 10^{-15}$$

$$-1.138924 \times 10^{-11} \leq F_j^{K_1} \leq 1.307303 \times 10^{-11}.$$

Here also we see that all the directional derivatives satisfy the first order conditions of optimality. The solution converged to the optimal p^* of the cell probabilities, which is given by (0.08526228, 0.11890648, 0.08113321, 0.29789578, 0.12303555, 0.29376671).

Chapter 5

Maximum Likelihood Estimation of Bradley-Terry Model for Paired Comparisons

5.1 Introduction

In this chapter, we apply our optimal design theory to solve another estimation problem. We consider the maximum likelihood estimation under Bradley-Terry model for paired comparisons. There are many problems in statistics in which categorical outcomes result from pairwise evaluations. The paired comparison method is a very old psychometric technique that has been used by many researchers in various fields. It is a well-developed method of ordering attributes or characteristics of a given set of items. See for example, Bradley and Terry (1952), Bradley and El-Helbawy (1976), Grasshoff and Schwabe (2008). Bradley and Terry (1952) proposed a logit model for paired evaluations. The Bradley-Terry model for paired comparisons has been broadly applied in many areas such as statistics, sports and machine learning. There has been some work in the literature in comparative experiment. Efron et al. (2001) used non-

parametric empirical Bayes analysis of microarrays in comparative experiment. Here our problem is in determining the maximum likelihood estimators of the parameters of the latent variable models such as the Bradley-Terry model where the data come from a paired comparisons experiment. We consider the parameters of these models in terms of a set of another parameters which we consider as weights in optimal design theory. These weights are positive and sum to one.

5.2 Formulation of the Proposed Problem

Suppose that we consider a paired comparison experiment in which J treatments, say, $T_1, T_2, \ldots, T_J$ are compared in pairs. In the simplest case, a subject or an individual is presented with two treatments and then asked to indicate which one she/he prefers or considers better. Let the number of comparisons of T_i to T_j be n_{ij}. Let π_{ij} be the probability that T_i is preferred to T_j ($i \neq j$; for i=1, 2,. . . , J; j=1, 2,. . . , J) in any single comparison of T_i and T_j, the same for all such pairwise comparisons, with the constraint

$$\pi_{ij} + \pi_{ji} = 1.$$

In this experiment, consider O_{ij} as the observed frequency that T_i is preferred to T_j and assume that there are no ties in the experiment so that, for $i < j$, $O_{ij} + O_{ji} = n_{ij}$. Also assume that there is no correlation between each pairwise comparison.

The general model is that the observed frequencies follow Binomial distribution with parameters n_{ij} and π_{ij}, i.e., $O_{ij} \sim Bin(n_{ij}, \pi_{ij})$ and the likelihood of the data is

given by

$$L_0(\pi) = \prod_{i<j} (\pi_{ij})^{o_{ij}} (\pi_{ji})^{o_{ji}}, \tag{5.1}$$

where π_{ij} is the probability that T_i is preferred to T_j which is of the form

$$\pi_{ij} \quad = \quad \frac{p_i}{p_i + p_j}, \; p_i > 0.$$

See, e.g., Bradley and Terry (1952), Bradley (1965), Wu et al. (2004), Torsney (2010), Torsney (2004) and Huang et al. (2006). Our focus is to estimate π_{ij} in terms of p_i and p_j. Here the fact is that we only have observations on comparisons between the treatments and π_{ij} is invariant to proportional changes in p_i and p_j. In consequence p_i's are only unique up to a constant multiple. As this relationship only defines the p_i's relative to each other, we can write

$$\pi_{ij} \quad = \quad \frac{c\, p_i}{c\, p_i + c\, p_j}. \tag{5.2}$$

A restriction must be imposed to find a particular set of p_i's corresponding to the maximum likelihood estimator of π_{ij}. A natural choice would be $\sum p_i = 1$. Finding the corresponding estimates of the p_i's requires solution of Problem (P1) of Chapter 2. Thus we can consider this maximum likelihood estimation problem as an application of our optimal design problem. Then we can consider p_i's as weights in our optimal design context.

Now, substituting for π_{ij}'s in (5.1), we wish to maximize the criterion function

$$\phi(p) = \frac{\prod_{i=1}^{J} p_i^{O_i}}{\prod_{i<j} (p_i + p_j)^{n_{ij}}}, O_i = \sum_{\substack{j=1 \\ i \neq j}}^{J} O_{ij} \tag{5.3}$$

over $\mathcal{P} \equiv \{p = (p_1, \ldots, p_j) : p_j \geq 0, \sum_{j=1}^{J} p_j = 1\}$. Naturally then this is an example of Problem (P1) of Chapter 2.

Thus we have transformed the maximum likelihood problem to a constrained optimization problem subject to the basic constraints on design weights. To determine the optimizing distribution, we consider a class of multiplicative algorithms, indexed by a function $f(.)$ which satisfies certain conditions (positive and strictly increasing). Thus we can apply the multiplicative algorithm that is mentioned in Chapter 2 and find the optimal solution.

5.3 Algorithms

In constructing optimal designs, explicit solutions are not possible, except in some simple cases. That is why we adopt numerical techniques such as multiplicative algorithm described in Section 2.6 to obtain optimal solution for the constraint optimization problem. We now formulate the multiplicative algorithm in our particular problem.

In order to use the algorithms, we need to find out the partial derivatives of the criterion function. The partial derivatives of the criterion function (5.3) are given by

$$d_j^{\phi} = \frac{\partial \phi}{\partial p_j} = \phi(p) \left\{ \frac{O_j}{p_j} - \sum_{s \neq j} \frac{n_{js}}{p_j + p_s} \right\}$$

and the corresponding directional derivatives of the above criterion function are given by

$$F_j^{\phi} = d_j^{\phi} - \sum_{i=1}^{J} p_i d_i^{\phi}.$$

Note that, in this likelihood estimation problem, our criterion function $\phi(p)$ is homogeneous function of degree zero, that is, $\phi(cp) = \phi(p)$ where c is a constant. So, in this case the partial derivatives are equal to the corresponding directional derivatives of the criterion function.

Since there are always positive and negative d_j, we require a function $f(d, \delta)$ which is defined for positive and negative d, where d represents a partial derivative. In order to apply the multiplicative algorithm (2.16), the function $f(d, \delta)$ should be positive and increasing. Keeping these in mind, we explore a suitable choice of the function in the following section, and apply the above maximum likelihood estimation problem using the Baseball and Tennis data sets of Agresti (2002), and the coffee data sets of Bradley and El-Helbawy (1976).

5.4 Applications and Results

There are several naturally arising examples of the above optimization problem. We can use Baseball and Tennis data sets in Agresti (2002) for the above problem. In the Baseball data set, there are 7 teams namely Milwaukee, Detroit, Toronto, New York, Boston, Cleveland and Baltimore. Each played with other team 13 times i.e, $n_{ij} = 13$. There are total of $N = 273$ observations; i.e. $\sum \sum O_{ij} = 273$. The observed frequencies are

$$
\begin{aligned}
O_{12}, O_{13}, O_{14}, O_{15}, O_{16}, O_{17} &= 7, 9, 7, 7, 9, 11 \\
O_{21}, O_{23}, O_{24}, O_{25}, O_{26}, O_{27} &= 6, 7, 5, 11, 9, 9 \\
O_{31}, O_{32}, O_{34}, O_{35}, O_{36}, O_{37} &= 4, 6, 7, 7, 8, 12 \\
O_{41}, O_{42}, O_{43}, O_{45}, O_{46}, O_{47} &= 4, 8, 6, 6, 7, 10 \\
O_{51}, O_{52}, O_{53}, O_{54}, O_{56}, O_{57} &= 6, 2, 6, 7, 7, 12 \\
O_{61}, O_{62}, O_{63}, O_{64}, O_{65}, O_{67} &= 4, 4, 5, 6, 6, 6 \\
O_{71}, O_{72}, O_{73}, O_{74}, O_{75}, O_{76} &= 2, 4, 1, 3, 1, 7
\end{aligned}
$$

with $J = 7$ as there were seven teams involved. We now apply our methodologies to this data set and run the algorithm (2.16). We start with taking $p_J^{(0)} = 1/J$. After 49 iterations with $f(d, \delta) = \Phi(\delta d)$ and $\delta = 0.004$, we obtain the optimal p^* as given by (0.18450185, 0.17343173, 0.16236162, 0.15129151, 0.14760148, 0.11439114, 0.06642066).

We apply the above optimization problem to another data set from Bradley and El-Helbawy (1976). This data-set nicely fits into the above context of multiple comparisons. In this data-set, there are 8 coffee types and 26 pairwise comparisons are made on each pair, i.e. $n_{ij} = 26$. The coffee types are the 8 combinations arising from a 2^3 factorial design. There are total of $N = 728$ observations; i.e. $\sum \sum O_{ij} = 728$. The

observed frequencies are given by

$$
\begin{aligned}
O_{12}, O_{13}, O_{14}, O_{15}, O_{16}, O_{17}, O_{18} &= 15, 15, 16, 19, 14, 19, 16 \\
O_{21}, O_{23}, O_{24}, O_{25}, O_{26}, O_{27}, O_{28} &= 11, 10, 15, 15, 14, 15, 12 \\
O_{31}, O_{32}, O_{34}, O_{35}, O_{36}, O_{37}, O_{38} &= 11, 16, 15, 15, 14, 18, 15 \\
O_{41}, O_{42}, O_{43}, O_{45}, O_{46}, O_{47}, O_{48} &= 10, 11, 11, 14, 11, 15, 13 \\
O_{51}, O_{52}, O_{53}, O_{54}, O_{56}, O_{57}, O_{58} &= 7, 11, 11, 12, 9, 14, 13 \\
O_{61}, O_{62}, O_{63}, O_{64}, O_{65}, O_{67}, O_{68} &= 12, 12, 12, 15, 17, 16, 18 \\
O_{71}, O_{72}, O_{73}, O_{74}, O_{75}, O_{76}, O_{78} &= 7, 11, 8, 11, 12, 10, 12 \\
O_{81}, O_{82}, O_{83}, O_{84}, O_{85}, O_{86}, O_{87} &= 10, 14, 11, 13, 13, 8, 14
\end{aligned}
$$

with $J = 8$ as 8 coffee types were compared.

We apply the methodologies discussed above in this data set and run the algorithm (2.16). We take the initial design to be $p_J^{(0)} = 1/J$. After only 22 iterations with $f(d, \delta) = \Phi(\delta d)$ and $\delta = 0.001$, we obtain the optimal p^* which is given by (0.15659341, 0.12637363, 0.14285714, 0.11675824, 0.10576923, 0.14010989, 0.09752747, 0.11401099).

We believe we have provided an easy and flexible methodology (using optimal design theory) to estimate the parameters of latent variable models such as the Bradley-Terry model for paired comparisons. As we have seen, the parameters of this model can be viewed as the optimal design weights. An extension of pairwise comparisons is to invite subjects to place three treatments in order of preference, i.e., to work in situations such as in models including triplets of treatments. Another possibility is to extend the model for ordinal comparisons in comparing the treatments and also to allow the model to consider the situation for ties.

Chapter 6

Optimal Structure (k) Designs for Comparing Test Treatments with a Control

6.1 Introduction

In this chapter, we introduce structure (k_1), structure (k_2) and structure (k_1k_2) properties of a factorial design. We establish properties of each of these structure designs in terms of the incidence and characteristic matrices of the designs. Furthermore, we develop methods of obtaining optimal R-type structure (k) designs and show that such designs are trace, A- and MV-optimal. The proposed methodologies are easy to follow and the construction of the designs comes out in a simple form.

Consider a factorial experiment with m factors such that ith factor has s_i levels for $i = 1, 2, \ldots, m$. Therefore the total number of treatment combinations in the experiment is $v = \prod_{i=1}^{m} s_i$. Let $N = (n_{ij})$ $(i = 1, 2, \ldots, v; j = 1, 2, \ldots, b)$ be the incidence matrix of a block design, where n_{ij}=1 or 0 if the ith treatment occurs in the jth block

or absent in the jth block respectively. The calculus for factorial arrangements has been applied to the analysis of several classes of experimental designs and has been addressed by several authors such as Mukerjee (1979), Mukerjee (1980), Kurkjian and Zelen (1962), Kurkjian and Zelen (1963), Zelen and Federer (1964), Zelen and Federer (1965), Paik and Federer (1973) and Cotter (1973). Kurkjian and Zelen (1963) applied the calculus for factorial arrangements to the analysis of block designs. They showed that the concurrence matrix $NN^{'}$ of the design with an incidence matrix $N_{(v\times b)}$ can be expressed as a linear combination of Kronecker products ($\otimes$) of I_i and E_i matrices, where I_i is an identity matrix of order s_i and E_i is a $s_i \times s_i$ matrix with each element unity. That is, $NN^{'}$ satisfies the property

$$NN^{'} = \sum_{s=0}^{m} \left\{ \sum_{\delta_1+\delta_2+\ldots+\delta_m=s} h(\delta_1, \delta_2, \ldots, \delta_m)(D_1^{\delta_1} \otimes D_2^{\delta_2} \otimes \ldots \otimes D_m^{\delta_m}) \right\}, \tag{6.1}$$

where $\delta_i = 0$ or 1, $h(\delta_1, \delta_2, \ldots, \delta_m)$ are constants depending on δ_i, and $D_i^{\delta_i}$ is a $s_i \times s_i$ matrix defined by

$$D_i^{\delta_i} = \begin{cases} I_i & \text{if } \delta_i = 0 \\ E_i & \text{if } \delta_i = 1. \end{cases}$$

A design satisfying (6.1) is called a property (A) design. The class of designs that have property (A) includes many designs that are used in practice such as randomized block designs, balanced incomplete block designs, group divisible designs and bulk of the Kronecker designs constructed by Vartak (1955) and those of Shah (1959) and Rao (1961). Let C and C^+ denote the characteristic matrix of the design and Moore-Penrose inverse of C respectively. Since C^+ is the Moore-Penrose inverse of C, it must satisfy

the conditions $CC^{+}C = C$ and $C^{+}CC^{+} = C^{+}$. Note that if NN' satisfies (6.1), i.e., if NN' has property (A), then C and C^{+} also have this property. This class of designs is particularly suitable for use in asymmetrical factorial experiments. Also, the analysis of the designs is simple and elegant even if there is no factorial structure underlying the treatment combinations.

Sia (1977) studied property (A) designs with respect to the A-optimality criterion. Zelen and Federer (1964) extended the idea of property (A) design to row-column designs. If the column incidence matrix $N_{(v\times b)}$ satisfies (6.1), it is still called a property (A) design. However, if the row incidence matrix $\tilde{N}_{(v\times b)}$ satisfies a similar property, then the design is called a property (B) design. Designs in which the row and column incidence matrices satisfy (6.1) are termed as property (AB) designs. Zelen and Federer (1964) derived the intra-block analysis for property (AB) designs. Paik and Federer (1973) showed that the property (A) design and property (B) design implies property (AB) design.

Mukerjee (1979) noted a major limitation of the previous work on factorial structure designs. The limitations are that the results are given in terms of a generalized inverse of a C-matrix. Results in terms of generalized inverse of a C-matrix are provided by Cotter (1973) and John and Smith (1972). In fact, Mukerjee (1979) introduced the notion of structure (k) design and determined a simple set of necessary and sufficient conditions for factorial structure which can be stated in terms of the C-matrix. On the basis of Mukerjee (1979), we define a property of a design and call it a structure (k) design.

A $v \times v$ matrix D, where $v = \prod_{i=1}^{m} s_i$, $s_i \geq 2$ for all i, is said to have structure (k),

if D can be expressed as a linear combination of Kronecker products of permutation matrices of order $s_1, s_2, \ldots, s_m$ (taken in that order), i.e., if D can be written as

$$D = \sum_{j=1}^{w} r_j (R_{j1} \otimes R_{j2} \otimes \ldots \otimes R_{jm}), \tag{6.2}$$

where w is some positive integer, $r_1, r_2, \ldots, r_w$ are some numbers, and for each j, R_{ji} is a $s_i \times s_i$ permutation matrix.

Clearly property (A) design is a special case of structure (k) design. The structure (k) property can be expressed in terms of $NN^{'}$ or a C-matrix.

In the following sections, we introduce structure (k_1), structure (k_2) and structure (k_1k_2) properties of a factorial design, and show that structure (k_1) design and structure (k_2) design implies structure (k_1k_2) design using the properties of the incidence and characteristic matrices of the designs. We also study the structure (k) designs with respect to the trace, A- and MV-optimality criteria. Starting from a structure (k) design and augmenting one control in each block, we develop methods of obtaining optimal R-type structure (k) designs and show that such designs are trace, A- and MV-optimal. The proposed methodologies are easy to follow. In addition, the construction of the designs comes out in a simple form. A summary of this work can be found in our paper Chowdhury et al. (2016).

6.2 Preliminaries

Definition 6.1. Proper Matrix: A square matrix where all row sums and column sums are equal is called a proper matrix.

Definition 6.2. Permutation Matrix: A square matrix with non-negative entries in which all row sums and column sums are equal to unity is called a permutation matrix.

Lemma 6.1. Any proper matrix can be expressed as a linear combination of permutation matrices of the same order.

Lemma 6.2. For any $v \times v$ permutation matrix R and for any x, $W^x R W^x$ has structure (k).

Lemma 6.3. A $v \times v$ matrix A, where $v = \prod_{i=1}^{m} s_i$, $s_i \geq 2$ for all i, has structure (k) if and only if A is expressible as a linear combination of Kronecker products of proper matrices of order $s_1, s_2, \ldots, s_m$ (taken in that order).

Lemma 6.4. For a connected block design, a necessary and sufficient condition for factorial structure is that column C-matrix has structure (k).

The above lemmas are due to Mukerjee (1979). Below we state a lemma which is due to Jacroux (1984).

Lemma 6.5. Let $d(v', b, k')$ be the semi-rectangular (SR) design obtained by reinforcing each block with a control treatment of Group Divisible (GD) design $\bar{d}$ having parameters $v = mn$, b, $r = bk/v$, $m = 2$, $n = v/2$ and $\lambda_2 = \lambda_1 + 1$, then the design d is trace optimal.

Theorem 6.1. There exists a hypercubic design (HCD), $\bar{d}$, having parameters $v = t^m$ $(t = 2)$, $r = \binom{m+n-1}{h+n-1}$, $n = k/t^h = 1$, $b = vr/k$, $k = t^h$ $(h = 1, 2, \ldots, m-1)$, $\lambda_i = \binom{m-i}{h-i}$, for $i \leq h$ and $\lambda_j = 0$ for $j > h$ if $\bar{d}$ satisfies the following conditions:

(i) the size of the ith group G_i $(i = 1, 2, \ldots, k)$ should be a multiple of block size k,

(ii) the block size (k) should be an even number,

(iii) the ith group G_i contains v/k treatment combinations,

(iv) each treatment combination occurs once and only once in each group G_i ($i = 1, 2, \ldots, k$).

Theorem 6.2. There exists a hypercubic design (HCD), $\bar{d}$, having parameters $v = t^m$ ($t = 3$), $r = 2^{m-k+2}$, $b = vr/k$, $k = n$ for $1 < n \leq m + 1$, and

$$\lambda_i = \begin{cases} 0 & \text{if } i \leq m - 1 \\ 1 & \text{otherwise} \end{cases}$$

if $\bar{d}$ satisfies the following conditions:

(i) the block size (k) is an integer, that is, $k = n$,

(ii) the size of the ith group G_i should be a multiple of t^n for $i = 1, 2, \ldots, t^{h-1}$, $h = 1, 2, \ldots, m - 1$,

(iii) the levels of the factor at ith position are not the same for all the factors while taking k treatment combinations from G_i different groups and keeping them in one block.

Theorems 6.1 and 6.2 are useful for the construction of SR-HCD designs.

Theorem 6.3. Let $d(v', b, k')$ be a SR-HCD design obtained by reinforcing each block with a control treatment of hypercubic design, $\bar{d}$, having parameters same as in Theorems 6.1 and 6.2. If d satisfies the conditions discussed in Lemma 6.5, then d is trace optimal.

Theorem 6.4. Let r_0 be the number of replication of the control treatment in d^*, where $d^* \in D_{r_0}(v + 1, b, k)$ is a Group Divisible Treatment Design (GDTD) having parameters $m = 2$, $n = v/2$, and $\lambda_2 = \lambda_1 + 1$. If $m_1(r_0) \leq m_2(r_0)$ and d^* is such that $trC^{-1}_{d^*11} \leq H_2(r_0)$, then d^* is A-optimal in $D_{r_0}(v + 1, b, k)$, where C_{d^*11} is the principal submatrix obtained

from C_d^* after deleting row 1 and column 1 and $H_2(r_0) = (1/m_1(r_0)) + ((v-1)/m_4(r_0))$ with $m_1(r_0) = b(k-1)/vk$, $m_4(r_0) = \{A - (2/k) - m_1(r_0)\}/(v-1)$, $A = b(k-1)^2/k$.

Theorem 6.5. For a given value of r_0, let $d^* \in D_{r_0}(v+1, b, k)$ be a Group Divisible Treatment Design (GDTD(s+1)) such that

$$r_{d^*0}k - \lambda_{d^*00} = r_0k - \lambda(r_0),\ \bar{\lambda}_0 = (r_0k - \lambda(r_0))/v,$$
$$r_{d^*i}k - \lambda_{d^*ii} = R(r_0)(k-1), \text{ for } i = 1, \ldots, v,$$
$$\bar{\lambda}_2 = \bar{\lambda}_1 + 1 \text{ where } \bar{\lambda}_1 = [(R(r_0)(k-1) - \bar{\lambda}_0)/(v-1)].$$

Also, for positive integers p and q, define $B(p,q) = (1 - ((1-\bar{m}p)(1-\bar{m}q))^{1/2})/\bar{m}$ and let $\bar{m}k = (k/v\bar{\lambda}_0) + ((\bar{v}-1)sk/v(\bar{v}(s-1)\bar{\lambda}_2 + v\bar{\lambda}_1 + \bar{\lambda}_0)) + ((s-1)k/v(v\bar{\lambda}_2 + \bar{\lambda}_0))$ where $\bar{v} = v/s$. Now, if

1. $r_0k - \lambda(r_0) - 2 < vB(r_0k - \lambda(r_0) - 2, R(r_0)(k-1))$

2. $(R(r_0)-1)(k-1)$ or $r_0k - \lambda(r_0)$ satisfies one of the appropriate inequalities $1/(R(r_0)-1)(k-1) > \bar{m}$ or $r_0k - \lambda(r_0) < (v-2)B(r_0k - \lambda(r_0), R(r_0)(k-1)) + B(r_0k - \lambda(r_0), (R(r_0)-1)(k-1)) + B(r_0k - \lambda(r_0), (R(r_0)+1)(k-1))$

3. $\bar{m}k < min\{(\bar{c}_{d00} + \bar{c}_{dii} + 2\bar{c}_{di0})/(\bar{c}_{d00}\bar{c}_{dii} - \bar{c}_{dio}^2), (\bar{c}_{d00} + \bar{c}_{djj} + 2\bar{c}_{dj0})/(\bar{c}_{d00}\bar{c}_{djj} - \bar{c}_{dj0}^2)\}$

4. $\bar{m}k < \{\bar{c}_{dpp}(\bar{c}_{d00} + \bar{c}_{dqq} + 2\bar{c}_{dpq}) - (\bar{c}_{dp0} + \bar{c}_{dpq})^2\}/\{\bar{c}_{d00}\bar{c}_{dpp}\bar{c}_{dqq} - \bar{c}_{d00}\bar{c}_{dpq}^2 - \bar{c}_{dpp}\bar{c}_{dq0}^2 - \bar{c}_{dqq}\bar{c}_{dp0}^2 + 2\bar{c}_{dp0}\bar{c}_{dq0}\bar{c}_{dpq}\}$

then d^* is MV-optimal in $D_{r_0}(v+1, b, k)$, where $N(r_0) = [\frac{r_0}{b}]$, $\lambda(r_0) = (r_0 - bN(r_0))(N(r_0)+1)^2 + (b - r_0 + bN(r_0))N^2(r_0)$, $R(r_0) = [(bk - r_0)/v)]$, r_{d^*i} is the i^{th} row sum of N_{d^*} (incidence matrix) which represent the number of times treatment i is replicated in

the design d^* and λ_{d^*ii} is the diagonal entries of i^{th} row and i^{th} column of the concurrence matrix $N_{d^*}N'_{d^*}$ for $i = 1, \ldots, v$, $\bar{c}_{d00} = (r_0k - \lambda(r_0))/k$, $\bar{c}_{dii} = R(r_0)(k-1)/k$, $\bar{c}_{dio} = -(\bar{\lambda}_0 - 1)/k$, $\bar{c}_{djj} = (R(r_0)(k-1)-2)/2$, $\bar{c}_{dj0} = -\bar{\lambda}_0/k$, $\bar{c}_{dpp} = \bar{c}_{dqq} = R(r_0)(k-1)/k$, $\bar{c}_{dp0} = \bar{c}_{dq0} = -\bar{\lambda}_0/k$, $-\bar{c}_{dpq} = (\bar{\lambda}_1 - 1)/k$ or $(\bar{\lambda}_1 + 2)/k$, and [.] denotes the greatest integer function.

For further details see Jacroux (1987a), Jacroux (1987b), Jacroux (1989). Theorems 6.1, 6.2, 6.3 are due to Thannipara (1992). Theorem 6.4 is due to Jacroux (1989). Theorem 6.5 is due to Jacroux (1987a).

6.3 Analysis of Structure (k_1), Structure (k_2) and Structure (k_1k_2) Designs for Two Way Elimination

Consider a block design with v treatments in b blocks such that each block contains k experimental units and each treatment is replicated r times. If we consider the design as an array with k rows and b columns where the entries in the array consist of the treatment numbers, the analysis of structure (k) designs follows from the work of Zelen and Federer (1964). Define the matrices $N = (n_{ij})$ and $\tilde{N} = (\tilde{n}_{ih})$ of dimensions $v \times b$ and $v \times k$ respectively, where n_{ij} = number of times treatment i occurs in block j and $\tilde{n}_{ih}$ = number of times treatment i occurs in row h. The matrix N is the incidence matrix for the design which relates the treatments to the (columns) blocks. We call N as the column incidence matrix and $\tilde{N}$ as the row incidence matrix. Using the matrices N and $\tilde{N}$, we can define column C-matrix and row C-matrix also. The column C-matrix is defined as

$$C = R - NK^{-1}N',$$

where, $R = diag(r_1, r_2, \ldots, r_v)$, and $K = diag(k_1, k_2, \ldots, k_b)$.

The row C-matrix is defined as

$$\tilde{C} = \tilde{R} - \tilde{N}K^{-1}\tilde{N}',$$

where, $\tilde{R} = diag(\tilde{r}_1, \tilde{r}_2, \ldots, \tilde{r}_v) = R$ and $\tilde{K} = diag(\tilde{k}_1, \tilde{k}_2, \ldots, \tilde{k}_v)$.

Let Y_{jh} $(j = 1, 2, \ldots, b; h = 1, 2, \ldots, k)$ denote the measurement made in the j^{th} block and h^{th} row. When treatment i is made in block j and row h, the random variable Y_{jh} is assumed to have expected value $E(Y_{jh}) = \mu + \tau_i + b_j + r_h$, where μ is a constant, and τ_i, b_j, and r_h are fixed effects associated with the treatments, blocks and rows respectively. These parameters satisfy the constraints $\sum_{i=1}^{v} \tau_i = \sum_{j=1}^{b} b_j = \sum_{h=1}^{k} r_h = 0$. We assume that Y_{jh}'s are uncorrelated with common variance σ^2.

When we analyze such design, our interest is usually focused on estimating the treatment effect τ_i. The estimates of the treatment effects can be obtained by solving a set of v simultaneous linear equations which depend on the incidence matrices N and $\tilde{N}$, and the adjusted treatment totals, which are functions of the observations. The adjusted treatment total for ith treatment is obtained by

$$Q_i = T_i - \sum_{j=1}^{b}(n_{ij}B_j)/k - \sum_{h=1}^{k}(\tilde{n}_{ih}R_h)/b + G/v, \tag{6.3}$$

where T_i = total for treatment i,

$B_j = \sum_{h=1}^{k} Y_{jh}$= total for jth block,

$R_h = \sum_{j=1}^{b} Y_{jh}$=total for hth row, and

$$G = \sum_{i=1}^{v} T_i = \sum_{j=1}^{b} B_j = \sum_{h=1}^{k} R_h = \sum_{j=1}^{b} \sum_{h=1}^{k} Y_{jh}.$$

The adjusted treatment totals Q_i's in (6.3) can be expressed in matrix notation as

$$Q = T - \frac{1}{k}(NB) - \frac{1}{b}(\tilde{N}R) + \frac{G}{v}\mathbf{1}, \tag{6.4}$$

where $T_{(v\times 1)}$, $B_{(b\times 1)}$, and $R_{(k\times 1)}$ are the column vectors of the treatment, block and row totals respectively, and $\mathbf{1}$ denotes a $v \times 1$ vector of 1's.

Now using Tocher (1952), the reduced normal equations for estimating the treatment effect vector $\tau' = (\tau_1, \tau_2, \ldots, \tau_v)$ can be written as

$$Q = \left[rI - \frac{1}{k}(NN') - \frac{1}{b}(\tilde{N}\tilde{N}') + \left(\frac{r}{v}\right)\mathbf{1}\mathbf{1}' \right]\hat{\tau}, \tag{6.5}$$

where I is an identity matrix of order v.

The estimate of the variance is

$$S^2 = [Y'Y - \hat{\tau}'Q - \frac{1}{b}(R'R) - \frac{1}{k}(B'B) + \frac{G^2}{vr}(\mathbf{1}'\mathbf{1})]/v_e,$$

where $v_e = (bk - b - v - k + 2)$, the degrees of freedom of S^2.

6.4 Structure (k_1), Structure (k_2) and Structure (k_1k_2) Properties and Factorial Structure

As mentioned earlier, a structural property of a design which is related to the block incidence matrix or column C-matrix of the design, is given by

$$NN^{'} = \sum_{j=1}^{w} \xi_j R_j, \tag{6.6}$$

where w is some positive integer, $\xi_1, \xi_2, \ldots, \xi_w$ are some numbers, $R_j = (R_{j1} \otimes R_{j2} \otimes \ldots \otimes R_{jm})$ and for each j, R_{ji} is a $s_i \times s_i$ permutation matrix. This structural property is termed as structure (k).

In the present work we call it a structure (k_1) property. We can define a similar property for the row incidence matrix $\tilde{N}$ and call it a structure (k_2) property which is given by

$$\tilde{N}\tilde{N}^{'} = \sum_{j=1}^{w} \tilde{\xi}_j R_j, \tag{6.7}$$

where $\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_w$ are some numbers.

When the structure properties (k_1) and (k_2) both hold, we have

$$\left[rI - \frac{1}{k}(NN^{'}) - \frac{1}{b}(\tilde{N}\tilde{N}^{'}) + \left(\frac{r}{v}\right) \mathbf{1}\mathbf{1}^{'} \right] = \sum_{j=1}^{w} \psi_j R_j \tag{6.8}$$

for some numbers $\psi_1, \psi_2, \ldots, \psi_w$.

We call this a structure (k_1k_2) property. Substituting (6.8) in (6.5), we can write the reduced normal equations as given by

$$\left(\sum_{j=1}^{w} \psi_j R_j\right)\hat{\tau} = Q, \tag{6.9}$$

that is,

$$D\hat{\tau} = Q, \tag{6.10}$$

where $D = \sum_{j=1}^{w} \psi_j R_j$.

On this basis we can define structure (k_1), structure (k_2) and structure (k_1k_2) properties as follows.

Definition 6.3. *Structure* (k_1) *property*: If a column C-matrix of connected block design satisfies the relation (6.6), then it is called a structure (k_1) property.

Definition 6.4. *Structure* (k_2) *property*: If a row C-matrix of connected block design satisfies the relation (6.7), then it is called a structure (k_2) property.

Definition 6.5. *Structure* (k_1k_2) *property*: If the column and row C-matrices satisfy the relation (6.8), then it is called a structure (k_1k_2) property.

From definitions 6.3 - 6.4, it is clear that structure (k_1) and structure (k_2) implies structure (k_1k_2) property.

From definitions 6.3 - 6.5, one sees that property (A), property (B) and property (AB) are special cases of structure (k_1), structure (k_2) and structure (k_1k_2) respectively.

Mukerjee (1979) has shown that a necessary and sufficient condition for factorial structure in connected block design is that C-matrix has structure (k) property. Interestingly, in this work, we see that a necessary and sufficient condition for factorial structure is that column C-matrix has structure (k_1) or row C-matrix has structure (k_2) or column and row C-matrices have structure (k_1k_2) property.

For a connected, equi-replicate and proper block design, factorial structure holds if and only if column incidence matrix has structure (k_1) or row incidence matrix has structure (k_2) or row and column incidence matrices have structure (k_1k_2).

6.5 An Example

Consider a hypercubic design (HCD) with parameters $v = 2^2$, $b = 4$, $r = 2$, $k = 2$, $\lambda_1 = 1$ and $\lambda_2 = 0$ whose blocks are

$$\begin{array}{cccc} 1 & 2 & 1 & 3 \\ 3 & 4 & 2 & 4 \end{array}$$

The structure of row-column designs are

$$NN' = \begin{pmatrix} N_1 & N_2 \\ N_2 & N_1 \end{pmatrix}$$

$$\tilde{N}\tilde{N}' = \begin{pmatrix} \tilde{N}_1 & \tilde{N}_2 \\ \tilde{N}_3 & \tilde{N}_4 \end{pmatrix}$$

$$C = \begin{pmatrix} C_1 & C_2 \\ C_2 & C_1 \end{pmatrix}$$

$$\tilde{C} = \begin{pmatrix} \tilde{C}_1 & \tilde{C}_2 \\ \tilde{C}_3 & \tilde{C}_4 \end{pmatrix}.$$

In this example,

$$NN' = \begin{pmatrix} 2 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \\ 1 & 0 & 2 & 1 \\ 0 & 1 & 1 & 2 \end{pmatrix}.$$

Now, we can write NN' as given by

$$\begin{aligned} NN' &= 1\left\{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right\} + 1\left\{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right\} \\ &\quad + 1\left\{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\right\} + 1\left\{\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right\} \\ &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \\ &= \sum_{j=1}^{4} \xi_j (R_{j1} \otimes R_{j2}), \end{aligned}$$

where $\xi_1 = 1, \xi_2 = 1, \xi_3 = 1$ and $\xi_4 = 1$ and $v = \prod_{i=1}^{2} s_i = 4$.

Also, the C-matrix can be written as

$$
\begin{aligned}
C &= \begin{pmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} & 0 \\ -\frac{1}{2} & 1 & 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 & 1 & -\frac{1}{2} \\ 0 & -\frac{1}{2} & -\frac{1}{2} & 1 \end{pmatrix} \\
&= \frac{1}{2}\left\{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right\} + \frac{1}{2}\left\{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right\} \\
&+ (-\frac{1}{2})\left\{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\right\} + (-\frac{1}{2})\left\{\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right\} \\
&= \sum_{j=1}^{4} \xi_j (R_{j1} \otimes R_{j2}),
\end{aligned}
$$

where $\xi_1 = \frac{1}{2}, \xi_2 = \frac{1}{2}, \xi_3 = -\frac{1}{2}$ and $\xi_4 = -\frac{1}{2}$.

Hence the above hypercubic design possesses structure (k_1) property. In a similar way, we can express $\tilde{N}\tilde{N}'$ and $\tilde{C}$ as linear combinations of Kronecker products of proper matrices of order s_1 and s_2. Now, it is clear that the hypercubic design possesses structure (k_2) and structure (k_1k_2) properties. Note that it also holds the relation that structure (k_1) and structure (k_2) implies structure (k_1k_2).

6.6 Optimal R-type Structure (k) Designs

We shall use $0, 1, \ldots, v$ to denote the $(v + 1)$ treatments being studied, with 0 representing the control treatment and $1, 2, \ldots, v$ representing the test treatments. In this section, we consider those designs that have equal block sizes for comparing several test treatments with a control. Assuming that homoscedasticity is satisfied, we study these designs with respect to the trace, A- and MV-optimality criteria.

Here we shall use $d(v^{'}, b, k^{'})$ to denote some particular block design that can be used in an experimental setting. The structure (k) design in the previous example is not optimal within the test treatments. However, if we augment one control in each block of such design, then we see that structure (k) design discussed in the example is optimal in the test treatment versus control treatment. It is interesting to see that augmented structure (k) design satisfies trace, A- and MV-optimality criteria. Here the augmented structure (k) design also satisfies the condition $r_0 = b$, that is, replication of the control treatment is equal to number of blocks of the design d. So we call it an optimal R-type structure (k) design. Thus, using the previous example, a R-type structure (k) design can be obtained as

$$\begin{array}{cccc} 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 3 \\ 3 & 4 & 2 & 4 \end{array}$$

In this example, we see that $v = 4$, $b = 4$, $k^{'} = 3$, $r_0 = 4$, $m = 2$, $n = 2$, $\lambda_0 = 2$, $\lambda_1 = 0$, $\lambda_2 = 1$, $trC_{d11}^{-1} = 3.5$,

$$m_1(4) = b(k-1)/vk = 0.666666,$$

$$m_2(4) = \{(A - m_1(4)) - [(v-1)/(v-2)]^{1/2} P_1\}/(v-1) = (4.666667 - 0.666667)/3 = 1.33334,$$

where

$$A = b(k-1)^2/k = 16/3 = 5.333333.$$

$$P_1 = [(B - (m_1(4))^2) - (A - m_1(4))^2/(v-1)]^{1/2} = [7.555556 - 7.259260]^{1/2} = 0.544331,$$

$$B = trC_{d11}^2 = 8,$$

and

$$H_2(4) = (1/m_1(4)) + ((v-1)/m_4(4)) = 1.500002 + 2.25001 = 3.750003,$$

where

$$m_4(4) = \{A - (2/k) - m_1(4)\}/(v-1) = \{5.333333 - 2/3 - 0.666666\}/3 = 1.333333.$$

Here $C_d = diag(r_{d0}, \ldots, r_{dv}) - \frac{1}{k} N_d N_d^{'}$ where $diag(r_{d0}, \ldots, r_{dv})$ denotes a $(v+1)\times(v+1)$ diagonal matrix and the i^{th} row sum of N_d is denoted by r_{di} which represents the number of times treatment i is replicated in the design. The matrix C_d is called the C-matrix of design d and is positive semi-definite with zero row sums. In the above example, the incidence matrix and C matrix of design d are

$$N_d = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

$$N_d N_d^{'} = \begin{pmatrix} 4 & 2 & 2 & 2 & 2 \\ 2 & 2 & 1 & 1 & 0 \\ 2 & 1 & 2 & 0 & 1 \\ 2 & 1 & 0 & 2 & 1 \\ 2 & 0 & 1 & 1 & 2 \end{pmatrix}$$

$$C_d = \begin{pmatrix} \frac{8}{3} & -\frac{2}{3} & -\frac{2}{3} & -\frac{2}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{3} & -\frac{1}{3} & -\frac{1}{3} & 0 \\ -\frac{2}{3} & -\frac{1}{3} & \frac{4}{3} & 0 & -\frac{1}{3} \\ -\frac{2}{3} & -\frac{1}{3} & 0 & \frac{4}{3} & -\frac{1}{3} \\ -\frac{2}{3} & 0 & -\frac{1}{3} & -\frac{1}{3} & \frac{4}{3} \end{pmatrix}.$$

Now we need to verify that the above structure (k) design satisfies trace, A- and MV-optimality criteria. Here we see that control treatment is replicated once in each block. So this is an SR(1) design and also it is obvious from the above example that in the original design all blocks have the same size, all the test treatments are replicated same number of times in blocks and $v = mn = 2 \times 2 = 4$ treatments are divided into 2 disjoint sets of size 2 such that the treatment in the same group occur in $\lambda_1 = 0$ blocks together whereas treatment occurring in different groups occur in $\lambda_2 = 1$ blocks together. So it is a Group Divisible Design. Clearly, the design d satisfies all the conditions of Lemma 6.5 and hence the design d is trace optimal.

We will now show that the design d constructed in the above example is A-optimal.

From the above structure (k) design, we observe that

$$m_1(4) = 0.666666 < m_2(4) = 1.33334$$

$$trC_{d11}^{-1} = 3.5 < H_2(4) = 3.750003$$

where C_{d11} is the principal submatrix obtained from C_d after deleting row 1 and column 1. So the design d satisfies all the conditions of Theorem 6.4 and hence the above structure (k) design is A-optimal.

We will now show that the design d is MV-optimal. We first calculate the following:

$$N(r_0) = [\tfrac{r_0}{b}] = 1$$

$$\lambda(r_0) = (r_0 - bN(r_0))(N(r_0) + 1)^2 + (b - r_0 + bN(r_0))N^2(r_0) = 4$$

$$R(r_0) = [(bk - r_0)/v)] = \tfrac{8}{4} = 2$$

$$r_0k - \lambda(r_0) = 12 - 4 = 8$$

$$\bar{\lambda}_0 = (r_0k - \lambda(r_0))/v = 2, \bar{\lambda}_1 = [(R(r_0)(k-1) - \bar{\lambda}_0)/(v-1)] = [0.7] = 0, \bar{\lambda}_2 = \bar{\lambda}_1 + 1 = 1$$

$$\bar{m}k = (k/v\bar{\lambda}_0) + ((\bar{v}-1)sk/v(\bar{v}(s-1)\bar{\lambda}_2 + v\bar{\lambda}_1 + \bar{\lambda}_0)) + ((s-1)k/v(v\bar{\lambda}_2 + \bar{\lambda}_0)) = \tfrac{7}{8} = 0.875$$

with $\bar{v} = v/s = 2$.

$$\bar{c}_{d00} = (r_0k - \lambda(r_0))/k = 2.666667$$

$$\bar{c}_{dii} = R(r_0)(k-1)/k = 1.333333$$

$$\bar{c}_{dio} = -(\bar{\lambda}_0 - 1)/k = -0.333333$$

$$\bar{c}_{djj} = (R(r_0)(k-1) - 2)/2 = 0.666667$$

$$\bar{c}_{dj0} = -\bar{\lambda}_0/k = -0.666667$$

$$\bar{c}_{dpp} = \bar{c}_{dqq} = R(r_0)(k-1)/k = 1.333333$$

$$\bar{c}_{dp0} = \bar{c}_{dq0} = -\bar{\lambda}_0/k = -0.666667$$

$$\text{-}\bar{c}_{dpq} = (\bar{\lambda}_1 - 1)/k = -0.333333.$$

Now we justify the four conditions of Theorem 6.5.

1. $r_0k - \lambda(r_0) - 2 = 8 - 2 = 6 < vB(r_0k - \lambda(r_0) - 2, R(r_0)(k-1)) = 8.8625$
2. $1/(R(r_0) - 1)(k-1) = 0.5 > \bar{m} = 0.2917$

3. $\bar{m}k = 0.875 < min\{(\bar{c}_{d00}+\bar{c}_{dii}+2\bar{c}_{di0})/(\bar{c}_{d00}\bar{c}_{dii}-\bar{c}^2_{dio}), (\bar{c}_{d00}+\bar{c}_{djj}+2\bar{c}_{dj0})/(\bar{c}_{d00}\bar{c}_{djj}-\bar{c}^2_{dj0})\} = 0.967742$

4. $\bar{m}k = 0.875 < \{\bar{c}_{dpp}(\bar{c}_{d00}+\bar{c}_{dqq}+2\bar{c}_{dpq})-(\bar{c}_{dp0}+\bar{c}_{dpq})^2\}/\{\bar{c}_{d00}\bar{c}_{dpp}\bar{c}_{dqq}-\bar{c}_{d00}\bar{c}^2_{dpq}-\bar{c}_{dpp}\bar{c}^2_{dq0}-\bar{c}_{dqq}\bar{c}^2_{dp0}+2\bar{c}_{dp0}\bar{c}_{dq0}\bar{c}_{dpq}\} = 1.145833.$

We see that the conditions given in Theorem 6.5 are verified. Thus, the above structure (k) design is MV-optimal. Hence we conclude that the design $d(5, 4, 3)$ is trace, A- and MV-optimal design .

Chapter 7

Conclusions and Future Work

7.1 Conclusions

The focus of this thesis was the construction of optimal designs with applications in estimation and factorial design. Throughout this thesis, we have emphasized consistently and developed methodologies for various types of optimal designs, estimation problems, and structure designs in the context of factorial design. Moreover, we have applied the methodologies in some real data sets.

In Chapter 1, we started with a basic introduction to optimal design theory. We discussed different types of optimal designs (such as exact and approximate designs) along with some of the fundamental concepts such as the design measure, variance function and the information matrix. By Caratheodory's theorem, we explained how an optimal design in a continuous design space can be replaced by at least one finite discrete probability distribution. We discussed how to discretize the continuous design space and how to obtain the optimal design in an induced design space. This chapter closes with the description of different optimality criteria and their properties.

In Chapter 2, we considered a class of optimization problems and determined the optimality conditions for our optimization problems. An important tool in this regard is the directional derivative of a criterion function. We extensively studied the properties of the directional derivatives and the General Equivalence Theorem. In order to determine the optimal designs, we considered a class of multiplicative algorithms indexed by a function which is positive and strictly increasing and may depend on one or more free parameters. We discussed some nice properties of the algorithms.

In Chapter 3, we focused on the construction of D-optimal design and its properties. We considered various polynomial regression models, namely, simple linear regression, quadratic regression, cubic regression and quartic regression models. We first constructed D-optimal design analytically using the Legnedre polynomial. We constructed D-optimal design which is unique, having a minimal support of k points which are the k roots of the Legnedre polynomial. We then constructed the D-optimal designs for various models by using a class of multiplicative algorithms. We also developed some useful strategies for better convergence of the algorithms by using the properties of the directional derivatives.

In Chapter 4, we solved an important estimation problem by using the theories of optimal design and simultaneous optimization techniques. We considered the problem of determining the maximum likelihood estimates of the cell probabilities under the hypothesis of marginal homogeneity in a square contingency table. We first formulated the problem for a general $n \times n$ contingency table. We approached this problem by initially formulating the Lagrangian function with constraints of marginal homogeneity. We then subsequently removed the Lagrange parameters by

substitution and transformed the problem to one of maximizing some functions of the cell probabilities simultaneously. We then considered two cases of the problem, namely, for 3×3 and 4×4 contingency tables. We applied our methodologies in some real data sets in which the hypothesis of marginal homogeneity is of interest, namely, the Migration data from Agresti (2002), a data on grading of the unaided distance vision of each eye given in Placket (1981).

In Chapter 5, we devoted our optimal design theory to the another estimation problem to determine the maximum likelihood estimators of the parameters of the latent variable models such as the Bradley Terry model where the data come from a paired comparisons experiment. We expressed the parameters of these models in terms of a set of another parameters which are regarded as weights in optimal design context. We approached this problem by considering the observed frequency having a binomial distribution and then replacing the binomial parameters in terms of optimal design weights. We then found out the maximum likelihood estimators of the parameters by transforming the problem to a constrained optimization problem. We applied our methodologies to some naturally arising data sets of this type, namely, a Baseball data set from Agresti (2002) and Coffee data set from Bradley and El-Helbawy (1976).

In Chapter 6, we worked on constructing optimal structure designs for comparing test treatments with a control. We introduced structure (k_1), structure (k_2) and structure (k_1k_2) designs and established their properties using the incidence and characteristic matrices. Starting from a structure (k) design and augmenting one control in each block, we developed methods of obtaining optimal R-type structure (k) designs and showed how such designs are trace, A- and MV-optimal. The proposed methodologies are easy to follow and the construction of the designs comes out in a simple form.

7.2 Future Work

We now list some potential topics that we would like to pursue in future.

1. We developed some useful strategies for the construction of D-optimal designs by using the properties of the directional derivatives in Chapter 3. We wish to develop further strategies for constructing some other optimal designs such as A-optimal and c-optimal designs. These two criteria are also important in the sense that in A-optimality we minimize the average variance of the parameter estimates whereas in c-optimality we minimize the variance of individual parameter estimates. We plan to do so using the properties of the partial and directional derivatives of the criterion functions.

2. In the context of the estimation problem of Chapter 4, we plan to work on finding the maximum likelihood estimator of cell probabilities in case of quasi independence in a contingency table (Morgan and Titterington (1977)). We can approach this problem by maximizing the criterion function of the form

$$\phi(p) \;=\; \prod_{\substack{i=1 \\ i\neq j}}^{J} \prod_{j=1}^{J} \left\{ \frac{p_j}{(1-p_i)} \right\}^{n_{ij}}.$$

   Quasi-independence states that only for some i and j, the cell probabilities p_{ij} ($\sum\sum p_{ij} = 1$) can be factorized into the form $p_{ij} = a_i b_j$, in contrast to full independence in which such factorization holds for all i and j.

   We can apply this problem to the mover-stayer model of Blumen et al. (1955) which is postulated for the transition probabilities of a J-state Markov Chain.

Note that this implies the conditional probabilities of state change are $p_{j|i} = p_j/(1 - p_i), i \neq j$.

3. In Chapter 5, we worked on the problem on determining maximum likelihood estimators of the parameters of Bradley-Terry model. A possible extension of this work can be done in terms of parameter estimation by considering the options of no preference or tie, for rankings.

   Another possible extension of pairwise comparisons is to invite subjects to place three treatments in order of preference, i.e., to work in situations such as in models including triplets of treatments.

Bibliography

Agresti, A. (2002). *Categorical Data Analysis, 2nd Edition*. New Jersey: John Wiley & Sons, Ltd. (Cited on pages 102, 112, 118 and 143.)

Andrews, D. F., P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey (1972). *Robust Estimates of Location*. Princeton: Princeton University Press. (Cited on page 33.)

Atkinson, A. C., A. N. Donev, and R. D. Tobias (2007). *Optimum Experimental Design, with SAS*. New York: Oxford University Press. (Cited on pages 1, 6, 17 and 51.)

Atwood, C. L. (1969). Optimal and Efficient Designs of Experiments. *Annals of Mathematical Statistics 40*, 1570–1602. (Cited on page 24.)

Atwood, C. L. (1976). Convergent Design Sequences, for Sufficiently Regular Optimality Criteria. *Annals of Statistics 4*, 1124–1138. (Cited on page 46.)

Atwood, C. L. (1980). Convergent Design Sequences for Sufficiently Regular Optimality Criteria, II: Singular Case. *Annals of Statistics 8*, 894–912. (Cited on page 46.)

Berger, M. P. F. and W. K. Wong (2009). *An Introduction to Optimal Designs for Social and Biomedical Research*. Chichester: John Wiley & Sons, Ltd. (Cited on pages 1, 17 and 51.)

Blumen, I., M. Kogan, and P. J. McCarthy (1955). *The Industrial Mobility of Labour as a Probability Process, Vol.6 of Cornell Studies of Industrial and Labour Relations*. New York: Cornell University. (Cited on page 144.)

Bradley, R. A. (1965). Another Interpretation of a Model for Paired Comparisons. *Psychometrika 30*, 315–318. (Cited on page 116.)

Bradley, R. A. and A. T. El-Helbawy (1976). Treatment Contrasts in Paired Comparisons: Basic Procedures with Application to Factorials. *Biometrika 63,2*, 255–262. (Cited on pages 114, 118, 119 and 143.)

Bradley, R. A. and M. E. Terry (1952). Rank Analysis of Incomplete Block Designs I. The Method of Paired Comparisons. *Biometrika 39*, 324–345. (Cited on pages 114 and 116.)

Chan, N. N. (1987). Schur-Convexity for *A*-optimal Designs. *Journal of Mathematical Analysis and Applications 122*, 1–6. (Cited on page 19.)

Chan, N. N. and K.-H. Li (1989). Majorization for *A*-optimal Designs. *Journal of Mathematical Analysis and Applications 142*, 101–107. (Cited on page 19.)

Chowdhury, M. and S. Mandal (2016). Maximum Likelihood Estimation of the Cell Probabilities under the Hypothesis of Marginal Homogeneity in Square Contingency Table. *(In Preparation)*. (Cited on page 85.)

Chowdhury, M., S. Mandal, D. K. Ghosh, and S. C. Bagui (2016). Optimal Structure (k) Designs for Comparing Test Treatments with a Control. *Journal of Statistical Theory and Applications (Accepted)*. (Cited on page 124.)

Cook, D. and V. Fedorov (1995). Constrained Optimization of Experimental Design (with discussion). *Statistics 26*, 129–178. (Cited on page 1.)

Cotter, S. C. (1973). *Confounding in Factorial Experiment*. Ph. D. thesis, Unpublished Ph.D. Thesis, University of Southamption, U K, Southamption. (Cited on pages 122 and 123.)

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion). *Journal of the Royal Statistical Society Series B 39*, 1–38. (Cited on page 46.)

Dette, H., A. P, and Z. Anatoly (2008). Improving Updating Rules in Multiplicative Algorithms for Computing *D*-optimal Designs. *Computational Statistics and Data Analysis 53*, 312–320. (Cited on page 51.)

Efron, B., R. Tibshirani, J. D. Storney, and V. Tusher (2001). Empirical Bayes Analysis of Microarray Experiment. *Journal of the American Statistical Association 96*, 1151–1160. (Cited on page 114.)

Elfving, G. (1952). Optimum Allocation in Linear Regression Theory. *Annals of Mathematical Statistics 23*, 255–262. (Cited on pages 19 and 26.)

Eplett, W. J. R. (1980). An Influence Curve for Two-Sample Rank Tests. *Journal of the Royal Statistical Society Series B 42*, 64–70. (Cited on page 33.)

Fedorov, V. V. (1972). *Theory of Optimal Experiments*. New York and London: Academic Press. (Cited on pages 14, 17, 25, 46, 51 and 57.)

Grasshoff, U. and R. Schwabe (2008). Optimal Design for the Bradley-Terry Paired Comparison Model. *Statistical Methods and Applications 17*, 275–289. (Cited on page 114.)

Hampel, F. R. (1968). *Contributions to the Theory of Robust Estimation*. Ph. D. thesis, University of California, Berkeley. (Cited on page 33.)

Hampel, F. R. (1971). A General Qualitative Definition of Robustness. *Annals of Mathematical Statistics 42*, 1887–1896. (Cited on page 33.)

Huang, T.-K., R. C. Weng, and C.-J. Lin (2006). Generalized Bradley-Terry Models and Multi-class Probability Estimates. *Journal of Machine Learning Research 7*, 85–115. (Cited on page 116.)

Ireland, C. T., H. H. Ku, and S. Kullback (1969). Symmetry and Marginal Homogeneity of an r $\times$ r Contingency Table. *Journal of the American Statistical Association 64*, 1323–1341. (Cited on page 85.)

Jacroux, M. (1984). On the Optimality and Usage of Reinforced Block Designs for Comparing Test Treatments with a Standard Treatment. *Journal of the Royal Statistical Society B 46*, 316–322. (Cited on page 125.)

Jacroux, M. (1987a). On the Determination and Construction of MV-optimal Designs for Comparing Test Treatments with a Standard Treatment. *Journal of Statistical Planning and Inference 15*, 205–225. (Cited on page 128.)

Jacroux, M. (1987b). On the Optimality of Block Designs for Comparing Test Treatments with a Control. Technical report. (Cited on page 128.)

Jacroux, M. (1989). The *A*-optimality of Block Designs for Comparing Test Treatments with a Control . *Journal of the American Statistical Association 84*, 310–317. (Cited on page 128.)

John, J. A. and T. M. F. Smith (1972). Two-Factor Experiments in Non-Orthogonal Designs. *Journal of the Royal Statistical Society B 34*, 401–409. (Cited on page 123.)

John, R. C. S. and N. R. Draper (1975). *D*-optimality for Regression Designs: A Review. *Technometrics 17*, 15–23. (Cited on page 17.)

Karlin, S. and W. J. Studden (1966). Optimal Experimental Designs. *Annals of Mathematical Statistics 37*, 783–815. (Cited on page 24.)

Kiefer, J. (1959). Optimum Experimental Designs (with discussion). *Journal of the Royal Statistical Society Series B 21*, 272–319. (Cited on pages 17 and 51.)

Kiefer, J. (1974). General Equivalence Theory for Optimum Designs (Approximate Theory). *Annals of Statistics 2*, 849–879. (Cited on pages 22, 27, 33 and 42.)

Kiefer, J. and J. Wolfowitz (1960). The Equivalence of Two Extremum Problems. *Canadian Journal of Mathematics 12*, 363–366. (Cited on pages 21 and 55.)

Kurkjian, B. M. and M. Zelen (1962). A Calculus for Factorial Arrangements. *Annals of Mathematical Statistics 33*, 600–619. (Cited on page 122.)

Kurkjian, B. M. and M. Zelen (1963). Applications of the Calculus of factorial

arrangements: I. Block and Direct Product Designs. *Biometrika 50*, 63–73. (Cited on page 122.)

Mandal, S. (2000). *Construction of Optimizing Distributions with Applications in Estimations and Optimal Designs*. Ph. D. thesis, University of Glasgow, Glasgow. (Cited on pages 34 and 51.)

Mandal, S. and B. Torsney (2000). Algorithms for the Construction of Optimizing Distributions. *Communications in Statistics - Theory and Methods 29*, 1219–1231. (Cited on pages 31, 45 and 85.)

Mandal, S. and B. Torsney (2006). Construction of Optimal Designs using a Clustering Approach. *Journal of Statistical Planning and Inference 136*, 1120–1134. (Cited on pages 1, 42, 45, 46 and 51.)

Mandal, S., B. Torsney, and K. C. Carriere (2005). Constructing Optimal Designs with Constraints. *Journal of Statistical Planning and Inference 128*, 609–621. (Cited on pages 45 and 51.)

Molchanov, I. and S. Zuyev (2000). Variational Calculus in the Space of Measures and Optimal Design. *Optimum Design 2000 51*, 79–90, Kluwer Academic Publishers. (Cited on page 46.)

Morgan, B. J. T. and D. M. Titterington (1977). A Comparison of Iterative Methods for Obtaining Maximum Likelihood Estimates in Contingency Tables with a Missing Diagonals. *Biometrika 64*, 265–269. (Cited on page 144.)

Mukerjee, R. (1979). Inter-effect Orthogonality in Factorial Experiments. *Calcutta Statistical Association Bulletin 28*, 83–108. (Cited on pages 122, 123, 125 and 132.)

Mukerjee, R. (1980). Further Results on the Analysis of Factorial Experiments. *Calcutta Statistical Association Bulletin 29*, 1–26. (Cited on page 122.)

Paik, U. B. and W. T. Federer (1973). Partially Balanced Designs Act Properties A and B. *Communications in Statistics 1*, 331–350. (Cited on pages 122 and 123.)

Placket, R. L. (1981). *The Analysis of Categorical Data*. New York: Macmillan Publishing Co. (Cited on pages 100, 101, 110 and 143.)

Pukelsheim, F. (1993). *Optimal Design of Experiments*. New York: Wiley. (Cited on pages 1, 17 and 51.)

Rao, P. V. (1961). Analysis of a Class of PBIB designs with More than Two Associate Classes. *Annals of Mathematical Statistics 32*, 800–808. (Cited on page 122.)

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton: Princeton University Press. (Cited on page 33.)

Rohde, C. A. (1965). Generalized Inverses of Partitioned Matrices. *Journal of the Society for Industrial and Applied Mathematics 13*, 1033–1035. (Cited on page 24.)

Shah, B. V. (1959). On a Generalization of the Kronecker Product Designs. *Annals of Mathematical Statistics 30*, 48–54. (Cited on page 122.)

Shah, K. R. and B. K. Sinha (1989). *Theory of Optimal Designs*, Volume Vol. 54, Springer-Verlag. Lecture Notes in Statistics. (Cited on pages 17 and 51.)

Sia, L. L. (1977). Optimum Spacings of Elementary Treatment Contrasts in Symmetrical 2-Factor PA Block Designs. *Canadian Journal of Statistics 5*, 227–234. (Cited on page 123.)

Sibson, R. (1974). D_A-optimality and Duality. Progress in Statistics. *Colloquia Mathematica Societatis János Bolyai 9*, 677–692. (Cited on pages 23 and 42.)

Silvey, S. D. (1978). Optimal Design Measures with Singular Information Matrices. *Biometrika 65*, 553–559. (Cited on page 56.)

Silvey, S. D. (1980). *Optimal Design*. London: Chapman and Hall. (Cited on pages 1, 14, 17, 24 and 51.)

Silvey, S. D. and D. M. Titterington (1973). A Geometric Approach to Optimal Design Theory. *Biometrika 60*, 21–32. (Cited on page 24.)

Silvey, S. D. and D. M. Titterington (1974). A Lagrangian Approach to Optimal Design. *Biometrika 61*, 299–302. (Cited on page 42.)

Silvey, S. D., D. M. Titterington, and B. Torsney (1978). An Algorithm for Optimal Designs on a Finite Design Space. *Communications in Statistics-Theory and Methods 7*, 1379–1389. (Cited on page 45.)

Thannipara, A. (1992). *On Optimal Block Designs for Comparing Test Treatments with a Control* . Ph. D. thesis, Unpublished Ph.D. thesis, Saurashtra University, Rajkot, India, Rajkot. (Cited on page 128.)

Titterington, D. M. (1975). Optimal Design: Some geometrical aspects of *D*-optimality. *Biometrika 62(2)*, 313–320. (Cited on page 51.)

Titterington, D. M. (1976). Algorithms for Computing *D*-optimal Designs on a Finite Design Space. *Proceedings of the 1976 Conference on Information Sciences and Systems pages 213-216, Department of Electrical Engineering, John Hopkins University, Baltimore, MD*. (Cited on page 45.)

Tocher, K. D. (1952). The Design and Analysis of Block Experiments. *Journal of the Royal Statistical Society B 14*, 45–100. (Cited on page 130.)

Torsney, B. (1977). Contribution to Discussion of "Maximum Likelihood from Incomplete Data via the EM algorithm" by Dempster et al. *Journal of the Royal Statistical Society Series B 39*, 26–27. (Cited on pages 1 and 45.)

Torsney, B. (1981). *Algorithms for a Constrained Optimization Problem with Applications in Statistics and Optimum Design*. Ph. D. thesis, University of Glasgow, Glasgow. (Cited on page 24.)

Torsney, B. (1983). A Moment Inequality and Monotonicity of an Algorithm. *Proceedings of International Symposium on Semi-Infinite Programming and Applications (Edited by Kortanek, K. O. and Fiacco, A. V.). Lecture Notes in Economics and Mathematical Systems vol. 215, pages 249-260, University of Texas, Austin*. (Cited on pages 45 and 51.)

Torsney, B. (1988). Computing Optimizing Distributions with Applications in Design, Estimation and Image Processing. *Optimal Design and Analysis of Experiments (Edited by Dodge, Y., Fedorov, V. V. and Wynn, H. P.) 361-370, Elsevier Science Publishers B. V., North Holland*. (Cited on pages 45 and 51.)

Torsney, B. (2004). Fitting Bradley Terry Models using a Multiplicative Algorithm. *Proceedings in Computational Statistics COMPSTAT. Physica-Verlag, Heidelberg*, 513–526. (Cited on page 116.)

Torsney, B. (2010). Estimation and Optimal Designing under Latent Variable Models for Paired Comparisons Studies via a Multiplicative Algorithm. *Contributions to Statistics*, 213–220. (Cited on page 116.)

Torsney, B. and A. M. Alahmadi (1992). Further Development of Algorithms for Constructing Optimizing Distributions. *Model Oriented Data Analysis. Proceedings of the 2nd IIASA Workshop in St. Kyrik, Bulgaria (Edited by Fedorov, V. V., Müller, W. G. and Vuchkov, I. N.) pages 121-129, Physica-Verlag*. (Cited on page 45.)

Torsney, B. and S. Mandal (2001). Construction of Constrained Optimal Designs. *Optimum Design 2000 141-152, Kluwer Academic Publishers*. (Cited on pages 90 and 100.)

Torsney, B. and S. Mandal (2006). Two classes of multiplicative algorithms for constructing optimizing distributions. *Computational Statistics and Data Analysis 51*, 1591–1601. (Cited on page 47.)

Tsay, J. Y. (1976). Linear Optimal Experimental Designs. *Proceedings of 1976 Conference on Information Sciences and Systems, Department of Electrical Engineering, John Hopkins University, Baltimore, MD*, 222–226. (Cited on page 25.)

Tsay, J. Y. (1977). A Convergence Theorem in *L*-optimal Design Theory. *Annals of Statistics 5*, 790–794. (Cited on page 25.)

Vartak, M. N. (1955). On an Application of Kronecker Product of Matrices to Statistical Designs. *Annals of Mathematical Statistics 26*, 420–438. (Cited on page 122.)

Von Mises, R. (1947). On the Asymptotic Distribution of Differentiable Statistical Functions. *Annals of Mathematical Statistics 18*, 309–348. (Cited on page 33.)

Wedderburn, R. W. M. (1974). Generalized Linear Models Specified in Terms of Constraints. *Journal of the Royal Statistical Society, B 36*, 449–454. (Cited on page 85.)

Whittle, P. (1971). *Optimization under Constraints*. London: Wiley-Interscience. (Cited on page 33.)

Whittle, P. (1973). Some General Points in the Theory of Optimal Experimental Designs. *Journal of the Royal Statistical Society Series B 35*, 123–130. (Cited on pages 32 and 42.)

Wu, C. F. J. (1976). *Contributions to Optimization Theory with Applications to Optimal Design of Experiments*. Ph. D. thesis, University of California, Berkeley. (Cited on page 42.)

Wu, C. F. J. (1978). Some Iterative Procedures for Generating Nonsingular Optimal Designs. *Communications in Statistics- Theory and Methods 7*, 1399–1412. (Cited on page 46.)

Wu, T.-F., C.-J. Lin, and R. C. Weng (2004). Probability Estimates for Multi-class Classification by Pairwise Coupling. *Journal of Machine Learning Research 5*, 975–1005. (Cited on page 116.)

Wynn, H. P. (1972). Results in the Theory and Construction of *D*-optimum Experimental Designs (with Discussion). *Journal of the Royal Statistical Society Series B 34*, 133–186. (Cited on pages 17 and 46.)

Yu, Y. (2010). Monotonic Convergence of a General Algorithm for Computing Optimal Designs. *The Annals of Statistics 38*, 1593–1606. (Cited on page 46.)

Zelen, M. and W. T. Federer (1964). Applications of the Calculus for Factorial Arrangements II: Two Way Elimination of Heterogeneity. *Annals of Mathematical Statistics 35*, 658–672. (Cited on pages 122, 123 and 128.)

Zelen, M. and W. T. Federer (1965). Application of the Calculus for Factorial Arrangements: III. Analysis of Factorials with Unequal Number of Observations. *Sankhya A 27*, 383–400. (Cited on page 122.)