Construction of Optimal Designs in Polynomial Regression Models

by

Chao Zhu

A Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in Partial Fulfillment of the Requirements for the Degree of

Master of Science

Department of Statistics
University of Manitoba
Winnipeg, Manitoba

Abstract

We consider a class of optimization problems in which the aim is to find some optimizing probability distributions. One particular example is optimal design. We first review the optimal design theory, and determine the optimality conditions using directional derivatives. We then construct optimal designs for various polynomial regression models by finding the analytic solutions and by using a class of algorithms. We consider a practical problem, namely a radiation-dosage example, and discuss important aspects of optimal design throughout this example.

We also construct optimal designs for various polynomial regression models with more than one design variable. We consider another practical problem, namely a vocabulary-growth study. We then construct D-optimal and c-optimal designs for various models with and without the interaction term and the second order terms in design variables. We also develop strategies for constructing designs by using the properties of the directional derivatives.

Acknowledgments

I am sincerely and heartily grateful to my advisor, Dr. Saumen Mandal, who has the attitude and the substance of a genius. I am thankful to him for supporting me financially from his NSERC grant. He continually and convincingly conveyed a spirit of adventure in regard to academic research, and an excitement in regard to teaching. Without his guidance and persistent help this thesis would not have been possible.

I would like to thank my committee members, Dr. Xikui Wang and Dr. Srimantoorao Appadoo for monitoring my work, spending time in reading and providing me with valuable comments.

I am thankful to Mr. Dave Gabrielson for helping me with LaTeX. I am also thankful to Dr. Theo Koulis for helping me with the graphics using R. In addition, I thank all the professors and instructors of the Department of Statistics at the University of Manitoba for teaching all the exciting courses.

At last but not least, I take this opportunity to express my profound gratitude to my parents. Without their confirmed support, altruistic care and insistent encouragement, I would never have enough abilities to finish my thesis.

Contents

Chapter 1

Introduction

An experiment is a process or study that results in the collection and analysis of data. An experiment can be designed to answer a variety of questions of interest. The results of experiments are not known in advance. 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.

Experimental design is the process of planning a study to meet specified objectives. Planning an experiment properly 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. A researcher's goal is to increase the efficiency of a properly designed experiment.

In regression experiments the inputs are often numerical and the values or levels of inputs must be chosen before running an experiment and observing

a measurement on some variables of interest. There will be a set of or several combinations of the inputs allowed. We must decide how many observations to take at each combination of inputs.

1.1 Stages of designing experiments

It is important and necessary for the researchers involved in the experiment to have a clear idea of exactly what to be studied, how the data are to be collected and how these data are to be analyzed. Usually, the experimental problems can be solved in six stages.

a. *Formulation of the problem.* In scientific research, many problems are expressed as a relationship between some explanatory variables X and the response variable Y. Therefore, the first purpose is to identify the correct set of variables in the study.

b. *Choice of the research design.* Choice of research design involves consideration of structuring the research, the sample sizes and the data collection. At this stage it is important to distinguish between qualitative versus quantitative variables, random versus fixed variables and a crossed or nested relation among variables and to select the number of measurements, time points and subjects within groups.

c. *Choice of statistical model.* A statistical model must be chosen before the

data are collected. The model is usually a formalization of relationships between variables in the form of mathematical equations. A statistical model describes how the change of explanatory variables will affect the response variables. In design context, when considering the factors that may influence the response variable, the experimenter should classify these factors as either potential design factors or nuisance factors.

d. *Data Collection and performing the experiment.* In this stage, all data are collected depending on the research design and the design variables. When running the experiment it is important to monitor the process carefully to ensure that everything is being done according to the proposed plans.

e. *Analysis of data.* In this stage, based on the chosen model, statistical methods are employed to analyze the data so that the results are objective rather than judgemental in nature. Sometimes the graphical methods play an important role in data analysis.

f. *Conclusions.* After the data have been properly analyzed the experimenter must carefully infer the practical conclusions about the results and recommend the course of actions. Sometimes the experimenter should perform some follow-up runs and confirmation testing to validate the conclusions.

1.2 The necessity and function of optimal design

Usually, an experiment can become more efficient by two directions. One is by working on finding best statistical model to fit the data. Another one is by looking for the most efficient method to collect the data. Some questions naturally arise in mind. How do you collect data to answer your research questions? What should your design be? Can we find an efficient design with a small or moderate sample size? These kinds of questions will be the most of the problems optimal design deal with.

Optimal design theory offers a possible way of finding an optimal or a highly efficient design using all current information. Once the objectives of the study are clearly mentioned, we can find some optimization methods that will generate the optimal design. Useful texts in optimal design theory are Atkinson et al. (2007), Atkinson and Donev (1992), Fedorov (1972), Pukelsheim (1993), Silvey (1980), Berger and Wong (2009).

In any regression model, the objective is good estimation of the parameters of the regression model. There are a variety of criteria defining good estimation. We choose a design to optimize a chosen criterion.

We start by considering the problem of selecting an experimental design to furnish information on models of the type:

$$y \sim p(y|\underline{x}, \underline{\theta}, \sigma) \tag{1.1}$$

where y is the response variable. In a particular experimental condition, y is considered as a sum of a real-valued response function evaluated at $\underline{x}$, and a random error. The vector $\underline{x} = (x_1, x_2, \ldots, x_m)^T$ consists of the design variables. These can be chosen by the experimenter, their values being restricted to a space $\mathcal{X}$, i.e. $\underline{x} \in \mathcal{X} \subseteq \mathbb{R}^m$. Thus $\mathcal{X}$ is the set of experimental conditions, and is called the design space. Typically this space is continuous but can be discrete. The vector $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)^T$ is a k-dimensional vector of unknown parameters. The true value of $\underline{\theta}$ is known to belong to a set $\Theta \in \mathbb{R}^k$. σ is a nuisance parameter. This is also fixed and unknown but is not of primary interest. $p(.)$ is a probability model.

The experimenter can freely choose the experimental conditions from the given experimental domain $\mathcal{X}$. In most applications, $\mathcal{X}$ is taken to be compact. That is, the design space is closed and bounded. For each $x \in \mathcal{X}$, an experiment can be performed whose outcome is a random variable $y = y(\underline{x})$, where $var(y(\underline{x})) = \sigma^2$. We generally suppose that σ does not depend on the experimental condition $\underline{x}$. In linear regression design the model is linear in the unknown parameters $\underline{\theta}$ but not necessarily linear in $\underline{x}$. So in linear models $y(\underline{x})$ has an expected value of the explicit form

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

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 regression functions $f_1, f_2, \ldots, f_k$ are known to the experimenter. In order to obtain an observation on y, a value for $\underline{x}$ must first be

selected from $\mathcal{X}$. It is assumed that $\underline{x}$ can be set to any chosen value in $\mathcal{X}$. Given this control over the selection of $\underline{x}$, a natural question to consider is at what values of $\underline{x}$ should observations, say n, on y be taken in order to obtain a 'best' inference or as reliable an inference as possible for all or some or a function of the parameters $\underline{\theta}$.

Such a 'best' selection of $\underline{x}$ values and/or allocation of the n observations to the elements of $\mathcal{X}$ is termed an optimal regression design.

The mode of inference must first be decided upon. For the moment let us suppose that it is point estimation. It will be seen that the solution proposed for this case will hold good for other modes of inference too. It is desired then 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 by some method of point estimation the estimator $\hat{\underline{\theta}}$ of $\underline{\theta}$ is obtained. Let $\hat{\underline{\theta}}$ be unbiased for $\underline{\theta}$. Typically the components $\hat{\theta}_j$ will be correlated. The $k \times k$ matrix $D(\hat{\underline{\theta}}) = E([\hat{\underline{\theta}} - \underline{\theta}][\hat{\underline{\theta}} - \underline{\theta}]^T)$, the dispersion matrix (or the variance-covariance matrix) of $\hat{\underline{\theta}}$ about $\underline{\theta}$, contains 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, but also in its off-diagonal cross product deviation terms. Generally speaking the 'smaller' is $D(\hat{\underline{\theta}})$ the better is the accuracy of $\hat{\underline{\theta}}$.

Suppose the model (1.2) is true. Let y_i denote the observation obtained

at $\underline{x}_i$ so that

$$E(y_i) = \underline{v}_i^T \underline{\theta}, \quad \underline{v}_i = (f_1(\underline{x}_i), f_2(\underline{x}_i), \ldots, f_k(\underline{x}_i))^T, i = 1, 2, \ldots, n. \tag{1.3}$$

Note that typically there will be several equalities between the $\underline{x}_i$'s, more than one observation being taken at the same $\underline{x}$ value. Suppose also that $y_1, y_2, \ldots, y_n$ are independent random variables with equal variance σ^2. The y_i's then satisfy the standard linear model:

$$E(Y) = X\underline{\theta}, \quad D(Y) = \sigma^2 I_n \tag{1.4}$$

where $Y = (y_1, y_2, \ldots, y_n)$, 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 $D(Y)$ denotes the dispersion matrix of Y. Least squares estimators are a conventional choice for this model having the optimality of being best linear unbiased estimators (BLUE). They are solutions of

$$(X^T X)\,\hat{\underline{\theta}} \;=\; X^T Y. \tag{1.5}$$

The $k \times k$ matrix $(X^T X)$ is the information matrix for $\underline{\theta}$. The larger $(X^T X)$, the greater is the information in the experiment. If all the parameters $\underline{\theta}$ are of interest, then the selection of $\underline{x}$ must at least ensure that the matrix $(X^T X)$ is non-singular, in which case the unique solution for (1.5) is given by:

$$\hat{\underline{\theta}} \;=\; (X^T X)^{-1} X^T Y \tag{1.6}$$

with

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

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

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

Clearly the dispersion matrix of $\hat{\underline{\theta}}$ does not depend on $\underline{\theta}$ and only depends proportionally on the parameter σ^2. We have to select $\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n\}$ to make the matrix $D(\hat{\underline{\theta}})$ 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:

$$E(y|\underline{v}, \underline{\theta}, \sigma) = \underline{v}^T\underline{\theta} \tag{1.7}$$

where

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

Clearly 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 $(f_1, f_2, \ldots, f_k)^T$. That is, $\mathcal{V}$ is the image under f of $\mathcal{X}$. So, $\mathcal{V}$ is an induced design space. Suppose that the design space $\mathcal{V}$ 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 to take this observation. Note that, originally the design space $\mathcal{X}$ is continuous, however, after discretization, we can deal with the induced design space $\mathcal{V}$. The design problem can now be expressed more precisely. At which of the 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 n observations, we must decide how many of these, say n_j, to take at $\underline{v}_j$, $\sum_{j=1}^{J} n_j = n$. Given these choices the matrix $(X^T X)$ can be expressed in the form:

$$X^T X = M(\underline{n}), \ \underline{n} = (n_1, n_2, \ldots, n_J)^T \tag{1.8}$$

where

$$\begin{aligned} M(\underline{n}) &= \sum_{j=1}^{J} n_j \underline{v}_j \underline{v}_j^T \\ &= V N V^T \end{aligned}$$

and $V = [\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_J]$, $N = diag(n_1, n_2, \ldots, n_J)$.

1.3.1 Exact versus approximate design

We now want to choose $\underline{n}$ to make the matrix $M(\underline{n})$ big in some sense. Given that the n_j's must be integer this is an integer programming problem and in the design context is described as an *exact design* problem. Typically integer

programming problems are difficult or at least laborious to solve even without additional constraints, mainly because the theory of calculus cannot be used to define the existence of or to identify optimal solutions. Furthermore, a solution would have to be worked out separately for different values of n. By the nature of the problem then, no formula for an optimal exact design could be devised that would express it as a function of n. Nevertheless one could not avoid having to solve such a problem if, for given n, one chose to seek optimal n_j's directly. However,

$$M(\underline{n}) \;=\; nM(p) \tag{1.9}$$

where

$$M(p) \;=\; \sum_{j=1}^{J} p_j \, \underline{v}_j \, \underline{v}_j^T \tag{1.10}$$

$$\;=\; VPV^T \tag{1.11}$$

and $P = \text{diag}(p_1, p_2, \ldots, p_J)$; $p_j = \frac{n_j}{n}$ and so is the proportion of observations taken at $\underline{v_j}$, so that $p_j \geq 0$, $\sum_{j=1}^{J} p_j = 1$; and $p = (p_1, p_2, \ldots, p_J)$ represents the resultant distribution on $\mathcal{V}$.

Thus our problem becomes that of choosing p to make $M(p)$ large subject to $p_j = \frac{n_j}{n}$. Relaxing the latter to $p_j \geq 0$ and $\sum_{j=1}^{J} p_j = 1$ yields an *approximate design* problem.

This is a simpler or more flexible problem to solve and yet one that is not

much visibly different from the original. Naturally an approximate solution that would be preferred to the original exact design problem would be np^*, rounded to a 'nearest' exact design. Hopefully this would be a near fully optimum exact design. Note that we can view p as defining a probability distribution on $\mathcal{V}$ to yield

$$M(p) \quad = \quad E_p[\underline{v}\,\underline{v}^T] \tag{1.12}$$

where $P(\underline{v} = \underline{v}_j) = p_j$.

Thus we can think of a design as defined by a set of weights or probabilities p_j, p_j being assigned to $\underline{v}_j \in \mathcal{V}$. Such a design may put weight $p_j = 0$.

1.3.2 Design Measure

We have referred to p above both as the vector $(p_1, p_2, \ldots, p_J)$ and as a probability distribution on $\mathcal{V}$. Of course this induces a distribution on the original design space $\mathcal{X}$. A full statement of this is

$$p \quad = \quad \left\{ \begin{array}{cccc} \underline{x}_1 & \underline{x}_2 & \ldots & \underline{x}_J \\ p_1 & p_2 & \ldots & p_J \end{array} \right\} \tag{1.13}$$

where the first line gives the locations of the design points with p_j the associated design weights, $\sum_{j=1}^{J} p_j = 1$ and $0 \leq p_j \leq 1$ for all j.

Support of a Design Measure

The support of the design measure p in the design space $\mathcal{V}$ is defined to be those vertices $\underline{v}_j$ with nonzero weighting under p. It is denoted by:

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

Often there will be an optimal design, say p^* such that $Supp(p^*)$ is a strict subset of $\mathcal{V}$.

Standardised Variance of the Predicted Response

The standardised variance of the predicted response on y at $\underline{x}$ for the design (1.13) is given by

$$d(\underline{x},\, p) \;=\; \underline{f}^T(\underline{x})\, M^{-1}(p)\, \underline{f}(\underline{x}), \tag{1.14}$$

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

1.3.3 The Information Matrix $M(p)$

Note that, from the definition of the information matrix $M(p)$ as given in (1.9), it can be written as

$$\begin{aligned} M(p) \;&=\; \sum_{j=1}^{J} p_j\, \underline{v}_j\, \underline{v}_j^T \\ &=\; VPV^T \end{aligned}$$

The information matrix $M(p)$ is symmetric and nonnegative definite. The symmetry of this matrix follows from its definition above. The nonnegativeness of the appropriate quadratic form is easy to verify.

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

The information matrix is widely used in optimal experimental design. Recall that the inverse of the variance-matrix (dispersion matrix) is actually the information matrix. Because of this reciprocity property of estimator-variance and information, minimizing the variance corresponds to maximizing the information. 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 in Optimal Design

It may be possible to obtain a best inference for all or some of the unknown parameters $\underline{\theta}$ by making the matrix $M(p)$ large in some sense. So we consider various ways in which to make the matrix $M(p)$ large, namely by maximizing some real valued function $\phi(p) = \psi\{M(p)\}$. Note that the 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.

We consider both the cases when interest is in inference about all of the parameters or a sub-set of the parameters $\underline{\theta}$ of a regression model. We first assume that the information matrix $M(p)$ is non-singular and hence positive definite. Now we consider different design criteria of interest and their properties.

***D*-optimality:** The most important and popular design criterion in applications is that of D-optimality, in which the criterion function is given by

$$\phi_D(p) = \psi_D\{M(p)\} = log\, det\{M(p)\} = -\, log\, det\{M^{-1}(p)\}.$$

Various motivations for D-optimality exist. A D-optimal design minimizes the volume of the conventional ellipsoidal confidence region for the parameters of the linear model. Other motivations for D-optimality lie in hypothesis testing under a normal linear model. This is the most extensively studied of all design criteria; see Kiefer (1959), Fedorov (1972), Silvey (1980), Atkinson and Donev (1992), Pukelsheim (1993) and Wynn (1972). Mandal and Torsney (2000) construct D-optimal designs in a variety of examples using a class of multiplicative algorithms [Torsney (1977)].

The criterion 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. The criterion is a concave function on $\mathbb{M}$. [Note that a function f is concave on a convex

set S if for any x, $y \in S$, and for any α, $0 < \alpha < 1$, $f[(1-\alpha)x + \alpha y] \geq (1-\alpha)f(x) + \alpha f(y)$.]

***A*-optimality:** A-optimality is defined by the following criterion function:

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

Thus an A-optimum design minimizes the sum of the variances of the parameter estimates or their average variance, but does not take correlations between the estimates into account.

The criterion is an increasing function over the set of positive definite symmetric matrices. Unlike D-optimality, linear transformations may not leave the A-optimum design unchanged.

***G*-optimality:** G-optimality is defined by the criterion function:

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

This criterion seeks to minimize the maximum value of $\underline{v}^T M^{-1}(p)\,\underline{v}$ which is proportional to the variance of $\underline{v}^T\hat{\underline{\theta}}$. Kiefer and Wolfowitz (1960) prove the equivalence of this criterion and the D-optimal criterion.

The criterion function is an increasing function over the set of positive definite symmetric matrices. ψ_G is concave on $\mathbb{M}$, and is invariant under a non-singular linear transformation of $\mathcal{V}$.

***E*-optimality:** In E-optimality the variance of the least well-estimated contrast $\underline{a}^T\underline{\theta}$ is minimized subject to the constraint $\underline{a}^T\underline{a} = 1$. Thus the E in the name of this criterion stands for extreme. This optimality criterion is defined

by the criterion function:

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

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

The criterion is an increasing function over the set of positive definite symmetric matrices. If λ_{max} is unique then ψ_E has unique partial derivatives corresponding to positive weights. Otherwise ϕ_E is not differentiable.

Sometimes interest is only in some of the unknown parameters or some combinations of the parameters. Suppose we are interested in s linear combinations which are elements of $A\underline{\theta}$, where A is an $s \times k$ matrix of rank $s \leq k$. In particular, when $A = [I_s : O]$ where I_s is the $s \times s$ identity matrix and O is the $s \times (k-s)$ zero matrix, we are interested only in estimating the first s parameters. If $M(p)$ is non-singular, the variance matrix of the least squares estimator of $A\underline{\theta}$ is proportional to $AM^{-1}(p)A^T$. So a good design will be one which makes the matrix $AM^{-1}(p)A^T$ as small as possible. Specific criteria which have been proposed include the following.

D_A-optimality: The criterion function for this optimality is defined by

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

To emphasize the dependence of the design on the matrix of coefficients A, Sibson (1974) called this criterion D_A-optimality.

The cirterion is an increasing function over the set of positive definite symmetric matrices. ψ_{D_A} is concave and has unique partial derivatives corre-

sponding to positive weights.

D_s-optimality: This optimality is an important special case of D_A-optimality. Note that if $A = [I_s : O]$ and we partition the matrix $M(p)$ as follows:

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

then the matrix $(AM^{-1}(p)A^T)^{-1}$ can be expressed as $(M_{11} - M_{12}\, M_{22}^{-1}\, M_{12}^T)$. So maximizing ϕ_{D_A} in this particular case is equivalent to maximizing $\phi_{D_s}(p) = log\, det\{M_{11} - M_{12}\, M_{22}^{-1}\, M_{12}^T\}$, which is known as the D_s-optimal criterion; see Karlin and Studden (1966) and Atwood (1969). The properties of this criterion are similar to those of D_A-optimality.

There is a vast statistical literature on optimal design criteria. The above is a list of the most common design criteria. A detailed list of different criteria can be found in Atkinson et al. (2007), Pukelsheim (1993), Fedorov (1972) and Silvey (1980). The alphabetical nomenclature for different design criteria was introduced by Kiefer (1959).

Chapter 2

Optimality Conditions

2.1 Introduction

Recall that when we discussed different optimality criteria, we expressed each of the criterion functions as a function of p [that is, $\phi(p)$] or as a function of the information matrix $M(p)$ [that is, $\psi\{M(p)\}$]. Note that we focussed on an approximate design rather than an exact design. Recall that if we have only one design variable (x), an approximate design is written as

$$p \;=\; \left\{ \begin{array}{cccc} x_1 & x_2 & \ldots & x_J \\ p_1 & p_2 & \ldots & p_J \end{array} \right\} \tag{2.1}$$

where the first line gives the locations of the design points (on the design space $\mathcal{X}$) with p_j, the associated design weights. Also note that as p_j's are the design weights, p_j's must satisfy the following constraints: $p_j \geq 0$, and $\sum_{j=1}^{J} p_j = 1$. Thus, in any of the above optimality criteria, we optimize a criterion function,

say $\phi(p)$, subject to the constraints $p_j \geq 0$, and $\sum_{j=1}^{J} p_j = 1$.

Our goal is to find an optimal design according to a criterion function. That is, we wish to choose the proportion p_j of observations, taken at x_j to ensure good estimation of $\underline{\theta}$ by optimizing some criterion. Recall that we can view p as a probability distribution on the induced design space $\mathcal{V}$, where $\mathcal{V}$ consists of the vertices $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_J$. This also induces a probability distribution on the original design space $\mathcal{X}$. Thus we consider the above general problem and wish to find an optimizing distribution (say, p^*). There are a variety of problems in statistics, which demand the calculation of such probability distributions. Optimal regression design is a particular example. Other examples include parameter estimation and stratified sampling.

In order to find an optimal design or an optimizing distribution, we first need to determine conditions for optimality for the above optimization problem. The emphasis is on a differential calculus approach. An important tool is the directional derivative $F_\phi\{p, q\}$ (will be defined in Section 2.3) of a criterion function $\phi(p)$ at p in the direction of q. This plays an important simplifying role in the calculus of optimization.

Before we determine the optimality conditions, we need to consider a class of optimization problems.

2.2 Class of Optimization Problems

Problem (P1)

Maximize a criterion $\phi(p)$ over $\mathcal{P} \equiv \{p = (p_1, p_2, \ldots, p_J) : p_j \geq 0, \sum_{j=1}^{J} p_j = 1\}$. The equality constraint $\sum p_j = 1$ renders the problem a constrained optimization problem, the full constraint region being a closed bounded convex set.

Problem (P2)

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

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

where $G(.)$ is a given one to one function and $\mathcal{V} = \{\underline{v}_1, \ldots, \underline{v}_J\}$ is a known set of vector (or matrix) vertices of fixed dimension.

Note that we could alternatively state that $x(p) = E_p[G(\underline{v})]$, where $G(\underline{v})$ is a random variable assuming the value $G(\underline{v}_j)$ with probability p_j.

That is, solve Problem (P1) for

$$\phi(p) \;=\; \psi\{E_p[G(\underline{v})]\},\;\; x \;=\; E_p[G(\underline{v})] \;=\; \sum_{j=1}^{J} p_j G(\underline{v}_j). \tag{2.2}$$

Clearly an example of Problem (P1) or Problem (P2) is a general optimal linear regression design problem. Note that, as with our design problem, a

generalization of Problem (P2) would be to seek a probability measure defined on the induced design space $\mathcal{V}$ to maximize a function $\phi(p)$.

Now we consider optimality conditions for the above optimization problems. Note that there are two approaches which we could adopt in solving the problems. We could seek out an optimizing p^* directly or first determine an x^* maximizing $\psi(x)$ over $\mathcal{CH}\{\mathcal{G}(\mathcal{V})\}$ and then find a p^* such that $x(p^*) = x^*$. The former approach, which is the main we will consider, would require conditions explicitly defining an optimizing p^*.

We define optimality conditions in terms of point to point directional derivatives. There are two derivatives of interest. We define these in terms of a function $\phi(p)$.

2.3 Directional Derivatives

Directional Derivatives: $F_\phi\{p, q\}$

Making use of differential calculus, we exploit 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

$$F_\phi\{p, q\} = \lim_{\varepsilon \downarrow 0} \frac{\phi\{(1-\varepsilon)p + \varepsilon q\} - \phi(p)}{\varepsilon}. \tag{2.3}$$

It is a derivative which can exist even if $\phi(.)$ is not differentiable.

Gâteaux Derivative: $G_\phi\{p, m\}$

Another directional derivative of a criterion function $\phi(.)$ is defined by the following

$$G_\phi\{p, m\} = \lim_{\varepsilon \downarrow 0} \frac{\phi\{p + \varepsilon m\} - \phi(p)}{\varepsilon}. \tag{2.4}$$

$G_\phi\{p, m\}$ 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

$$F_\phi\{p, q\} \quad = \quad G_\phi\{p, m\} \tag{2.5}$$

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.

Kiefer (1974) used the concept of Gâteaux derivative in his design theory though he did not call it a directional derivative. Certainly it does not benefit from concavity of $\phi(.)$. However, we will see that this representation of $F_\phi\{p, q\}$ in terms of $G\{p, m\}$ is useful for studying their properties and for deriving the partial derivatives of a criterion function.

Now we discuss some properties of the above directional derivatives.

1. If $p, q \in S$, where S is a convex set, then so does $\{(1-\varepsilon)p+\varepsilon q\}$, which is

clearly an advantage if one wishes $F_\phi\{p, q\}$ only for $p, q \in S$. In contrast, $G_\phi(p, q)$ does not particularly benefit from such convexity.

2. $F_\phi\{p, q\} \geq \phi(q)$ - $\phi(p)$ if $\phi(.)$ is concave. This can be shown by using the definition of a concave function.

3. $F_\phi\{p, p\} = 0$, a desirable property since no change is effected in $\phi(.)$ if one does not move from p. In contrast, $G_\phi\{p, p\} = F_\phi\{p, 2p\} \neq 0$.

We have not so far made any assumptions about differentiability of the criterion function ϕ. When the criterion function ϕ is differentiable, it plays an important simplifying role in the optimization of ϕ. Now we proceed to redefine the concept in terms of $F_\phi\{p, q\}$. Consider the form of the directional derivative of a linear function $L(p) = a^T p + b$. Clearly,

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

Similarly,

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

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

Thus for $\phi(.)$ to be differentiable at p it must be that

$$F_\phi\{p,q\} \;=\; (q-p)^T \frac{\partial \phi}{\partial p} \;=\; (q-p)^T d \quad \text{for all } q$$

$$= \sum_{i=1}^{J} (q_i - p_i) d_i, \;\; d_i \,=\, \partial\phi/\partial p_i, \; i = 1, \ldots, J, \;\; d = \frac{\partial \phi}{\partial p}$$

or

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

It is worth mentioning that the condition on $G_\phi\{p,q\}$ is a familiar definition of differentiability.

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

$$F_\phi\{p,e_j\} \;=\; d_j \;-\; \sum_{i=1}^{J} p_i \, d_i \tag{2.6}$$

where e_j is the j^{th} unit vector in $\mathbb{R}^J$. Let us denote $F_\phi\{p,e_j\}$ by the simple notation F_j. We call F_j the vertex directional derivative of $\phi(.)$ at p. We determine the optimality conditions using this simplified form of vertex directional derivative F_j.

2.4 Vertex Direction Optimality Theorem

Recall that in Problem (P1) our problem is to maximize a criterion function $\phi(p)$ subject to $p_j \geq 0$, $j = 1, 2, \ldots, J$, $\sum p_j = 1$. If $\phi(p)$ is differentiable at an optimizing distribution p^*, then the first-order conditions for $\phi(p^*)$ to be a local maximum of $\phi(p)$ in the feasible region of the problem are

$$F_j^* = F_\phi\{p^*, e_j\} \begin{cases} = 0 & \text{for} \quad p_j^* > 0 \\ \leq 0 & \text{for} \quad p_j^* = 0. \end{cases} \tag{2.7}$$

If $\phi(p)$ is concave on its feasible region, then the first-order stationarity condition (2.7) is both necessary and sufficient for optimality, a result known as the general equivalence theorem in optimal design [Kiefer (1974)].

Other authors also derived or considered the theorem in general optimal design problem or in a more general setting than the design problem. See, for example, Whittle (1973). Some authors too have derived the theorem but using Lagrangian theory. See, for example, Sibson (1974) and Silvey and Titterington (1974). The general equivalence theorem plays an important role in constructing optimal designs, specifying a finite set of optimality conditions. It should be easy to check whether or not these are satisfied by a solution obtained by numerical techniques. Differentiability though is an essential requirement.

Next, we will consider construction of optimal designs and check that the optimal design satisfies the above general equivalence theorem.

Chapter 3

Construction of Optimal Designs for Polynomial Regression in One Variable

3.1 Introduction

As we discussed in Chapter 1, there are two ways of constructing an optimizing probability distribution. In optimal design context, one is by means of an exact design which finds the exact integer values of the numbers of trials (observations) (n_j's) at different design points or at the vertices of the induced design space. The other one is by means of an approximate design, which finds the corresponding proportions (p_j's) of observations so that $p_j \geq 0$, $\sum p_j = 1$; and $p = (p_1, p_2, \ldots, p_J)$ represents the resultant distribution on the induced design space $\mathcal{V}$. We focus on the latter approach as this is a simpler or more flexible problem to solve and yet one that is not much visibly different from

the original.

We first construct such approximate optimal designs by maximizing the D-optimal criterion subject to the basic constraints on the design weights. As mentioned before, this optimality criterion seeks to maximize the criterion function:

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

Because of this reciprocity property of the covariance matrix and the information matrix, maximizing the determinant of the information matrix is equivalent to minimizing the determinant of the covariance matrix. That is, in D-optimality, the generalized variance of the parameter estimates is minimized.

There are various motivations for D-optimality. These extend beyond the idea of point estimation and joint inference of the parameters $\underline{\theta}$. There is an interesting statistical interpretation of D-optimal design. If we assume normality of the errors in the linear model, then the general form of the joint confidence region for the vector of unknown parameters $\underline{\theta} \in \Theta$ is described by an ellipsoid of the form:

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

where $\underline{\hat{\theta}}$ is the least squares estimate or the maximum likelihood estimate of $\underline{\theta}$. The D-optimal criterion chooses $M(p)$ to make the volume of the above ellipsoid as small as possible because it is the case that this volume is proportional

to $[det\{M(p)\}]^{-\frac{1}{2}}$. The value of $[log\, det\{M(p)\}]$ is finite if and only if $M(p)$ is non-singular, i.e. when all the unknown parameters are estimable. This optimality criterion can be expressed in terms of the eigenvalues of the information matrix $M(p)$. Let the eigenvalues of $M(p)$ be $\lambda_1, \lambda_2, \ldots, \lambda_k$. Then, the eigenvalues of $M^{-1}(p)$ are $1/\lambda_1, 1/\lambda_2, \ldots, 1/\lambda_k$. These eigenvalues of $M^{-1}(p)$ are proportional to the squares of the lengths of the axes of the confidence ellipsoid. Thus, the D-optimal design minimizes the product of the eigenvalues of $M^{-1}(p)$, that is, $\prod_{i=1}^{k} 1/\lambda_i$.

This is the most important and popular design criterion in applications, and is most extensively studied of all design criteria. The references include Kiefer (1959), Fedorov (1972), Silvey (1980), Berger and Wong (2009), Atkinson and Donev (1992), Atkinson et al. (2007), Shah and Sinha (1989), Pukelsheim (1993), Mandal and Torsney (2006), Mandal et al. (2005), and Torsney (1983, 1988).

The D-optimality criterion $[\phi_D(p)]$ has several useful properties. First, note that, the D-optimality criterion is a concave function of the positive definite symmetric matrices.

The criterion function ϕ_D is differentiable whenever it is finite, and the first partial derivatives are given by

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

In many situations, we need to compare a certain design to the D-optimal design. Suppose that we have a design p for a given model of k parameters. Suppose the D-optimal design is p^*. Then, the relative efficiency of the design p with respect to the D-optimal design p^* is given by:

$$D_{eff} = \left\{ \frac{det\, M(p)}{det\, M(p^*)} \right\}^{1/k}. \tag{3.4}$$

We call this as D-efficiency of the design p. Taking the kth root of the above ratio of the determinants gives us an efficiency measure that is proportional to design size, irrespective of the dimension of the model.

Recall that we discussed a class of optimization problems in Chapter 2 and we referred Problem (P1) as our general problem. We also determined the optimality conditions using the directional derivatives and the general equivalence theorem. Now we will construct optimal designs satisfying the first-order conditions as discussed before. An analytic solution of the problem of constructing optimal designs is possible only in simple cases.

We first consider some polynomial regression problems for which explicit solutions can be obtained.

3.2 Constructing Designs in Polynomial Regression: Analytic Approach

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

$$E(y|x) \;=\; \theta_0 + \theta_1 x + \theta_2 x^2 + \ldots + \theta_{k-1} x^{k-1} = \underline{v}_x^T \underline{\theta} \tag{3.5}$$

where $\underline{v}_x = (1,\, x,\, x^2,\, \ldots, x^{k-1})^T$, $x \in [-1, 1]$ and $\underline{\theta} = (\theta_0,\, \theta_1,\, \ldots, \theta_{k-1})^T$. Also, $\underline{v}_x \in \mathcal{V} = \{\underline{v}_x \,:\, \underline{v}_x = (1,\, x,\, x^2,\, \ldots, x^{k-1})^T,\; -1 \leq x \leq 1\}$, the induced design space.

Here we have a standardized continuous design space. We construct D-optimal designs using the Legendre polynomial of Fedorov (1972). We construct the discrete D-optimal designs which is unique, having 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

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

Note 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 start with an example of simple linear regression model to demonstrate the method. Then, we will apply the method to quadratic and cubic model also. In the simple linear regression $k = 2$ in model (3.5). The model is

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

Simplifying the Legendre polynomial of (3.6), we obtain

$$(1 - x^2)P_1'(x) = (1 - x^2). \tag{3.7}$$

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

$$x = \pm 1. \tag{3.8}$$

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

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

Similarly, we can derive for the quadratic and cubic regression model. For the quadratic model, we take $k = 3$ in model (3.5) (on page 30). The model is

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

Therefor, the equation (3.6) simplifies to

$$(1 - x^2)P_2'(x) = 3x(1 - x^2). \tag{3.9}$$

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

$$x = \pm 1, 0. \tag{3.10}$$

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

$$p^* = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \end{array} \right\}.$$

For the cubic model, we take $k = 4$ in (3.5). The model is given by

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

So, the Legendre polynomial simplifies to

$$(1 - x^2)\, P_3'(x) \;\; = \;\; \tfrac{(15x^2-3)\,(1-x^2)}{2}. \tag{3.11}$$

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

$$x \;\; = \;\; \pm 1, \;\; \pm 1/\sqrt{5} \;\approx\; \pm 0.45. \tag{3.12}$$

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

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

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

3.3 A Radiation Dosage Example

We consider a practical problem of Berger and Wong (2009), in which a radiologist is interested in studying the linear effect of radiation dosage (X) on tumour shrinkage (Y). A simple linear regression model is considered for the relationship between these two variables. The mathematical equation of the simple linear regression is $y_i = \theta_0 + \theta_1 x_i + \epsilon_i$, where θ_0 is the intercept parameter, θ_1 is the slope parameter and ϵ_i are the random errors with zero mean and constant variance σ^2. We assume that the observations are independent. The tumour shrinkage for the ith patient is y_i for the radiation dose x_i. The dosage levels are given by 1 to 8 and the total number of patients (n) in the experiment is 16. Table 3.1 shows 8 dosage levels and the number of patients (n_i) assigned to each of levels. The summation of all n_i should equal to n. We wish to obtain an optimal design in this problem, that is, we wish to obtain the optimal dose levels and the corresponding number of patients.

Table 3.1: A radiation dosage design.

Dosage levels							
1	2	3	4	5	6	7	8
n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8

The design issue for this experiment is that we only have small sample of $n = 16$ patients. This means that we need to decide the number of patients assigned to each or some of the dosage levels so that the experiment becomes

more efficient. An optimal design is first determined by the choice of design points. In this example, we have 8 different design points, say, d_1, d_2, ..., d_8. Many designs could be considered with respect to different numbers of patients at each of the design points. For example, a simple design could be that each of the design points receives equal number of patients, that is, a balanced design. However, we need to find out an optimal design among all the designs.

It will be convenient if we first standardize the design space. First note that the design points d_1, d_2, ..., d_8 have a lower and upper limit, that is, $d_{min} \leq d_j \leq d_{max}$. Naturally the design points have to be selected within these two limits. The set of all design points is called the design space. We usually use a linear transformation to convert all design points to a standard region, such as between -1 and 1. Let the new transformed variable (design variable) be x. This can be done by the following linear transformation of the original design points $d_{min} \leq d_j \leq d_{max}$ to $x_{min} \leq x_j \leq x_{max}$.

$$x_j = \frac{d_j - \tilde{d}}{d_{max} - \tilde{d}}, \text{ where } \tilde{d} = \frac{d_{min} + d_{max}}{2}. \tag{3.13}$$

Now, applying the above transformation, we obtain the design points x_j's as: -1, -0.7143, -0.4286, -0.1429, 0.1429, 0.4286, 0.7143, 1.

Note that it is possible to convert the design points by some other transformations. However, some characteristics of the design may change after the transformation. Naturally then the optimal design may be different from the

original design. However, in case of D-optimality, the design is invariant under such transformation. We prove this in the following theorem.

Theorem 3.3. The D-optimal criterion is invariant under a non-singular linear transformation of the design space.

Proof. We know that the D-optimal criterion function is given by $\phi_D(p) = log\, det\{M(p)\}$, where $M(p)$ is the information matrix.

Let the design points in the original design be $d_1, d_2, \ldots, d_J$, and the corresponding vertices be $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_J$. Thus, the information matrix can be written as $M(p) = VPV^T$ [see (1.11) on page 10].

Let $\mathcal{V} = [\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_J]$ and $\mathcal{W} = [\underline{\omega}_1, \underline{\omega}_2, \ldots, \underline{\omega}_J]$ be the induced design spaces corresponding to the design points d_j's and x_j's respectively. That is, $\mathcal{V} = [\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_J]$ is transformed to $\mathcal{W} = [\underline{\omega}_1, \underline{\omega}_2, \ldots, \underline{\omega}_J]$ under the linear transformation $\underline{\omega}_j = A\underline{v}_j$, where A is a $k \times k$ non-singular matrix, and k is the number of parameters in the regression model.

Then a design assigning weight p_j to $\underline{\omega}_j$ has information matrix given by

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

Now, the D-optimal criterion for the transformed variable can be written as

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

Thus, maximizing $\phi_D\{M_\omega(p)\}$ is equivalent to maximizing $\phi_D\{M(p)\}$. Hence the theorem.

3.3.1 Estimation of parameters and optimal design

In this section, we determine the optimal design to estimate the parameters θ_0 and θ_1. We discuss the role of optimal design to estimate the parameters as efficiently as possible. It is well known that the least squares estimators $\hat{\theta}_0$ and $\hat{\theta}_1$ can be found by the following two mathematical equations:

$$
\hat{\theta}_0 = \bar{y} - \hat{\theta}_1\bar{x} \;\text{ and }\; \hat{\theta}_1 = \frac{cov(x,y)}{var(x)} = \frac{SS_{xy}}{SS_x}, \tag{3.14}
$$

where the sum of squares between x and y is $SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$ and the sum of squares of x is $SS_x = \sum (x_i - \bar{x})^2$. Moreover, the least squares estimators $\hat{\theta}_0$ and $\hat{\theta}_1$ is unbiased, which means $E(\hat{\theta}_0) = \theta_0$ and $E(\hat{\theta}_1) = \theta_1$. The variances of the two estimators $\hat{\theta}_0$ and $\hat{\theta}_1$ are

$$
var(\hat{\theta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right) \text{ and } var(\hat{\theta}_1) = \frac{\sigma^2}{SS_x} = \frac{\sigma^2}{n\, var(x)}. \tag{3.15}
$$

Usually, The estimators will be more efficient as the variances of the estimators get smaller. In particular, we are interested in minimizing the variance of the estimate of the slope parameter θ_1. Now, we try to make a connection between this variance and the D-optimal criterion. For the above model, the covariance matrix of $\hat{\underline{\theta}}$ can be obtained as:

$$cov(\hat{\underline{\theta}}) = \frac{\sigma^2}{n\, SS_x} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}. \tag{3.16}$$

The determinant of this covariance matrix can be simplified as

$$det\, cov(\hat{\underline{\theta}}) = \frac{\sigma^4}{n\, SS_x}. \tag{3.17}$$

As we mentioned before, because of the reciprocity property, maximizing the determinant of the information matrix is equivalent to minimizing the determinant of the covariance matrix. In other words, in D-optimality we can minimize the determinant of the covariance matrix. In this example, we consider the above determinant $[det\, cov(\hat{\underline{\theta}})]$ as our D-criterion.

Now, looking at the expressions of (3.15) and (3.17) we observe that the variance $var(\hat{\theta}_1)$ or the D-optimal criterion decreses when the variance of the error terms σ^2 decreases and/or the variation SS_x increases and/or the sample size n increases.

From the above statement, a simple way to increase the efficiency of the estimator or the D-optimal criterion is to increase the sample size n. This is usually done in practice. However, this will lead to additional costs of collecting

the data. Another way is to reduce the measurement error by having more precise measurement of the variables to reduce σ^2. This will lead to additional costs too.

Based on the total control of the sampling and selection process, we have yet another way to improve efficiency of the estimators without additional cost. We can increase SS_x by selecting the values of the independent variable to make the estimator be more efficient. This is one of the roles of optimal design by selecting the optimal values of the independent variable.

In Table 3.2 (on page 39), we have four different designs of the radiation-dosage example for studying the effect of radiation dosage on the reduction of tumours in breast cancer patients. In this table, Design 1 is the original design and the 16 patients are randomly assigned to the 8 different dosage levels. In Design 2, each dosage level is assigned to two patients. In Design 3, more patients are assigned to smaller and larger dosages. Design 4 assigns patients only at the extreme dosage levels. For each of these designs we calculate SS_x, $var(\hat{\theta}_1)$ and the D-optimal criterion. We see that Design 4 has the highest SS_x (196) among all the four designs. Also Design 4 has the smallest $var(\hat{\theta}_1)$ ($0.0051\ \sigma^2$) and smallest D-optimal criterion ($3.1875\ 10^{-4}\ \sigma^4$) among these four designs. Thus, Design 4 is the optimal design among these four designs. This design consists of two distinct design points (dosage levels) and assigns $n/2$ (50%) patients to each of these two dosage levels. Thus, this optimal

Table 3.2: Four designs assigning to eight different radiation dosage levels.

Design points d_j	*Design* 1	*Design* 2	*Design* 3	*Design* 4
1	2	2	3	8
2	1	2	2	0
3	2	2	2	0
4	3	2	1	0
5	4	2	1	0
6	1	2	2	0
7	1	2	2	0
8	2	2	3	8
n	16	16	16	16
SS_x	70	84	108	196
$var(\hat{\theta}_1)$	0.0143	0.0119	0.0093	$0.0051(\sigma^2)$
D-criterion	$8.9375\ 10^{-4}$	$7.4375\ 10^{-4}$	$5.8125\ 10^{-4}$	$3.1875\ 10^{-4}\ (\sigma^4)$

design can be written as

$$p^* = \left\{ \begin{array}{cc} 1 & 8 \\ 0.5 & 0.5 \end{array} \right\}. \tag{3.18}$$

In terms of the transformed variable x, this optimal design can be written as

$$p^* = \left\{ \begin{array}{cc} -1 & 1 \\ 0.5 & 0.5 \end{array} \right\}. \tag{3.19}$$

At this point, a natural question arises that if we consider some other designs in this problem, is Design 4 still optimal? The answer is 'yes'. Note that optimality of the design depends on the class of designs under consideration. We will show that (later in this chapter) the Design 4 is optimal among all designs on the given design space for estimating the slope parameter as well as in terms of D-optimality. A quick check would be to calculate the vertex

directional derivatives at the two design points and see whether they satisfy the first-order conditions [as given in (2.7) on page 25].

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

To do this, we first calculate the partial derivatives

$$d_1 = \frac{\partial \phi_D}{\partial p_1} = \underline{v}_1^T \, M^{-1}(p) \, \underline{v}_1 = 2, \;\; d_2 = \frac{\partial \phi_D}{\partial p_2} = \underline{v}_2^T \, M^{-1}(p) \, \underline{v}_2 = 2.$$

Then we calculate the directional derivatives

$$F_1 = d_1 - \sum p_i \, d_i = 0, \;\; F_2 = d_2 - \sum p_i \, d_i = 0.$$

We see that the directional derivatives at the two points are zero. This indicates that Design 4 is optimal.

Typically, the optimal support must in a sense be computed, as a prelude to determining the optimal weights and this is essentially done by some algorithms. Hence we consider the following class of algorithms, indexed by a function which satisfies certain conditions and depends on one or more free parameters.

3.4 A Class of Algorithms

Problems (P1) and (P2) (see Chapter 2) have a distinctive set of constraints, namely the variables $p_1, p_2, \ldots, p_J$ must be nonnegative and sum to 1. An

iteration which neatly submits to these and has some suitable properties is the following multiplicative algorithm:

$$p_j^{(r+1)} = \frac{p_j^{(r)} f(x_j^{(r)})}{\sum_{i=1}^{J} p_i^{(r)} f(x_i^{(r)})} \tag{3.20}$$

where $x_j^{(r)} = d_j^{(r)}$,

$$d_j^{(r)} = \left.\frac{\partial \phi}{\partial p_j}\right|_{p=p^{(r)}} \text{(partial derivatives at } r^{th} \text{ iteration, i.e., at } p = p^{(r)}),$$

and the function $f(x)$ satisfies the following conditions:

(i) $f(x)$ is positive;

(ii) $f(x)$ is strictly increasing in x.

The function $f(x)$ may depend on a free positive parameter δ.

3.4.1 Properties of the Algorithms

Under the conditions imposed on $f(x)$, iterations under (3.20) possess the following properties.

(a) $p^{(r)}$ is always feasible. We find the optimal solution in the feasible region.

(b) Under the above iteration $supp(p^{(r+1)}) \subseteq supp(p^{(r)})$, but some weights can converge to zero.

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

This type of iteration was first proposed by Torsney (1977), taking $f(d) = d^\delta$, with $\delta > 0$. Subsequent empirical studies include Silvey et al. (1978), which is a study of the choice of δ when $f(d) = d^\delta$, $\delta > 0$; Torsney (1988), which mainly considered $f(d) = e^{\delta d}$ in a variety of applications, for which one criterion $\phi(p)$ could have negative derivatives; Mandal et al. (2005) and Torsney and Mandal (2001) consider constrained optimal design problems.

Other iterations for problems like (P2) have been proposed in the literature. Vertex direction algorithms which perturb one p_j and change the others proportionately were proposed by Fedorov (1972) and Wynn (1972). These are useful when many of the p_j are zero at the optimum as happens in regression design problems. Certainly some modification would be needed if there are many zero optimal weights. Mandal and Torsney (2006) explore one such modification based on a clustering approach. This is related to the fact that the support points of a discretized design space can be viewed as consisting of some clusters of points. These clusters begin to emerge in early iterations of algorithm (3.20). At this point the current set of weights are transformed to weights within clusters and total cluster weights. Optimal values of these are then sought using a modified version of the algorithm. They explore this idea through several regression models and enjoy improved convergence.

3.5 Constructing Designs in Polynomial Regression: Algorithmic Approach

Polynomial regression models in one design variable was given in (3.5) and discussed in Section 3.2. In that section we constructed D-optimal designs by some analytic approach. In this section, we construct those designs using the class of multiplicative algorithms. We also propose some useful strategies for better convergence of the algorithms.

Recall that Problem (P1) provides some constrained optimization problems having the single linear equality constraint $\sum p_j = 1$ as well as $p_j \geq 0$, $j = 1, 2, \ldots, J$, the full constraint region being a closed bounded convex set. Therefore we have an example of problem (P1) with

- $\phi_D(p) = log\, det\{M(p)\}$
- $M(p) = \sum_{j=1}^{J} p_j \underline{v}_j \underline{v}_j^T$.

It is also an example of Problem (P2) where

(i) $G(\underline{v}) = \underline{v}\,\underline{v}^T$;

(ii) $\underline{v} \in \mathcal{V} \subseteq \mathbb{R}^k$, $\mathcal{V}$ is the induced design space;

(iii) $x\ [= x(p) = M(p)]$ is a symmetric $k \times k$ matrix and

(iv) $\psi(x) = log\{det(x)\}$ $[= \phi(p)]$.

We now propose some important strategies for constructing the optimal designs and for better convergence of the algorithms (3.20). Convergence of the algorithm (3.20) could be slow if we do not choose the function $f(.)$ and its arguments in an objective way. Motivated by this fact, we now attempt to improve convergence by considering some choices of $f(.)$ for which we replace d_j by F_j. Some of the choices of $f(.)$ may not be good because D-optimal derivatives are positive and centred on k. For example, note that 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' on 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; in particular $\delta = 1$ for D-optimality and $\delta = 1/2$ for c-optimality yield monotonic iterations. Thus, we need to be careful in choosing the values of δ as well. Now, in algorithm (3.20), we replace d_j by F_j and prove the following important result.

Theorem 3.5. Under the conditions imposed on $f(.)$, we always have $F_{\phi}\{p^{(r)}, p^{(r+1)}\} \geq 0$, where $F_{\phi}\{p^{(r)}, p^{(r+1)}\}$ is the directional derivative of $\phi(.)$ at the current iteration $p^{(r)}$ in the direction of the next iteration $p^{(r+1)}$.

Proof. The inequality property can be seen by letting a random variable F take

the value F_j, the vertex directional derivatives with probability p_j ($p_j = p_j^{(r)}$).

$$
\begin{aligned}
F_\phi\{p^{(r)},\, p^{(r+1)}\} &= \Big[p^{(r+1)} - p^{(r)}\Big]^T \underline{d} \;=\; \sum_{i=1}^{J} \Big[p_i^{(r+1)} - p_i^{(r)}\Big] d_i \\
&= \sum_{i=1}^{J} \Big[p_i^{(r+1)} - p_i^{(r)}\Big](d_i - \bar{d}) \;=\; \sum_{i=1}^{J} \Big[p_i^{(r+1)} - p_i^{(r)}\Big] F_i \\
&= \frac{\sum_{i=1}^{J} p_i\, f(F_i, \delta)\; F_i}{\sum_{i=1}^{J} p_i\, f(F_i, \delta)} - \sum_{i=1}^{J} p_i\, F_i \\
&= \frac{\Big[\sum_{i=1}^{J} p_i\, f(F_i, \delta)\; F_i\Big] \;-\; \Big[\sum_{i=1}^{J} p_i\, F_i\Big]\Big[\sum_{i=1}^{J} p_i\, f(F_i, \delta)\Big]}{\sum_{i=1}^{J} p_i\, f(F_i, \delta)} \\
&= \frac{Cov\Big[F, f(F, \delta)\Big]}{E\Big[f(F, \delta)\Big]}.
\end{aligned}
$$

Now, we know that the function $f(F, \delta)$ is increasing in F. So the covariance between F and $f(F, \delta)$ is positive. We also know that the function $f(F, \delta)$ is positive. Hence the theorem.

Note that any criterion has both positive and negative vertex directional derivatives. So the function $f(F, \delta)$ needs to be defined for positive and negative values of F. From equation (2.6), 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 a suitable function is one that is centred at zero and changes reasonably quickly around $F = 0$. It should also be desirable to treat positive and negative F_j's symmetrically, at least when all p_j^*'s are positive.

Based on the above, we now construct the optimal design for various polynomial regression models in one design variable.

3.5.1 Simple linear regression

We revisit the simple linear regression model of the radiation-dosage example of Section 3.3. The model again is given by

$$E(y|x) \; = \; \theta_0 + \theta_1 x. \tag{3.21}$$

We consider the standardized design interval as $-1 \leq x \leq 1$. We first discretize the design space to be in some form of uniform grid on the continuous design space. In particular, we approximate the design interval by a grid of points equally spaced at intervals of 0.1.

We first take the argument x as the partial derivative of the criterion function and consider five choices of $f(x)$, such as $f(d) = d^{\delta}$, $f(d) = exp(\delta d)$, $f(d) = \ln(e + \delta d)$, $f(d) \; = \; \exp(\delta d) \, / \, [1 + \exp(\delta d)]$, the logistic c.d.f. and $f(d) = \Phi(\delta d)$, the normal c.d.f. We consider appropriate choices of the

parameter δ in each case and run the algorithm (3.20) until the first-order conditions are satisfied. In Tables 3.3 we report the number of iterations needed to achieve $\underset{1\leq j\leq J}{max}\{F_j\} \leq 10^{-n}$, for $n = 1, 2, 3, 4, 5$, where F_j are the directional derivatives. We started with the equal initial weights, that is, $1/J,\ j = 1, 2, \ldots, J$. The results were obtained by using the software R version 2.15.0 . Under each choice of $f(x)$, the best choices of δ are given in bold font. The best choices δ correspond to the least number of iterations.

Among the choices of $f(d)$ we considered, the two choices $f(d) = d^{\delta}$ and $f(d) = \exp(\delta d)$ come out to be better than the others. For example, with $f(d) = d^{\delta}$, $\delta = 2.0$, the number of iterations need to achieve $n = 5$ is 50. Look the contrast for choosing the value of δ. For example, with $f(d) = \ln(e + \delta d)$, $\delta = 0.1$, the number of iterations need to achieve $n = 5$ is 1545, whereas for $\delta = 2.0$, the number of iterations need to achieve $n = 5$ is 312. We see similar things happen for other choices as well.

Now we attempt to improve the convergence of the algorithm by objectively choosing the function $f(x)$ and its argument. Improving the convergence is always better to save the time and costs of the experiment. As we proposed before, we replace the partial derivatives d_j by the directional derivatives F_j for some suitable choices of $f(x)$. One of the reasons for this replacement is that $\sum p_j F_j = 0$. That is, when we replace d_j by F_j, we should choose the function $f(x)$ in such a way that the function is centred at zero and changes reasonably

Table 3.3: Number of iterations needed to achieve the first-order conditions, with $x = d$ (Simple linear regression)
$f(d) = d^{\delta}$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.1	95	283	508	738	968
0.2	48	142	255	370	485
0.8	13	36	65	93	122
0.9	12	33	58	83	109
1.0	11	29	52	75	98
1.1	10	27	47	68	89
1.2	9	25	44	63	82
1.7	7	18	31	45	58
1.8	7	17	29	42	55
1.9	6	16	28	40	52
2.0	**6**	**15**	**27**	**38**	**50**

$f(d) = \exp(\delta d)$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.1	47	138	254	375	496
0.2	24	69	128	188	248
0.3	16	47	85	126	166
0.7	7	20	37	54	72
0.8	7	18	33	48	63
0.9	6	16	29	42	56
1.0	**5**	**15**	**26**	**38**	**50**

$f(d) = [\exp(\delta d)] / [1 + \exp(\delta d)]$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.1	103	307	565	831	1099
0.2	59	174	318	466	615
0.5	37	106	191	278	365
0.6	36	104	185	268	352
0.7	**36**	**105**	**186**	**268**	**351**
0.8	38	108	191	276	360
1.0	43	123	215	308	402
1.1	47	134	233	333	434
1.7	94	266	454	642	830
1.8	107	304	518	731	945
1.9	123	349	593	836	1079
2.0	142	402	681	960	1238

$f(d) = \ln(e + \delta d)$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.1	145	431	794	1168	1545
0.2	83	246	451	663	875
0.8	38	112	202	294	387
0.9	37	107	194	282	371
1.0	36	104	187	273	358
1.1	35	101	182	265	348
1.2	34	99	178	259	340
1.7	32	93	167	242	318
1.8	32	92	165	240	315
1.9	32	92	164	239	313
2.0	**31**	**91**	**164**	**238**	**312**

quickly about $F = 0$. With this change, the algorithm (3.20) becomes

$$p_j^{(r+1)} = \frac{p_j^{(r)} f(F_j^{(r)})}{\sum_{i=1}^{J} p_i^{(r)} f(F_i^{(r)})} \tag{3.22}$$

where $F_j^{(r)}$ are the vertex directional derivatives at $p_j^{(r)}$.

We consider two choices of the function $f(x)$ with the potential to satisfy these requirements. The choices are $f(F) = \Phi(\delta F)$ and $f(F) = \frac{\exp(\delta F)}{1+\exp(\delta F)}$. The results are reported in Table 3.4. Comparing Table 3.3 and Table 3.4, we see that we improve the convergence quite a lot. For example, with $f(d) = [\exp(\delta d)] / [1 + \exp(\delta d)]$, for $\delta = 0.7$ and $n = 5$, the number of iterations needed is 351 (see Table 3.3), whereas using the directional derivatives, for $\delta = 2.0$ and $n = 5$, the number of iterations needed is only 50 (see Table 3.4). We even get better improvement for the other choice. With $f(F) = \Phi(F)$, for $\delta = 2.0$ and $n = 5$, the number of iterations needed is only 31.

The solutions converged to the optimal design

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

which is exactly the same as we obtained using the analytic approach in Section 3.2 as well as the radiation dosage example in Section 3.3.

Now we confirm this D-optimal design by plotting the standardized variance of the predicted response on y at x ($d(x, p^*)$) versus the the design variable x. This is given in Figure 3.1. We see the variance function attains the

maximum at 2, the number of parameters in the SLR model.

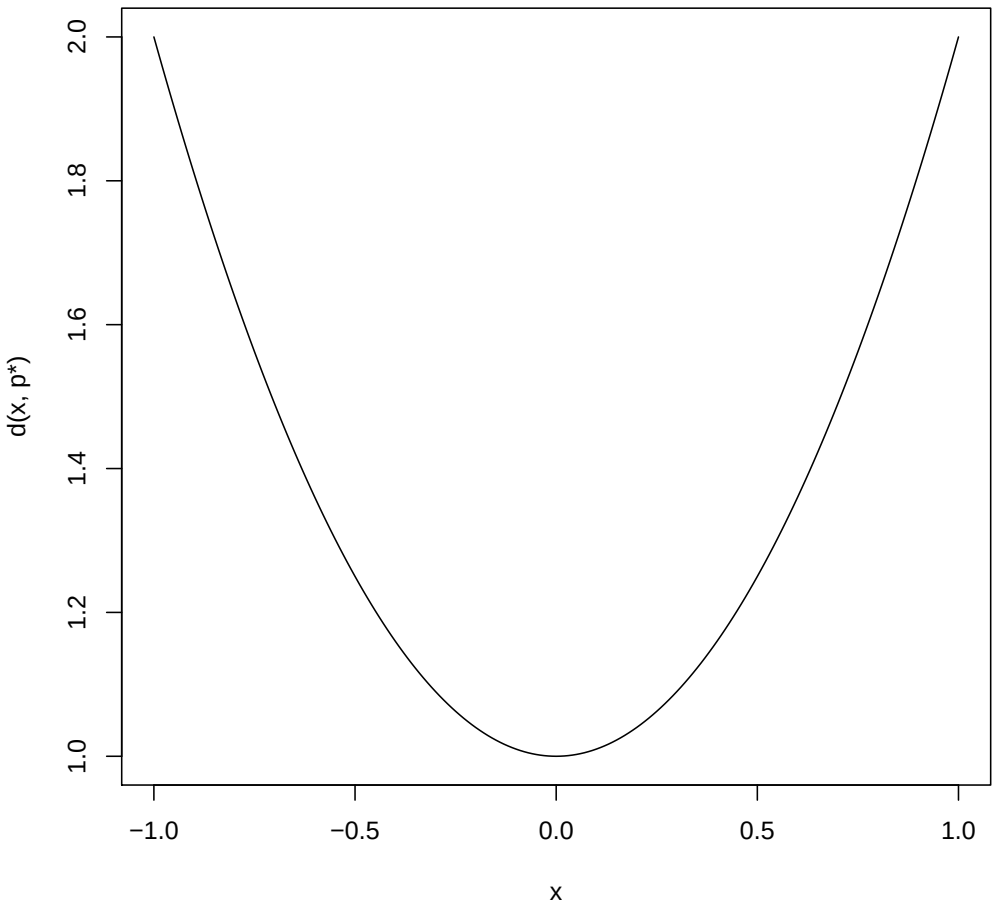
Table 3.4: Number of iterations needed to achieve the first-order conditions, with $x = F$ (simple linear regression)
$f(F) = \exp(\delta F)/[1 + \exp(\delta F)]$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.1	93	276	508	748	989
0.2	47	139	255	375	494
0.8	13	37	65	94	124
0.9	12	33	58	84	110
1.0	11	30	53	76	99
1.1	10	28	48	69	90
1.7	8	19	32	45	58
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
0.2	30	88	160	235	310
0.8	9	24	42	60	77
0.9	9	22	37	53	69
1.0	8	20	34	48	62
1.1	8	19	31	44	56
1.7	6	13	21	29	37
1.8	6	13	20	27	35
1.9	6	12	19	26	33
2.0	**6**	**12**	**18**	**25**	**31**

Figure 3.1: The standardized variance of the predicted response on y at x ($d(x, p^*)$) for the SLR model.



3.5.2 Quadratic regression model

We take $k = 3$ in (3.5) and consider the quadratic regression model as given by

$$E(y|x) \;=\; \theta_0 + \theta_1 x + \theta_3 x^2, \tag{3.23}$$

with the design interval $-1 \leq x \leq 1$. Similar to the previous model, here also we first discretize the design interval by a grid of points equally spaced at intervals of 0.1 between -1 and 1. We first take the argument x as the partial derivative of the criterion function and consider five choices of $f(x)$ as considered in the previous section. We consider appropriate choices of the parameter δ in each case again and run the algorithm (3.20) until the first-order conditions are satisfied.

We start with the equal initial weights, that is, $1/J,\ j = 1, 2, \ldots, J$. In Tables 3.5 we report the number of iterations needed to achieve $\underset{1\leq j\leq J}{max}\{F_j\} \leq 10^{-n}$, for $n = 1, 2, 3, 4, 5$. Among the choices of $f(d)$ we considered, the two choices $f(d) = d^{\delta}$ and $f(d) = \exp(\delta d)$ come out to be better. For example, with $f(d) = d^{\delta}$, $\delta = 1.9$, the number of iterations need to achieve $n = 5$ is 320. Here again note the contrast for choosing the value of δ. For example, with $f(d) = \exp(\delta d)/[1 + \exp(\delta d)]$, $\delta = 2.0$, the number of iterations need to achieve $n = 5$ is 39722, whereas for $\delta = 0.4$, the number of iterations need to achieve $n = 5$ is 2181. We can see similar results for other choices of $f(d)$ as well.

Table 3.5: Number of iterations needed to satisfy the required directional derivatives with $x = d$ (Quadratic model)
$f(d) = \Phi(\delta d)$

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**
0.4	54	490	1125	1711	2293
0.5	64	621	1421	2158	2890
0.8	150	2369	5385	8150	10893
0.9	273	4516	10254	15503	20708

$f(d) = \frac{\exp(\delta d)}{1+\exp(\delta d)}$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.1	119	1001	2320	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
0.6	56	505	1160	1766	2367
0.9	77	760	1739	2642	3537
1.0	88	908	2076	3152	4220
1.1	101	1101	2513	3815	5106
1.8	320	5288	12035	18225	24366
1.9	405	6745	15348	23238	31063
2.0	514	8630	19634	29719	39722

$f(d) = \exp(\delta d)$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.1	51	426	988	1509	2026
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**

$f(d) = d^{\delta}$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.1	145	1285	2962	4515	6055
0.2	73	643	1482	2258	3028
0.7	22	184	424	646	866
0.8	19	161	371	565	758
0.9	17	144	330	502	674
1.0	16	129	297	452	606
1.1	14	118	270	411	551
1.2	13	108	248	377	505
1.6	10	81	186	283	379
1.7	9	76	175	266	357
1.8	7	72	165	252	337
1.9	**9**	**69**	**157**	**239**	**320**

$f(d) = ln(e + \delta d)$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.1	168	1420	3292	5027	6750
0.2	100	849	1965	3000	4027
0.7	54	463	1069	1631	2188
0.8	52	447	1032	1574	2111
0.9	51	436	1005	1532	2055
1.0	50	427	985	1502	2014
1.1	49	421	970	1479	1984
1.2	48	416	959	1462	1961
1.7	47	406	935	1424	1910
1.8	47	405	933	1422	1907
1.9	**47**	**405**	**933**	**1421**	**1906**

In the quadratic model, we also replace d_j by F_j to reduce the number of iterations, that is, to obtain the better convergence of the algorithm. We again focus on the two choices $f(F) = \Phi(\delta F)$ and $f(F) = \frac{\exp(\delta F)}{1+\exp(\delta F)}$. The results are reported in Table 3.6. We now compare Table 3.5 and Table 3.6 and see that we improve the convergence compared to choosing the partial derivatives. For example, with $f(d) = [\exp(\delta d)] / [1 + \exp(\delta d)]$, for $\delta = 0.4$ and $n = 5$, the number of iterations needed is 2181 (see Table 3.5), whereas using the directional derivatives, for $\delta = 1.3$ and $n = 5$, the number of iterations needed is 311 (see Table 3.6). Similar things happen in the results for $f(F) = \Phi(\delta F)$.

The solutions converged to the optimal design

$$p^* = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \end{array} \right\},$$

the same solution as we obtained using the analytic approach in Section 3.2.

We also confirm this D-optimality and plot the standardized variance of the predicted response versus the the design variable x. This is given in Figure 3.2. We see the variance function attains the maximum at 3, the number of parameters in this model.

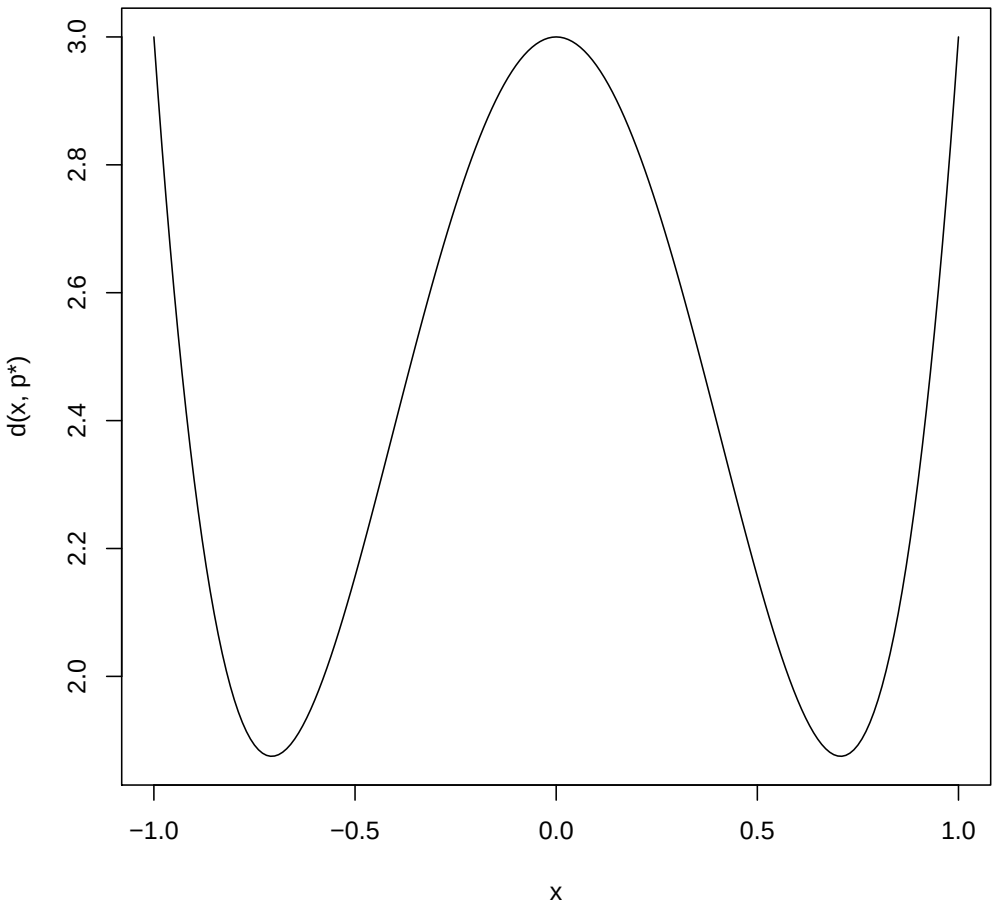
Table 3.6: Number of iterations needed to satisfy the required directional derivatives with $x = F$ (Quadratic model)
$f(F) = \frac{\exp(\delta F)}{1+\exp(\delta F)}$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.1	101	852	1975	3016	4049
0.2	51	427	988	1508	2024
0.3	34	285	659	1005	1349
0.4	26	214	494	754	1011
0.5	21	172	396	603	809
0.6	17	144	330	503	674
0.7	15	123	283	431	577
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
0.2	32	268	619	945	1268
0.3	22	179	413	630	845
0.4	16	135	310	472	633
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**

Figure 3.2: The standardized variance of the predicted response on y at x ($d(x, p^*)$) for the quadratic model.



3.5.3 Cubic regression model

The cubic model is given by

$$E(y|x) = \theta_0 + \theta_1 x + \theta_3 x^2 + \theta_4 x^3, \tag{3.24}$$

by taking $k = 4$ in the polynomial regression model (3.5). We consider the design interval to be $-1 \leq x \leq 1$. Similar to the previous models, we first discretize the design space to be in some form of uniform grid on the continuous design space. However, for this model, we change the increment of x from 0.1 to 0.01 for more accuracy.

We first take the argument x as the partial derivative of the criterion function and consider five choices of $f(x)$ as considered in the previous models. We consider appropriate choices of the parameter δ in each case again and run the algorithm (3.20) until the first-order conditions are satisfied. We start with the equal initial weights, and in Table 3.7 we report the number of iterations needed to achieve the first-order conditions for the choice of the argument as the partial derivatives. Here again we improve the convergence by replacing d_j by F_j. In Table 3.8 we report the corresponding iteration results.

Now if we look at the results of the above two tables, it is clear that the convergence is improved a great deal by replacing the argument with the directional derivatives. For example, with $f(d) = \Phi(\delta d)$, for $\delta = 0.2$ and $n = 5$, the number of iterations needed is 94543, whereas using the directional derivatives, for $\delta = 0.62$ and $n = 5$, the number of iterations needed is only

14051. We find similar results if we look at the results for other choices as well.

Interestingly we notice that the support points are viewed as consisting of clusters of points. For this particular model, there are two clusters centred 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:

$$p = \left\{ \begin{array}{cccccc} -1.00 & -0.45 & -0.44 & 0.44 & 0.45 & 1.00 \\ 0.25 & 0.224186 & 0.025816 & 0.025816 & 0.224186 & 0.25 \end{array} \right\}.$$

This suggests that the solution for the continuous design 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

$$\frac{(-0.45)(0.224186) + (-0.44)(0.025816)}{0.224186 + 0.025816} \approx -0.45$$

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

$$p^* = \left\{ \begin{array}{cccc} -1.00 & -0.45 & 0.45 & 1.00 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{array} \right\}.$$

We again confirm this D-optimality and plot the standardized variance of the predicted response versus the the design variable x. This is given in Figure

3.3. We see the variance function attains the maximum at 4, the number of parameters in this model.

Figure 3.3: The standardized variance of the predicted response on y at $\underline{x}$ ($d(x, p^*)$) for the cubic model.

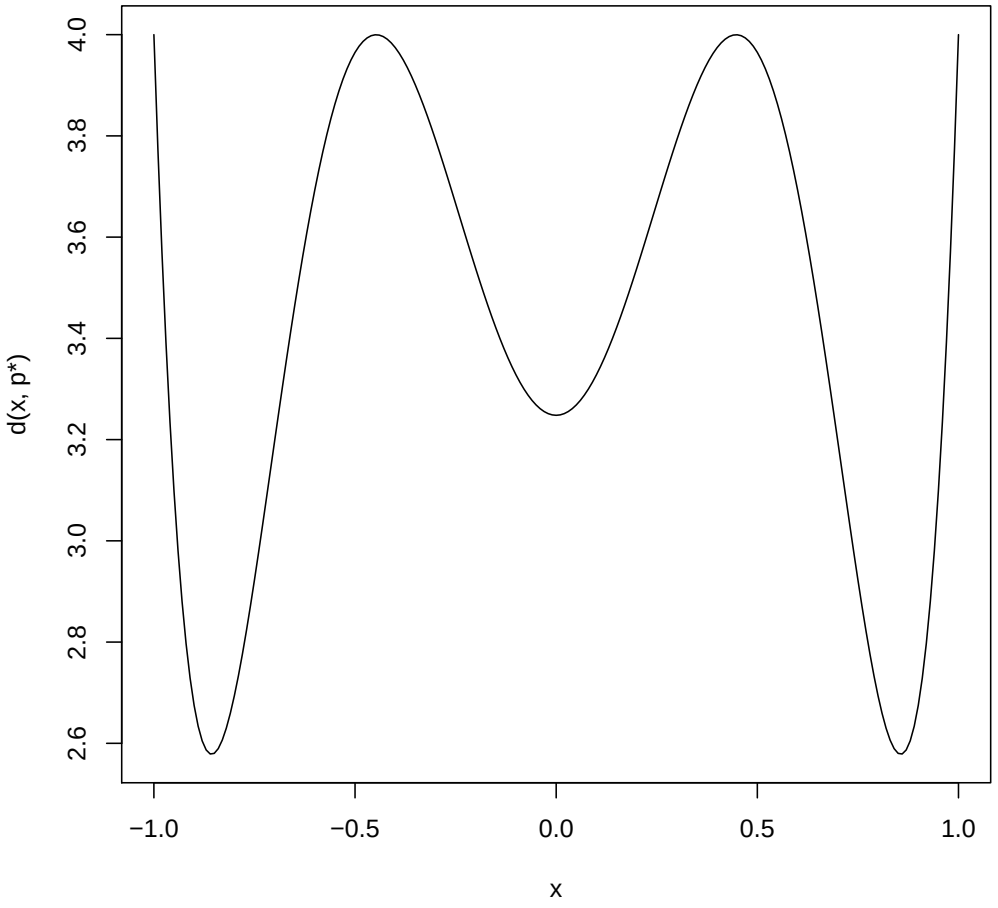


Table 3.7: Number of iterations needed to satisfy the required directional derivatives with $x = d$ (Cubic model)
$f(d) = d^{\delta}$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
1.9	19	104	961	5148	14632
1.92	23	103	951	5095	14480
1.96	37	87	932	4991	14185
1.98	65	165	923	4941	14041
1.99	**115**	**317**	**533**	**4916**	**13971**
2.0	9777	112715			

$f(d) = \frac{\exp(\delta d)}{1+\exp(\delta d)}$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.2	80	785	7353	39428	112085
0.3	**73**	**702**	**6568**	**35208**	**100084**
0.4	79	726	6790	36391	103439

$f(d) = \exp(\delta d)$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.2	26	243	2279	12222	34747
0.3	17	162	1519	8148	23165
0.31	17	157	1470	7885	22417
0.35	14	138	1302	6983	19855
0.36	14	134	1265	6789	19303
0.37	**13**	**130**	**1230**	**6605**	**18781**

$f(d) = \Phi(\delta d)$

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

$f(d) = ln(e + \delta d)$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
1.4	62	614	5741	30776	87482
1.5	**62**	**614**	**5741**	**30773**	**87473**
1.6	62	614	5746	30798	87546

Table 3.8: Number of iterations needed to satisfy the required directional derivatives with $x = F$ (Cubic model)
$f(F) = \Phi(\delta F)$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.6	14	103	954	5109	14519
0.61	14	101	939	5026	14281
0.62	**14**	**95**	**924**	**4945**	**14051**

$f(F) = \frac{\exp(\delta F)}{1+\exp(\delta F)}$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.7	17	140	1304	6987	19858
0.8	16	123	1142	6114	17377
0.9	**14**	**109**	**1015**	**5435**	**15446**
1.0	14	79	3305	98785	

$f(F) = exp(\delta F)$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.3	17	162	1519	8148	23165
0.35	14	138	1302	6983	19855
0.36	14	134	1265	6789	19303
0.37	**13**	**130**	**1230**	**6605**	**18781**

Chapter 4

Construction of Optimal Designs for Polynomial Regression in Two Variables

4.1 Introduction

So far we have considered regression models in one design variable, namely, the simple linear regression, the quadratic regression and the cubic regression models. However, in many regression models we need to deal with more than one design variables such as multiple linear regression or polynomial regression in two or more variables.

In optimal design context, values or levels of inputs (the design variables) must be chosen before running an experiment and observing a measurement on some variable (the response variable) of interest. There will be a set of or several combinations of the inputs allowed. We must decide how many

observations to take at each combination of inputs. So, there are two important questions for the design of these kind of models.

The first question relates to the number of levels for each of the independent variables, that is, the design variables. Now the question is, how many combinations of levels of the variables should we choose and how do we choose the combinations of levels to have an efficient design.

Naturally the second question is, how many units should be assigned to the combinations of levels of the variables and how do we choose the number of units to have an efficient design.

We answer these two questions by considering several regression models in two variables in this chapter. We first introduce the models.

4.1.1 Polynomial regression in m variables

The second-order polynomial in m design variables is

$$E(Y|\underline{x}) = \theta_o + \sum_{j=1}^{m} \theta_j x_j + \sum_{j=1}^{m-1} \sum_{k=j+1}^{m} \theta_{jk} x_j x_k + \sum_{j=1}^{m} \theta_{jj} x_j^2, \tag{4.1}$$

where Y is the response variable and $\underline{x}$ are design variables. From the model we see that this is a $(m+1)(m+2)/2$ parameter model.

We consider $m = 2$ and take the standardized case of the design space to be the cube, $-1 \leq x_i \leq 1$, $i = 1, 2$. We consider some combinations of this

model and construct the optimal design later this chapter.

4.1.2 Multiple linear regression

If we do not consider the interaction and the second-order terms in the above model, we obtain the multiple linear regression model. The multiple linear regression model with the response variable Y, and the (k-1) predictor variables $x_1, x_2, \ldots, x_{k-1}$ is given by

$$E(Y|x_1, \ldots, x_{k-1}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_{k-1} x_{k-1}. \tag{4.2}$$

The total number of parameters in the above model is k. The parameter θ_0 is the intercept and the remaining $(k-1)$ parameters θ_i's $(i = 1, 2, \ldots, k-1)$ are the coefficient of the $(k-1)$ independent variables.

In this chapter, we construct D and c-optimal designs for the above two types of models. We start with a practical example.

4.2 A Vocabulary-growth Study

We consider a practical problem of Berger and Wong (2009), which is taken from educational research. The response variable (Y) is the rate of increase in vocabulary growth among pupils. The goal of this research is to study how vocabulary growth is related to the school grade (x_1) of the pupil as well as to the income class (x_2) of the pupil's parents. The pupils from the 8th, 9th, 10th

and 11th school grades are sampled in this study. The pupils from each grade are then divided into three groups depending on the family income levels. We code the three family income levels as 1, 2 and 3 with 1 being the lowest level and 3 being the highest level. Thus, it gives a total of 12 different groups of pupils. Then, the growth in vocabulary of these pupils can be described using a multiple linear regression model with two predictors (x_1, x_2), as given by

$$E(Y|x_1, x_2) = \theta_0 + \theta_1 x_1 + \theta_2 x_2. \tag{4.3}$$

Although there is an intercept term θ_0, we are often interested in making inference for the two parameters θ_1 and θ_2, that is, on the effect of school grade and income category on vocabulary growth. So we are interested in studying the variance-covariance matrix of these two parameter estimates $\hat{\theta} = (\hat{\theta}_1,\ \hat{\theta}_2)$, which is given by

$$Cov(\hat{\theta}) = \begin{bmatrix} var(\hat{\theta}_1) & cov(\hat{\theta}_1,\ \hat{\theta}_2) \\ cov(\hat{\theta}_1,\ \hat{\theta}_2) & var(\hat{\theta}_2) \end{bmatrix}. \tag{4.4}$$

After some simplification on these four elements, the matrix can be written as

$$Cov(\hat{\theta}) = \frac{\sigma^2}{n} \begin{bmatrix} \frac{1}{(1-r_{12}^2)var(x_1)} & -\frac{cov(x_1,x_2)}{(1-r_{12}^2)var(x_1)var(x_2)} \\ -\frac{cov(x_2,x_1)}{(1-r_{12}^2)var(x_1)var(x_2)} & \frac{1}{(1-r_{12}^2)var(x_2)} \end{bmatrix}, \tag{4.5}$$

where σ^2 is the error variance, n is the total number of observations, $var(x_1)$ and $var(x_2)$ are the variances of x_1 and x_2, $cov(x_1, x_2)$ is their covariance, and r_{12} is the correlation between the two variables x_1 and x_2.

Looking at the expressions of the elements of the above variance-covariance matrix, we observe that the efficiency of the estimators θ_1 and θ_2 increases

when one or more of the following conditions hold, namely, the variance of the error terms σ^2 decreases; the sample size n increases; the $var(x_1)$ and $var(x_2)$ increase; and the squared correlation r_{12}^2 decreases.

We will see that the above conditions are satisfied when we maximize the D-optimal criterion for this study. In D-optimality, we maximize the determinant of the information matrix $M(p)$, or its logarithm $logdetM(p)$. So, we seek to maximize the criterion function:

$$\phi_D(p) = \psi_D\{M(p)\} = logdet\{M(p)\} = -logdet\{M^{-1}(p)\}. \tag{4.6}$$

In Table 4.1 (on page 69) we list the four designs for the vocabulary-growth study, namely, Design 1, Design 2, Design 3 and Design 4. This table also gives the twelve design points and the combinations of the levels of x_1 and x_2. As described above, we calculate the variances of the predictor variables x_1 and x_2, the correlation r_{12} between x_1 and x_2, variances of the parameter estimates $\hat{\theta_1}$ and $\hat{\theta_2}$ and the D-optimal criterion.

We see that, among the four designs considered, *Design* 4 has the lowest variances $var(\hat{\theta_1})$, $var(\hat{\theta_2})$, the largest variances $var(x_1)$, $var(x_2)$, the lowest correlation r_{12}, and the largest D-optimal criterion $\phi_D(p)$. Therefore, the *Design* 4 is the *optimal design* for the vocabulary-growth study.

Now we study the above model along with other models with a more general set-up in the following sections.

Table 4.1: Vocabulary-growth study: The 4 designs and the combinations of the levels of x_1 and x_2.

	Design 1		*Design* 2		*Design* 3		*Design* 4	
Design points	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2
d_1	8	1	8	1	8	1	8	1
d_2	8	2	8	2	8	1	8	1
d_3	8	3	8	3	8	3	8	1
d_4	9	1	8	1	8	1	8	3
d_5	9	2	8	2	8	2	8	3
d_6	9	3	8	3	8	3	8	3
d_7	10	1	11	1	11	1	11	1
d_8	10	2	11	2	11	2	11	1
d_9	10	3	11	3	11	3	11	1
d_{10}	11	1	11	1	11	1	11	3
d_{11}	11	2	11	2	11	3	11	3
d_{12}	11	3	11	3	11	3	11	3
$var(x_l)$	1.250	0.667	2.250	0.667	2.250	0.833	2.250	1.000
r_{12}	0.000		0.000		0.183		0.000	
$var(\hat{\theta}_l)$	0.067	0.125	0.037	0.125	0.038	0.103	0.037	0.083
$\phi_D(p)$	-0.1823		0.4055		0.5947		0.8109	

4.3 Construction of D-optimal designs in two variable models

4.3.1 Two variable model without interaction and second order terms

We first revisit the multiple linear regression model of the previous section. The model is

$$\begin{aligned} E(Y|x_1, x_2) &= \theta_0 + \theta_1 x_1 + \theta_2 x_2 \\ &= \underline{v}_x^T \underline{\theta} \end{aligned} \tag{4.7}$$

where $\underline{v}_x = (1,\ x_1,\ x_2)^T,\ -1 \leq x_i \leq 1,\ i = 1, 2,$

$\underline{v}_x \in \mathcal{V} = \{\underline{v}_x : \underline{v}_x = (1,\ x_1,\ x_2)^T,\ -1 \leq x_i \leq 1,\ i = 1, 2\}$, the induced design space.

Unlike the models considered in Chapter 3, as we have more than one design variable, we consider the discretized design space consisting of all pairs (x_1, x_2) arising when the values for each x_i, $i = 1, 2$ are those between -1 and +1 taken at steps of 0.1. That is, the space consists of $(21)^2 = 441$ pairs of (x_1, x_2).

Thus, we are considering designs of the type:

$$p = \left\{ \begin{array}{ccccc} x_{1\,1} & x_{1\,2} & . & . & x_{1\,441} \\ x_{2\,1} & x_{2\,2} & . & . & x_{2\,441} \\ p_1 & p_2 & . & . & p_{441} \end{array} \right\}.$$

Here also we make use of the Problems (P1) and (P2) that have a distinctive set of constraints as we had before. Thus, in order to construct the optimal design, we use the algorithms (3.20). However, note that we now have two design variables, and we need to use the vertices $\underline{v}_x$ as given in model (4.7). Then accordingly we need to calculate the information matrix, the design criteria, the partial and directional derivatives.

We consider the use of the functions $f(d) = exp(\delta d)$, $f(d) = \frac{exp(\delta d)}{1+exp(\delta d)}$ and $f(d) = \Phi(\delta d)$. Results are reported in Table 4.2. We record the number of iterations needed to achieve $\underset{1\leq j\leq J}{max}\{F_j\} \leq 10^{-n}$, for $n = 1, 2, 3, 4, 5$. The initial design is $p_j^{(0)} = 1/J$, $j = 1, 2, \ldots, J$ with $J = 441$ as we approximated the design space consisting of 441 pairs of (x_1, x_2). Iteration counts show that convergence is slow especially for the choice $f(d) = \frac{exp(\delta d)}{1+exp(\delta d)}$. For example, for $\delta = 2.0$, the number of iterations needed to achieve the above condition at $n = 5$ is 10260. On the other hand, the choice $f(d) = exp(\delta d)$ is not bad as we see the number of iterations to satisfy the first-order conditions.

We also attempt to improve the convergence of the algorithm by objectively choosing the argument in the function $f(x)$. In particular, we replace the partial derivatives d_j by the corresponding directional derivatives F_j. The results are reported in Table 4.3. Looking at the results we see that we improve the convergence of the algorithm. For example, with $f(d) = \Phi(\delta d)$, for $\delta = 0.3$ and $n = 5$, the number of iterations needed is 548, whereas using the

directional derivatives, for $\delta = 1.7$ and $n = 5$, the number of iterations needed is only 42. We get better improvement for the other choices as well.

The design converged to the optimal solution

$$p^* = \left\{ \begin{array}{cccc} -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{array} \right\}. \tag{4.8}$$

where the first two rows are the values of x_1, x_2 respectively while the third row gives the corresponding weights.

Finally we confirm this D-optimal design by plotting the variance function versus the two design variables. This is given in Figure 4.1. We see the variance function attains the maximum at 3, the number of parameters in the model. In Figure 4.2 we also plot the optimal weights against the support points as given in (4.8).

4.3.2 Two variable model with interaction

Now we consider the two-variable model and add the interaction between the two variables x_1 and x_2. Note that interaction is the failure of levels of one variable (factor) to behave consistently across the levels of the another variable. In such a case we need to consider the model with the interaction term. The model is given by

$$\begin{aligned} E(Y|x_1, x_2) &= \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 \\ &= \underline{v}_x^T \underline{\theta} \end{aligned} \tag{4.9}$$

Table 4.2: Number of iteration times needed to achieve $max\{F_j\} \leq 10^{-n}$.

$f(d_j) = \frac{exp(\delta d_j)}{1+exp(\delta d_j)}$

δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.1	165	408	686	969	1252
0.3	87	208	343	480	617
0.4	**85**	**199**	**324**	**451**	**578**
0.5	89	205	332	459	587
0.9	158	346	543	741	938
1.0	192	417	651	884	1117
1.5	588	1235	1878	2515	3152
2.0	2032	4177	6225	8244	10260

$f(d_j) = exp(\delta d_j)$

δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.1	69	172	291	411	532
0.5	15	35	59	83	107
0.9	8	20	33	46	60
1.0	**8**	**18**	**30**	**42**	**54**

$f(d_j) = \Phi(\delta d_j)$

δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.1	116	284	475	669	864
0.2	83	197	324	453	583
0.3	**83**	**191**	**309**	**429**	**548**
0.4	98	220	350	480	611
0.5	132	286	448	609	771

Table 4.3: Number of iteration times needed to achieve $max\{F_j\} \leq 10^{-n}$.

$f(F_j) = \frac{exp(\delta F_j)}{1+exp(\delta F_j)}$

δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.1	140	347	584	824	1065
0.5	32	74	121	168	216
0.9	20	44	70	95	121
1.0	19	40	63	86	110
1.5	15	29	44	59	74
2.0	12	23	34	45	56
2.6	11	19	28	36	44
2.7	**11**	**19**	**27**	**35**	**43**
2.8	10	18	26	34	

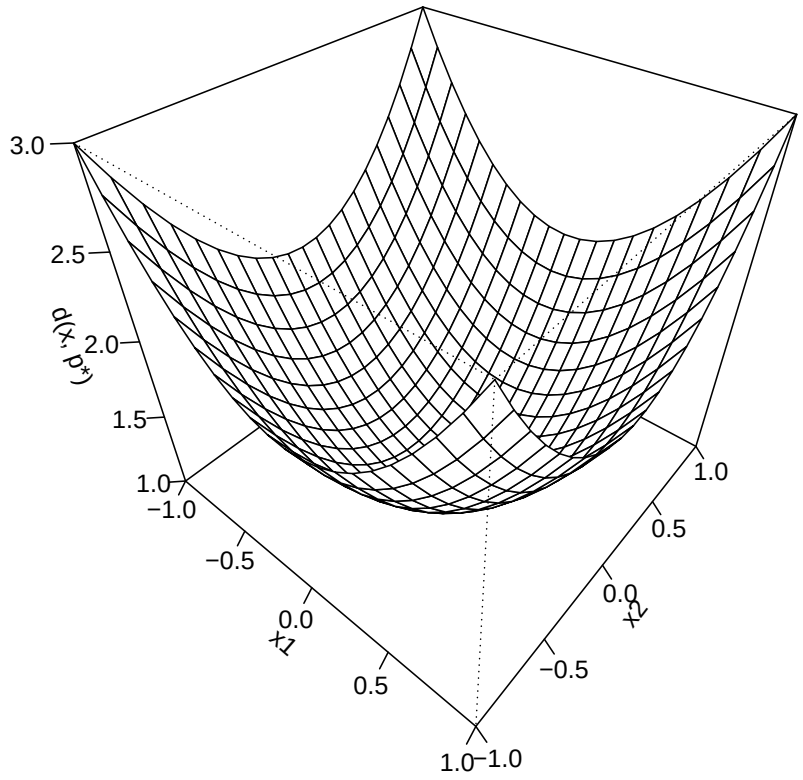
$f(F_j) = exp(\delta F_j)$

δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.1	69	172	291	411	532
0.5	15	35	59	83	107
0.9	8	20	33	46	60
1.0	**8**	**18**	**30**	**42**	**54**

$f(F_j) = \Phi(\delta F_j)$

δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.1	89	220	367	518	668
0.5	22	49	78	107	136
0.9	15	30	45	61	77
1.0	14	27	41	56	70
1.5	11	20	29	38	47
1.6	11	19	28	36	45
1.7	**10**	**19**	**27**	**34**	**42**
1.8	10	18	25	33	

Figure 4.1: The standardized variance of the predicted response on y at $\underline{x}$ ($d(\underline{x}, p^*)$) for model (4.7).

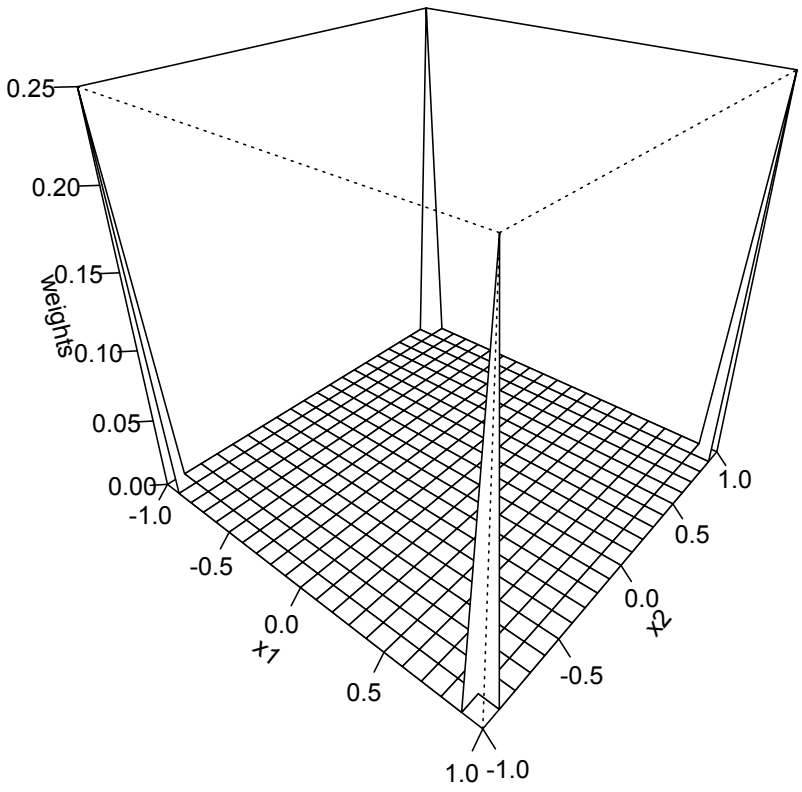


where $\underline{v}_x = (1,\ x_1,\ x_2,\ x_1x_2)^T,\ -1 \leq x_i \leq 1,\ i = 1, 2,$

$\underline{v}_x \in \mathcal{V} = \{\underline{v}_x : \underline{v}_x = (1,\ x_1,\ x_2,\ x_1x_2)^T,\ -1 \leq x_i \leq 1,\ i = 1, 2\}$, the induced design space.

As we did before, we consider the discretized design space consisting of all pairs (x_1, x_2) arising when the values for each x_i, $i = 1, 2$ are those between

Figure 4.2: D-optimal design (p^*) for model (4.7).



-1 and +1 taken at steps of 0.1. That is, the space consists of $(21)^2 = 441$ pairs of (x_1, x_2). We consider similar choices of the functions $f(d)$ and the free positive parameter δ like before, make use of the algorithm (3.20).

The results are given in Table 4.4. We see that convergence is slow especially for the choices $f(d) = \frac{exp(\delta d)}{1+exp(\delta d)}$ and $f(d) = \Phi(\delta d)$, whereas the choice $f(d) = exp(\delta d)$. We again try to improve the convergence by considering the argument x as the directional derivatives of the criterion function. The results are given in Table 4.5. From the results we see that the convergence is quite improved and we obtain the optimal design by 46 or 47 iterations for all of the choices of $f(F)$. The optimal support points with corresponding design weights for this model are obtained as

$$p^* = \left\{ \begin{array}{cccc} -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{array} \right\}. \tag{4.10}$$

We also confirm this D-optimal design by plotting the standardized variance of the predicted response on y at $\underline{x}$ $(d(\underline{x}, p^*))$ versus the two design variables. This is given in Figure 4.3. Note the difference from the plot in Figure 4.1. As we have 4 parameters in our model, the variance function attains the maximum at 4. In Figure 4.4 we also plot the optimal weights against the support points as given in (4.10).

Table 4.4: Number of iteration times needed to achieve $max\{F_j\} \leq 10^{-n}$.
$f(d_j) = \frac{exp(\delta d_j)}{1+exp(\delta d_j)}$

δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.1	114	250	398	547	696
0.2	80	169	263	358	453
0.3	**77**	**156**	**239**	**322**	**406**
0.4	84	165	249	334	419
0.5	99	190	283	376	470
0.9	266	485	699	913	1126
1.0	357	645	925	1204	1482

$f(d_j) = exp(\delta d_j)$

δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.1	43	97	157	217	278
0.4	11	24	39	54	70
0.5	9	19	31	43	55
0.6	**7**	**16**	**26**	**36**	**46**

$f(d_j) = \Phi(\delta d_j)$

δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.1	85	183	288	393	499
0.2	**74**	**149**	**227**	**306**	**385**
0.3	91	174	259	344	428
0.4	137	251	365	478	591

Table 4.5: Iteration results by considering the argument x as the directional derivatives
$f(F_j) = \frac{exp(\delta F_j)}{1+exp(\delta F_j)}$

δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.1	91	200	319	439	559
0.5	24	46	69	92	115
0.9	16	28	40	53	65
1.0	15	26	37	48	59
1.2	13	22	32	41	50
1.3	**13**	**21**	**29**	**38**	**46**
1.4	12	20	28	35	

$f(F_j) = exp(\delta F_j)$

δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.1	43	97	157	217	278
0.4	11	24	39	54	70
0.5	9	19	31	43	55
0.6	**7**	**16**	**26**	**36**	**46**

$f(F_j) = \Phi(\delta F_j)$

δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.1	59	128	202	277	352
0.5	17	31	45	59	73
0.7	14	24	33	43	53
0.8	**13**	**21**	**30**	**38**	**47**
0.9	12	19	27	34	

Figure 4.3: The standardized variance of the predicted response on y at $\underline{x}$ ($d(\underline{x}, p^*)$) for model (4.9).

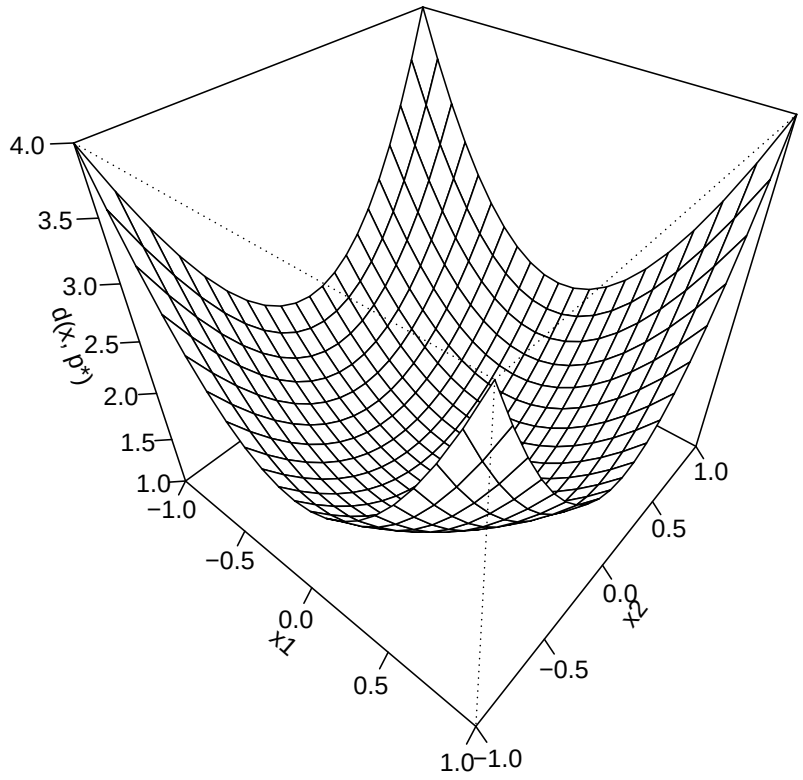
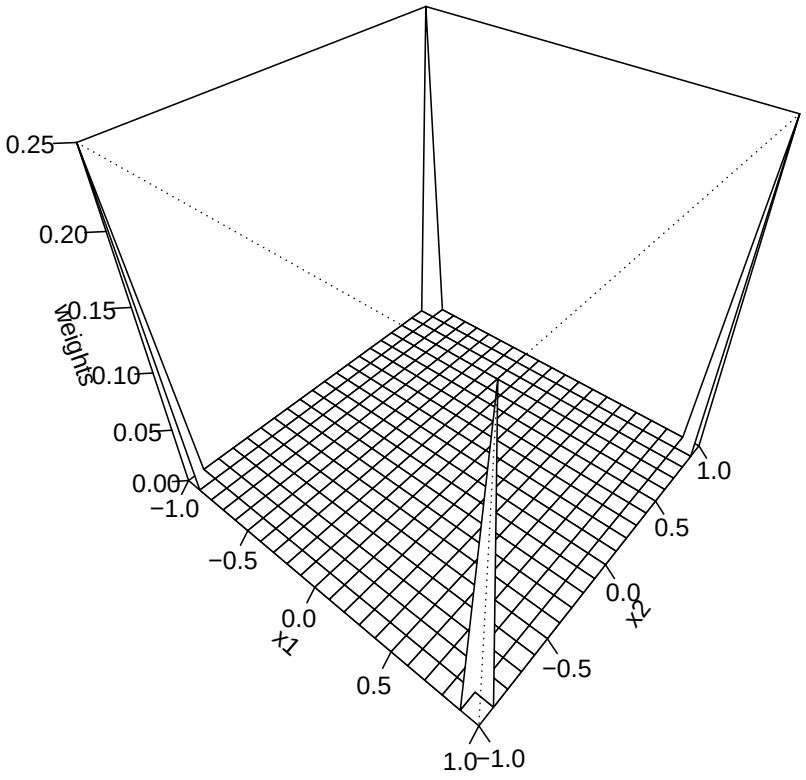


Figure 4.4: D-optimal design (p^*) for model (4.9).



4.3.3 The second-order model in two variables with interaction and second order terms

We now consider the full model taking into consideration of the interaction term as well as the second-order terms in the two design variables x_1 and x_2. We consider $m = 2$ in model (4.1) and take the standardized version of the design space to be the cube, $-1 \leq x_i \leq 1$, $i = 1, 2$.

Thus the model is, with a revised parameterization,

$$\begin{aligned} E(Y|x_1, x_2) &= \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 + \theta_4 x_1^2 + \theta_5 x_2^2 \\ &= \underline{v}_x^T \underline{\theta} \end{aligned} \tag{4.11}$$

where $\underline{v}_x = (1,\ x_1,\ x_2,\ x_1x_2,\ x_1^2,\ x_2^2)^T,\ -1 \leq x_i \leq 1,\ i = 1, 2.$

$\underline{v}_x \in \mathcal{V} = \{\underline{v}_x : \underline{v}_x = (1,\ x_1,\ x_2,\ x_1x_2,\ x_1^2,\ x_2^2)^T,\ -1 \leq x_i \leq 1,\ i = 1, 2\}$, the induced design space.

We report the results in Table 4.6. We see that convergence is slow especially for the choices $f(d) = \frac{exp(\delta d)}{1+exp(\delta d)}$ and $f(d) = \Phi(\delta d)$. We again try to improve the convergence by considering the argument x as the directional derivatives of the criterion function. The results are given in Table 4.7. From the results we see that the convergence is quite improved by considering the directional derivatives.

The optimal design converged to the 9 support points (4 at corners, 4 at

mid-sides and 1 at the centre) along with the optimal weights as given by

$$p^* = \begin{Bmatrix} -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 \\ -1 & -1 & -1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0.146 & 0.08 & 0.146 & 0.08 & 0.096 & 0.08 & 0.146 & 0.08 & 0.146 \end{Bmatrix} \quad (4.12)$$

We confirm this D-optimal design by plotting the standardized variance of the predicted response on y at $\underline{x}$ ($d(\underline{x}, p^*)$) versus the two design variables. This is given in Figure 4.5. Note the difference of this variance plot. As we have 6 parameters in our model, the variance function attains the maximum at 6. In Figure 4.6 we also plot the optimal weights against the support points as given in (4.12).

4.4 Construction of c-optimal designs in two variable models

We start by explaining the theories of c-optimal design. As c-optimality is a special case of linear optimality, it better to introduce this optimality first.

Linear Optimality

Let L be a $k \times k$ matrix of coefficients. The maximization of the criterion function

$$\phi_L(p) = \psi_L\{M(p)\} = -\,tr\{M^{-1}(p)L\} \quad (4.13)$$

leads to a linear, or L-optimum design. It is linear in the elements of the covariance matrix $M^{-1}(p)$.

Table 4.6: Number of iteration times needed to achieve $max\{F_j\} \leq 10^{-n}$.
$f(d_j) = \frac{exp(\delta d_j)}{1+exp(\delta d_j)}$

δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.1	278	1543	2806	4029	5247
0.2	**212**	**1185**	**2150**	**3083**	**4014**
0.3	229	1292	2340	3353	4362
0.5	400	2326	4200	6008	7808

$f(d_j) = exp(\delta d_j)$

δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.1	99	546	994	1428	1860
0.2	50	273	497	714	930
0.3	**33**	**182**	**331**	**476**	**620**

$f(d_j) = \Phi(\delta d_j)$

δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.09	223	1244	2261	3245	4225
0.1	**214**	**1193**	**2167**	**3109**	**4048**
0.2	221	1254	2269	3251	4229

Table 4.7: Continuing to Table 4.6 after considering the directional derivatives $f(F_j) = \frac{exp(\delta F_j)}{1+exp(\delta F_j)}$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.1	197	1093	1989	2855	3719
0.5	40	221	400	572	744
0.6	34	185	334	477	620
0.7	**30**	**159**	**287**	**410**	**532**
0.8	27	141			

$f(F_j) = exp(\delta F_j)$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.1	99	546	994	1428	1860
0.2	50	273	497	714	930
0.3	**33**	**182**	**331**	**476**	**620**

$f(F_j) = \Phi(\delta F_j)$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.1	124	686	1247	1789	2330
0.2	62	344	624	895	1165
0.3	42	231	417	597	777
0.4	**32**	**174**	**314**	**448**	**583**
0.5	27	141	252		

Figure 4.5: The standardized variance of the predicted response on y at $\underline{x}$ ($d(\underline{x}, p^*)$) for two variables with interaction terms and second order terms.

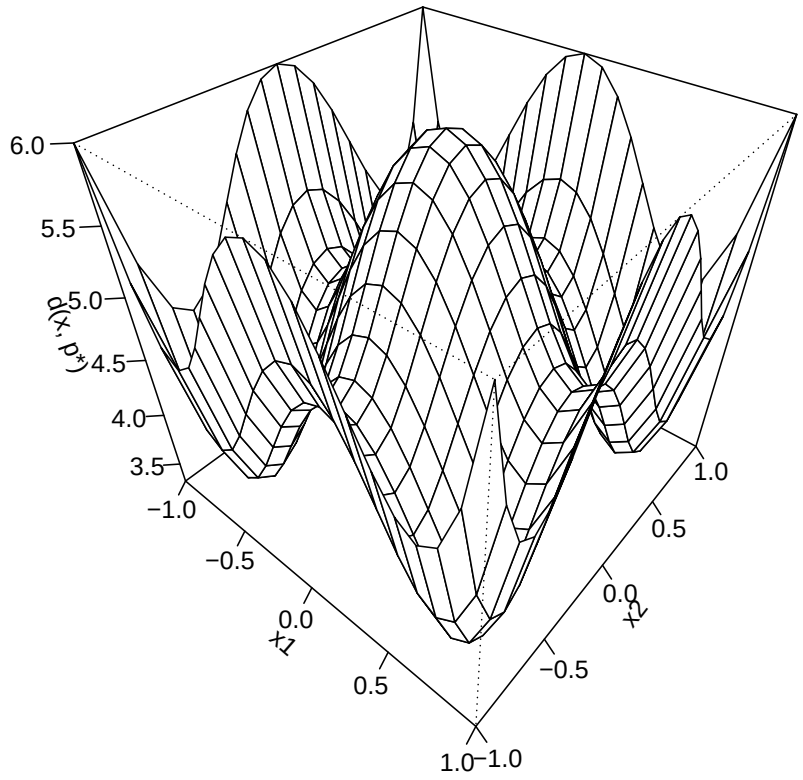
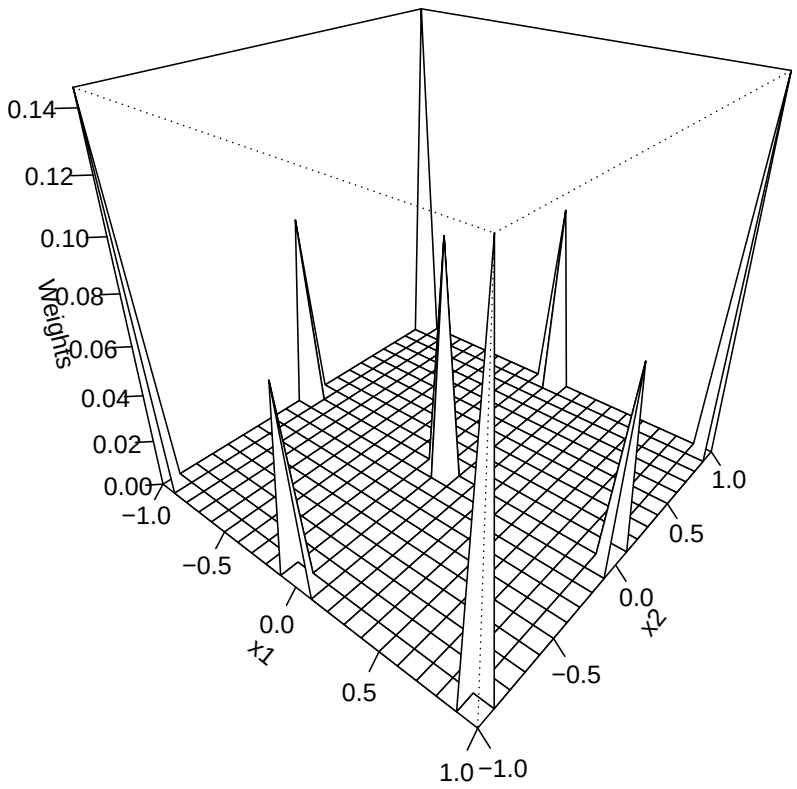


Figure 4.6: D-optimal design (p^*) for the model with two variables with interaction terms and second order terms.



If L is of rank $s \leq k$ it can be expressed in the form $L = A^T A$ where A is a $s \times k$ matrix of rank s. Then the criterion function (4.13) can be expressed as

$$\phi_L(p) = -tr\{M^{-1}(p)L\} = -tr\{M^{-1}(p)A^T A\} = -tr\{AM^{-1}(p)A^T\}. \quad (4.14)$$

This form stresses the relationship with the D_A-optimum design, where the determinant, rather than the trace, of $\{-AM^{-1}(p)A^T\}$ is maximized.

An alternative name for this design criterion would therefore be A_A-optimality, with A-optimal recovered when $L = I$, the identity matrix.

The partial derivatives of ϕ_L are given by

$$\frac{\partial \phi_L}{\partial p_j} \;=\; \underline{v}_j^T M^{-1}(p)\, A^T A M^{-1}(p) \underline{v}_j. \quad (4.15)$$

c-optimality:

This is an important special case of the above linear optimality. Note that the case $A = \underline{c}^T$, where $\underline{c}$ is a $k \times 1$ vector, corresponds to another standard criterion known in the literature as the c-optimality criterion [Elfving (1952)]. This criterion seeks to maximize the criterion function:

$$\phi_c(p) \;=\; -\,\underline{c}^T M^{-1}(p) \underline{c}. \quad (4.16)$$

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.

Using the partial derivatives (4.15) of linear optimality, we obtain the the partial derivatives of ϕ_c as given by

$$\frac{\partial \phi_c}{\partial p_j} = [\underline{c}^T M^{-1}(p)\, \underline{v}_j]^2. \tag{4.17}$$

We only consider two cases of the second-order model in two variables, which are the model with interaction terms only and the full model with the second order terms.

First of all, we give the two variables model with interaction terms only. The model is

$$E(Y|x_1, x_2) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2, \tag{4.18}$$

where x_1 and x_2 can taken values from -1 to 1 at steps of 0.1. We use similar δ values before, and work with three choices of $f(d)$, which are $f(d) = exp(\delta d)$, $f(d) = \frac{exp(\delta d)}{1+exp(\delta d)}$ and $f(d) = \Phi(\delta d)$, in constructing the c-optimal design. In this case, note that we use the partial derivatives of ϕ_c calculated by equation (4.17).

First, we are interested in constructing the c-optimal design with respect to the interaction term $x_1 x_2$. Thus, we take the c-vector as $c = (0\ 0\ 0\ 1)$ in this model. Table 4.8 (using both the partial and directional derivatives) shows the number of iteration times needed to achieve $max\{F_j\} \leq 10^{-n}$. The optimal support points with the corresponding design weights for this model

are obtained as

$$p^* = \left\{ \begin{array}{cccc} -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{array} \right\}. \tag{4.19}$$

Next, we consider the full model, which includes interaction terms as well as the second order terms.

$$E(Y|x_1, x_2) \;\; = \;\; \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 + \theta_4 x_1^2 + \theta_5 x_2^2$$

Here again we are interested in constructing the c-optimal design with respect to the interaction term $x_1 x_2$ only. Thus, here we take the c-vector as $c = (0\ 0\ 0\ 1\ 0\ 0)$ in this model. Table 4.9 (using both the partial and directional derivatives) shows the number of iteration times needed to achieve $max\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4, 5$.

We obtain the optimal support points with corresponding design weights as given by

$$p^* = \left\{ \begin{array}{cccc} -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{array} \right\}. \tag{4.20}$$

Table 4.8: Number of iteration times needed to achieve $max\{F_j\} \leq 10^{-n}$.

$f(d_j) = \frac{exp(\delta d_j)}{1+exp(\delta d_j)}$

δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.1	103	302	547	800	1054
0.5	41	93	154	216	279
0.9	42	81	124	168	212
1.0	44	82	123	165	207
1.1	**46**	**83**	**124**	**164**	**205**
1.2	49	86	125	165	205
1.3	52	89	128	168	207
1.5	60	98	137	176	216

$f(d_j) = exp(\delta d_j)$

δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.1	32	123	239	359	480
0.5	2	2	10	34	58
0.9	2	2	2	2	9
1.0	**2**	**2**	**2**	**2**	**3**

$f(d_j) = \Phi(\delta d_j)$

δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.1	73	202	360	523	687
0.5	40	80	124	170	215
0.6	43	80	120	162	203
0.7	**47**	**82**	**121**	**160**	**199**
0.8	52	87	125	163	200
0.9	59	95	132	169	206
1.0	69	105	142	180	217
1.5	172	225	272	318	364
2.0	582	682	758	833	907

$f(F_j) = \frac{exp(\delta F_j)}{1+exp(\delta F_j)}$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
1.0	15	35	58	81	104
1.5	12	25	40	55	70
2.0	10	20	31	42	53
2.1	10	19	30	40	51
2.2	10	19	29	39	49
2.3	10	18	28	37	47
2.77	9	16	24	31	40
2.78	**9**	**16**	**24**	**31**	**39**
2.79	9	16	24	31	

$f(F_j) = exp(\delta F_j)$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.1	32	123	239	359	480
0.5	2	2	10	34	58
0.9	2	2	2	2	9
1.0	**2**	**2**	**2**	**2**	**3**

$f(F_j) = \Phi(\delta F_j)$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.1	63	182	328	478	628
0.5	17	42	71	100	129
0.9	12	26	41	57	73
1.0	11	24	38	52	66
1.2	10	21	32	44	55
1.3	10	20	30	41	51
1.6	9	17	25	34	42
1.7	9	16	24	32	40
1.71	9	16	24	32	40
1.72	**9**	**16**	**24**	**32**	**40**
1.73	9	16	24	31	

Table 4.9: Number of iteration times needed to achieve $max\{F_j\} \leq 10^{-n}$.
$f(d_j) = \frac{exp(\delta d_j)}{1+exp(\delta d_j)}$

δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.8	41	81	127	173	219
0.9	42	81	124	168	212
1.0	44	82	123	165	207
1.1	**46**	**83**	**124**	**164**	**205**
1.2	49	86	125	165	205
1.3	52	89	128	168	207

$f(d_j) = exp(\delta d_j)$

δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.1	32	123	239	359	480
0.5	2	2	10	34	58
0.8	2	2	2	2	17
0.9	2	2	2	2	9
1.0	**2**	**2**	**2**	**2**	**3**

$f(d_j) = \Phi(\delta d_j)$

δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.1	73	202	360	523	687
0.3	41	94	156	220	283
0.5	40	80	124	170	215
0.6	43	80	120	162	203
0.7	**47**	**82**	**121**	**160**	**199**
0.8	52	87	125	163	200
0.9	59	95	132	169	206
1.0	69	105	142	180	217
1.1	80	119	157	195	233

Table 4.10: Continuing to Table 4.9

$f(F_j) = \frac{exp(\delta F_j)}{1+exp(\delta F_j)}$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.1	95	284	517	757	998
0.2	52	147	263	383	503
0.9	16	38	63	89	115
1.1	14	32	53	74	95
1.5	12	25	40	55	70
2.7	9	16	24	32	40
2.8	**9**	**16**	**24**	**31**	**39**
2.81	9	16	23	31	

$f(F_j) = exp(\delta F_j)$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.1	32	123	239	359	480
0.5	2	2	10	34	58
0.8	2	2	2	2	17
0.9	2	2	2	2	9
1.0	**2**	**2**	**2**	**2**	**3**

$f(F_j) = \Phi(\delta F_j)$

δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.1	63	182	328	478	628
0.2	35	95	167	242	317
0.5	17	42	71	100	129
1.0	11	24	38	52	66
1.5	9	18	27	36	45
1.6	9	17	25	34	42
1.7	9	16	24	32	40
1.71	9	16	24	32	40
1.72	**9**	**16**	**24**	**32**	**40**
1.73	9	16	24	31	
1.8	9	16	23	30	

Chapter 5

Conclusion

In this thesis we have considered constructing optimal designs for various polynomial regression models. The goal was to have a best inference for all or some of the unknown parameters in a regression model by making the information matrix $M(p)$ large in some sense. So we considered various ways in which to make the matrix $M(p)$ large, namely by maximizing some real valued function $\phi(p) = \psi\{M(p)\}$.

We started with a basic introduction of optimal design theory in Chapter 1. The designing experiments divided into six stages. The optimal design would affect the second stage which is the choice of the research design and help experimenters setting up an optimal design to reduce the experimental cost. We discussed how we discretize a continuous design space and how we obtain an optimal design in two ways, namely an exact design and an approximate design. We discussed different concepts of optimal design theory, such as

a design measure, the information matrix and variance function. Moreover, we discussed different criteria in optimal design by maximizing a real valued function of the information matrix.

We consider a class of optimization problems and determine optimality conditions of these problems in Chapter 2. Problem (P1) and Problem(P2) are general optimal linear regression design problems. The emphasis was on a differential calculus approach. An important tool was the directional derivative of a criterion function. We have seen that this plays an important simplifying role in the calculus of optimization. By the vertex directional optimality theorem, we determined the support points along with the corresponding optimal weights by calculating the vertex directional derivatives.

We constructed optimal designs for various polynomial regression models in one variable in Chapter 3. We approached this by analizing D-optimality. We first constructed such designs analytically by using the Legendre polynomials. We constructed the discrete D-optimal designs which are unique, having a minimal support of k points which are the k roots of the polynomials. Then we considered a practical problem - A Radiation Dosage Example. We discussed several important aspects of optimal design throughout this example. We then constructed optimal designs for various models in one variable, namely, the simple linear regression model, the quadratic regression model and the cubic regression model. We did this by using a class of algorithms indexed by a function $f(.)$ which satisfies certain conditions. Our optimization problems

have a distinctive set of constraints, namely the variables $p_1, p_2, \ldots, p_J$ must be nonnegative and sum to 1. The above class of algorithms neatly submit to these constraints and have some suitable properties. Using the properties of the directional derivatives and by choosing suitable functions $f(.)$ and its arguments, we investigated techniques for improving the convergence of the algorithms.

However, in many regression models we needed to deal with more than one design variable such as multiple linear regression or polynomial regression in two or more variables. We constructed such optimal designs in Chapter 4. We start with a practical problem - A Vocabulary-growth Study. We used multiple linear regression model in two design variables for this problem as well as discussed some important issues for constructing optimal designs. We then constructed D-optimal designs for various polynomial regression models in two variables. The models were divided into three groups, namely the model without interaction and second order terms, the model with interaction terms only, and the model with interaction and second order terms. Sometimes we need to focus on a particular parameter in a model. This motivated us to construct c-optimal designs for some of the models. We constructed such c-optimal designs for the interaction term in two models in two design variables.

It seems reasonable to say that we have identified some good strategies for constructing optimal designs for various regression models. The next step would be to solve some optimization problems especially if we have some addi-

tional constraints. For future research, we would like to solve some estimation problems using our optimal design techniques. We would like to work on one practical problem by computing the cell probabilities under an independence model in a two-way classification table of two attributes when the data can be viewed as incomplete, that is, some of the cell frequencies are missing. When there is no missing data maximum likelihood estimates of the cell probabilities can be obtained explicitly. But, certainly numerical techniques are needed to compute the maximum likelihood estimates when there are missing cell frequencies. In that case some of the terms in the likelihood may be linear functions of several probabilities. We would like to work on this problem.

Bibliography

Atkinson, A., Donev, A., and Tobias, R. (2007). *Optimum experimental designs, with SAS*. Oxford University Press, Oxford.

Atkinson, A. C. and Donev, A. N. (1992). *Optimum Experimental Designs*. Oxford Statistical Science Series-8. Oxford University Press, Oxford.

Atwood, C. L. (1969). Optimal and efficient designs of experiments. *Annals of Mathematical Statistics*, 40:1570–1602.

Berger, M. P. F. and Wong, W. K. (2009). *An introduction to optimal designs for social and biomedical research*. John Wiley & Sons.

Elfving, G. (1952). Optimum allocation in linear regression theory. *Annals of Mathematical Statistics*, 23:255–262.

Fedorov, V. V. (1972). *Theory of optimal experiments*. Academic Press, New York and London.

Karlin, S. and Studden, W. J. (1966). Optimal experimental designs. *Annals Mathematical Statistics*, 37:783–815.

Kiefer, J. (1959). Optimum experimental designs (with discussion). *J. Roy. Statist. Soc. Series B*, 21:272–319.

Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *Annals of Statistics*, 2:849–879.

Kiefer, J. and Wolfowitz, J. (1960). The equivalence of two extremum problems. *Canadian J. Math.*, 12:363–366.

Mandal, S. and Torsney, B. (2000). Algorithms for the construction of optimizing distributions. *Communications in Statistics - Theory and Methods*, 29:1219–1231.

Mandal, S. and Torsney, B. (2006). Construction of optimal designs using a clustering approach. *Journal of Statistical Planning and Inference*, 136:1120–1134.

Mandal, S., Torsney, B., and Carriere, K. C. (2005). Constructing optimal designs with constraints. *Journal of Statistical Planning and Inference*, 128:609–621.

Pukelsheim, F. (1993). *Optimal design of experiments*. Wiley, New York.

Shah, K. R. and Sinha, B. K. (1989). Theory of optimal designs. *Lecture Notes in Statistics*, Vol. 54, Springer-Verlag.

Sibson, R. (1974). D_A-optimality and duality. Progress in Statistics. *Colloq. Math. Soc. Janos. Bolyai*, 9:677–692.

Silvey, S. D. (1980). *Optimal Design*. Chapman and Hall, London.

Silvey, S. D. and Titterington, D. M. (1974). A lagrangian approach to optimal design. *Biometrika*, 61:299–302.

Silvey, S. D., Titterington, D. M., and Torsney, B. (1978). An algorithm for optimal designs on a finite design space. *Communications in Statistics A*, 7:1379–1389.

Torsney, B. (1977). Contribution to discussion of "Maximum likelihood from incomplete data via the EM algorithm" by Dempster et al. *J. Roy. Statist. Soc. Series B*, 39:26–27.

Torsney, B. (1983). A moment inequality and monotonicity of an algorithm. *Proc. Internat. Symp. 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.

Torsney, B. (1988). Computing optimising 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.

Torsney, B. and Mandal, S. (2001). Construction of constrained optimal designs. *Optimum Design 2000*, 141-152, Kluwer Academic Publishers.

Whittle, P. (1973). Some general points in the theory of optimal experimental designs. *J. Roy. Statist. Soc. Series B*, 35:123–130.

Wynn, H. P. (1972). Results in the theory and construction of D-optimum experimental designs (with discussion). *J. Roy. Statist. Soc. Series B*, 34:133–147, 170–186.