Optimal Designs for Minimizing Variances of Parameter Estimates in Linear Regression Models

by

Manqiong Chen

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfilment of the requirements of the degree of

MASTER OF SCIENCE

Department of Statistics

University of Manitoba

Winnipeg

ABSTRACT

In statistical inference, it is important to estimate the parameters of a regression model in such a way that the variances of the estimates are as small as possible. Motivated by this fact, we have tried to address this important problem using optimal design theory.

We start with some optimal design theory and determine the optimality conditions in terms of a directional derivative. We construct the optimal designs for minimizing variances of the parameter estimates in two ways. The first one is the analytic approach, in which we derive the derivatives of our criterion and solve the resulting equations. In another approach, we construct the designs using a class of algorithms.

We also construct designs for minimizing the total variance of some parameter estimates. This is motivated by a practical problem in Chemistry. We attempt to improve the convergence of the algorithm by using the properties of the directional derivatives.

ACKNOWLEDGEMENTS

I would like to express my heartfelt and deepest appreciation to my supervisor, Dr. Saumen Mandal, for his encouragement, constant guidance, persistent help and great effort he put into my graduate training, without which this dissertation would not have been possible.

I am also thankful to my committee members, Dr. Aerambamoorthy Thavaneswaran and Dr. Yang Zhang for monitoring my work and spending time reading this thesis and providing valuable suggestions to deepen my understanding of the research topic.

I am also grateful to all the professors and instructors of the Department of Statistics at University of Manitoba for their help and wonderful teaching of all courses. I would also like to extend my appreciation to all the support staff of the department for taking all of our administrative work. I am also very thankful to my classmates and friends for their constant encouragement.

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

The last but not the least, I would like to take this opportunity to extend my profound appreciation from my deep heart to my beloved parents for their unceasing encouragement and endless support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In an experiment, we face several questions of interest. In order to answer these questions, we need to carefully design an experiment. First, before we run an experiment, we need to select the values or levels of inputs, and then we observe a measurement on some variables of interest. Usually there will be a set of (or some combinations of) the inputs allowed. We must decide how many observations we take at each value or level of the inputs.

In regression experiments the inputs are usually numerical. Our objective is good estimation of the parameters of the model. There are many criteria defining good estimation in the literature. We choose a design to optimize a certain criterion function. Some useful texts in optimal design are Fedorov (1972), Atkinson et al. (2007), Pukelsheim (1993), Silvey (1980),

and Berger and Wong (2009).

First, we will go through some fundamental concepts of linear model and optimal design theory, such as the definition of a design measure, variance function, information matrix, various criterion functions along with their properties.

Suppose we have a model of the type:

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

where y is the response variable, $\underline{x} = (x_1, x_2, \cdots, x_m)^T$ are design variables which belong to a m dimension real design space $\mathcal{X}$, that is, $\underline{x} \in \mathcal{X}$. Normally, we assume that $\mathcal{X}$ is closed and bounded, that is, $\mathcal{X}$ is compact. $\underline{\theta} = (\theta_1, \theta_2, \cdots, \theta_k)^T$ are unknown parameters with dimension k. σ is a fixed and unknown nuisance parameter. $p(.)$ is a probability model. For each $\underline{x}$ in the design space, we can perform an experiment and observe a measurement on a random variable $y = y(\underline{x})$, where $var(y(\underline{x})) = \sigma^2$.

In linear regression, we can write down the model as:

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

where $\underline{f}(\underline{x}) = (f_1(\underline{x}), \cdots, f_k(\underline{x}))^T$ is a vector of k real valued functions. The function $\underline{f}$ is known to us.

For more details, let us first consider the mode of inference such as point estimation. In order to obtain good estimates of some or all the parameters $\underline{\theta}$, we first need to select n support points $(x_1, x_2, \cdots, x_n) \in \mathcal{X}$.

How do we know that the estimator $\hat{\underline{\theta}}$ is the optimal one? We first check whether it is unbiased or not. If it is unbiased, that is, $E(\hat{\underline{\theta}}) = \underline{\theta}$, then the next step is to consider whether the variance is as small as possible. The smaller the variance the better is the accuracy of $\hat{\underline{\theta}}$. The $k \times k$ covariance matrix of $\hat{\underline{\theta}}$ can be write as $D(\hat{\underline{\theta}}) = E([\hat{\underline{\theta}} - \underline{\theta}][\hat{\underline{\theta}} - \underline{\theta}]^T)$. We also call it as dispersion matrix. The dispersion matrix not only contains the variance parts of $\hat{\underline{\theta}}$ in diagonal elements, but also contain the covariance parts of $\hat{\underline{\theta}}$ in off-diagonal elements.

If the model (1.2) is true, and if the variances of the random variables $y_1, y_2, \cdots, y_n$ are all equal to σ^2, then the standard linear model can be written as:

$$E(Y) = X\underline{\theta},\ D(Y) = \sigma^2 I_n \tag{1.3}$$

where $Y = (y_1, y_2, \cdots, y_n)$, X is the design matrix, I_n is identity matrix of dimension n and $D(Y)$ is the covariance matrix of Y.

We can use least squares estimation method which is a standard choice for this model having the optimality of being best linear unbiased estimators

(BLUE) of the parameters. These are solutions of:

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

where $X^T X$ is the information matrix of dimension $k \times k$. In other words, it is the inverse of the dispersion matrix of $\hat{\underline{\theta}}$. If we are interested in all of the parameters, then the selection of the values of the design variable must ensure that the matrix $(X^T X)$ is non-singular. In that case the unique solution is given by:

$$\hat{\underline{\theta}} \;=\; (X^T X)^{-1} X^T Y. \tag{1.4}$$

Here $E(\hat{\underline{\theta}}) = \underline{\theta}$, and $D(\hat{\underline{\theta}}) = \sigma^2 (X^T X)^{-1}$. As the model is linear, the covariance matrix of $\hat{\underline{\theta}}$ does not depend on $\underline{\theta}$. If $X^T X$ is singular, then we need to deal with a generalized inverse of $X^T X$, in which case, we will have more than one solution of $\underline{\theta}$.

The predicted value of the response at $\underline{x}$ can be written as:

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

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

1.1 Exact and Approximate Designs

Usually the design space $\mathcal{X}$ is continuous. We need to discretize this space, in order to make our model more useful. To discretize the design space, we set $\underline{v} = (f_1(\underline{x}), f_2(\underline{x}), \cdots, f_k(\underline{x}))^T$, $\underline{v} \in \mathcal{V}$, and $\mathcal{V} = \{\underline{v} \in \mathbf{R}^k : \underline{v} = (f_1(\underline{x}), f_2(\underline{x}), \cdots, f_k(\underline{x}))^T, \underline{x} \in \mathcal{X}\}$. Then the linear model becomes:

$$E(y|\underline{v}, \underline{\theta}, \sigma) \quad = \quad \underline{v}^T \underline{\theta}. \tag{1.5}$$

Apparently, selecting a vector $\underline{x}$ in the design space $\mathcal{X}$ is the same as selecting a k-vector $\underline{v}$ in the closed bounded k-dimensional space $\mathcal{V} = \underline{f}(\mathcal{X})$, where $\underline{f}$ is the vector with k real valued functions.

We assume that the discretized design space $\mathcal{V}$ consists of J distinct vectors $\underline{v}_1, \underline{v}_2, \cdots, \underline{v}_J$. We call $\underline{v}_j$'s as the vertices of the feasible region of our problem. A value for $\underline{v}$ must first be chosen from the J elements of $\mathcal{V}$ to obtain an observation on y.

Now we can precisely present the design problem as follows. Suppose we have a total number of n observations. Our goal is to obtain best estimators of $\underline{\theta}$. We must determine how many observations, say n_j, to be taken at $\underline{v}_j$, where $\sum_{j=1}^{J} n_j = n$. The matrix $X^T X$ can be written as:

$$X^T X = M(\underline{n}), \underline{n} = (n_1, n_2, \cdots, n_J)^T \tag{1.6}$$

where

$$\begin{aligned} M(\underline{n}) &= \sum_{j=1}^{J} n_j \underline{v}_j \underline{v}_j^T \\ &= VNV^T \end{aligned} \tag{1.7}$$

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

We now want to select the values of the vector $\underline{n}$ in order to make the matrix $M(\underline{n})$ large in some sense. Finding the exact integer values n_j is an integer programming problem, and is known as an exact design problem in the optimal design context. However, since the theory of calculus cannot be accustomed to define the existence of or to identify optimal solutions, even without additional constraints, interger programming problems are hard or at least strenuous to solve. In addition, a solution should have to be formulated individually for different values of n. Therefore, there is no general formula for an optimal exact design, so we need to figure out a convenient method to solve this problem. For this reason, we seek for an approximate design as follows.

Let $p_j = n_j/n$, $j = 1, 2, \cdots J$. Then we can write (1.7) as

$$M(\underline{n}) = nM(p). \tag{1.8}$$

where

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

and $P = diag(p_1, p_2, \cdots, p_J)$.

Note that p_j is the proportion of observations taken at $\underline{v}_j$, so that $p_j \geq 0$, $\sum_{j=1}^{J} p_j = 1$; and $p = (p_1, p_2, \cdots, p_J)$ is the resultant distribution on $\mathcal{V}$. Thus our problem becomes that of choosing the design p in order to make our information matrix $M(p)$ large subject to $p_j{=}n_j/n$. Relaxing this problem subject to the constraints $p_j \geq 0$ and $\sum_{j=1}^{J} p_j = 1$ yields an approximate design problem.

The difference between exact design and approximate design is as follows. First, for approximate design, the corresponding weight is continuous, whereas, for exact design, it is discrete. The sum of corresponding weights of an approximate design is 1, but sum of corresponding numbers of subjects of an exact design is n. It is quite natural that an approximate design is preferred to the original exact design and is given by np^*, rounded to a nearest exact design.

In addition, p can be defined as a probability distribution on $\mathcal{V}$ such that $M(p) = E_p[\underline{v}\underline{v}^T]$, where p_j the probability corresponding to the vertex $\underline{v}_j$, that is, $P(\underline{v} = \underline{v}_j) = p_j$.

1.2 Design Measure

The form of a design measure can be written as follows:

$$\xi = \left\{ \begin{array}{cccc} \underline{x}_1 & \underline{x}_2 & \cdots & \underline{x}_J \\ p_1 & p_2 & \cdots & p_J \end{array} \right\}$$

where p_j's are the weights corresponding to the support or design point x_j where $x_j \in \mathcal{X}$ and $\sum_{j=1}^{J} p_j = 1$, $0 \leq p_j \leq 1$ for all j. We can also use the notation p instead of ξ.

1.3 Support of a Design Measure

The support of a design measure ξ in the design space $\mathcal{V}$ is defined as a set of vectors $\underline{v}_j$ which have nonzero weights. This is given by

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

If p^* is an optimal design, then the support of the design measure under the design p^* may be a strict subset of $\mathcal{V}$.

Note that, if we have k parameters in our model, we should have at least k support points in our design. If it is less than k, we will lose some efficiency

of the design. From Carathéodory's theorem (Silvey 1980, Appendix 2), there is an upper bound for the number of support points, which is $k(k+1)/2$. Likewise, if we have support points which is greater than $k(k+1)/2$, we will lose some efficiency as well.

1.4 Standardized Variance of the Predicted Response

Standardized variance of the predicted response on y at $\underline{x}$ can be written as:

$$d(\underline{x}, p) = \underline{f}^T(\underline{x}) M^{-1}(p) \underline{f}(\underline{x}) \tag{1.9}$$

where $M^{-1}(p)$ is covariance matrix, that is the inverse of the information matrix. We know the predicted response is $\hat{y}(\underline{x}) = \underline{f}^T(\underline{x})\hat{\underline{\theta}}$. So its variance is:

$$\begin{aligned} Var(\hat{y}(\underline{x})) &= \underline{f}^T(\underline{x}) Var(\hat{\underline{\theta}}) \underline{f}(\underline{x}) \\ &= \underline{f}^T(\underline{x}) \sigma^2 (X^T X)^{-1} \underline{f}(\underline{x}) \\ &= \frac{\sigma^2}{n} \underline{f}^T(\underline{x}) M^{-1}(p) \underline{f}(\underline{x}). \end{aligned}$$

When we compare designs it is often helpful to scale the variance for

σ^2 and the number of trials n. Thus we consider the standardized variance $d(\underline{x}, p)$ as given in (1.9).

1.5 Properties of the Information Matrix $M(p)$

We have already defined the information matrix $M(p)$, which is given by

$$\begin{aligned} M(p) &= \sum_{j=1}^{J} p_j \underline{v}_j \underline{v}_j^T \qquad (1.10) \\ &= V P V^T. \qquad (1.11) \end{aligned}$$

Two important properties of this matrix is that it is symmetric and nonnegative definite. The symmetry of this matrix follows by its definition, whereas the nonnegativeness property can be verified by

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

We widely use this matrix in optimal experimental design. Because of the reciprocity property, the information matrix is the inverse of the covariance matrix of the parameter estimates.

1.6 Criteria in Optimal Designs

1.6.1 Criteria of a Design

Our goal is to seek for the best inference for some or all of the unknown parameters $\underline{\theta} \in \Theta$. In that case we need to make the information matrix $M(p)$ large in some sense. This can be done by maximizing some real valued function $\phi(p) = \psi\{M(p)\}$. The function ϕ is called the criterion function. Naturally, a design maximizing the criterion $\phi(p)$ is called a ϕ-optimal design.

Suppose the information matrix $M(p)$ is non-singular, that is, it is positive definite. We now consider the different design criteria and their properties as follows.

1.6.2 D-optimality

D-optimality is the most important design criterion in applications. In D-optimality, the determinant of the information matrix $M(p)$ or its logarithm $logdet\{M(p)\}$ is maximized. Therefore, the criterion function is given by

$$\phi_D(p) = \psi_D\{M(p)\} = log\, det\{M(p)\} = -log\, det\{M^{-1}(p)\}. \qquad (1.12)$$

To maximize the determinant of the information matrix is the same as to minimize the determinant of the covariance matrix because of the reciprocity property. Hence, in D-optimality, we minimize the generalized variance of the parameter estimates.

The relationship between D-optimal design and standardized variance of the predicted response is of interest. Suppose we have a design variable x in a given model, and we have a D-optimal design, say p^*. Then it follows that

$$sup_x\, d(x, p^*) \;\; = \;\; k \tag{1.13}$$

where $d(x, p)$ is the standardized variance of the predicted response as defined in (1.9).

In a linear model, if the errors are normally distributed, then the general form of the joint confidence region for the vector of unknown parameters $\underline{\theta} \in \Theta$ can be explained by an ellipsoid. The joint confidence region is:

$$\{\underline{\theta} \,:\, (\underline{\theta} - \underline{\hat{\theta}})^T M(p)(\underline{\theta} - \underline{\hat{\theta}}) \leq c\}, \text{ for some critical value } c, \tag{1.14}$$

where $\underline{\hat{\theta}}$ is the least squares estimate or the maximum likelihood estimate of $\underline{\theta}$. D-optimality seeks to minimize the volume of the ellipsoid since this volume is proportional to $[det\{M(p)\}]^{-\frac{1}{2}}$.

We can also explain the D-optimality criterion in terms of the eigenvalues of the information matrix $M(p)$ or its inverse. Suppose the eigenvalues of $M(p)$ are λ_i, $i = 1, 2, \cdots, k$. Then the eigenvalues of $M^{-1}(p)$ are $1/\lambda_i$, $i = 1, 2, \cdots, k$, which are proportional to the squares of the lengths of the axes of the confidence ellipsoid. Therefore, the product of the eigenvalues of $M^{-1}(p) : \prod_{i=1}^{k} 1/\lambda_i$ is minimized by the D-optimal design.

There are many other properties of D-optimality criterion. Above all, it is a concave function of positive definite symmetric matrices. When the criterion function ϕ_D is finite, it is always differentiable, and the first partial derivatives are given by:

$$\frac{\partial \phi_D}{\partial p_j} = \underline{v}_j^T M^{-1}(p) \underline{v}_j. \tag{1.15}$$

Moreover, the D-optimality criterion is invariant under a non-singular linear transformation of $\mathcal{V}$. This can be easily proved by considering a non-singular transformation of the vertices $(\underline{v}_j)$ in the induced design space $\mathcal{V}$.

This is the most extensively studied criterion in the literature. The references include Kiefer (1959), Fedorov (1972), Silvey (1980), Berger and Wong (2009), Atkinson et al. (2007), Pukelsheim (1993), Mandal and Torsney (2006), and Mandal et al. (2005).

1.6.3 D_A-optimality

Instead of all the k parameters, we may be interested in some of the parameters or some linear combinations of the parameters of the linear model. Suppose that we are interested in s linear combinations of the parameters θ_i, where $i = 1, 2, \cdots, k$, which are elements of $A\underline{\theta}$, where $A^{s\times k}$ is a matrix of rank $s \leq k$. Assuming that $M(p)$ is non-singular, the covariance matrix of the least squares estimator of $A\underline{\theta}$ is proportional to $AM^{-1}(p)A^T$. Then D_A-optimality seeks to maximize the following criterion:

$$\phi_{D_A}(p) = \psi_{D_A}\{M(p)\} = -log\, det\{AM^{-1}(p)A^T\}. \tag{1.16}$$

Sibson (1974) called this as D_A-optimality in order to emphasize the dependence of the design on the matrix of coefficients A. The partial derivative of D_A-optimality criterion is:

$$\frac{\partial \phi_{D_A}}{\partial p_j} = \underline{v}_j^T M^{-1}(p)A^T[AM^{-1}(p)A^T]^{-1}AM^{-1}(p)\underline{v}_j. \tag{1.17}$$

As we can see, this is criterion has very similar properties as D-optimality criterion.

1.6.4 D_s-optimality

This optimality is an important special case of D_A-optimality. We adopt D_s-optimality when we are interested in estimating a subset of s parameters. Note that, in D_A-optimality, if $A = [I_s : O]$, where I_s is the identity matrix with dimension s and O is the zero matrix with dimension $s \times (k-s)$, then the criterion becomes a D_s-optimality criterion. In this situation, without loss of generality, we are interested only in estimating the first s parameters θ_1 to θ_s in $\underline{\theta} \in \Theta$.

As $A = [I_s : O]$, we can partition this information matrix $M(p)$ as:

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

Then the inverse of $AM^{-1}(p)A^T$ can be written as:

$$(AM^{-1}(p)A^T)^{-1} \quad = \quad M_{11} - M_{12}M_{22}^{-1}M_{12}^T$$

(Rhode, 1965). Therefore, in this special case, maximizing ϕ_{D_A} criterion is equivalent to maximizing the following criterion

$$\phi_{D_s}(p) \quad = \quad log\, det\{M_{11} - M_{12}M_{22}^{-1}M_{12}^T\} \tag{1.18}$$

which is known as the D_s-optimality. For further details we refer to the fol-

lowing: Karlin and Studden (1966), Atwood (1969), Silvey and Titterington (1973) and Silvey (1980).

1.6.5 *A*-optimality

In *A*-optimality, we maximize the criterion function as follows:

$$\phi_A(p) = \psi_A\{M(p)\} = -Trace\{M^{-1}(p)\}. \tag{1.19}$$

Thus we see that *A*-optimum design minimizes the sum of the variances of all of the parameter estimates because of the reciprocity property of the covariance matrix and the information matrix. Note that it does not take correlations between the parameter estimates into account.

The simplicity of *A*-optimality criterion is that it only requires to calculate the k diagonal entries of the matrix $M^{-1}(p)$. However, it still has some disadvantages. Unlike *D*-optimality, the *A*-optimality criterion may not be invariant under linear transformations of the scale of the independent variables. That is, each scale of the independent variable may lead to another optimal design. This is a drawback of *A*-optimality. However, this problem of scale does not arise for designs with the factors qualitative such as in block designs. In such a case, *A*-optimality is frequently used. The partial

derivatives of A-optimality criterion are given by:

$$\frac{\partial \phi_A}{\partial p_j} = \underline{v}_j^T M^{-2}(p) \underline{v}_j. \tag{1.20}$$

1.6.6 G-optimality

In G-optimality, we maximize the following criterion function:

$$\phi_G(p) = \psi_G\{M(p)\} = -Max_{\underline{v} \in \mathcal{V}}\, \underline{v}^T M^{-1}(p) \underline{v}. \tag{1.21}$$

We see that the maximum value of $\underline{v}^T M^{-1}(p) \underline{v}$ (which is proportional to the variance of $\underline{v}^T \hat{\underline{\theta}}$) is minimized by this criterion. If we are interested in predicting the response variable as efficiently as possible, we can benefit from this optimality criterion.

This criterion is equivalent to the D-optimality criterion, which is proved by Kiefer and Wolfowitz (1960). The connection between G-optimality and D-optimality is as follows. It is known that the standardized variance for a G-optimal design, say p^*, is less than or equal to k, where k is the number of parameters. That is, $d(\underline{x}, p^*) \leq k$. The equality occurs at the design points. This connects to the link between a D-optimal design and the standardized variance of the predicted response. We can use this inequality to check whether a design is D-optimal or not.

Similar to D-optimality, G-optimality is also invariant under a non-singular linear transformation of the scale of the independent variables. If $\underline{v}_j^T M^{-1}(p)\underline{v}_i = Max_t\, \underline{v}_t^T M^{-1}(p)\underline{v}_t$, then the partial derivative of ϕ_G can be written as follows:

$$\frac{\partial \phi_G}{\partial p_j} \quad = \quad [\underline{v}_j^T M^{-1}(p)\underline{v}_j]^2. \tag{1.22}$$

1.6.7 Linear Optimality

Suppose L is a matrix of coefficients of dimension $k \times k$. A linear or L-optimal design maximizes the criterion function

$$\phi_L(p) = \psi_L\{M(p)\} = -tr\{M^{-1}(p)L\}. \tag{1.23}$$

This function is linear in the elements of the covariance matrix $M^{-1}(p)$. For a detailed description of this criterion, we refer to Fedorov (1972).

The relationship between L-optimal design and D_A-optimal design can be expresses as follows. Suppose L is of rank $s \leq k$. Then L can be written as $L = A^T A$ where A is a matrix of dimension $s \times k$ with rank s. Thus, the above criterion function can be written as

$$\phi_L(p) = -tr\{M^{-1}(p)L\} = -tr\{M^{-1}(p)A^T A\} = -tr\{AM^{-1}(p)A^T\}. \tag{1.24}$$

As we know, the determinant, rather than the trace, of $\{-AM^{-1}(p)A^T\}$ is maximized in D_A-optimality. Therefore, it has an alternative name as A_A-optimality.

The partial derivatives of Linear optimal design are:

$$\frac{\partial \phi_L}{\partial p_j} = \underline{v}_j^T M^{-1}(p) A^T A M^{-1}(p) \underline{v}_j. \tag{1.25}$$

We organize the rest of the thesis as follows. In Chapter 2, we determine the optimality conditions for our optimization problems in terms of a point-to-point directional derivative. We also study the properties of these directional derivatives. In Chapter 3, we formulate our optimization problem for minimizing variances of the least-squares estimates of the parameters of a linear model. We derive the partial derivatives of our criterion for different regression models, then solve the resulting equations and construct the optimal designs. We then construct the optimal designs by an algorithmic approach in Chapter 4. We consider a class of multiplicative algorithms, indexed by a function which satisfies certain conditions. We study the properties of this algorithm, and then construct our optimal designs using such algorithms. In Chapter 5, we construct optimal designs by minimizing the average or total variance of some of the parameter estimates in some regression models. This is motivated by a practical problem in Chemistry. We first formulate the general problem in terms of a set of parametric functions and then reduce

the problem in terms of some parameters by choosing the coefficient vectors as unit vectors according to the parameters of interest. We also attempt to improve the convergence rate of the algorithm by using the properties of the directional derivatives. Finally, in Chapter 6, we conclude the thesis with a brief review of the main findings and a discussion of some possible future work.

Chapter 2

Optimality Conditions

2.1 Introduction

In order to obtain a best inference for the parameters $\underline{\theta}$, we need to make the matrix $M(p)$ large in some sense. Making the matrix $M(p)$ large means what kind of criterion function we are interested in. For example, we may be interested in all of the parameters, a set of the parameters, or some linear combinations of the parameters. As we discussed in the previous chapter, we do this by maximizing a real valued function of the information matrix, that is, $\phi(p) = \psi\{M(p)\}$. The function $\phi(p)$ is called the criterion function. There is a vast literature on optimal design criteria. The criteria we discussed in Chapter 1 are the most common ones. For a detailed list of different criteria

we refer to Atkinson et al. (2007), Pukelsheim (1993), Fedorov (1972) and Silvey (1980). Kiefer (1959) introduced the alphabetical nomenclature for different design criteria. In this thesis, we focus on the criterion functions that are the variances of the parameter estimates of a linear model, and a linear combination of these variances.

The general problem that we wish to consider is the following:

$$\text{Maximize a criterion } \phi(p) \text{ over } \mathcal{P} \equiv \left\{p = (p_1, p_2, \cdots, p_J) : p_j \geq 0, \Sigma p_j = 1\right\}. \quad (2.1)$$

There are a variety of applications of this general problem. Each problem seeks to calculate a probability distribution on the space $\mathcal{P}$. The examples include optimal regression design, estimation, stratified sampling etc.

Let us consider the conditions for optimality for this optimization problem. We focus on a differential calculus approach. In particular, an important tool is the directional derivative (denoted by $F_\phi\{p, q\}$) of a criterion function of $\phi(p)$ at p in the direction of q. We will define this directional derivative in the following section, and will discuss some of its properties. We also define another type of directional derivative, which will be helpful for deriving the partial derivatives of our criterion functions. These directional derivatives play an important simplifying role in our calculus of optimization.

There are two different approaches for us to solve the problem. We

could find out an optimizing p^* directly or first determine an x^* maximizing a criterion over $\mathcal{P}$ and then find a p^* such that $x(p^*) = x^*$. We focus on the former approach. This would require conditions explicitly defining an optimizing p^*. We determine the optimality conditions in the following.

2.2 Directional Derivatives

There are two derivatives of interest. We define these in terms of the criterion function $\phi(p)$.

2.2.1 Definition 1

Using differential calculus, we exploit the directional derivative of Whittle (1973) and determine our optimality conditions. The directional derivative of a criterion function $\phi(.)$ at p in the direction of q is defined as:

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

Wether $\phi(.)$ can be differentiable or not, the above derivative can always

exist. If $\phi(.)$ is differentiable, then (2.2) can be simplified as:

$$F_\phi(p,q) \;=\; (q-p)^T \frac{\partial \phi}{\partial p} \;=\; \sum_{i=1}^{J} (q_i - p_i) d_i \tag{2.3}$$

where $d_i = \partial\phi/\partial p_i \,, i = 1, \cdots, J$.

If $q = e_j$, the j^{th} unit vector in $\mathbb{R}^J$, then the directional derivative can be written as:

$$F_j \;=\; F_\phi(p, e_j) = d_j - \sum_{i=1}^{J} p_i d_i. \tag{2.4}$$

These derivatives correspond to the vertices of the feasible region of our problem, and thus we call these derivatives as the vertex directional derivatives.

2.2.2 Definition 2

Another similar directional derivative could be very useful and is defined as follows:

$$G_\phi\{p,m\} \;=\; lim_{\varepsilon \downarrow 0} \frac{\phi\{p+\varepsilon m\} - \phi(p)}{\varepsilon}. \tag{2.5}$$

There are interesting relationships between the derivatives given in (2.2) and (2.5). First note that, $G_\phi\{p,m\} = F_\phi\{p,q\}$ when $m = q - p$, and $G_\phi\{p,m\} = F_\phi\{p, p+m\}$. This concept also have been used by Kiefer

(1974) in his design theory. Note also that, $G_\phi\{p, e_j\} = \partial^+\phi/\partial p_j$, the right hand partial derivative of $\phi(.)$ with respect to the j^{th} component of p. We take advantage of this for deriving the partial derivatives of our criterion functions.

We define our optimality conditions based on Definition 1 whereas we use Definition 2 for finding partial derivatives of our criterion functions in this thesis.

2.2.3 Properties

- If $p, q \in S$, where S is a convex set, then the convex combination $\{(1-\varepsilon)p + \varepsilon q\}$ also belongs to S. This is apparently good point for us if we want $F_\phi\{p, q\}$ only for $p, q \in S$. However $G_\phi\{p, q\}$ does not have this property.

- If $\phi(.)$ is concave, then $F_\phi\{p, q\}$ satisfies the inequality: $F_\phi\{p, q\} \geq \phi(q) - \phi(p)$. We can prove this by using Jensen's inequality.

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

- $F_\phi\{p, p\} = 0$, a appropriate property if one does not move from p, that is, there is no change effected in $\phi(.)$. Reversely, $G_\phi\{p, p\} = F_\phi\{p, 2p\} \neq 0$.

- A converse concept may be of interest. Consider the directional derivatives of $\phi(.)$ at p as p is approached from the direction of q

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

2.3 Optimality Conditions

Suppose that the criterion function $\phi(.)$ 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 general problem are given by

$$F_j^* = F_\phi\{p^*, e_j\} \begin{cases} = 0 & \text{if } p_j^* > 0 \\ \leq 0 & \text{if } p_j^* = 0 \end{cases} \tag{2.6}$$

where F_j's are the vertex directional derivatives as given in (2.4).

If the criterion function $\phi(.)$ is concave on the feasible region, then the above first-order conditions (2.6) are both necessary and sufficient for optimality. This is known as General Equivalence Theorem in optimal design theory (Kiefer, 1974).

In fact, the above is a simplified version of the general equivalence theorem. We will check whether or not these conditions are satisfied by a postulated solution obtained by our numerical techniques. After we determine the optimality conditions for an optimization problem, we often need to find an appropriate algorithm for finding the optimal design. We will consider a class of multiplicative algorithms in Chapter 4.

Chapter 3

Minimizing Variances of Parameter Estimates in Linear Models: Analytic Approach

3.1 Introduction

As we discussed earlier, our general problem is to maximize a criterion function $\phi(p)$ subject to the constraints $p_j \geq 0$, $j = 1, 2, \cdots, J$ and $\Sigma p_j = 1$. We stated this problem in (2.1). We wish to find an optimizing probability distribution according to a chosen criterion.

In statistical inference, it is important to determine the estimates of the parameters of a regression model in such a way that the variances of the parameter estimates are as minimum as possible. For example, when we construct a UMVUE (Uniformly Minimum-Variance Unbiased Estimator) or a BLUE (Best Linear Unbiased Estimator), we always minimize the variance of the estimators. Motivated by this fact, we construct optimal designs for minimizing variances of the least-squares estimates of the parameters (or linear functions of the parameters) of a linear model. This ensures that the parameters are estimated with minimum variance. In this chapter, we construct such designs by some analytic approach. Then, in the following chapter, we will construct the designs using some algorithmic approach.

3.2 Formulation of the Problem

The criterion function for minimizing variances is a special case of linear optimality. The criterion function for linear optimality is $-tr\{M^{-1}(p)L\}$, where L is a matrix of coefficients of dimension $k \times k$. Note that L can be written as $L = A^T A$, where A is a matrix of dimension $s \times k$ with rank s. Thus, the above criterion function can be written as $-tr\{AM^{-1}(p)A^T\}$.

Suppose that A is a $1 \times k$ vector, say $\underline{c}^T$. This reduces the criterion

function as given by

$$\phi_c(p) \;=\; -\,\underline{c}^T M^{-1}(p)\underline{c}. \tag{3.1}$$

Suppose that we would like to minimize the variance of the estimate of a parametric function $\underline{c}^T\underline{\theta}$. Let $\hat{\underline{\theta}}$ be the least squares estimator of $\underline{\theta}$. Then the variance of the least squares estimator $\underline{c}^T\hat{\underline{\theta}}$ of $\underline{c}^T\underline{\theta}$ is proportional to $\underline{c}^T M^{-1}(p)\underline{c}$. That is, to minimize the variance of $\underline{c}^T\hat{\underline{\theta}}$, we minimize $\underline{c}^T M^{-1}(p)\underline{c}$. Equivalently, we can maximize $-\underline{c}^T M^{-1}(p)\underline{c}$, which is the same criterion as given in (3.1).

Thus, in our general problem (2.1), we will consider $\phi(p) = \phi_c(p) = -\underline{c}^T M^{-1}(p)\underline{c}$. Now, if we choose $\underline{c}$ to be a unit vector, then maximizing $\phi_c(p)$ is equivalent to minimizing the variance of a particular parameter estimate according to the components of the vector $\underline{c}$.

Note that the function $\underline{c}^T M^{-1}(p)\underline{c}$ is a convex function on $\mathcal{M}$, where is $\mathcal{M}$ is the set of all positive definite symmetric matrices. That is, our criterion function $\phi_c(p)$ in (3.1) is a concave function.

3.3 Analytic Solutions for Some Polynomial Regression Models

We now construct optimal designs for minimizing variances of parameter estimates in some polynomial regression models. The polynomial regression model with one variable of order $k-1$ is given by:

$$E(Y|x) = \underline{v}^T \underline{\theta} \tag{3.2}$$

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

In particular, we will consider the quadratic ($k = 3$) and cubic ($k = 4$) regression models. We will also consider a practical problem in Chemistry later in this chapter.

In the following sections, we construct the optimal designs analytically. We first derive the derivatives of the criterion functions with respect to p_j's, and then solving the equations by equating them to zero. Note that some of the equations are too complicated to solve by hand. We solved those equations by using the package Maple.

3.3.1 Quadratic Regression

The model is given by

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

with the design space $\mathcal{X} = [-1, 1]$. We would like to minimize the variance of the estimate of the parameter corresponding to the square term, that is, θ_2. So, in this case our $\underline{c}$ vector is $\underline{c} = (0, 0, 1)^T$. We consider the support points $\{-1, 0, 1\}$. We need to find out the corresponding probabilities or weights of these support points by using the above criterion (3.1).

Here $\phi_c(p) = -\underline{c}^T M^{-1}(p)\underline{c}$ and $M(p) = \sum_{j=1}^{J} p_i v_j v_j^T = \underline{V} P \underline{V}^T$.

For θ_2: $x = \{-1, 0, 1\}$, $c = (0, 0, 1)^T$. So, the V matrix is:

$$V = \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}.$$

Thus, the information matrix is obtained as

$$\begin{aligned} M(p) &= \underline{V} P \underline{V}^T \\ &= \begin{bmatrix} 1 & -2p_1 + 1 - p_2 & 1 - p^2 \\ -2p_1 + 1 - p_2 & 1 - p_2 & -2p_1 + 1 - p_2 \\ 1 - p_2 & -2p_1 + 1 - p_2 & 1 - p_2 \end{bmatrix} \end{aligned}$$

and its determinant is given by

$$det[M(p)] = -4p_1^2 p_2 - 4p_1 p_2^2 + 4p_1 p_2.$$

Then we can obtain $M^{-1}(p)$, which is given by:

$$M^{-1}(p) = \begin{bmatrix} \frac{1}{p_2} & 0 & -\frac{1}{p_2} \\ 0 & \frac{p_2 - 1}{4p_1(p_1 + p_2 - 1)} & -\frac{2p_1 + p_2 - 1}{4p_1(p_1 + p_2 - 1)} \\ -\frac{1}{p_2} & -\frac{2p_1 + p_2 - 1}{4p_1(p_1 + p_2 - 1)} & \frac{4p_1^2 + 4p_1 p_2 + p_2^2 - 4p_1 - p_2}{4p_1 p_2(p_1 + p_2 - 1)} \end{bmatrix}.$$

So, our criterion function becomes

$$\begin{aligned} \phi_c(p) &= -\underline{c}^T M^{-1}(p) \underline{c} \\ &= -(4p_1^2 + 4p_1 p_2 + p_2^2 - 4p_1 - p_2)/D_1 \end{aligned}$$

where $D_1 = 4p_1 p_2(p_1 + p_2 - 1)$. Note that we have the constraint $\sum p_i = 1$, that is, $p_1 + p_2 + p_3 = 1$. So, considering $p_3 = 1 - p_1 - p_2$, we can express

the criterion as a function of p_1 and p_2 only. We now obtain the partial derivatives of the criterion with respect to p_1 and p_2. They are given by

$$\begin{aligned}\frac{\partial \phi_c(p)}{\partial p_1} &= -(8p_1 + 4p_2 - 4)/D_1 + (4p_1^2 + 4p_1p_2 + p_2^2 - 4p_1 - p_2)/D_1p_1 \\ &\quad +4p_1p_2(4p_1^2 + 4p_1p_2 + p_2^2 - 4p_1 - p_2)/D_1^2\end{aligned}$$

$$\begin{aligned}\frac{\partial \phi_c(p)}{\partial p_2} &= -(4p_1 + 2p_2 - 1)/D_1 + (4p_1^2 + 4p_1p_2 + p_2^2 - 4p_1 - p_2)/D_1p_2 \\ &\quad +4p_1p_2(4p_1^2 + 4p_1p_2 + p_2^2 - 4p_1 - p_2)/D_1^2.\end{aligned}$$

Both derivatives are zero at the maximum. Solving the resulting equations, we obtain $p_1 = 0.25$, $p_2 = 0.5$. Hence, $p_3 = 0.25$.

Hence the optimal design for minimizing the variance of $\hat{\theta}_2$ is:

$$\xi^* = \begin{Bmatrix} -1 & 0 & 1 \\ 0.25 & 0.5 & 0.25 \end{Bmatrix}.$$

3.3.2 Cubic Regression

The model is given by

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

with the design space $\mathcal{X} = [-1, 1]$. Here we wish to minimize the variance of the estimate of the parameter corresponding to the cubic term, that is, for θ_3. So, in this case our $\underline{c}$ vector is $\underline{c} = (0, 0, 0, 1)^T$. We consider the support points $\{-1, -0.5, 0.5, 1\}$. We need to find out the corresponding probabilities or weights of these support points by using the above criterion (3.1).

Here $\phi_c(p) = -\underline{c}^T M^{-1}(p)\underline{c}$ and $M(p) = \sum_{j=1}^{J} p_i v_j v_j^T = \underline{V} P \underline{V}^T$.

For θ_3: $x = \{-1, -0.5, 0.5, 1\}$, $c = (0, 0, 0, 1)^T$. Then, we have

$$V = \begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & -0.5 & 0.5 & 1 \\ 1 & 0.25 & 0.25 & 1 \\ -1 & -0.125 & 0.125 & 1 \end{bmatrix}.$$

Thus, we obtain the information matrix as given by

$$
\begin{aligned}
M(p) &= \underline{V} P \underline{V}^T \\
&= \begin{bmatrix}
1 & -2p_1 - 1.5p_2 - 0.5p_3 + 1 \\
-2p_1 - 1.5p_2 - 0.5p_3 + 1 & -0.75p_2 - 0.75p_3 + 1 \\
-0.75p_2 - 0.75p_3 + 1 & -2p_1 - 1.125p_2 - 0.875p_3 + 1 \\
-2p_1 - 1.125p_2 - 0.875p_3 + 1 & -0.9375p_2 - 0.9375p_3 + 1
\end{bmatrix} \\
&\quad \begin{bmatrix}
-0.75p_2 - 0.75p_3 + 1 & -2p_1 - 1.125p_2 - 0.875p_3 + 1 \\
-2p_1 - 1.125p_2 - 0.875p_3 + 1 & -0.9375p_2 - 0.9375p_3 + 1 \\
-0.9375p_2 - 0.9375p_3 + 1 & -2p_1 - 1.03125p_2 - 0.96875p_3 + 1 \\
-2p_1 - 1.03125p_2 - 0.96875p_3 + 1 & -0.98438p_2 - 0.98438p_3 + 1
\end{bmatrix}
\end{aligned}
$$

and its determinant is given by

$$
\begin{aligned}
det[M(p)] &= -1.2 \times 10^{-9} p_2^4 + 1.15 \times 10^{-9} p_2^3 - 4 \times 10^{-9} p_1 p_2^3 - 2 \times 10^{-9} p_1^2 p_2^2 \\
&\quad -7 \times 10^{-10} p_2^3 p_3 - 1 \times 10^{-10} p_2^2 p_3^2 + 2 \times 10^{-9} p_1 p_2^2 - 1.26563 p_1 p_2 p_3^2 \\
&\quad +1.26563 p_2 p_1 p_3 - 1.26563 p_1 p_2^2 p_3 - 1.26563 p_2 p_1^2 p_3.
\end{aligned}
$$

We do not write down $M^{-1}(p)$ here since the expressions are too long. Then,

our criterion function becomes

$$\begin{aligned}\phi_c(p) &= -\underline{c}^T M^{-1}(p)\underline{c} \\ &= -[-2.25p_2p_3 - 9p_1p_2 - 9p_1p_3 + 2.25p_2^2p_3 + 2.25p_2p_3^2 \\ &\quad +9p_2p_1^2 + 9p_1p_2^2 + 9p_3p_1^2 + 9p_1p_3^2 + 18p_2p_1p_3]/D_2\end{aligned}$$

where $D_2 = 5.0625p_3p_1p_2(p_1 + p_2 + p_3 - 1)$.

By considering $p_4 = 1 - p_1 - p_2 - p_3$, we express the criterion as a function of p_1, p_2 and p_3 only. We now obtain the partial derivatives of the criterion with respect to p_1, p_2 and p_3. They are given by

$$\begin{aligned}\frac{\partial\phi_c(p)}{\partial p_1} &= \frac{1}{D_2}[(-2.25p_2p_3 - 9p_1p_2 - 9p_1p_3 + 2.25p_2^2p_3 + 2.25p_2p_3^2 \\ &\quad +9p_2p_1^2 + 9p_1p_2^2 + 9p_3p_1^2 + 9p_1p_3^2 + 18p_2p_1p_3)/p_1 \\ &\quad -(-9p_2 - 9p_3 + 18p_1p_2 + 9p_2^2 + 18p_1p_3 + 9p_3^2 + 18p_2p_3)] \\ &\quad +\frac{1}{D_2^2}[5.0625p_3p_1p_2(-2.25p_2p_3 - 9p_1p_2 - 9p_1p_3 + 2.25p_2^2p_3 \\ &\quad +2.25p_2p_3^2 + 9p_2p_1^2 + 9p_1p_2^2 + 9p_3p_1^2 + 9p_1p_3^2 + 18p_2p_1p_3)]\end{aligned}$$

$$
\begin{aligned}
\frac{\partial \phi_c(p)}{\partial p_2} &= \frac{1}{D_2}[(-2.25p_2p_3 - 9p_1p_2 - 9p_1p_3 + 2.25p_2^2p_3 + 2.25p_2p_3^2 \\
&\quad +9p_2p_1^2 + 9p_1p_2^2 + 9p_3p_1^2 + 9p_1p_3^2 + 18p_2p_1p_3)/p_2 \\
&\quad -(-2.25p_3 - 9p_1 + 4.5p_2p_3 + 2.25p_3^2 + 9p_1^2 + 18p_1p_2 + 18p_1p_3)] \\
&\quad +\frac{1}{D_2^2}[5.0625p_3p_1p_2(-2.25p_2p_3 - 9p_1p_2 - 9p_1p_3 + 2.25p_2^2p_3 \\
&\quad +2.25p_2p_3^2 + 9p_2p_1^2 + 9p_1p_2^2 + 9p_3p_1^2 + 9p_1p_3^2 + 18p_2p_1p_3)]
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \phi_c(p)}{\partial p_3} &= \frac{1}{D_2}[(-2.25p_2p_3 - 9p_1p_2 - 9p_1p_3 + 2.25p_2^2p_3 + 2.25p_2p_3^2 \\
&\quad +9p_2p_1^2 + 9p_1p_2^2 + 9p_3p_1^2 + 9p_1p_3^2 + 18p_2p_1p_3)/p_3 \\
&\quad -(-2.25p_2 - 9p_1 + 2.25p_2^2 + 4.5p_2p_3 + 9p_1^2 + 18p_1p_3 + 18p_1p_2)] \\
&\quad +\frac{1}{D_2^2}[5.0625p_3p_1p_2(-2.25p_2p_3 - 9p_1p_2 - 9p_1p_3 + 2.25p_2^2p_3 \\
&\quad +2.25p_2p_3^2 + 9p_2p_1^2 + 9p_1p_2^2 + 9p_3p_1^2 + 9p_1p_3^2 + 18p_2p_1p_3)].
\end{aligned}
$$

Those derivatives are zero at the maximum. Solving the resulting equations, we obtain $p_1 = 0.167$, $p_2 = 0.333$ and $p_3 = 0.333$. Hence, $p_4 = 0.167$.

Hence the optimal design for minimizing the variance of $\hat{\theta}_3$ is:

$$
\xi^* = \begin{Bmatrix} -1 & -0.5 & 0.5 & 1 \\ 0.167 & 0.333 & 0.333 & 0.167 \end{Bmatrix}.
$$

3.3.3 A Practical Model in Chemistry

This is a practical problem arising in Chemistry, considered by Torsney and Alahmadi (1995). The regression model describes the relationship between the viscosity Y and the concentration x of a chemical solution. Viscosity is the response. The model is

$$E(Y|x) \quad = \quad \theta_0 x + \theta_1 x^{1/2} + \theta_2 x^2,$$

with the design interval restricted to (0, 0.2]. We would like to minimize the variance of the estimate of the parameter corresponding to the x, $x^{1/2}$ and x^2 term, that is, for the parameters θ_0, θ_1 and θ_2. So, in this case our $\underline{c}$ vectors are $(1,0,0)^T$, $(0,1,0)^T$ and $(0,0,1)^T$ respectively. We consider the support points $\{0.02, 0.12, 0.2\}$. We choose these points as we know these are the support points for this model. We will confirm this in the next chapter. We need to find out the corresponding probabilities or weights of these support points by using the above criterion (3.1).

For θ_0

We have $\phi_c(p) = -\underline{c}^T M^{-1}(p)\underline{c}$ and $M(p) = \sum_{j=1}^{J} p_i v_j v_j^T = \underline{V} P \underline{V}^T$.

$x = \{0.02, 0.12, 0.2\}$, $c = (1, 0, 0)^T$. So, we have the V matrix:

$$V = \begin{bmatrix} x \\ x^{0.5} \\ x^2 \end{bmatrix} = \begin{bmatrix} 0.02 & 0.12 & 0.2 \\ 0.14142 & 0.34641 & 0.44721 \\ 0.0004 & 0.0144 & 0.04 \end{bmatrix}.$$

Thus, the information matrix is obtained as

$$\begin{aligned} M(p) &= \underline{V} P \underline{V}^T \\ &= \begin{bmatrix} -0.0396p_1 - 0.0256p_2 & -0.08661p_1 - 0.04787p_2 & -0.007992p_1 - 0.006272p_2 \\ +0.04 & +0.08944 & +0.008 \\ -0.08661p_1 - 0.04787p_2 & -0.18p_1 - 0.08p_2 & -0.01783p_1 - 0.01290p_2 \\ +0.08944 & +0.2 & +0.01788 \\ -0.00799p_1 - 0.00627p_2 & -0.01783p_1 - 0.01290p_2 & -0.00160p_1 - 0.00139p_2 \\ +0.008 & +0.01789 & +0.0016 \end{bmatrix} \end{aligned}$$

and its determinant is given by

$$\begin{aligned} det[M(p)] =& 1.5 \times 10^{-15} p_1 - 6 \times 10^{-15} p_2 + 1.67560 \times 10^{-8} p_1 p_2 - 1.67560 \times 10^{-8} p_1^2 p_2 \\ & - 1.67560 \times 10^{-8} p_1 p_2^2 - 5 \times 10^{-16} p_2^3 - 1.1 \times 10^{-14} p_1^2 + 6 \times 10^{-15} p_2^2 \\ & + 9.3 \times 10^{-15} p_1^3. \end{aligned}$$

Then we obtain: $M^{-1}(p) =$

$$\begin{bmatrix}
\begin{array}{c}(3.0008\times10^{-5}p_1^2+8.1411\times10^{-5}p_1p_2\\ -3.0008\times10^{-5}p_1+5.5005\times10^{-5}p_2^2\\ -5.5005\times10^{-5}p_2)/D_3\end{array} &
\begin{array}{c}-(3.9441\times10^{-6}p_1^2+1.7728\times10^{-5}p_1p_2\\ -3.9441\times10^{-6}p_1+1.4240\times10^{-5}p_2^2\\ -1.4240p_2\times10^{-5})/D_3\end{array} &
\begin{array}{c}-(1.0594\times10^{-4}p_1^2+2.0270\times10^{-4}p_1p_2\\ -1.0594\times10^{-4}p_1+1.1582\times10^{-4}p_2^2\\ -1.1582\times10^{-4}p_2)/D_3\end{array} \\
\begin{array}{c}-(3.9441\times10^{-6}p_1^2+1.7728\times10^{-5}p_1p_2\\ -3.9441\times10^{-6}p_1+1.4240\times10^{-5}p_2^2\\ -1.4240\times10^{-5}p_2)/D_3\end{array} &
\begin{array}{c}(5.184\times10^{-7}p_1^2+4.1472\times10^{-6}p_1p_2\\ -5.184\times10^{-7}p_1+3.6864\times10^{-6}p_2^2\\ -3.6864\times10^{-6}p_2)/D_3\end{array} &
\begin{array}{c}(1.3925\times10^{-5}p_1^2+4.1498\times10^{-5}p_1p_2\\ -1.3925\times10^{-5}p_1+2.9984\times10^{-5}p_2^2\\ -2.9984\times10^{-5}p_2)/D_3\end{array} \\
\begin{array}{c}-(1.0594\times10^{-4}p_1^2+2.0270\times10^{-4}p_1p_2\\ -1.0594\times10^{-4}p_1+1.1582\times10^{-4}p_2^2\\ -1.1582\times10^{-4}p_2)/D_3\end{array} &
\begin{array}{c}(1.3925\times10^{-5}p_1^2+4.1498\times10^{-5}p_1p_2\\ -1.3925\times10^{-5}p_1+2.9984\times10^{-5}p_2^2\\ -2.9984\times10^{-5}p_2)/D_3\end{array} &
\begin{array}{c}(3.7404\times10^{-4}p_1^2+5.1706\times10^{-4}p_1p_2\\ -3.7404\times10^{-4}p_1+2.4387\times10^{-4}p_2^2\\ -2.4387\times10^{-4}p_2)/D_3\end{array}
\end{bmatrix}$$

So, our criterion function becomes

$$\begin{aligned}
\phi_c(p) &= -\underline{c}^T M^{-1}(p)\underline{c} \\
&= -(3.00081\times10^{-5}p_1^2+8.14110\times10^{-5}p_1p_2-3.00081\times10^{-5}p_1 \\
&\quad +5.50049\times10^{-5}p_2^2-5.50049\times10^{-5}p_2)/D_3
\end{aligned}$$

where $D_3 = 1.1\times10^{-14}p_1^3+1.67560\times10^{-8}p_1^2p_2+1.6756\times10^{-8}p_1p_2^2+1.5\times10^{-14}p_1^2-2\times10^{-15}p_2^2-1.67560\times10^{-8}p_1p_2-8\times10^{-15}p_2-6\times10^{-15}p_1$. Note that we have the constraint $\sum p_i = 1$, that is, $p_1+p_2+p_3 = 1$. So, considering $p_3 = 1-p_1-p_2$, we can express the criterion as a function of p_1 and p_2 only. We now obtain the partial derivatives of the criterion with respect to p_1 and

p_2. They are given by

$$\begin{aligned}
\frac{\partial \phi_c(p)}{\partial p_1} &= -(6.00163 \times 10^{-5} p_1 + 8.14110 \times 10^{-5} p_2 - 3.00081 \times 10^{-5})/D_3 \\
&\quad +(3.00081 \times 10^{-5} p_1^2 + 8.14110 \times 10^{-5} p_1 p_2 - 3.00081 \times 10^{-5} p_1 \\
&\quad +5.50049 \times 10^{-5} p_2^2 - 5.50049 \times 10^{-5} p_2)(3.3 \times 10^{-14} p_1^2 \\
&\quad +3.35120 \times 10^{-8} p_1 p_2 + 1.6756 \times 10^{-8} p_2^2 + 3 \times 10^{-14} p_1 \\
&\quad -1.67560 \times 10^{-8} p_2 - 6 \times 10^{-15})/D_3^2
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \phi_c(p)}{\partial p_2} &= -(8.14110 \times 10^{-5} p_1 + 1.10010 \times 10^{-4} p_2 - 5.50049 \times 10^{-5})/D_3 \\
&\quad +(3.00081 \times 10^{-5} p_1^2 + 8.14110 \times 10^{-5} p_1 p_2 - 3.00081 \times 10^{-5} p_1 \\
&\quad +5.50049 \times 10^{-5} p_2^2 - 5.50049 \times 10^{-5} p_2)(1.67560 \times 10^{-8} p_1^2 \\
&\quad +3.3512 \times 10^{-8} p_1 p_2 - 4 \times 10^{-15} p_2 - 1.67560 \times 10^{-8} p_1 \\
&\quad -8 \times 10^{-15})/D_3^2.
\end{aligned}$$

Both derivatives are zero at the maximum. Solving the resulting equations, we obtain $p_1 = 0.501$, $p_2 = 0.370$. Hence, $p_3 = 0.129$.

Hence the optimal design for minimizing the variance of $\hat{\theta}_0$ is:

$$\xi^* = \begin{Bmatrix} 0.02 & 0.12 & 0.2 \\ 0.501 & 0.370 & 0.129 \end{Bmatrix}.$$

For θ_1

Now we would like to minimize the variance of the estimate of the parameter corresponding to $x^{1/2}$ term, that is, θ_1. Here $c = (0, 1, 0)^T$ and we already know $M^{-1}(p)$. Therefore our criterion function becomes

$$\begin{aligned}
\phi_c(p) &= -\underline{c}^T M^{-1}(p)\underline{c} \\
&= -(5.184 \times 10^{-7} p_1^2 + 4.1472 \times 10^{-6} p_1 p_2 - 5.184 \times 10^{-7} p_1 \\
&\quad +3.6864 \times 10^{-6} p_2^2 - 3.6864 \times 10^{-6} p_2)/D_3.
\end{aligned}$$

Similar to before, we express the criterion as a function of p_1 and p_2 only. Then the partial derivatives are given by

$$\begin{aligned}
\frac{\partial \phi_c(p)}{\partial p_1} &= -(1.0368 \times 10^{-6} p_1 + 4.1472 \times 10^{-6} p_2 - 5.184 \times 10^{-7})/D_3 \\
&\quad +(5.184 \times 10^{-7} p_1^2 + 4.1472 \times 10^{-6} p_1 p_2 - 5.184 \times 10^{-7} p_1 \\
&\quad +3.6864 \times 10^{-6} p_2^2 - 3.6864 \times 10^{-6} p_2)(3.3 \times 10^{-14} p_1^2 \\
&\quad +3.35120 \times 10^{-8} p_1 p_2 + 1.6756 \times 10^{-8} p_2^2 + 3 \times 10^{-14} p_1 \\
&\quad -1.67560 \times 10^{-8} p_2 - 6 \times 10^{-15})/D_3^2
\end{aligned}$$

$$\begin{aligned}\frac{\partial \phi_c(p)}{\partial p_2} &= -(4.1472 \times 10^{-6} p_1 + 7.3728 \times 10^{-6} p_2 - 3.6864 \times 10^{-6})/D_3 \\ &\quad +(5.184 \times 10^{-7} p_1^2 + 4.1472 \times 10^{-6} p_1 p_2 - 5.184 \times 10^{-7} p_1 \\ &\quad +3.6864 \times 10^{-6} p_2^2 - 3.6864 \times 10^{-6} p_2)(1.67560 \times 10^{-8} p_1^2 \\ &\quad +3.3512 \times 10^{-8} p_1 p_2 - 4 \times 10^{-15} p_2 - 1.67560 \times 10^{-8} p_1 \\ &\quad -8 \times 10^{-15})/D_3^2.\end{aligned}$$

Both derivatives are zero at the maximum. Solving the resulting equations, we obtain $p_1 = 0.667$, $p_2 = 0.250$. Hence, $p_3 = 0.083$.

Hence the optimal design for minimizing the variance of $\hat{\theta}_1$ is:

$$\xi^* = \begin{Bmatrix} 0.02 & 0.12 & 0.2 \\ 0.667 & 0.250 & 0.083 \end{Bmatrix}.$$

For θ_2

Finally, we would like to minimize the variance of the estimate of the parameter corresponding to the square term, that is, θ_2. Here $c = (0, 0, 1)^T$. Therefore our criterion function becomes

$$\begin{aligned}\phi_c(p) &= -\underline{c}^T M^{-1}(p) \underline{c} \\ &= -(3.74036 \times 10^{-4} p_1^2 + 5.17059 \times 10^{-4} p_1 p_2 - 3.74036 \times 10^{-4} p_1 \\ &\quad +2.43872 \times 10^{-4} p_2^2 - 2.43872 \times 10^{-4} p_2)/D_3.\end{aligned}$$

Here also we express the criterion as a function of p_1 and p_2 only. Then the partial derivatives are given by

$$\begin{aligned}
\frac{\partial \phi_c(p)}{\partial p_1} &= -(7.48071 \times 10^{-4} p_1 + 5.17059 \times 10^{-4} p_2 - 3.74036 \times 10^{-4})/D_3 \\
&\quad +(3.74036 \times 10^{-4} p_1^2 + 5.17059 \times 10^{-4} p_1 p_2 - 3.74036 \times 10^{-4} p_1 \\
&\quad +2.43872 \times 10^{-4} p_2^2 - 2.43872 \times 10^{-4} p_2)(3.3 \times 10^{-14} p_1^2 \\
&\quad +3.35120 \times 10^{-8} p_1 p_2 + 1.6756 \times 10^{-8} p_2^2 + 3 \times 10^{-14} p_1 \\
&\quad -1.67560 \times 10^{-8} p_2 - 6 \times 10^{-15})/D_3^2
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \phi_c(p)}{\partial p_2} &= -(5.17059 \times 10^{-4} p_1 + 4.87744 \times 10^{-4} p_2 - 2.43872 \times 10^{-4})/D_3 \\
&\quad +(3.74036 \times 10^{-4} p_1^2 + 5.17059 \times 10^{-4} p_1 p_2 - 3.74036 \times 10^{-4} p_1 \\
&\quad +2.43872 \times 10^{-4} p_2^2 - 2.43872 \times 10^{-4} p_2)(1.67560 \times 10^{-8} p_1^2 \\
&\quad +3.3512 \times 10^{-8} p_1 p_2 - 4 \times 10^{-15} p_2 - 1.67560 \times 10^{-8} p_1 \\
&\quad -8 \times 10^{-15})/D_3^2.
\end{aligned}$$

Both derivatives are zero at the maximum. Solving the resulting equations, we obtain $p_1 = 0.347$, $p_2 = 0.430$. Hence, $p_3 = 0.223$.

Hence the optimal design for minimizing the variance of $\hat{\theta}_2$ is:

$$\xi^* = \left\{ \begin{array}{ccc} 0.02 & 0.12 & 0.2 \\ 0.347 & 0.430 & 0.223 \end{array} \right\}.$$

3.4 Graphical Display

Recall that in Chapter 1, we considered the standardized variance of the predicted response in (1.9). For one design variable x, this is given by $d(x,p) = \underline{f}^T(x)M^{-1}(p)\underline{f}(x)$. We simply call this as variance function. Note that, one property of D-optimality is that the supremum of $d(x,p^*)$ is the number of parameters (Atkinson et al., 2007). Since we consider minimizing variance of one parameter estimate, we need to adjust this variance function in our case.

For any model $E(Y|x) = \underline{f}^T(x)\underline{\theta}$, where $\underline{\theta}$ has k parameters, we divide it into two groups, such as

$$E(Y|x) = \underline{f}_1^T(x)\underline{\theta}_{(1)} + \underline{f}_2^T(x)\underline{\theta}_{(2)} \tag{3.3}$$

where $\underline{\theta}_{(1)}$ has $k-s$ parameters and $\underline{\theta}_{(2)}$ has s parameters.

We will first divide information matrix into four parts, which can be written as

$$M(p) = \begin{bmatrix} M_{11}(p) & M_{12}(p) \\ M_{21}(p) & M_{22}(p) \end{bmatrix}. \tag{3.4}$$

Suppose we are interested in $\underline{\theta}_{(2)}$. Let $C = [O : I_s]$, where O is the $s \times (k-s)$ zero matrix and I_s is the $s \times s$ identity matrix. Then we can easily find out

covariance matrix related to $\underline{\theta}_{(2)}$ is

$$M^{22}(p) = CM^{-1}(p)C^T = [M_{22}(p) - M_{12}(p)M_{11}^{-1}(p)M_{12}(p)]^{-1}. \qquad (3.5)$$

As we are interested in $\underline{\theta}_{(2)}$, the determinant which should be maximized is

$$\frac{|M(p)|}{|M_{11}(p)|} = |M_{22}(p) - M_{12}(p)M_{11}^{-1}(p)M_{21}(p)|. \qquad (3.6)$$

Then the expression of the variance function for $\underline{\theta}_{(2)}$ can be written as

$$d(x,p) = \underline{f}^T(x)M^{-1}(p)\underline{f}(x) - \underline{f}_{-1}^T(x)M_{11}^{-1}(p)\underline{f}_{-1}(x) \qquad (3.7)$$

where $d(x,p) \leq s$, s is the number of parameters in $\underline{\theta}_{(2)}$. In our case, we are interested in minimizing the variance of one parameter estimate, that is, here, $s = 1$.

For example, suppose we are interested in θ_2 in our Quadratic Regression Model $E(Y|x) = \theta_0 + \theta_1 x + \theta_2 x^2$. We obtained the design measure as given by

$$\text{For } \theta_2\text{: } \xi^* = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ 0.25 & 0.5 & 0.25 \end{array} \right\}.$$

We first divide the form into two parts

$$E(Y|x) \;=\; \underline{f}_1^T(x)\underline{\theta}_{(1)} + \underline{f}_2^T(x)\underline{\theta}_{(2)}$$

where $\underline{f}_1(x) = (1\ x)^T$, $\underline{\theta}_{(1)} = (\theta_0\ \theta_1)^T$, $\underline{f}_2(x) = x^2$ and $\underline{\theta}_{(2)} = \theta_2$. Then from (3.7), we have at the optimum p^*,

$$\begin{aligned} d(x,p^*) &= \underline{f}^T(x)M^{-1}(p^*)\underline{f}(x) - \underline{f}_1^T(x)M_{11}^{-1}(p^*)\underline{f}_1(x) \\ &= \begin{bmatrix} 1 & x & x^2 \end{bmatrix} \begin{bmatrix} 2 & 0 & -2 \\ 0 & 2 & 0 \\ -2 & 0 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix} - \begin{bmatrix} 1 & x \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix} \\ &= 1 - 4x^2 + 4x^4. \end{aligned}$$

Similarly we obtain the expressions of the variance functions for all of the above optimal designs as given by

- Quadratic Regression for θ_2: $d(x,p^*) = 1 - 4x^2 + 4x^4$
- Cubic Regression for θ_1: $d(x,p^*) = 9.00007x^2 - 23.95252x^4 + 15.93663x^6$
- Cubic Regression for θ_3: $d(x,p^*) = 1.20144 \times 10^{-14}x + 9.01201x^2 - 3.20172 \times 10^{-14}x^3 - 2.40160 \times 10x^4 + 2.13306 \times 10^{-14}x^5 + 1.6 \times 10x^6$
- Chemistry Model for θ_1: $d(x,p^*) = 495.0116x - 5086.3462x^{1.5} + 13065.813x^2 + 15479.8444x^{2.5} - 79529.296x^3 + 121020.186x^4$

- Chemistry Model for θ_2: $d(x, p^*) = 495.0299x - 5084.9711x^{1.5} + 13058.2685x^2 + 15468.9617x^{2.5} - 79448.97x^3 + 120845.63x^4$.

In Figures 3.1-3.5 we plot of these variance functions. The red horizontal dash line indicates $s = 1$ as we are interested in only one parameter. The blue vertical dash lines indicate the support points of the design. We can see that the maximum value of $s = 1$ occurs at the support points for each graph.

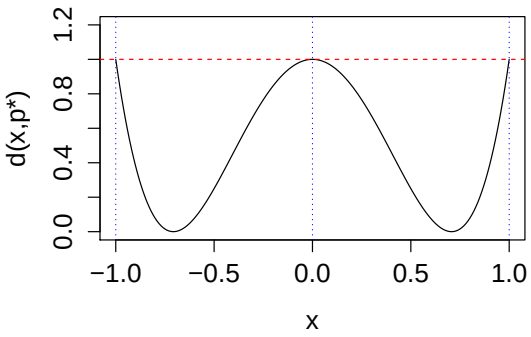

Figure 3.1: Quadratic Regression for θ_2.

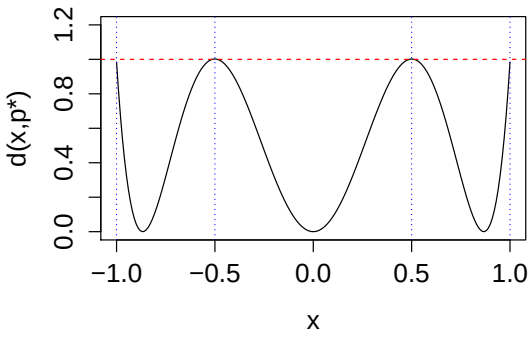

Figure 3.2: Cubic Regression for θ_1.

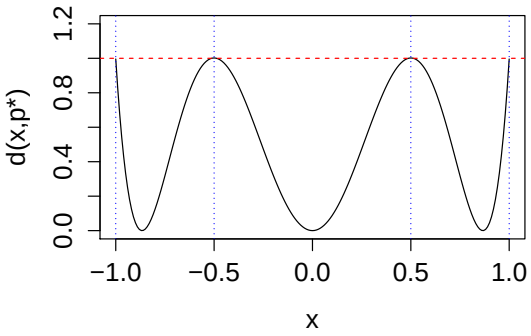

Figure 3.3: Cubic Regression for θ_3.

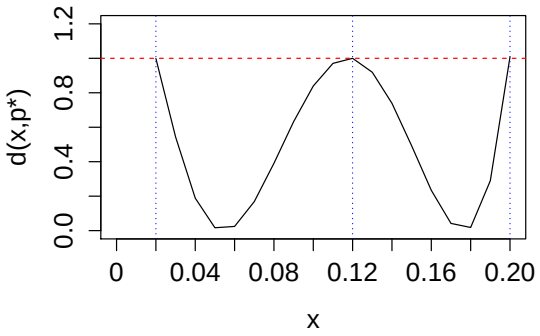

Figure 3.4: Chemistry Model for θ_1.

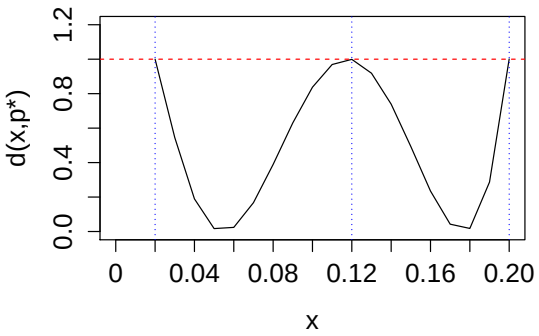

Figure 3.5: Chemistry Model for θ_2.

Chapter 4

Minimizing Variances of Parameter Estimates in Linear Models: Algorithmic Approach

4.1 Introduction

An analytic solution of an optimization problem for constructing optimal designs is possible only in simple cases. Sometime it may be possible to find the optimal weights when the support points are known in advance as we did in the previous chapter. It is typically not possible to evaluate an explicit solution p^* to problem (2.1) or in particular to derive an optimal

regression design explicitly. Iterative techniques must be needed to determine the optimal design.

We mentioned earlier why there is a need for special algorithms. From Caratheodory's Theorem (Silvey, 1980) we know that there always exists an optimal design measure with finite support. We wish to identify an optimizing p^*. As we mentioned in Chapter 1, typically we have a continuous design space $\mathcal{X}$. We need to discretize this space, in order to find out the optimal design. Selecting a value of x in the design space $\mathcal{X}$ is the same as selecting a k-vector $\underline{v}$ in the closed bounded k-dimensional space $\mathcal{V}$, the induced design space as we discussed in Chapter 1.

Assume that the discretized design space $\mathcal{V}$ consists of J distinct vectors $\underline{v}_1, \underline{v}_2, \cdots, \underline{v}_J$. We call $\underline{v}_j$'s as the vertices of the feasible region of our problem. A value for $\underline{v}$ must first be chosen from the J elements of $\mathcal{V}$ to obtain an observation on y.

An ideal discretization would be some form of uniform grid on a continuous space $\mathcal{X}$. Thus, we approximate the design space $\mathcal{X}$ by a grid of J points equally spaced at intervals between the end points of $\mathcal{X}$. After discretizing this space $\mathcal{X}$, we then find the induced design space $\mathcal{V} = \{\underline{v}_1, \underline{v}_2, \cdots, \underline{v}_J\}$. As we consider many points in the design space, there will be many zero weights at the optimum. Hence we consider the following class of algorithms, indexed by a function which depends on derivatives of the criterion function.

4.2 A Class of Algorithms

The general problem (2.1), mentioned in Chapter 2, has a particular set of constraints, such as nonnegative variables $p_1, p_2, \cdots, p_J$, and $\sum p_i = 1$. An iteration which neatly submits to these and has many nice properties is the following algorithm:

$$p_j^{(r+1)} \propto p_j^{(r)} f(d_j^{(r)}) \tag{4.1}$$

where $d_j^{(r)} = \left.\frac{\partial \phi}{\partial p_j}\right|_{p=p^{(r)}}$ and the function $f(.)$ is positive, strictly increasing and may depend on a free positive parameter δ.

This type of iteration was proposed by Torsney (1977) by taking the function $f(d)$ as d^δ, where $\delta > 0$. Silvey et al. (1978) did some empirical study for the choice of δ. Torsney (1988) considered the function $f(d) = e^{\delta d}$ in a variety of applications. Torsney (1983) explored monotonicity of particular values of the free parameter δ. Mandal and Torsney (2000) explored some systematic choices of $f(.)$. Torsney and Mandal (2001) and Mandal et al. (2005) explored this algorithm for some constrained optimal design problems. In the case of D-optimality, Titterington (1976) described a proof of monotonicity of $f(d, \delta) = d$. Mandal and Torsney (2006) further developed the algorithm using a clustering approach.

4.2.1 Properties of the Algorithm

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

- Under the conditions imposed on $f(.)$, algorithm (4.1) guarantees that $F_\phi\{p^{(r)}, p^{(r+1)}\} \geq 0$, 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 $p^{(r+1)}$. Equality occurs when the partial derivatives d_j's corresponding to nonzero weights p_j's have a common value, d. Therefore, in this case,

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

  with $x = d$ or 0.

- Under the above iteration, the support of the design measure is such that $supp(p^{(r+1)}) \subseteq supp(p^{(r)})$. Moreover, some weights can converge to zero.

- If the partial derivatives $\partial\phi/\partial p_j^{(r)}$ corresponding to nonzero $p_j^{(r)}$ have a common value, then an iterate $p^{(r)}$ is a fixed point of the iteration. In such a case the corresponding directional derivatives $F_j^{(r)}$ are zero.

4.3 Construction of Optimal Designs for Minimizing Variances of Parameter Estimates

We now construct optimal designs for minimizing variances of parameter estimates in the polynomial regression models that we considered in the previous chapter. In our general problem (2.1), we will consider the criterion function $\phi(p) = \phi_c(p) = -\underline{c}^T M^{-1}(p)\underline{c}$. If we choose $\underline{c}$ to be a unit vector, then maximizing $\phi_c(p)$ is equivalent to minimizing the variance of a particular parameter estimate according to the components of the vector $\underline{c}$.

We now construct the optimal designs for various regression models in the following sections. We take the design space $\mathcal{X} = [-1, 1]$. As we discussed earlier, we discretize this design space in some form of uniform grid between the two end points -1 and 1. In particular, we approximate the design interval by a grid of points equally spaced at intervals of 0.01.

In order to find the optimal design by using an algorithm, we first need to find the partial derivatives d_j of our criterion function. We derive this in the following theorem.

Theorem 4.3. For $\phi(p) = \phi_c(p) = -Var(\underline{c}^T\underline{\hat{\theta}}) = -\underline{c}^T M^{-1}(p)\underline{c}$, d_j, $j = 1, 2, \ldots, J$, are given by

$$d_j = (\underline{v}_j^T M^{-1}(p)\,\underline{c})^2$$

where $\underline{v}_j$'s are the vertices in the induced design space.

Proof. Since $\phi(p)$ is a function of p as well as a function of the information matrix $M(p)$, let $\phi(p) = \psi(M(p))$. Let N be another matrix similar to M. From the definition of Gâteaux derivative of Section 2,

$$G_\psi\{M, N\} = \lim_{\varepsilon \downarrow 0} \frac{\psi(M + \varepsilon N) - \psi(M)}{\varepsilon}. \tag{4.2}$$

$G_\psi\{M, N\}$ is called the Gâteaux derivative of ψ at M in the direction of N. Now,

$$\begin{aligned} (M + \varepsilon N)^{-1} &= (I + \varepsilon M^{-1} N)^{-1} M^{-1} \\ &= M^{-1/2}(I + \varepsilon M^{-1/2} N M^{-1/2})^{-1} M^{-1/2} \\ &= M^{-1/2}(I + \varepsilon B)^{-1} M^{-1/2} \end{aligned} \tag{4.3}$$

where $B = M^{-1}N$ or $M^{-1/2}NM^{-1/2}$ and

$$\begin{aligned} (I + \varepsilon B)^{-1} &= I - \varepsilon B + \varepsilon^2 B^2 - \varepsilon^3 B^3 + \varepsilon^4 B^4 - \cdots\cdots \\ &= I - \varepsilon B \quad \text{for } \varepsilon \text{ small.} \end{aligned} \tag{4.4}$$

Thus, by using (4.3) and (4.4) above, $\psi(M+\varepsilon N)$ would be

$$\begin{aligned}
\psi(M+\varepsilon N) &= -\underline{c}^T(M+\varepsilon N)^{-1}\underline{c} \\
&= -\underline{c}^T M^{-1/2}(I-\varepsilon B)M^{-1/2}\underline{c} \\
&= -\underline{c}^T M^{-1}\underline{c} + \varepsilon\, \underline{c}^T M^{-1/2} B M^{-1/2}\underline{c} \\
&= -\underline{c}^T M^{-1}\underline{c} + \varepsilon\, \underline{c}^T M^{-1} N M^{-1}\underline{c}. \qquad (4.5)
\end{aligned}$$

Then, from (4.5) and (4.2), with $M = M(p)$ and $N = \underline{v}_j\underline{v}_j^T$, we obtain

$$\begin{aligned}
\frac{\partial\phi(p)}{\partial p_j} &= G_\psi\{M(p),\, \underline{v}_j\underline{v}_j^T\} \\
&= (\underline{v}_j^T M^{-1}(p)\,\underline{c})^2.
\end{aligned}$$

This completes the proof of the theorem.

As we already have the partial derivatives of our criterion function, we can use an algorithm to construct the optimal designs for our optimization problem. Note that, choice of the function $f(.)$ and its argument play an important role in the convergence of the algorithm (4.1). Thus, we write the full form of the algorithm as given by:

$$p_j^{(r+1)} = \frac{p_j^{(r)} f(x_j^{(r)}, \delta)}{\sum_{j=1}^{J} p_j^{(r)} f(x_j^{(r)}, \delta)} \qquad (4.6)$$

with the appropriate choice of the first argument x in the function $f(x, \delta)$.

We first report the performance of algorithm (4.6) by taking $f(x, \delta)$ as $x = d$, the partial derivatives of our variance criterion (3.1). In particular, we choose $f(d, \delta) = exp(\delta d)$, $f(d, \delta) = \Phi(\delta d)$, $f(d, \delta) = exp(\delta d)/(1 + exp(\delta d))$. We chose these functions because they satisfy the required conditions, that is, they are positive and increasing.

We also attempt to improve convergence of the algorithm by considering some objective choices of $f(x, \delta)$. Note that our criterion (3.1) has both positive and negative vertex directional derivatives. So, if we consider $x = F$, the function $f(F, \delta)$ needs to be defined for positive and negative F. From equation (2.4), 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

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

These properties suggest that a suitable function is one that is centred at zero and changes quickly about $F = 0$.

Two choices of the function $f(x, \delta)$ with the potential to satisfy these requirements are $f(x, \delta) = \Phi(\delta x)$, the normal cumulative distribution function; and $f(x, \delta) = exp(\delta x)/(1 + exp(\delta x))$, the logistic cumulative distribution function. These functions change quickly at zero.

Note also that the choice of $f(x, \delta)$ depends on a free parameter δ which

should be positive. Clearly the value of δ is crucial in this function. Keeping this in mind we explore various values of δ and report the values that achieve the fastest convergence.

We record, for $n = 1, 2, \cdots, 6$, the number of iterations needed to achieve $\underset{1\le j\le J}{max}\{F_j\} \le 10^{-n}$, where F_j's are the vertex directional derivatives. In all the cases we take the initial design to be $p_j^{(0)} = 1/J$, $j = 1, 2, \cdots, J$.

4.3.1 Quadratic Regression Model

For this model, we take $k = 3$ in (3.2), and $\underline{v} = \underline{\eta}(x) = (1, x, x^2)^T$, $x \in [-1, 1]$, $\underline{\theta} = (\theta_0, \theta_1, \theta_2)^T$. We would like to construct our minimum variance design for the parameters θ_1 and θ_2.

We first run the algorithm (4.6) by taking $f(x, \delta)$ as $x = d$, the partial derivatives of our variance criterion (3.1). We choose $f(d, \delta) = exp(\delta d)$, $f(d, \delta) = \Phi(\delta d)$, $f(d, \delta) = exp(\delta d)/(1 + exp(\delta d))$.

As we discussed before, we attempt to improve the convergence of the algorithm by replacing $x = d$ by $x = F$. However, note that, for exponential function $f(x, \delta) = exp(\delta x)$, replacing d by F does not make any difference. This can be easily proved by the expression of $p_j^{(r+1)}$ in the algorithm. We report the iteration results in Tables 4.3.1, 4.3.2, 4.3.3, and 4.3.4.

Table 4.3.1: Quadratic Regression for θ_1 with $x = d$

Number of iterations need to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$						
$f(x,\delta) = exp(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.93	5	51	154	276	400	524
0.97	5	48	147	264	383	502
0.99	5	47	144	259	375	492
1	5	47	143	256	372	487
1.01	**4**	**46**	**141**	**254**	**368**	**485**
$f(x,\delta) = \Phi(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.75	40	196	528	913	1305	1698
0.8	42	197	527	909	1298	1688
0.825	**42**	**198**	**527**	**909**	**1297**	**1687**
0.85	**43**	**199**	**528**	**910**	**1298**	**1687**
0.875	44	201	530	912	1301	1690
0.95	46	207	540	925	1317	1710
$f(x,\delta) = exp(\delta x)/(1+exp(\delta x))$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
1.05	40	206	562	975	1397	1818
1.225	42	207	555	959	1370	1782
1.275	**43**	**208**	**555**	**959**	**1370**	**1781**
1.325	44	209	557	961	1372	1784
1.4	45	212	562	968	1380	1793

We see some advantages of using $f(x,\delta)$ with $x = F$ instead of $x = d$ by looking at Tables 4.3.1-4.3.4. For instance, in this model, for $f(x,\delta) = \Phi(\delta x)$ with $x = d$, $\delta = 0.825$ or 0.85 and $n = 6$, the number of iterations needed is 1687 (see Table 4.3.1), whereas by using $x = F$ and $\delta = 1.25$, this number reduces to 494 (see Table 4.3.2). Let us also look at the results for the choice $f(x,\delta) = exp(\delta x)/(1+exp(\delta x))$. For example, when $x = d$, $\delta = 0.325$ and

Table 4.3.2: Quadratic Regression for θ_1 with $x = F$

Number of iterations need to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$						
$f(x,\delta) = \Phi(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
1.05	12	65	180	314	451	588
1.1	12	62	172	300	431	561
1.2	11	57	158	276	395	515
1.225	11	56	155	270	387	504
1.25	**11**	**55**	**152**	**265**	**379**	**494**
$f(x,\delta) = exp(\delta x)/(1+exp(\delta x))$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
1.8	11	61	168	293	420	548
1.9	11	58	159	278	398	519
1.95	11	57	156	271	388	506
2	11	55	152	264	379	493
2.025	**11**	**55**	**150**	**261**	**374**	**492**

$n = 6$, the number of iterations needed is 72465 (see Table 4.3.3). However this number reduces to 21233 for $\delta = 0.475$ by using $x = F$ (see Table 4.3.4). As we mentioned earlier, replacing d by F in the exponential function $f(x,\delta) = exp(\delta x)$ does not make any difference. So we did not consider $x = F$ for this function.

We obtain the optimal designs as given by:

$$\text{For } \theta_1\text{: } \xi^* = \left\{ \begin{array}{cc} -1 & 1 \\ 0.5 & 0.5 \end{array} \right\}, \text{ with minimum variance} = 1$$

Table 4.3.3: Quadratic Regression for θ_2 with $x = d$

Number of iterations need to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$						
$f(x,\delta) = exp(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.07	62	698	7119	30343	51423	72034
0.08	53	610	6229	26550	44994	63029
0.0825	51	591	6039	25745	43630	61118
0.085	49	572	5861	24987	42346	59320
0.0875	**44**	**553**	**5690**	**24269**	**41133**	**57622**
$f(x,\delta) = \Phi(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.125	74	778	7845	33389	56571	79238
0.15	71	722	7253	30859	52281	73227
0.2	**73**	**682**	**6803**	**28919**	**48988**	**68611**
0.225	77	686	6819	28978	49084	68743
0.275	89	733	7228	30686	51969	72778
$f(x,\delta) = exp(\delta x)/(1+exp(\delta x))$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.225	74	764	7681	32679	55364	77546
0.3	75	721	7200	30613	51858	72631
0.325	**77**	**721**	**7186**	**30545**	**51740**	**72465**
0.375	82	737	7316	31083	52648	73733
0.425	89	771	7629	32396	54867	76839

For θ_2: $\xi^*=\left\{ \begin{array}{ccc} -1 & 0 & 1 \\ 0.25 & 0.5 & 0.25 \end{array} \right\}$, with minimum variance = 4.

In all the cases, the directional derivatives are zero for the support points but negative for other points. That is, the directional derivatives satisfy the first-order conditions (2.6). By using the properties of the directional derivatives, we improve the performance of convergence of the algorithm quite a lot.

Table 4.3.4: Quadratic Regression for θ_2 with $x = F$

Number of iterations need to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$						
$f(x,\delta) = \Phi(\delta F)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.125	43	491	5000	21300	36093	50558
0.175	32	351	3572	15215	25782	36113
0.25	24	246	2501	10652	18048	25280
0.275	25	224	2274	9684	16408	22982
0.3	**59**	**207**	**2086**	**8878**	**15041**	**21068**
$f(x,\delta) = exp(\delta x)/(1+exp(\delta x))$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.275	32	357	3627	15451	26181	36673
0.375	25	262	2661	11332	19200	26894
0.425	24	232	2348	9999	16942	23730
0.45	29	219	2218	9444	16001	22412
0.475	**47**	**208**	**2102**	**8948**	**15159**	**21233**

4.3.2 Cubic Regression Model

For this model, we take $k = 4$ in (3.2), and $\underline{v} = \underline{\eta}(x) = (1, x, x^2, x^3)^T$, $x \in [-1, 1]$, $\underline{\theta} = (\theta_0, \theta_1, \theta_2, \theta_3)^T$. We would like to construct our minimum variance design for the parameters θ_1, θ_2 and θ_3.

We first run the algorithm (4.6) by taking $f(x, \delta)$ as $x = d$, the partial derivatives of our variance criterion (3.1). We choose $f(d, \delta) = exp(\delta d)$, $f(d, \delta) = \Phi(\delta d)$, $f(d, \delta) = exp(\delta d)/(1 + exp(\delta d))$.

We again attempt to improve the convergence of the algorithm by replacing $x = d$ by $x = F$ for this model as well. We report the iteration

results in Tables 4.3.5, 4.3.6, 4.3.7, 4.3.8, 4.3.9 and 4.3.10.

Table 4.3.5: Cubic Regression for θ_1 with $x = d$

Number of iterations need to achieve $\underset{1\le j\le J}{max}\{F_j\} \le 10^{-n}$						
$f(x,\delta) = exp(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.075	62	660	3664	6615	9469	12314
0.095	49	521	2893	5222	7475	9722
0.1025	45	482	2681	4840	6928	9010
0.1075	64	460	2556	4615	6606	8590
0.11	**128**	**433**	**2498**	**4510**	**6455**	**8395**
$f(x,\delta) = \Phi(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.025	290	3005	16657	30061	43025	55949
0.05	184	1862	10286	18555	26552	34524
0.1	**160**	**1542**	**8458**	**15241**	**21801**	**28340**
0.125	177	1660	9069	16331	23354	30355
0.175	266	2386	12921	23234	33206	43147
$f(x,\delta) = exp(\delta x)/(1+exp(\delta x))$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.1	173	1728	9528	17182	24585	31965
0.135	166	1624	8924	16085	23011	29915
0.14	**167**	**1622**	**8912**	**16062**	**22978**	**29871**
0.145	167	1623	8914	16065	22982	29876
0.16	170	1641	9000	16218	23197	30155

We again see some advantages of using $f(x,\delta)$ with $x = F$ instead of $x = d$ by looking at Tables 4.3.5-4.3.10. For instance, in this model, for $f(x,\delta) = \Phi(\delta x)$ with $x = d$, $\delta = 0.1$ and $n = 6$, the number of iterations needed is 28340 (see Table 4.3.5), whereas by using $x = F$ and $\delta = 0.1375$, this number reduces to 8421 (see Table 4.3.6). Let us also look at the results for the choice $f(x,\delta) = exp(\delta x)/(1+exp(\delta x))$. For example, when $x = d$,

Table 4.3.6: Cubic Regression for θ_1 with $x = F$

Number of iterations need to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$						
$f(x,\delta) = \Phi(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.11	57	568	3137	5657	8095	10525
0.12	53	522	2876	5186	7421	9648
0.13	48	482	2655	4788	6850	8906
0.135	63	464	2557	4611	6597	8577
0.1375	**111**	**454**	**2510**	**4527**	**6477**	**8421**
$f(x,\delta) = exp(\delta x)/(1+exp(\delta x))$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.18	56	554	3059	5517	7895	10264
0.2	50	500	2754	4966	7105	9238
0.21	46	476	2623	4730	6767	8798
0.215	59	465	2562	4620	6610	8594
0.22	**121**	**442**	**2504**	**4515**	**6460**	**8398**

$\delta = 0.325$ and $n = 6$, the number of iterations needed is 72465 (see Table 4.3.7). However this number reduces to 20278 for $\delta = 0.4975$ by using $x = F$ (see Table 4.3.8).

We obtain the optimal designs as given by:

For θ_1: $\xi^*=\begin{Bmatrix} -1 & -0.5 & 0.5 & 1 \\ 0.056 & 0.444 & 0.444 & 0.056 \end{Bmatrix}$, with minimum variance = 9

For θ_2: $\xi^*=\begin{Bmatrix} -1 & 0 & 1 \\ 0.25 & 0.5 & 0.25 \end{Bmatrix}$, with minimum variance = 4.

Table 4.3.7: Cubic Regression for θ_2 with $x = d$

Number of iterations need to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$						
$f(x,\delta) = exp(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.065	67	752	7667	32678	55379	77575
0.07	62	698	7119	30343	51423	72034
0.08	53	610	6229	26550	44994	63029
0.085	49	572	5861	24987	42346	59320
0.0875	**44**	**553**	**5690**	**24269**	**41133**	**57622**
$f(x,\delta) = \Phi(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.125	74	778	7845	33389	56571	79238
0.15	71	722	7253	30859	52281	73227
0.2	**73**	**682**	**6803**	**28919**	**48988**	**68611**
0.225	77	686	6819	28978	49084	68743
0.25	82	703	6964	29577	50094	70156
$f(x,\delta) = exp(\delta x)/(1+exp(\delta x))$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.25	74	741	7432	31613	53556	75011
0.3	75	721	7200	30613	51858	72631
0.325	**77**	**721**	**7186**	**30545**	**51740**	**72465**
0.35	79	726	7226	30710	52017	72851
0.4	85	752	7451	31649	53603	75070

For θ_3: $\xi^*=\left\{ \begin{array}{cccc} -1 & -0.5 & 0.5 & 1 \\ 0.167 & 0.333 & 0.333 & 0.167 \end{array} \right\}$,
with minimum variance = 16.

In all the cases, the directional derivatives are zero for the support points but negative for other points. That is, the directional derivatives satisfy the first-order conditions. Again, by using the properties of the di-

Table 4.3.8: Cubic Regression for θ_2 with $x = F$

Number of iterations need to achieve $\max_{1\leq j\leq J}\{F_j\} \leq 10^{-n}$						
$f(x,\delta) = \Phi(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.175	32	351	3572	15215	25782	36113
0.2	29	308	3126	13314	22559	31599
0.25	24	246	2501	10652	18048	25280
0.275	25	224	2274	9684	16408	22982
0.3	**59**	**207**	**2086**	**8878**	**15041**	**21068**
$f(x,\delta) = exp(\delta x)/(1+exp(\delta x))$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.47	41	210	2124	9043	15320	21459
0.48	57	207	2081	8855	15002	21012
0.49	103	165	2040	8676	14697	20585
0.495	183	311	2021	8590	14551	20379
0.4975	**323**	**581**	**2013**	**8549**	**14479**	**20278**

rectional derivatives, here also we improve the performance of convergence of the algorithm.

Table 4.3.9: Cubic Regression for θ_3 with $x = d$

Number of iterations need to achieve $\max_{1\leq j\leq J}\{F_j\} \leq 10^{-n}$						
$f(x,\delta) = exp(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.016	297	3099	11675	19329	26842	34343
0.0174	272	2848	10734	17772	24681	31578
0.018	262	2753	10376	17179	23857	30524
0.0184	255	2692	10149	16804	23336	29858
0.0186	**250**	**2661**	**10038**	**16620**	**23082**	**29532**
$f(x,\delta) = \Phi(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.03	316	3195	12001	19858	27571	35270
0.04	283	2835	10634	17592	24422	31241
0.05	**275**	**2720**	**10188**	**16849**	**23388**	**29916**
0.06	282	2759	10317	17058	23675	30280
0.07	303	2924	10919	18046	25043	32028
$f(x,\delta) = exp(\delta x)/(1+exp(\delta x))$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.05	319	3214	12067	19966	27721	35461
0.07	291	2900	10870	17979	24958	31924
0.08	**291**	**2873**	**10759**	**17793**	**24697**	**31590**
0.09	296	2904	10865	17966	24936	31894
0.1	305	2983	11154	18441	25593	32733

Table 4.3.10: Cubic Regression for θ_3 with $x = F$

Number of iterations need to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$						
$f(x,\delta) = \Phi(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.055	111	1133	4260	7050	9789	12523
0.06	101	1039	3905	6463	8973	11480
0.07	87	891	3348	5540	7692	9840
0.075	76	832	3125	5171	7179	9184
0.0775	**223**	**805**	**3024**	**5004**	**6948**	**8888**
$f(x,\delta) = exp(\delta x)/(1+exp(\delta x))$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.08	121	1243	4673	7734	10739	13739
0.09	108	1105	4154	6875	9546	12213
0.1125	87	885	3324	5501	7637	9770
0.12	74	830	3117	5157	7160	9160
0.1225	**114**	**813**	**3053**	**5052**	**7014**	**8973**

4.3.3 Chemistry Model

For this model, we have $\underline{v} = \underline{\eta}(x) = (x, x^{1/2}, x^2)^T$, $x \in (0, 0.2]$, $\underline{\theta} = (\theta_0, \theta_1, \theta_2)^T$. We would like to construct our minimum variance design for the parameters θ_1 and θ_2.

We first run the algorithm (4.6) by taking $f(x, \delta)$ as $x = d$, the partial derivatives of our variance criterion (3.1). We choose $f(d, \delta) = exp(\delta d)$, $f(d, \delta) = \Phi(\delta d)$, $f(d, \delta) = exp(\delta d)/(1 + exp(\delta d))$. We then attempt to improve the convergence of the algorithm by replacing $x = d$ by $x = F$. We report the iteration results in Tables 4.3.11, 4.3.12, 4.3.13, and 4.3.14.

We see some advantages of using $f(x, \delta)$ with $x = F$ instead of $x = d$ by looking at Tables 4.3.11-4.3.14. For instance, in this model, for $f(x, \delta) = \Phi(\delta x)$ with $x = d$, $\delta = 1.7 \times 10^{-3}$ or 1.75×10^{-3} and $n = 6$, the number of iterations needed is 1798 (see Table 4.3.11), whereas by using $x = F$ and $\delta = 2.4 \times 10^{-3}$, this number reduces to 544 (see Table 4.3.12). Let us also look at the results for the choice $f(x, \delta) = exp(\delta x)/(1 + exp(\delta x))$. For example, when $x = d$, $\delta = 1 \times 10^{-5}$ and $n = 6$, the number of iterations needed is 2607 (see Table 4.3.13). However this number reduces to 748 for $\delta = 1.6 \times 10^{-5}$ by using $x = F$ (see Table 4.3.14).

We obtain the optimal designs as given by:

Table 4.3.11: Chemistry for θ_1 with $x = d$

Number of iterations need to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$						
$f(x,\delta) = exp(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
5.35×10^{-4}	625	916	1206	1497	1787	2077
5.55×10^{-4}	608	888	1168	1448	1728	2007
5.65×10^{-4}	601	876	1151	1426	1701	1975
5.75×10^{-4}	595	866	1136	1406	1676	1946
$\mathbf{5.85\times10^{-4}}$	**593**	**859**	**1125**	**1390**	**1656**	**1921**
$f(x,\delta) = \Phi(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
1.5×10^{-3}	520	781	1041	1301	1562	1822
1.6×10^{-3}	515	773	1031	1289	1547	1805
$\mathbf{1.7\times10^{-3}}$	**512**	**770**	**1027**	**1284**	**1541**	**1798**
$\mathbf{1.75\times10^{-3}}$	**512**	**769**	**1026**	**1283**	**1541**	**1798**
1.85×10^{-3}	513	771	1029	1287	1545	1803
1.95×10^{-3}	516	776	1037	1297	1557	1817
$f(x,\delta) = exp(\delta x)/(1+exp(\delta x))$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
2×10^{-3}	573	857	1140	1424	1707	1991
2.4×10^{-3}	550	824	1097	1371	1644	1917
$\mathbf{2.6\times10^{-3}}$	**547**	**820**	**1092**	**1363**	**1635**	**1907**
2.8×10^{-3}	548	821	1094	1367	1639	1912
3.8×10^{-3}	603	905	1207	1508	1809	2111

For θ_1: $\xi^*=\left\{\begin{array}{ccc} 0.02 & 0.12 & 0.20 \\ 0.667 & 0.250 & 0.083 \end{array}\right\}$, with minimum variance = 495.011

For θ_2: $\xi^*=\left\{\begin{array}{ccc} 0.02 & 0.12 & 0.20 \\ 0.347 & 0.430 & 0.223 \end{array}\right\}$, with minimum variance = 120845.605.

Table 4.3.12: Chemistry for θ_1 with $x = F$

Number of iterations need to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$						
$f(x,\delta) = \Phi(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
2.1×10^{-3}	180	272	363	455	546	638
2.28×10^{-3}	159	244	328	412	496	581
2.32×10^{-3}	151	234	317	400	483	565
2.36×10^{-3}	143	225	306	387	469	550
$\mathbf{2.4 \times 10^{-3}}$	**144**	**224**	**304**	**384**	**464**	**544**
$f(x,\delta) = exp(\delta x)/(1+exp(\delta x))$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
2.6×10^{-3}	230	348	467	585	703	822
3.5×10^{-3}	173	261	349	436	524	612
3.8×10^{-3}	162	243	324	404	485	566
$\mathbf{3.9 \times 10^{-3}}$	**158**	**237**	**315**	**394**	**472**	**551**
4×10^{-3}	316	432	546	660	774	888

In all the cases, the directional derivatives are zero for the support points but negative for other points. That is, the directional derivatives satisfy the first-order conditions.

Note also that, in all of the above models, we improve the performance of convergence of the algorithm quite a lot by by using the properties of the directional derivatives.

Table 4.3.13: Chemistry for θ_2 with $x = d$

Number of iterations need to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$						
$f(x,\delta) = exp(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
2.5×10^{-6}	1147	1401	1656	1910	2165	2419
3×10^{-6}	955	1167	1379	1591	1803	2015
3.2×10^{-6}	895	1094	1293	1492	1690	1889
3.5×10^{-6}	818	1000	1182	1363	1545	1727
$\mathbf{3.7\times10^{-6}}$	**773**	**945**	**1117**	**1289**	**1461**	**1632**
$f(x,\delta) = \Phi(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
4×10^{-6}	1383	1687	1992	2297	2602	2907
6×10^{-6}	1190	1452	1713	1975	2236	2497
$\mathbf{7\times10^{-6}}$	**1173**	**1430**	**1687**	**1944**	**2201**	**2458**
8×10^{-6}	1189	1449	1710	1970	2230	2490
1.1×10^{-5}	1421	1730	2040	2349	2658	2967
$f(x,\delta) = exp(\delta x)/(1+exp(\delta x))$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
4×10^{-6}	1878	2293	2708	3123	3538	3953
8×10^{-6}	1298	1584	1869	2155	2440	2726
$\mathbf{1\times10^{-5}}$	**1243**	**1516**	**1789**	**2061**	**2334**	**2607**
1.2×10^{-5}	1252	1526	1801	2075	2349	2623
2.2×10^{-5}	1962	2387	2813	3238	3664	4089

Table 4.3.14: Chemistry for θ_2 with $x = F$

Number of iterations need to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$						
$f(x,\delta) = \Phi(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
9.4×10^{-6}	379	463	546	630	714	797
9.8×10^{-6}	364	444	524	604	685	765
1×10^{-5}	357	435	514	592	671	750
$\mathbf{1.01 \times 10^{-5}}$	**353**	**431**	**509**	**587**	**665**	**742**
1.02×10^{-5}	409	477	547	615	683	753
$f(x,\delta) = exp(\delta x)/(1+exp(\delta x))$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
8×10^{-6}	713	871	1029	1187	1345	1503
1×10^{-5}	570	696	822	948	1074	1201
1.4×10^{-5}	406	496	586	676	765	855
1.5×10^{-5}	379	463	546	630	714	798
$\mathbf{1.6 \times 10^{-5}}$	**356**	**434**	**513**	**591**	**669**	**748**

Chapter 5

Optimal Designs for Minimizing the Average or Total Variance of the Parameter Estimates

5.1 Introduction

In the previous two chapters we constructed designs for minimizing individual variances of the parameter estimates in some regression models. In some regression designs, it may be of interest to construct optimal designs for min-

imizing the average variance of some of these parameter estimates. Sometime it may also be of interest to construct designs subject to having the equality of variances or covariances of the estimates of some parametric functions. In these directions, some work has been done by Torsney and Alahmadi (1995) and Mandal et al. (2005). They considered the problem of constructing optimal design by minimizing the covariances between the estimates of two linear parametric functions in linear regression design. Mandal et al. (2005) transformed the problem to a constrained optimization problem in which case they minimized some functions of the design weights simultaneously.

In this chapter we consider the case of minimizing the average variance of these parameter estimates. Note that minimizing the average variance is the same as minimizing the total variance.

5.2 Formulation of the Problem

Consider the polynomial regression model with one variable as given in (3.2). However, for notational convenience, let us consider the parameters as $\theta_1, \theta_2, \cdots, \theta_k$ instead of $\theta_0, \theta_1, \cdots, \theta_{k-1}$. The model is then given by

$$E(Y|x) = \underline{v}^T \underline{\theta} \tag{5.1}$$

where $\underline{v} = \underline{\eta}(x) = (1, x, x^2, \cdots, x^k)^T$, $x \in [-1, 1]$, $\underline{\theta} = (\theta_1, \theta_2, \cdots, \theta_k)^T$, $\underline{v} \in \mathcal{V} = \{\underline{v} : \underline{v} = \underline{\eta}(x) = (1, x, x^2, \cdots, x^k)^T, -1 \leq x \leq 1\}$, the induced design space.

Suppose that we are interested in a set of s ($s \leq k$) parameters or parametric functions in the model. That is, we wish to minimize the total variance of the estimates of these parameters or some parametric functions. We first formulate the problem in terms of a set of parametric functions. We then reduce the problem in terms of some parameters by choosing the coefficient vectors as unit vectors according to the parameters of interest. Without loss of generality, let us consider s parametric functions $\underline{a}_i^T\underline{\theta}$, $i = 1, 2, \cdots, s$. That is, we would like to minimize $\sum_{i=1}^{s} Var(\underline{a}_i^T\hat{\underline{\theta}}) = \sum_{i=1}^{s} \underline{a}_i^T M^{-1}(p)\underline{a}_i$. Equivalently, we can maximize the negative of this function, that is, we maximize $-\sum_{i=1}^{s} \underline{a}_i^T M^{-1}(p)\underline{a}_i$.

That is, in our general problem (2.1), we are interested in maximizing

$$\phi(p) = -\sum_{i=1}^{s} \underline{a}_i^T M^{-1}(p)\underline{a}_i \tag{5.2}$$

subject to the constraints $p_j \geq 0$ and $\sum p_j = 1$.

Now, let $\phi_i(p) = -\underline{a}_i^T M^{-1}(p)\underline{a}_i$, $i = 1, 2, \cdots, s$. Note that each function $\underline{a}_i^T M^{-1}(p)\underline{a}_i$ is a convex function on $\mathcal{M}$, where is $\mathcal{M}$ is the set of all positive definite symmetric matrices. That is, each $\phi_i(p)$ is a concave function.

Note also that, maximizing $\phi_i(p)$ will give us the optimal design for minimizing the variance of $\underline{a}_i^T\underline{\theta}$, $i = 1, 2, \cdots, s$. So, according to these $\phi_i(p)$'s, we would like to maximize

$$\phi(p) = \sum_{i=1}^{s} \phi_i(p) \tag{5.3}$$

subject to the constraints $p_j \geq 0$ and $\sum p_j = 1$.

Note that when $\underline{a}_i$'s are unit vectors, we minimize the total or average variance of s parameters. When $s = k$ we minimize the total or average variance of all the parameters. So A-optimality is recovered when $s = k$.

Let the partial derivatives of $\phi_i(p)$ be $d_j^{\phi_i} = \partial\phi_i(p)/\partial p_j$, $i = 1, 2, \cdots, s$, $j = 1, 2, \cdots, J$. These derivatives can be obtained by using Theorem 4.3 of Chapter 4. Let the corresponding directional derivatives be $F_j^{\phi_i}$.

We now need to find out an appropriate algorithm to construct the optimal design for the above optimization problem. We can use algorithm (4.6), but with an appropriate choice of x in the function $f(x, \delta)$. If we choose the argument x as the partial derivatives, then choice of x for our criterion function (5.3) is

$$x_j \;=\; d_j^{\phi_1} + d_j^{\phi_2} + \cdots + d_j^{\phi_s}. \tag{5.4}$$

Using the partial derivatives makes the convergence of the algorithm quite slow depending on the nature of the derivatives as we have seen in the previous chapter. We again attempt to improve convergence by replacing the partial derivatives $d_j^{\phi_i}$'s by the corresponding directional derivatives $F_j^{\phi_i}$'s in the argument of the function $f(x, \delta)$. That is, in this case, we consider

$$x_j = F_j^{\phi_1} + F_j^{\phi_2} + \cdots + F_j^{\phi_s}. \tag{5.5}$$

When we replace the partial derivatives by the directional derivatives, we have seen that the two choices of the function $f(x, \delta)$ perform well, which are $f(x, \delta) = \Phi(\delta x)$, the normal cumulative distribution function; and $f(x, \delta) = exp(\delta x)/(1 + exp(\delta x))$, the logistic cumulative distribution function. These functions change quickly at zero. This was the motivation for replacing partial derivatives by the directional derivatives. Also, the value of δ is crucial in this function. Keeping this in mind we explore various values of δ and report the values which achieve the fastest convergence.

We consider again the three regression models, that is, Quadratic, Cubic and Chemistry models, and report the results in the following sections.

5.2.1 Quadratic Regression Model

For the Quadratic regression model, we have

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

Suppose that we are interested in minimizing the average or total variance of the estimates of the parameters θ_2 and θ_3. We then maximize

$$\phi(p) = \phi_2(p) + \phi_3(p) \tag{5.6}$$

subject to the constraints $p_j \geq 0$ and $\sum p_j = 1$. Here $\phi_2(p) = -\underline{a}_2^T M^{-1}(p)\underline{a}_2$, $\underline{a}_2 = (0,\, 1,\, 0)^T$, $\phi_3(p) = -\underline{a}_3^T M^{-1}(p)\underline{a}_3$, $\underline{a}_3 = (0,\, 0,\, 1)^T$.

As we discussed earlier, we discretize this design space in some form of uniform grid between the two end points -1 and 1. In particular, we approximate the design interval by a grid of points equally spaced at intervals of 0.01. We first run the algorithm (4.6) by taking $f(x, \delta)$ with x as the partial derivatives as given in (5.4), according to the criterion function (5.6). We choose $f(d, \delta) = \Phi(\delta d)$, $f(d, \delta) = exp(\delta d)/(1 + exp(\delta d))$.

We record, for $n = 1, 2, \cdots, 6$, the number of iterations needed to achieve $\underset{1 \leq j \leq J}{max}\{F_j\} \leq 10^{-n}$, where F_j are the vertex directional derivatives. In all the cases we take the initial design to be $p_j^{(0)} = 1/J$, $j = 1,\, 2,\, \cdots,\, J$.

Table 5.2.1: Quadratic Regression for $-(Var(\hat{\theta}_2) + Var(\hat{\theta}_3))$ with $x = d$

Number of iterations need to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$ $f(x,\delta) = \Phi(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.12	95	1003	10094	41229	69286	96755
0.13	96	990	9953	40645	68302	95381
0.134	97	987	9919	40504	68065	95049
0.138	97	986	9897	40413	67910	94833
0.14	**98**	**985**	**9891**	**40385**	**67862**	**94766**
$f(x,\delta) = exp(\delta x)/(1+exp(\delta x))$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.17	101	1076	10846	44305	74457	103978
0.18	100	1062	10693	43677	73399	102500
0.19	99	1052	10582	43219	72628	101423
0.195	99	1048	10541	43047	72340	101019
0.2	**100**	**1045**	**10508**	**42911**	**72111**	**100699**

As we discussed earlier, we then attempt to improve the convergence of the algorithm by replacing $x = d$ by $x = F$, and consider x as given in (5.5).

We report the iteration results in Tables 5.2.1-5.2.2. We can see the advantages of using the directional derivatives, and convergence of the algorithm is improved by great deal. For example, when we choose $f(x,\delta) = \Phi(\delta x)$, with $x = d$, $\delta = 0.14$ and $n = 6$, 94766 iterations are needed to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$ (see Table 5.2.1). Whereas only 28041 iteration numbers are needed when we choose $x = F$ and $\delta = 0.214$ (see Table 5.2.2). Similar things happen for the choice $f(x,\delta) = exp(\delta x)/(1+exp(\delta x))$.

The optimal design measure and variance for $\phi(p) = -(Var(\hat{\theta}_2) +$

Table 5.2.2: Quadratic Regression for $-(Var(\hat{\theta}_2) + Var(\hat{\theta}_3))$ with $x = F$

Number of iterations need to achieve $\max_{1\leq j\leq J}\{F_j\} \leq 10^{-n}$						
$f(x,\delta) = \Phi(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.19	31	325	3291	13449	22603	31566
0.2	41	310	3128	12779	21475	29989
0.21	91	141	2984	12175	20458	28567
0.214	**321**	**565**	**2937**	**11956**	**20083**	**28041**
0.215	4065	11317	18491	25627	32759	39891
$f(x,\delta) = exp(\delta x)/(1+exp(\delta x))$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.33	59	301	3027	12360	20771	29005
0.332	67	300	3009	12286	20646	28831
0.335	85	133	2984	12178	20463	28574
0.34	171	299	2943	12002	20165	28158
0.343	**1521**	**4299**	**7075**	**9817**	**19991**	**27921**

$Var(\hat{\theta}_3))$ are:

$$\xi^*=\left\{\begin{array}{ccc} -1 & 0 & 1 \\ 0.293 & 0.414 & 0.293 \end{array}\right\}, \text{ with minimum total variance} = 5.828427.$$

Suppose now that we are interested in minimizing the total variance of the estimates of all the parameters θ_1, θ_2 and θ_3 in the model. We then maximize

$$\phi(p) = \phi_1(p) + \phi_2(p) + \phi_3(p) \tag{5.7}$$

subject to the constraints $p_j \geq 0$ and $\sum p_j = 1$. Here $\phi_1(p) = -\underline{a}_1^T M^{-1}(p)\underline{a}_1$,

$\underline{a}_1 = (1,\, 0,\, 0)^T$, $\phi_2(p) = -\underline{a}_2^T M^{-1}(p)\underline{a}_2$, $\underline{a}_2 = (0,\, 1,\, 0)^T$, $\phi_3(p) = -\underline{a}_3^T M^{-1}(p)\underline{a}_3$, $\underline{a}_3 = (0,\, 0,\, 1)^T$.

We run the algorithm (4.6) by considering the same set-up as before. In $f(x, \delta)$, we take x as given in (5.4) and (5.5), the partial and directional derivatives of our criterion function (5.7) respectively. We report the results in Tables 5.2.3-5.2.4.

Table 5.2.3: Quadratic Regression for $-(Var(\hat{\theta}_1) + Var(\hat{\theta}_2) + Var(\hat{\theta}_3))$ with $x = d$

Number of iterations need to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$ $f(x,\delta) = \Phi(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.085	131	1384	13923	50789	83596	115821
0.09	130	1368	13749	50151	82544	114362
0.1	130	1352	13565	49469	81419	112802
0.105	**130**	**1351**	**13544**	**49390**	**81288**	**112620**
0.11	131	1354	13567	49470	81418	112799
$f(x,\delta) = exp(\delta x)/(1+exp(\delta x))$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.144	137	1435	14417	52583	86546	119907
0.15	137	1430	14359	52367	86190	119412
0.16	**137**	**1428**	**14325**	**52239**	**85977**	**119116**
0.168	138	1432	14350	52326	86119	119312
0.176	140	1439	14418	52567	86515	119860

We again see the improvement in convergence here. For instance, for the choice $f(x, \delta) = exp(\delta x)/(1 + exp(\delta x))$, with $x = d$, $\delta = 0.16$ and $n = 6$, 119116 iterations are needed to be obtain the optimal design (see Table 5.2.3), whereas only 33305 iterations are needed when we choose $x = F$ and

Table 5.2.4: Quadratic Regression for $-(Var(\hat{\theta}_1) + Var(\hat{\theta}_2) + Var(\hat{\theta}_3))$ with $x = F$

Number of iterations need to achieve $\max_{1 \le j \le J}\{F_j\} \le 10^{-n}$						
$f(x, \delta) = \Phi(\delta x)$						
δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$
0.14	39	441	4459	16273	26786	37113
0.15	59	412	4162	15189	25001	34639
0.153	99	405	4082	14892	24512	33961
0.155	201	311	4034	14704	24200	33527
0.1565	**1513**	**2633**	**3747**	**14581**	**23986**	**33224**
$f(x, \delta) = exp(\delta x)/(1 + exp(\delta x))$						
δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$
0.22	41	448	4528	16525	27201	37688
0.23	39	428	4331	15807	26019	36049
0.24	59	411	4151	15149	24935	34548
0.245	105	404	4068	14841	24427	33844
0.249	**397**	**689**	**4008**	**14608**	**24040**	**33305**

$\delta = 0.249$ (see Table 5.2.4).

We obtain optimal design for $\phi(p) = -(Var(\hat{\theta}_1) + Var(\hat{\theta}_2) + Var(\hat{\theta}_3))$ as given by:

$$\xi^* = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ 0.25 & 0.5 & 0.25 \end{array} \right\}, \text{ with minimum total variance} = 8.$$

The directional derivatives are zero for all the support points but negative for others, that is, they satisfy the first-order conditions (2.6).

5.2.2 Cubic Regression Model

For the Cubic regression model, we have

$$E(Y|x) \quad = \quad \theta_1 + \theta_2 x + \theta_3 x^2 + \theta_4 x^3.$$

Suppose that we are interested in minimizing the total variance of the estimates of the parameters θ_2 and θ_4. We then maximize

$$\phi(p) = \phi_2(p) + \phi_4(p) \tag{5.8}$$

subject to the constraints $p_j \geq 0$ and $\sum p_j = 1$. Here $\phi_2(p) = -\underline{a}_2^T M^{-1}(p)\underline{a}_2$, $\underline{a}_2 = (0,\, 1,\, 0,\, 0)^T$, $\phi_4(p) = -\underline{a}_4^T M^{-1}(p)\underline{a}_4$, $\underline{a}_4 = (0,\, 0,\, 0\,, 1)^T$.

We again approximate the design interval by a grid of points equally spaced at intervals of 0.01. We then run the algorithm (4.6) by taking $f(x, \delta)$ with x as both the partial and directional derivatives as given in (5.4) and (5.5) respectively, according to our criterion function (5.8). We choose $f(d, \delta) = \Phi(\delta d)$, $f(d, \delta) = exp(\delta d)/(1 + exp(\delta d))$.

We report the iteration results in Tables 5.2.5-5.2.6. Here also we see that the convergence of the algorithm is improved by great deal. For example, when we choose $f(x, \delta) = \Phi(\delta x)$, with $x = d$, $\delta = 0.0315$ and $n = 6$, 49486 iterations are needed to achieve $\underset{1 \leq j \leq J}{max}\{F_j\} \leq 10^{-n}$ (see Table 5.2.5),

Table 5.2.5: Cubic Regression for $-(Var(\hat{\theta}_2) + Var(\hat{\theta}_4))$ with $x = d$

Number of iterations need to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$						
$f(x,\delta) = \Phi(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.02	496	4471	16111	29287	42501	55704
0.025	460	4118	14831	26953	39110	51257
0.0315	**448**	**3982**	**14325**	**26027**	**37761**	**49486**
0.033	450	3987	14342	26056	37803	49540
0.04	475	4167	14973	27193	39445	51688
$f(x,\delta) = exp(\delta x)/(1+exp(\delta x))$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.03	525	4731	17047	30991	44975	58948
0.04	480	4295	15461	28101	40776	53441
0.048	**473**	**4210**	**15146**	**27522**	**39933**	**52334**
0.054	480	4249	15277	27756	40269	52773
0.06	494	4358	15663	28452	41278	54092

whereas only 14705 iteration numbers are needed when we choose $x = F$ and $\delta = 0.047$ (see Table 5.2.6). Similar things happen for the choice $f(x,\delta) = exp(\delta x)/(1 + exp(\delta x))$.

The optimal design measure and variance for $\phi(p) = -(Var(\hat{\theta}_2) + Var(\hat{\theta}_4))$ are:

$$\xi^* = \left\{ \begin{array}{cccc} -1 & -0.51 & 0.51 & 1 \\ 0.136 & 0.364 & 0.364 & 0.136 \end{array} \right\},$$

with minimum total variance = 26.46344.

The directional derivatives again are zero for all the support points and negative for the other points. That is, the first-order conditions (2.6) are satisfied. We also see the improvement of convergence of algorithm.

Table 5.2.6: Cubic Regression for $-(Var(\hat{\theta}_2) + Var(\hat{\theta}_4))$ with $x = F$

Number of iterations need to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$						
$f(x,\delta) = \Phi(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.03	203	1844	6657	12099	17556	23008
0.04	153	1384	4994	9075	13168	17256
0.042	146	1319	4757	8644	12541	16435
0.045	136	1231	4441	8068	11706	15340
0.047	**339**	**1180**	**4267**	**7743**	**11225**	**14705**
$f(x,\delta) = exp(\delta x)/(1+exp(\delta x))$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0.04	242	2206	7965	14478	21008	27534
0.05	195	1766	6374	11584	16808	22028
0.06	163	1472	5312	9654	14008	18357
0.07	140	1263	4554	8276	12007	15735
0.075	**325**	**1179**	**4261**	**7736**	**11219**	**14698**

5.2.3 Chemistry Model

For the Chemistry model, we have

$$E(Y|x) \quad = \quad \theta_1 x + \theta_2 x^{1/2} + \theta_3 x^2.$$

As we mentioned earlier, this regression model describes the relationship between the viscosity Y and the concentration x of a chemical solution (Torsney and Alahmadi, 1995). Viscosity is the response. First, note the design interval $(0, 0.2]$ is very short here. The chemists who supplied this model were more concerned with the terms $x^{1/2}$ and x^2. We already constructed designs for minimizing the variances of the estimates of the corresponding parame-

ters θ_2 and θ_3. Now we would like to minimize the average or total variance of the estimates of these two parameters. Accordingly, we then maximize

$$\phi(p) = \phi_2(p) + \phi_3(p) \tag{5.9}$$

subject to the constraints $p_j \geq 0$ and $\sum p_j = 1$. Here $\phi_2(p) = -\underline{a}_2^T M^{-1}(p)\underline{a}_2$, $\underline{a}_2 = (0,\, 1,\, 0)^T$, $\phi_3(p) = -\underline{a}_3^T M^{-1}(p)\underline{a}_3$, $\underline{a}_3 = (0,\, 0,\, 1)^T$.

According to this design space, we approximate the design interval by a grid of points equally spaced at intervals of 0.01 between 0.02 and 0.2. We then run the algorithm (4.6) by taking $f(x, \delta)$ with x as given in (5.4) and (5.5), the partial and directional derivatives of our criterion function (5.9) respectively. We choose $f(d, \delta) = \Phi(\delta d)$, $f(d, \delta) = exp(\delta d)/(1 + exp(\delta d))$. We record, for $n = 1, 2, \cdots, 6$, the number of iterations needed to achieve $\underset{1 \leq j \leq J}{max}\{F_j\} \leq 10^{-n}$, where F_j are the vertex directional derivatives. In all the cases we take the initial design to be $p_j^{(0)} = 1/J$, $j = 1, 2, \cdots, J$. Note that, in this model, we have $J = 19$.

We report the results in Tables 5.2.7-5.2.8. We also see the convergence of the algorithm is improved by great deal. For example, when we choose $f(x, \delta) = exp(\delta x)/(1 + exp(\delta x))$, with $x = d$, $\delta = 1.05 \times 10^{-5}$ and $n = 6$, 2650 iterations are needed to achieve $\underset{1 \leq j \leq J}{max}\{F_j\} \leq 10^{-n}$ (see Table 5.2.7). However, only 754 iterations are needed when we choose $x = F$ and $\delta = 1.61 \times 10^{-5}$ (see Table 5.2.8). Similar things happen for the choice $f(x, \delta) = \Phi(\delta x)$.

Table 5.2.7: Chemistry Model for $-(Var(\hat{\theta}_2) + Var(\hat{\theta}_3))$ with $x = d$

Number of iterations need to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$						
$f(x,\delta) = \Phi(\delta x)$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
4.5×10^{-6}	1327	1620	1913	2206	2499	2791
6.3×10^{-6}	1202	1466	1730	1995	2259	2523
6.9×10^{-6}	**1194**	**1457**	**1719**	**1981**	**2244**	**2506**
7.3×10^{-6}	1196	1459	1722	1984	2247	2510
7.9×10^{-6}	1209	1475	1740	2005	2270	2535
$f(x,\delta) = exp(\delta x)/(1+exp(\delta x))$						
δ	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
4.5×10^{-6}	1758	2147	2536	2925	3314	3703
9.3×10^{-6}	1275	1556	1837	2117	2398	2678
1.05×10^{-5}	**1263**	**1540**	**1818**	**2095**	**2373**	**2650**
1.17×10^{-5}	1272	1551	1830	2109	2388	2667
1.29×10^{-5}	1298	1583	1867	2151	2436	2720

The optimal design measure and variance for $\phi(p) = -(Var(\hat{\theta}_2) + Var(\hat{\theta}_3))$ are:

$$\xi^*=\left\{\begin{array}{ccc} 0.02 & 0.12 & 0.2 \\ 0.349 & 0.429 & 0.223 \end{array}\right\}, \text{ with minimum total variance} = 121565.6.$$

In summary, for all of the above models, the directional derivatives are zero for all the support points, however, they are negative for others points. Hence, the directional derivatives satisfies the first-order conditions (2.6). Considerable improvements in convergence of the algorithm are seen for each of these models.

Table 5.2.8: Chemistry Model for $-(Var(\hat{\theta}_2) + Var(\hat{\theta}_3))$ with $x = F$

Number of iterations need to achieve $\max_{1 \le j \le J} \{F_j\} \le 10^{-n}$						
$f(x, \delta) = \Phi(\delta x)$						
δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$
4.5×10^{-6}	805	984	1162	1341	1519	1697
7.5×10^{-6}	482	588	695	801	908	1015
8.5×10^{-6}	424	518	612	706	800	894
9.5×10^{-6}	380	464	548	632	715	799
1.01×10^{-5}	**339**	**431**	**515**	**595**	**673**	**753**
$f(x, \delta) = exp(\delta x)/(1 + exp(\delta x))$						
δ	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$
4.5×10^{-6}	1287	1572	1858	2144	2429	2714
7.7×10^{-6}	750	917	1083	1250	1416	1582
1.25×10^{-5}	461	563	665	767	869	971
1.45×10^{-5}	397	485	573	661	748	836
1.61×10^{-5}	**353**	**437**	**517**	**595**	**675**	**754**

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this thesis we have tried to address an important problem of estimation and optimal regression design. The problem is that of estimating a regression model, parameters or linear combinations of these in such a manner that they are estimated with minimum variance. This is always the goal in statistical inference.

We first reviewed some basic optimal design theory including properties of the least squares estimation and the information matrix of the design. We discussed how to discretize a design space, and then to find a design measure

and the support of the design. We discussed the standardized variance of the predicted response which play a key role in verifying the optimal designs. We discussed two types of designs, namely, approximate and exact designs, and took advantage of the approximate designs in terms of its flexibility in constructing the optimal designs. We then focussed on several design criteria and their properties. One particular criterion is of interest is the linear optimality. This criterion plays an important role in our optimization problems. We then determined the optimality conditions for our optimization problem. For this purpose, we considered two types of directional derivatives. We discussed the properties of these directional derivatives and some links between them. The first one is used to determine the optimality conditions, and the second one is used to obtain the partial derivatives of our criterion function for minimizing variances of parameter estimates.

We formulated our optimization problem for minimizing variances of the least-squares estimates of the parameters (or linear functions of the parameters) of a linear model. We approached constructing the designs in two ways. The first one is the analytic approach, in which we derived the partial derivatives of our criterion for different regression models, solved the equations by equating them to zero, and then constructed the optimal designs. We considered the models, namely, the Quadratic and Cubic Regression models and a practical model in Chemistry.

The second approach is the algorithmic approach. In order to find the

optimal design by using an algorithm, we first derived the partial derivatives of our variance function. Then we considered a class of multiplicative algorithms, indexed by a function which satisfies certain conditions. In particular, the function has to be positive and strictly increasing and may depend on a free positive parameter. We also discussed the properties of this algorithm. We then investigated some useful strategies for better convergence of the algorithm by using the properties of the vertex directional derivatives.

In some regression designs, it was also of interest to construct optimal designs for minimizing the average or total variance of some of the parameter estimates. Note that minimizing the average variance is the same as minimizing the total variance. We first formulated the problem in terms of a set of parametric functions in general. We then reduced the problem in terms of some parameters by choosing the coefficient vectors as unit vectors according to the parameters of interest. We derived the partial and directional derivatives of the criterion function of total variance. Again we attempted to improve the convergence rate of the algorithm by using the properties of the directional derivatives. Considerable improvements in convergence are seen for each of these models we considered in the thesis.

6.2 Future Work

The focus of this thesis was to construct optimal designs for minimizing the variances of parameter estimates in linear models. We also constructed designs for minimizing the average or total variance of some of these parameter estimates. A possible future work would be to construct optimal designs such that the variances of the estimates of some parameters or some linear functions of the parameters are equal. This optimization problem may be of practical interest when we are equally interested in some parameters or parametric functions in the model.

Suppose that we wish to minimize the maximum of two variance functions $\phi_1(p)$ and $\phi_2(p)$. Let $\phi(p)$ be the maximum of these two functions. Now two cases may arise. In the first case the minimum of $\phi(p)$ occurs at the point where one of the functions is minimized, say, $\phi_2(p)$. In another case the minimum of $\phi(p)$ occurs at a design where both functions are equal. If this design is to be the optimal design, it must minimize both variance functions subject to them being equal. We would like to consider this problem as our possible future work.

Bibliography

Atkinson, A., A. Donev, and R. Tobias (2007). *Optimum experimental designs, with SAS*, Volume 34. Oxford University Press.

Atwood, C. L. (1969). Optimal and efficient designs of experiments. *Ann. Math. Statist. 40*, 1570–1602.

Berger, M. P. and W. K. Wong (2009). *An introduction to optimal designs for social and biomedical research*, Volume 83. John Wiley & Sons.

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

Karlin, S. and W. J. Studden (1966). Optimal experimental designs. *Ann. Math. Statist. 37*(4), 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 J. Wolfowitz (1960). The equivalence of two extremum problems. *Canadian Journal of Mathematics 12*(5), 363–365.

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

Mandal, S. and B. Torsney (2006). Construction of optimal designs using a clustering approach. *Journal of statistical planning and inference 136*(3), 1120–1134.

Mandal, S., B. Torsney, and K. Carriere (2005). Constructing optimal designs with constraints. *Journal of statistical planning and inference 128*(2), 609–621.

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

Rohde, C. A. (1965). Generalized inverses of partitioned matrices. *Journal of the Society for Industrial and Applied Mathematics 13*(4), 1033–1035.

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. *London: Chapman and Hall*.

Silvey, S. D. and D. M. Titterington (1973). A geometric approach to optimal design theory. *Biometrika 60*(1), 21–32.

Silvey, S. D., D. M. Titterington, and B. Torsney (1978). An algorithm for optimal designs on a finite design space. *Communications in Statistics A 14*, 1379–1389.

Titterington, D. M. (1976). Algorithms for computing D-optimal designs on a finite design space. In *Proc. 1976 Conf. on Information Sciences and Systems*, pp. 213–216. Dept. of Elect. Eng., John Hopkins Univ., Baltimore, MD.

Torsney, B. (1977). Constribution to discussion of "maximum likelihood from incomplete data via the EM algorithm" by dempster et al. *Journal of the Royal Statistical Society B 39*, 26–27.

Torsney, B. (1983). A moment inequality and monotonicity of an algorithm. In *Semi-Infinite Programming and Applications*, Volume 215, pp. 249–260. Lecture Notes in Economics and Mathematical Systems.

Torsney, B. (1988). Computing optimising distributions with applications in design, estimation and image processing. optimal design and analysis of experiments (edited by y. dodge, v. v. fedorov and h. p. wynn). *Elsevier Science Publishers B. V., North Holland*, 361–370.

Torsney, B. and A. M. Alahmadi (1995). Designing for minimally dependent observations. *Statistica Sinica*, 499–514.

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

Whittle, P. (1973). Some general points in the theory of optimal experimental design. *Journal of the Royal Statistical Society. Series B (Methodological)*, 123–130.