OPTIMAL DESIGNS WITH APPLICATIONS IN ESTIMATION

By

Stella Leung

A Thesis

Submitted to the Faculty of Graduate Studies

in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

Department of Statistics

University of Manitoba

Winnipeg, Manitoba

2004

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION

OPTIMAL DESIGNS WITH APPLICATIONS IN ESTIMATION

BY

Stella Leung

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirement of the degree

Of

MASTER OF SCIENCE

Table of Contents

Abstract

The quest of how to optimally design experiments originally extends back to 1918 where Smith was one of the first to state a criterion and obtain optimal designs for regression problems. Many years later, Kiefer (1959) contributed tremendously to this subject which included the equivalence theorem and optimality criteria as well as the construction of various optimal designs using algorithms.

We first introduce basic linear design theory and discuss their properties. We determined optimality conditions based on directional derivatives along with the properties of these derivatives.

This thesis mainly explores constructing optimizing distributions with applications in estimation by exploring a class of algorithms, indexed by a function $f(.)$, where $f(.)$ is positive and strictly increasing. The function may depend on a free positive parameter δ.

Estimation problems and their properties are studied and their results are reported, namely for the 3×3 case and 4×4 case. We also consider another estimation problem, namely, constructing optimizing distributions with equality of variances of the estimates of two parametric functions of interest.

This thesis goes further by discussing how we can improve convergence rates of the algorithm by choosing the function $f(.)$ and the parameter δ.

Acknowledgements

I would like to express my deepest sense of gratitude to my supervisor Dr. Saumen Mandal whose guidance, encouragement, patience and grace was instrumental in making this thesis a reality. Without his contribution and support, this thesis would have never been possible.

My sincere thanks to my committee members, Dr. M. Samanta of the Department of Statistics and Dr. P. Irani of the Department of Computer Science for their valuable contributions.

I am grateful to the various members of the Department of Statistics for their assistance and opening my eyes to the field of statistics that has changed my life forever. A special thanks to our department secretary, Margaret Smith for her administrative assistance and her smile that was always contagious.

I would like to thank my boyfriend, Ludwig Lee, for his inspiration, moral support, and editorial help. Thank you for always being there for me even when there were times I thought I could not finish.

I would also like to thank my colleagues of the M.Sc students of 2004 class at the University of Manitoba for sharing knowledge and experiences during our time of study. You know who you are!

I genuinely express my appreciation to my friend, James Tsang, who helped me significantly on my programming part of my thesis.

Finally, I take this opportunity to express my profound gratitude to my parents, Norman and Linda Leung. Without their constant support, love and understanding, this day would have never happened. The completion of my thesis was entirely because of them and I would like to dedicate this M.Sc thesis to them.

Winnipeg, Manitoba
August 2004

Stella Leung

Chapter 1

Optimal Design Theory

1.1 Introduction

Optimal designs originally date back to 1918 where Smith's paper consisted of mathematical work of designed experiments. In this paper, she calculated optimal designs for polynomial regression models. Many years later, the next well-known individual to study optimal design to great lengths was Kiefer (1959). Kiefer's tremendous contributions to this subject include his extensive and fundamental work on equivalence theorem and optimality criteria as well as the construction of various optimal designs using algorithms. Well-known sources that most books cite from in the field of optimal design are Atkinson and Donev (1992), Silvey (1980), Fedorov (1972), and Pukelsheim (1993).

A well-designed experiment is an essential technique that must be performed in order to answer problems of interest. Experiments must be conducted and designed in a procedure with statistical methods to collect outcomes in an efficient way. The goal is to reduce expenses, effort and time, while trying to minimize any random errors that may incur.

Whenever a problem with the need of accepting or not accepting a set of alternative decisions is encountered, specific experiments that consist of chosen values or levels of outputs to gather observations on which the decision has to be based must be designed. These experiments, in some sense, have to be optimum to select an optimum decision, thus arising the theory of optimal experimental design.

The focus of this chapter is to give a description of optimal design theory for linear models. Provided are some basic concepts of optimal design theory such as the definition of a design, variance function, an information matrix, and various criterion functions and their properties.

First, consider the problem of selecting an experimental design to accommodate information on models of the type: $y \sim p(y \mid \underline{x}, \underline{\theta}, \sigma)$

where

y is the response variable.

$\underline{x} = (\underline{x}_1, \underline{x}_2, ..., \underline{x}_n)^T$ are the design variables. These values can be chosen by the experimenter and are restricted to a space χ, i.e. $\underline{x} \in \chi \subseteq \mathbb{R}^m$. Therefore, the set of experimental conditions are χ. The design space is χ, although sometimes discrete, will generally be continuous.

$\underline{\theta} = (\underline{\theta}_1, \underline{\theta}_2, ..., \underline{\theta}_k)^T$ is a k-dimensional vector of unknown parameters. $\underline{\theta}$ is known to belong to the set $\theta \in \mathbb{R}^k$.

σ is a nuisance parameter which is fixed but unknown. This parameter is not of fundamental interest.

$p(.)$ is a probability model.

In most cases, χ is assumed to be compact. The experimental conditions from the given domain χ can be utterly chosen by the experimenter.

For every $x \in \chi$, an experiment can be conducted whose outcome is a random variable $y = y(\underline{x})$, where $var(y(\underline{x})) = \sigma^2$ assuming σ does not depend on the experimental condition $\underline{x}$.

In linear regression design, the model is linear in the unknown parameters $\underline{\theta}$ but is not necessarily linear in $\underline{x}$. As a result, in linear models, $y(\underline{x})$ has an expected value of the explicit form:

$$E(y \mid \underline{x}, \underline{\theta}, \sigma) = \underline{f}^T(\underline{x})\,\underline{\theta} \tag{1.1.1}$$

where

$\underline{f}(\underline{x}) = (f_1(\underline{x}), f_2(\underline{x}), ..., f_k(\underline{x}))^T$ is a vector of k real-valued functions defined on χ, what are known to the experimenter before-hand are the regression functions $f_1, f_2, ..., f_k$.

A value for $\underline{x}$ must always be chosen from χ in order to acquire an observation on y. It is understood that $\underline{x}$ can be set to any chosen value in χ. This leads to the consideration of at what value of $\underline{x}$ should observations, say n, on y be taken

in order to attain a 'best' inference for all or some of the parameters $\underline{\theta}$. Obtaining this reliable inference, or allocating n observations to the elements of χ is termed an optimal regression design.

At this time, presume that this is point estimation for the mode of inference. The projected solution for this example will hold well for other future modes of inference as well.

Deciding what n values ($\underline{x}_1$, $\underline{x}_2$, ..., $\underline{x}_n$) to produce 'best' point estimation $\hat{\underline{\theta}}$ of some or all of the parameters $\underline{\theta}$ is something to consider.

Let the estimator $\hat{\underline{\theta}}$ of $\underline{\theta}$ be obtained by some method of point estimation. Let $\hat{\underline{\theta}}$ be unbiased for $\underline{\theta}$. The components $\hat{\theta}_j$ will be correlated. Debatably then, the $k \times k$ matrix $D(\hat{\underline{\theta}}) = E([\hat{\underline{\theta}} - \underline{\theta}][\hat{\underline{\theta}} - \underline{\theta}]^T)$ which is the *dispersion matrix* of $\hat{\underline{\theta}}$ about $\underline{\theta}$, holds information about the accuracy of $\hat{\underline{\theta}}$ not only in its diagonal elements, which measures the mean square deviation of $\hat{\theta}_j$, but also in its off-diagonal cross product deviation terms. For the most part, the smaller $D(\hat{\underline{\theta}})$ gets, the greater the accuracy of $\hat{\underline{\theta}}$.

Consider model 1.1.1 to be true and let y_i represent the observation obtained at $\underline{x}_i$ where,

$$E(y_i) = \underline{v}_i^T \underline{\theta},\ \underline{v}_i = (f_1(\underline{x}_i),\ f_2(\underline{x}_i),\ ...,\ f_k(\underline{x}_i))^T,\ i = 1, 2, ..., n. \tag{1.1.2}$$

Suppose y_1, y_2, ..., y_n are independent random variables with equal variance σ^2. Also, there will be several equalities between the $\underline{x}_i$'s, where more than one observations are

being taken at the same $\underline{x}$ value. Therefore, $\underline{y}_i$'s then satisfy the following standard linear model:

$$E(Y) = X\,\underline{\theta},\ D(Y) = \sigma^2 I_n \tag{1.1.3}$$

where

$Y = (y_1, y_2, ..., y_n)$,

X is an $n \times k$ matrix whose $(i,j)th$ element is $f_j(\underline{x}_i)$,

θ is a $k \times 1$ vector of unknown parameters,

σ^2 is the constant error variance (usually unknown),

I_n is the identity matrix of order n,

$D(Y)$ symbolizes the dispersion matrix of Y.

Model 1.1.3 can also be referred to as a *fixed-effects* linear model.

Least squares estimators are a predictable choice for a model having the optimality of being best linear unbiased estimators (BLUE). Solutions are of:

$$(X^T X)\hat{\underline{\theta}} = X^T Y \tag{1.1.4}$$

where $(X^T X)$ is the information matrix for $\underline{\theta}$ of order $k \times k$.

When $(X^T X)$ gets larger, the information will become more superior in the experiment. If all parameters $\underline{\theta}$ are of interest, then the selection of $\underline{x}$ must at least substantiate the matrix $(X^T X)$ is non-singular. In this case, the unique solution for 1.1.4 is given by:

$$\hat{\underline{\theta}} = (X^T X)^{-1} X^T Y \tag{1.1.5}$$

and

$$E(\hat{\underline{\theta}}) = \underline{\theta}$$
$$D(\hat{\underline{\theta}}) = \sigma^2(X^TX)^{-1}$$

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

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

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

As a result, it can be seen that the dispersion matrix of $\hat{\underline{\theta}}$ does not have to depend on $\underline{\theta}$ and only depends proportionally on the parameter σ^2. We select $\{\underline{x}_1,\ \underline{x}_2,\ ...,\ \underline{x}_n\}$ to make the matrix $D(\hat{\underline{\theta}})$ as small as possible. That is, we to select $\{\underline{x}_1,\ \underline{x}_2,\ ...,\ \underline{x}_n\}$ which makes the matrix (X^TX) large in some sense.

1.2 Discretizing the Design Space

The model 1.1.1 can also be written as:

$$E(y \mid \underline{v}, \underline{\theta}, \sigma) = \underline{v}^T\underline{\theta} \tag{1.2.1}$$

where

$\underline{v} = (f_1(\underline{x}),\ f_2(\underline{x}),\ ...,\ f_k(\underline{x}))^T$, $\underline{v} \in \mathcal{V}$,

$\mathcal{V} = \{\underline{v} \in \mathbb{R}^k : \underline{v} = (f_1(\underline{x}),\ f_2(\underline{x}),\ ...,\ f_k(\underline{x}))^T,\ \underline{x} \in \chi\}$.

Choosing a vector $\underline{x}$ in the design space χ is equivalent to choosing a k-vector $\underline{v}$ in the closed k-dimensional space $\mathcal{V} = \underline{f}(\chi)$, where $\underline{f}$ is the vector valued function $(f_1, f_2, ..., f_k)^T$. $\mathcal{V}$ is the image under f of χ and uses an induced design space. Generally, the design space is continuous, but we can assume that $\mathcal{V}$ is discrete.

Let $\mathcal{V}$, the discrete design space consists of J distinct vectors $\underline{v}_1, \underline{v}_2, ..., \underline{v}_J$. To obtain an observation on y, we must choose a value for $\underline{v}$ from the J elements of $\mathcal{V}$ to be the point at which to take this observation. Using Caratheodory's theorem, this design space, $\mathcal{V}$ is taken to be discrete, suggesting that it can be done without error.

At what points $\underline{v}_j$ should observations be taken and, if the total observations are allowed, how many of these observations can be taken at these points in order to obtain 'best' least squares estimators of $\underline{\theta}$? With this in mind, the design problem can be expressed exactly.

With n observations, we have to decide how many observations, say n_j to take at $\underline{v}_j$, $\sum_{j=1}^{J} n_j = n$. With this in mind, the matrix (X^TX) can be written in the following form:

$$X^TX = M(\underline{n}),\ \underline{n} = (\underline{n}_1, \underline{n}_2, ..., \underline{n}_J)^T \tag{1.2.2}$$

where

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

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

By choosing $\underline{n}$ now, we can make the matrix $M(\underline{n})$ large. Since n_j's must be integers it triggers an integer programming problem and in the design context it is described as an *exact design* problem.

Integer programming problems are generally tedious to solve mainly because the theory of calculus cannot be used to define of or to identify optimal solutions. Therefore, a solution has to be worked out completely separately for different values of $\underline{n}$. However, there is a simpler way to solve the problem. We can write the information matrix as:

$$M(\underline{n}) = nM(p) \tag{1.2.3}$$

where

$$M(p) = \sum_{j=1}^{J} p_j \underline{v}_j \underline{v}_j^T \tag{1.2.4}$$

$$= VPV^T \tag{1.2.5}$$

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

Hence, choosing p to make $M(p)$ large subject to $p_j = \frac{n_j}{n}$ becomes our new problem. Relaxing the latter to $p_j \geq 0$ and $\sum_{j=1}^{J} p_j = 1$ generates an *approximate design* problem. Indeed, this is a more flexible problem to solve and visibly not much different from the original.

Design Measure:

Previously we have referred to p as both the vector $(p_1, p_2, ..., p_J)$ and as a probability distribution on $\mathcal{V}$. A full statement of this could possibly be:

$$p = \left\{ \begin{array}{cccc} \underline{x}_1 & \underline{x}_2 & \cdots, & \underline{x}_J \\ p_1, & p_2, & \cdots, & p_J \end{array} \right\} \tag{1.2.6}$$

where

x_j's are the values of the factors, that is, the design points. p_j's are the associated design weights, $\sum_{j=1}^{J} p_j = 1$ and $0 \leq p_j \leq 1$ for all j.

A more suitable and less confusing notation is:

$$\xi = \left\{ \begin{array}{cccc} \underline{x}_1 & \underline{x}_2 & \cdots, & \underline{x}_J \\ p_1, & p_2, & \cdots, & p_J \end{array} \right\} \tag{1.2.7}$$

with ξ defined to be the design measure.

Exact designs have a specific number of trails, $\underline{n}$. The design measures for an exact design is written as:

$$\xi = \left\{ \begin{array}{cccc} \underline{x}_1 & \underline{x}_2 & \cdots, & \underline{x}_J \\ \frac{n_1}{n}, & \frac{n_2}{n}, & \cdots, & \frac{n_J}{n} \end{array} \right\} \tag{1.2.8}$$

where

n_j is the integer number of trials at x_j and $\sum_{j=1}^{J} n_j = n$.

Support of a Design Measure:

The support of a design measure (p) in $\mathcal{V}$ is defined by:

$$supp(p) = \{v_j \in \mathcal{V} : p_j > 0, i = 1, 2, ..., J\} \tag{1.2.9}$$

That is, $supp(p)$ is the collection of those v_j which has non-zero p_j.

1.3 Optimality Criteria

By making the matrix $M(p)$ large, it may be possible to obtain a best inference for all or some of the unknown parameters $\underline{\theta} \in \Theta$. Therefore, to make the matrix $M(p)$ large, that is to say, by maximizing some real valued function $\phi(p) = \psi\{M(p)\}$, various methods are considered.

There are many design criteria and they are mostly labeled after the letters of the alphabet. These criteria are sometimes called *alphabetic optimality.* The function ϕ is identified as the criterion function. A criterion defined by the function ϕ is called ϕ-optimality. A design maximizing $\phi(p)$ is called a ϕ-optimal design. There are two types of criteria. One type is when 'all' the parameters in the model are of interest and the other is when 'not all' k parameters are of interest.

Consider the case when interest is in inference about all of the parameters $\underline{\theta}$. Therefore, we must have $M(p)$ as non-singular, and thus positive definite. When all parameters are of interest, then there are several possible criteria which include:

$D-$optimality, $A-$optimality, and $G-$optimality. Mandal (2000) extensively researched on many of these criteria.

$D-$optimality

The most important and popular design criterion in applications is that of $D-$optimality. $D-$optimality seeks to maximize the value of $|X^TX|$, the determinant of the information matrix X^TX. $D-$optimality results in minimizing the generalized variance of the estimates of the parameter based on a pre-specified model. This criterion is also known as the determinant criterion. The criterion function of $D-$optimality is given by:

$$\phi_D(p) = \psi_D\{M(p)\} = \log\det\{M(p)\} = -\log\det\{M^{-1}(p)\} \tag{1.3.1}$$

Other motivations for $D-$optimality exist, and they extend way beyond ideas of point estimation and all fall into the field of explicit joint inference. If we assume normality of the errors in linear models, then the general form of the joint confidence region for the vectors of unknown parameters $\underline{\theta} \in \Theta$ is described by an ellipsoid of the form:

$$\{\underline{\theta} : (\underline{\theta} - \hat{\underline{\theta}})^T(\underline{\theta} - \hat{\underline{\theta}}) \leq c)\}, \text{ for some critical value c} \tag{1.3.2}$$

where $\hat{\underline{\theta}}$ is the least squares estimates or the maximum likelihood estimate of $\underline{\theta}$.

The $D-$optimality criterion chooses the information matrix $M(p)$ to make the volume of the above ellipsoid as small as possible because this volume is proportional

to $[\det\{M(p)\}]^{-1/2}$. The value of $[\log\det\{M(p)\}]$ is finite if and only if $M(p)$ is non-singular, meaning when all the unknown parameters are estimable. See Keifer (1959), Farrell et al. (1967), Fedorov (1972), Silvey (1980), Pazman (1986), Shah and Sinha (1989), Atkinson and Donev (1992), Pukelsheim (1993). Mandal (2000) considered the construction of $D-$optimal designs in a variety of examples. This is the most extensively studied of all the design criteria.

$A-$optimality

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

$$\phi_A(p) = \psi_A\{M(p)\} = -Trace\{M^{-1}(p)\} \tag{1.3.3}$$

$A-$optimality minimizes the trace of the inverse of the information matrix. It minimizes the sum (or the averages) of the variances of the parameter estimates based on a pre-specified model, but does not take correlations between these estimates into account. $A-$optimum design is also known as *trace criterion.* This criterion was considered by Elfving (1952) and Chernoff (1953).

$G-$optimality

$G-$optimality seeks to minimize the maximum prediction variance over a specified set of design points. That is to say, it minimizes the maximum value of $\underline{v}^T M^{-1}(p)\underline{v}$ which is proportional to the variance of $\underline{v}^T\underline{\hat{\theta}}$. Kiefer and Wolfowitz (1960) proved the equivalence of this criterion with the $D-$optimal criterion.

The criterion function for $G-$optimality is defined by:

$$\phi_G(p) = \psi_G\{M(p)\} = -\max_{\underline{v}\in\mathcal{V}} \underline{v}^T M^{-1}(p)\underline{v} \tag{1.3.4}$$

Now consider the case when interest is not in all k parameters, but only in some of the unknown parameters or some combinations of the parameters of the linear model 1.1.1.

Say we are interested in s linear combinations of the parameters θ_1, θ_2, ..., θ_k, namely those s linear combinations which are elements of the vector $\underline{\alpha} = A\underline{\theta}$, where A is a $s \times k$ matrix of rank $s \leq k$.

If $M(p)$ is non-singular, then the variance matrix of the least squares estimator of $A\underline{\theta}$ is proportional to the matrix $AM^{-}(p)A^T$. However, if $M(p)$ is singular, then the basic requirement for estimating the vector $\underline{\alpha} = A\underline{\theta}$ is that the row space of A is in the range space of $M(p)$ which results in the invariance of the matrix $AM^{-}(p)A^T$ to the choice of generalized inverse $M^{-}(p)$ of $M(p)$.

A good design in this case would be one that makes the matrix $AM^{-}(p)A^T$ (or $AM^{-1}(p)A^T$ if M(p) is non-singular) as small as possible. Criteria include D_A-, D_S- and Linear ($L-$) optimality.

D_A-optimality

D_A-optimality is used when we are interested in s linear combinations of $\underline{\theta}$, that is, the elements of the vector $A^T\underline{\theta}$. If X^TX is non-singular, this criterion maximizes the determinant of $[A^T(X^TX)^{-1}A]^{-1}$.

The criterion function is given by:

$$\phi_{D_A}(p) = \psi_{D_A}\{M(p)\} = -\log\det\{AM^{-1}(p)A^T\} \tag{1.3.5}$$

Consider the special case of D_A-optimality, which is D_S-optimality.

D_S-optimality is used when we are interested in s parameters $A = [I_S : O]$ and we partition the matrix $M(p)$ as follows:

$$M(p) = \begin{bmatrix} M_{11}^{s\times s} & M_{12}^{s\times k-s} \\ M_{12}^T & M_{22}^{s\times k-s} \end{bmatrix} \tag{1.3.6}$$

Using algebra we can express the matrix $(AM^{-1}(p)A^T)^{-1}$ as $(M_{11} - M_{12}M_{22}^{-1}M_{12}^T)$ [see Rhode (1965) and Torsney (1981)]. Our design criterion becomes that of selecting p to maximize the determinant of this matrix. So maximizing ϕ_{D_A} in this case is equivalent to maximizing:

$$\phi_{D_S}(p) = \log\det\{M_{11} - M_{12}M_{22}^{-1}M_{12}^T\} \tag{1.3.7}$$

which is known as the D_S-optimal criterion. See Karlin and Studden (1966), Atwood (1969), Silvey and Titterington (1973) and Silvey (1980).

Linear Optimality

Let L be a systematic and positive definite $k \times k$ matrix of coefficients. The function for $L-$optimality is defined as:

$$\phi_L(p) = \psi_L\{M(p)\} = -Trace\{M^{-1}(p)L\} \tag{1.3.8}$$

It is linear in the elements of the covariance matrix $M^{-1}(p)$. There is a relationship between $L-$optimum and D_A-optimum designs. In D_A-optimality, the determinant rather than the trace of $A(X^TX)^{-1}A^T$ is minimized. The form that stressed this relationship is when L is of rank $s \leq k$. L is expressed as $L = A^TA$, where A is a $s \times k$ matrix with rank s.

Then the criterion function 1.3.8 can be defined as:

$$\phi_L(p) = -Trace\{M^{-1}(p)L\} = -Trace\{M^{-1}(p)A^TA\} = -Trace\{AM^{-1}(p)A^T\} \tag{1.3.9}$$

There is a special case of $L-$optimality when A is a column vector. This special case is called $c-$**optimality**, which minimizes the variance of a linear combination $c^T\underline{\hat{\theta}}$.

Thus, the design criterion will be minimizing $c^T\{M(p)\}^{-1}c$. An important reference of this criterion is Elfving (1952).

Chapter 2

Optimization Problems and Optimality Conditions

2.1 Introduction

First, we determine optimality conditions for which p^* will be optimal for an optimization problem in this chapter. We determine optimality conditions in terms of point to point directional derivatives. Then we consider some optimization problems in estimations. The directional derivative $F_\phi\{p, q\}$ of a criterion function $\phi(.)$ at p in the direction of q is an important tool. This has a significant simplifying role in the calculus of optimization.

First consider a class of optimization problems in which we wish to find an optimizing distribution. Particular examples are optimal regression, maximum likelihood estimation, stratified sampling and image processing problems.

2.2 Optimization Problems

Consider the following problems.

Problem 1

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

Having the equality constraint $\sum_j p_j = 1$ presents the problem of a nondegenerate constraint optimization problem, the full constraint region being a closed bounded convex set.

Problem 2

Maximize $\Phi(\theta)$ over $\Theta = \{\theta = (\theta_1, \theta_2, ..., \theta_t) : \theta_j \geq 0, C\theta = a\}$ where C is a $s \times t$ matrix of rank s, and a is in the range space of C.

As we can see, Problem 2 is a generalized form of Problem 1. One occurrence of Problem 2 arises when testing the linear hypothesis about the parameters in multinomial models for categorical data. These parameters are probabilities so that the constraint $C\theta = a$ must either include as a component that $\underline{1}^T\theta = 1$, where $\underline{1}$ is a vector of 1's, or proclaim that various subsets of the components of θ should sum to unity. We will consider an example of such linear hypothesis in Chapter 4.

2.3 Directional Derivatives

In order to maximize a criterion $\phi(p)$, we need to characterize optimality conditions on p. We define optimality conditions in terms of point to point directional derivatives.

Let

$$g(p,\, q,\, \varepsilon,\,) = \phi\{(1-\varepsilon)p + \varepsilon q\} \tag{2.3.1}$$

$$F_\phi\{p,\, q\} = \lim_{\varepsilon \downarrow 0} \frac{g(p,q,\varepsilon) - \phi(p)}{\varepsilon} = \left.\frac{dg(p,g,\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0^+} \tag{2.3.2}$$

$F_\phi\{p,\, q\}$ is called the directional derivative of $\phi(.)$ at p in the direction of q as stated by Whittle (1973). This derivative can exist even if $\phi(.)$ is not differentiable.

2.4 Properties of $F_\phi\{p,\, q\}$

Some general properties of the directional derivative $F_\phi\{p,\, q\}$ are as follows.

Property 1:

If $p, q \in S$, where S is a convex set, then $\{(1-\varepsilon)p + \varepsilon q\} \in S$ for all $0 < \varepsilon < 1$. This would be an advantage if one wishes $F_\phi\{p, q\}$ only for $p, q \in S$

Property 2:

$F_\phi\{p, q\} \geq \phi(q) - \phi(p)$ if $\phi(.)$ is concave.

Proof:

$$\begin{aligned} F_\phi\{p,q\} &= \lim_{\varepsilon \downarrow 0}[\frac{\phi\{(1-\varepsilon)p+\varepsilon q\}-\phi(p)}{\varepsilon}] \\ &\geq \lim_{\varepsilon \downarrow 0}[\frac{(1-\varepsilon)\phi(p)+\varepsilon\phi(q)-\phi(p)}{\varepsilon}] \\ &= \phi(q)-\phi(p) \end{aligned} \tag{2.4.1}$$

Up to this point, any assumptions about differentiability of the criterion function ϕ has not been made. A function does not have to be differentiable at a point p in order to have well defined directional derivatives in all directions.

Despite that, when the criterion function ϕ is differentiable, it plays a vital simplifying role in the optimization of ϕ. Mandal (2000) studied the properties of $F_\phi\{p,q\}$ extensively.

Have in mind that at point p, $\phi(.)$ should be smoothly changing in all directions. A more precise definition is that at point p, the $\phi(.)$-surface should just touch or possibly cross in parallel a unique linear hyper-plane, the tangent plane to $\phi(.)$ at p, or the supporting hyperplane at p if the two surfaces do not cross. This plane will provide a linear approximation to $\phi(.)$ at p in any direction, so that the linear approximation to $\phi(.)$ at p which it suggests in the direction of q and in the opposite direction will be the same apart from a difference in sign.

If two surfaces occur at the same time, they will obviously have some common characteristics at the point of contrast p. They must have common first derivatives, partial, or directional derivatives, and hence whatever properties are enjoyed by the derivatives of one function at p, must be enjoyed by those of the other function.

For $\phi(.)$ to be differentiable at p, it must be that

$$\begin{aligned} F_\phi\{p,q\} &= (q-p)^T\frac{\partial\phi}{\partial p} = (q-p)^T d \text{ for all } q \\ &= \sum_{i=1}^{J}(q_i-p_i)d_i, \quad d_i = \frac{\partial\phi}{\partial p_i}, \quad i=1,\ ...,\ J, \quad d = \frac{\partial\phi}{\partial p} \end{aligned} \tag{2.4.2}$$

In Problem 1, when $p \in P$ we have,

$$F_\phi\{p,e_j\} = d_j - \sum_{i=1}^{J} p_i d_i = F_j, \quad \text{say.} \tag{2.4.3}$$

We call F_j a vertex directional derivative of ϕ.

2.5 Optimality Conditions: The General Equivalence Theorem

The General Equivalence Theorem is the central development on the theory of optimum design of experiments of which it depends upon.

This theorem can be seen as an application of the result where the derivatives are zero at a minimum of a function. Keep in mind that the function depends on the measure p through the information matrix $M(p)$.

Recall the derivative of $\phi(.)$ in the direction of q is

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

In optimal design, the main concern is to minimize the convex function $\psi\{M(p)\}$ by using the directional derivative of $F_\phi\{p,q\}$ in the direction of q.

D-optimality is an example in which $\psi_D\{M(p)\} = \log\det\{M^{-1}(p)\}$ is minimized so that the determinant of the information matrix, $M(p)$, is maximized. By taking the logarithm of the determinant it leads to minimization of a convex function. Thus, the General Equivalence Theorem can be viewed as an application of the result that the derivatives are zero at a minimum of a function. Nonetheless, the function depends on the measure p through the information matrix $M(p)$. Let the measure $\bar{p}$ put unit mass at the point x and let the measure p' be given by,

$$p' = (1-\alpha)p + \alpha\bar{p}$$

Then,

$$M(p') = (1-\alpha)M(p) + \alpha M(\bar{p}). \tag{2.5.1}$$

Thus, the derivative of ψ in the direction of $\bar{p}$ is,

$$\Delta(p) = \lim_{\alpha \downarrow 0} = \frac{\psi\{(1-\alpha)M(p) + \alpha M(\bar{p})\} - \psi\{M(p)\}}{\alpha} \tag{2.5.2}$$

The General Equivalence Theorem states the equivalence of the following three conditions on p^*: [Atkinson and Donev (1992)]

(1) the design p^* minimizes $\psi\{M(p)\}$;
(2) the minimum of $\Delta(p) \geq 0$;
(3) the derivative $\Delta(p)$ achieves its minimum at the points of design.

This theorem is very important in the theory of optimal design. According to this theorem, it provides methods for the construction and checking of optimum designs.

Our problem is maximizing a criterion $\phi(p)$ in Problem 1. We write the optimality conditions in terms of this problem:

If $S = P$, $\phi(p)$ is concave on P and p^* is a differentiable point of $\phi(.)$ on P, then p* maximizes $\phi(.)$ on P iff

$$\frac{\partial \phi}{\partial p_j^*} = \sum_{i=1}^{J} p_i^* \frac{\partial \phi}{\partial p_i^*} \qquad \text{when } p_j^* > 0 \tag{2.5.3}$$

$$\frac{\partial \phi}{\partial p_j^*} \leq \sum_{i=1}^{J} p_i^* \frac{\partial \phi}{\partial p_i^*} \qquad \text{when } p_j^* = 0 \tag{2.5.4}$$

In terms of directional derivatives F_j, the optimality conditions are:

$$F_j^* = 0 \qquad \text{when } p_j^* > 0 \tag{2.5.5}$$

$$\leq 0 \qquad \text{when } p_j^* = 0 \tag{2.5.6}$$

General equivalence theorem plays an important role in constructing optimal designs. It specifies a finite set of optimality conditions. It is easy to check whether or not these conditions are satisfied by a postulated solution obtained by numerical techniques. We use these optimality conditions to construct the optimizing distributions in Chapters 4 and 5.

Chapter 3

Algorithms

3.1 Introduction

It is typically not possible to evaluate an explicit solution p^* to optimal designs. An analytic solution to the problem of forming optimal designs is possible only in simple cases. Generally, it is not possible to evaluate an exact solution p^* to Problem 1 and 2 or to derive an optimal regression design explicitly. Iterative techniques must be needed and consequently, certain algorithms have been devised for a constrained optimization problem (particularly for the design problem) which requires the calculation of an optimizing probability distribution.

It can be seen that there always exists an optimal measure with finite support (Caratheodory's Theorem). We wish to identify an optimizing p^*. Of course, this will be the case if $\mathcal{V}$ is a discretization of a continuous space. The implication of this is that at the optimum there will be zero weights. Hence, we consider the following class of algorithms, indexed by a function which depends on derivatives and one or more free parameters.

An algorithm for an optimization problem is a sequence of successive approximations to a solution p^*. First, we make an initial guess $p^{(0)}$ to p^* and try by some means to derive from $p^{(0)}$, an improved approximation $p^{(1)}$. Then by the same means a further improvement $p^{(2)}$ is derived from $p^{(1)}$ and we carry on this way. A sequence $p^{(0)}, p^{(1)}, ...$ is thus generated in the belief that the sequence will converge to the optimum p^*.

3.2 A Class of Algorithms

Problem 1 contains a unique set of constraints, specifically the variables p_1, p_2, ..., p_J must be nonnegative and add up to 1. An iteration which neatly submits to these and has some suitable properties is the multiplicative algorithm:

$$p_j^{(r+1)} = \frac{p_j^{(r)} f(x_j^{(r)})}{\sum_{i=1}^{J} p_i^{(r)} f(x_i^{(r)})} \tag{3.2.1}$$

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

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

$F_j^{(r)} = d_j^{(r)} - \sum_{i=1}^{J} p_i^{(r)} d_i^{(r)}$ [a directional derivative of ϕ at $p = p^{(r)}$],

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

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

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

$f(x)$ may depend on one or more free parameters. We use only one free parameter δ. The value of δ is positive.

Therefore, as a result of the conditions for (local) optimality, a solution to Problem 1 is a fixed point of the iteration and the partial derivatives (d_j) share a common value. This is a necessary but not a sufficient condition for $p^{(r)}$ to solve problem 1.

Torsney (1977) first proposed this type of iteration, taking $x = d$, $f(d) = d^{\delta}$, with $\delta > 0$. This requires derivatives to be positive. Following empirical studies include Silvey, Titterington and Torsney (1978), which is a study of the choice of δ when $f(d) = d^{\delta}$, $\delta > 0$; Torsney (1988), which mainly considers $f(d) = exp\{d\delta\}$ in a variety of applications, for which one criterion $\phi(.)$ could have negative derivatives. Mandal and Torsney (2000) considers systematic choices of $f(.)$. Torsney and Alahmadi (1992) explore other choices of $f(.)$. Mandal (2000) uses this algorithm in a variety of problems. Torsney and Mandal (2001) and Mandal et al. (2004) use this algorithm for constrained optimization problems. Mandal and Torsney (2004) considered a clustering approach to improve the convergence rates considerably.

Titterington (1976) describes a proof of monotonicity of $f(d) = d$ in the case of D-optimality. Torsney (1983) explores monotonicity of particular values of δ for particular $\phi(p)$. Torsney (1983) also establishes a sufficient condition for monotonicity of $f(d) = d^{\delta}$, $\delta = 1/(t+1)$ when the criterion $\phi(p)$ is homogeneous of degree $-t$, $t > 0$ with positive derivatives and proves this condition to hold in the case of linear design criteria such as c-optimal and A-optimal criteria when $t = 1$ so that $\delta = 1/2$. In other

cases the value $\delta = 1$ can be shown to yield an EM algorithm which is known to be monotonic and convergent. See Dempster et al (1977). The EM algorithm is known to have slow convergence.

Convergence results depend on properties of the criterion function $\phi(.)$, on the function $f(.)$ and on δ. In the later chapters we have tried to explore variety of choices of $f(.)$ and of its argument for constructing optimal designs with applications in estimation. In Chapter 4 we consider the problem of determining maximum likelihood estimates under a hypothesis of of marginal homogeneity for data in a square contingency table. In Chapter 5 we consider the problem of finding optimal design with equality of variances of the estimates of two linear parametric functions. We use Minitab statistical package for the programming purposes and for the running of the algorithm in Chapters 4 and 5.

3.3 Properties of the Iteration

Under the conditions imposed on $f(.)$, iterations under (3.2.1) possess the following properties considered by Torsney (1988), Torsney and Alahmadi (1992), Mandal (2000) and Mandal and Torsney (2000).

Property 1: $p^{(r)}$ is always feasible.

Property 2: $F_\phi\{p^{(r)}, p^{(r+1)}\} \geq 0$ with equality when the d_j's corresponding to nonzero p_j's have a common value, d, in which case $x_j = d_j = d$ or $x_j = F_j = 0$ and so, with $x = d$ or 0,

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

Consider the case $x_j = d_j$.

The inequality property can be seen by letting a positive random variable D take the value $\frac{\partial \phi}{\partial p_j}$ with probability p_j ($p_j = p_j^{(r)}$).

Then

$$F_\phi\{p^{(r)}, p^{(r+1)}\} = Cov[D, f(D)]/E[f(D)]. \tag{3.3.1}$$

Proof:

$$\begin{aligned}
F_\phi\{p^{(r)}, p^{(r+1)}\} &= [p^{(r+1)} - p^{(r)}]^T \underline{d} \\
&= \sum_{i=1}^{J} [p_i^{(r+1)} - p_i^{(r)}] d_i \\
&= \sum_{i=1}^{J} p_i^{(r+1)} d_i - \sum_{i=1}^{J} p_i^{(r)} d_i \\
&= \frac{\sum_{i=1}^{J} p_i f(d_i) d_i}{\sum_{i=1}^{J} p_i f(d_i)} - \sum_{i=1}^{J} p_i d_i \qquad (3.3.2) \\
&= \frac{[\sum_{i=1}^{J} p_i f(d_i) d_i] - [\sum_{i=1}^{J} p_i d_i][\sum_{i=1}^{J} p_i f(d_i)]}{\sum_{i=1}^{J} p_i f(d_i)} \qquad (3.3.3) \\
&= \frac{Cov[D, f(D)]}{E[f(D)]}
\end{aligned}$$

The covariance between D and $f(D)$ must be nonnegative if $f(D)$ is increasing in D. Thus an increase in the criterion can be obtained by a partial but possibly not a full step from $p^{(r)}$ in the direction of $p^{(r+1)}$.

Property 3: Under the above iteration $supp(p^{(r+1)}) \subseteq supp(p^{(r)})$, but weights can converge to zero.

Property 4: An iterate $p^{(r)}$ is a fixed point of the iteration if the derivatives $\frac{\partial \phi}{\partial p_j^{(r)}}$ corresponding to nonzero $p_j^{(r)}$ are all equal. Equivalently if the corresponding vertex directional derivatives $F_j^{(r)}$ are zero. This is a necessary but not a sufficient condition for $p^{(r)}$ to solve Problem 1.

There are other algorithms for finding optimizing distributions. These vary in attribute. Some are simple computationally. Some are highly efficient but heavy in computation.

Vertex direction algorithms were first proposed by Fedorov (1972) and Wynn (1972). These are useful when many weights (p_j) are zero at the optimum. When all weights (p_j) are positive at the optimum or when it has been found which are positive, constrained steepest ascent or Newton type iterations may be appropriate. [see Wu (1978) and Atwood (1976, 1980)]

Each algorithm has advantages and disadvantages depending on the optimization problem under consideration. We use the above multiplicative algorithm for constructing optimizing distributions. As mentioned earlier, this algorithm neatly submits to our problems of interest. Also, the performance of the algorithm is investigated in finding one optimizing distribution for each problem. We improve the convergence rates of the algorithm by subjectively choosing the function $f(.)$ and the free parameter δ. Convergence rates also vary according to the choice of the argument of $f(.)$.

Chapter 4

Optimization Problems in Estimation

4.1 Introduction

In Chapter 2, we considered Problems 1 and 2 which are examples of more problems in statistics which call on the calculations of one or many optimizing distributions or measures.

Take into consideration three examples of Problem 1:

Example 1: One of the elementary examples is that of finding the maximum likelihood estimators of the probabilities of a multinomial likelihood

$$\phi(p) = c(x)p_1^{x_1}p_2^{x_2}\cdots p_J^{x_J} \tag{4.1.1}$$

It is well known that the optimum choice of p_j is $p_j^* = \frac{x_j}{n}$, $n = \sum_{j=1}^{J} x_j$.

Example 2: Estimating the mixing parameters (probabilities) of a mixture distribution given data $\underline{y}_1$, $\underline{y}_2$, ..., $\underline{y}_n$ is another example. This will originate when the component probability models $f_i(y)$ of the mixture are themselves free of any unknown parameters, this would produce the likelihood function.

$$\phi(p) = \prod_{i=1}^{n} \left\{ \sum_{j=1}^{J} p_j f_j(\underline{y}_i) \right\} \tag{4.1.2}$$

A useful text on this is Titterington, Smith and Makov (1985). Other references include Smith and Makov (1978), Dempster, Laird and Rubin (1977).

Properties of Examples 1 and 2:

(i) Since independence is a common assumption in the formulation of probability models, the two functions are all homogeneous.

(ii) The functions have positive derivatives as is obvious from the following respective expressions for $\frac{\partial \phi}{\partial p_j}$:

Example 1: $\frac{\partial \phi}{\partial p_j} = \phi(p)[\frac{x_j}{p_j}]$

Example 2: $\frac{\partial \phi}{\partial p_j} = \phi(p) \left[\sum_i \frac{f_j(\underline{y}_i)}{\sum_s p_s f_s(\underline{y}_i)} \right]$

(iii) In some instances the functions are concave.

With Property (iii), it ensures the existence of a unique maximum while properties (i) and (ii) are useful in the formulation of an algorithm.

Now we consider the problem of determining maximum likelihood estimates under the hypothesis of marginal homogeneity for data in a square $n \times n$ contingency table, Torsney (1988) was the first to consider this problem. Mandal and Torsney (2000) also considers a standardized version of this problem.

Given observed frequencies, O_{ij}, $i = 1, 2, ..., n$, $j = 1, 2, ..., n$ and assuming a single multinomial distribution conditional on $N = \sum_{i=1}^{n} \sum_{j=1}^{n} O_{ij}$, with p_{ij} being cell probabilities, we want to solve the following version of Problem 1. The likelihood function is proportional to $\sum_{i=1}^{n} \sum_{j=1}^{n} O_{ij}\, ln(p_{ij})$.

Maximize $\psi(p) = \sum_{i=1}^{n} \sum_{j=1}^{n} O_{ij}\, ln(p_{ij})$ subject to

$$
\begin{aligned}
& p_{ij} \geq 0, i = 1, 2, ..., n, \ j = 1, 2, ..., n, \\
& \sum_{i=1}^{n} \sum_{j=1}^{n} p_{ij} = 1, \\
& \sum_{j=1}^{n} p_{rj} = \sum_{j=1}^{n} p_{jr} \text{ for } r = 1, 2, ..., n.
\end{aligned}
$$

The latter conditions are the conditions for marginal homogeneity. We can make some simplification of the problem in view of the fact that at the solution

$$p_{ii} = \frac{O_{ii}}{N}, \ \ i = 1, 2, ..., n,$$

and also that one of the linear constraints, e.g. that corresponding to $r = n$, can be

taken away since they are linearly dependent.

Let us look at the case $n = 3$, i.e. a 3×3 contingency table.

4.2 3×3 Case - Maximum Likelihood Estimation

For simplicity, let $(u_1, u_2, u_3, u_4, u_5, u_6) = (O_{12}, O_{31}, O_{23}, O_{21}, O_{13}, O_{32})$ and $(x_1, x_2, x_3, x_4, x_5, x_6) = (E_{12}, E_{31}, E_{23}, E_{21}, E_{13}, E_{32})$, where $E_{ij} = Np_{ij}$, $i = 1, 2, 3$, $j = 1, 2, 3$ and therefore are expected frequencies.

Hence, our problem in terms of x_i's and u_i's is now just simply,

Maximize $\psi(x) = \sum_{t=1}^{6} u_t \, ln(x_t)$ subject to

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

$$\sum_{t=1}^{6} x_t = b = (N - \sum_{i=1}^{3} O_{ii}), \tag{4.2.1}$$

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

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

Last two equations come from the marginal homogeneity conditions. Actually there are three equations from this condition. However, we can consider the above two equations because three equations together become linearly dependent.

For a standardized version of this problem, it can be given by the transformation $z_t = \frac{x_t}{b}$. Accordingly,

$$\psi = \psi(z) = \sum_t u_t ln(z_z) + \sum_t u_t ln(b) \tag{4.2.2}$$

where $\sum_t z_t = 1$.

Thus our problem is to maximize $\psi(z) = \sum_t u_t ln(z_t)$ subject to

$$z_t \geq 0, \ t = 1, 2, ..., 6,$$

$$\sum_{t=1}^{6} z_t = 1, \tag{4.2.3}$$

$$z_1 - z_2 - z_4 + z_5 = 0,$$

$$-z_1 + z_3 + z_4 - z_6 = 0.$$

We can also write as:

$$\underline{z} \in Z = \{\underline{z} : \underline{z} \in \mathbb{R}^6, z_t \geq 0, t = 1, 2, ..., 6, C\underline{z} = \underline{a}\}$$

$$\text{where } C = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 0 & -1 & 1 & 0 \\ -1 & 0 & 1 & 1 & 0 & -1 \end{bmatrix} \text{ and } \underline{a} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

This is a convex polygon. The vertices are given by,

$$\underline{v}_1 = (1/2,\ 0,\ 0,\ 1/2,\ 0,\ 0)^T$$

$$\underline{v}_2 = (0,\ 1/2,\ 0,\ 0,\ 1/2,\ 0)^T$$

$$\underline{v}_3 = (0,\ 0,\ 1/2,\ 0,\ 0,\ 1/2)^T$$

$$\underline{v}_4 = (1/3,\ 1/3,\ 1/3,\ 0,\ 0,\ 0)^T$$

$$\underline{v}_5 = (0,\ 0,\ 0,\ 1/3,\ 1/3,\ 1/3)^T$$

Now we can solve a similar version of Problem 1 with $J = 5$. Also $\underline{z} = E_p\{G(\underline{v})\} = E_p\{\underline{v}\} = \sum_{j=1}^{5} p_j \underline{v}_j$ (As G($\underline{v}$) = $\underline{v}$) and $\phi(p) = \sum_{t=1}^{6} u_t \ ln\{\sum_{j=1}^{5} p_j (\underline{v}_j)_t\}$ where $(\underline{v}_j)_t = t^{th}$ element of $\underline{v}_j$.

Let $V = (\underline{v}_1, \underline{v}_2, ..., \underline{v}_5)$. It can be shown that the partial derivatives are:

$$d_j = \frac{\partial \phi}{\partial p_j} = \underline{v}_j^T \underline{w} \tag{4.2.4}$$

where $\underline{w} = (w_1, w_2, ..., w_6)^T$, $w_i = \frac{u_i}{z_i}$, $i = 1, 2, ..., 6$

Hence, in vector notation:

$$\underline{d} = \frac{\partial \phi}{\partial p} = V^T \underline{w}.$$

Now, from the definition of directional derivatives, the vertex directional derivatives are given by

$$\begin{aligned} F_j &= \frac{\partial \phi}{\partial p_j} - \sum_{i=1}^{5} p_i \frac{\partial \phi}{\partial p_i} \\ &= d_j - p^T \underline{d} \\ &= d_j - p^T V^T \underline{w} \end{aligned} \tag{4.2.5}$$

Thus, in vector notation:

$$\underline{F} = \underline{d} - p^T V^T \underline{w}.$$

Now we apply the above optimization problem to an example. For an interest in the hypothesis of marginal homogeneity, look at a specific example of data in Placket (1974). A grading of the unaided distance vision of each eye of 7477 women had the following frequencies:

$(O_{12}, O_{31}, O_{23}, O_{21}, O_{13}, O_{32}) = (266, 153, 510, 234, 190, 444)$, $b = 1797$.

We use algorithm (3.2.1) to find the optimizing distribution. Using algorithm (3.2.1), we record for $n = 1, 2, 3, 4$ the number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$, for $n = 1, 2, 3, 4$ starting from equal initial weights $p_j^{(0)} = 1/J$, $j = 1, 2, ..., J$, where F_j are the vertex directional derivatives. Refer to §4.4 for tables of results computed for various choices of $f(.)$ and δ.

Table 4.4 shows the numbers of iterations for the best choices of δ (i.e. achieving fastest convergence) for each of $f(.)$.

The results clearly show that the numbers of iterations depend on the choice of $f(.)$. In our case, $f(d) = d^\delta$ and $f(d) = exp\{d\delta\}$ are better.

4.3 4×4 Case - Maximum Likelihood Estimation

Now we consider 4×4 contingency table. The procedure is the same as with the 3×3 case.

In 4×4 case, our problem is to maximize $\phi(p) = \sum_{i=1}^{n}\sum_{j=1}^{n} O_{ij}\ ln(p_{ij})$ subject to

$$p_{ij} \geq 0, i = 1, 2, 3, 4,\ j = 1, 2, 3, 4$$

$$\sum_{i=1}^{4}\sum_{i=4}^{4} p_{ij} = 1, \tag{4.3.1}$$

$$p_{12} + p_{13} + p_{14} - p_{21} - p_{31} - p_{41} = 0,$$

$$-p_{12} + p_{21} + p_{23} - p_{24} - p_{32} - p_{42} = 0,$$

$$-p_{13} - p_{23} + p_{31} + p_{32} + p_{34} - p_{43} = 0.$$

Again, for simplicity consider the following notations:

Let $(u_1, u_2, u_3, u_4, ..., u_{12}) = (O_{12}, O_{31}, O_{24}, O_{43}, O_{13}, O_{21}, O_{34}, O_{42}, O_{14}, O_{41}, O_{23}, O_{32})$ and $(x_1, x_2, x_3, x_4, ..., x_{12}) = (E_{12}, E_{31}, E_{24}, E_{43}, E_{13}, E_{21}, E_{34}, E_{42}, E_{14}, E_{41}, E_{23}, E_{32})$, where $E_{ij} = Np_{ij}$, $i = 1, 2, 3, 4$, $j = 1, 2, 3, 4$ are expected frequencies.

At the solutions $p_{ii} = \frac{O_{ii}}{N}, i = 1, 2, 3, 4$ and in terms of x_i's and u_i's, the simplified problem is to maximize $\psi(x) = \sum_{t=1}^{12} u_t \; ln(x_t)$ subject to

$$x_t \geq 0, \; t = 1, 2, ..., 12,$$

$$\sum_{t=1}^{12} x_t = b = (N - \sum_{i=1}^{4} O_{ii}), \tag{4.3.2}$$

$$x_1 - x_2 + x_5 - x_6 + x_9 - x_{10} = 0,$$

$$x_1 - x_3 - x_6 + x_8 - x_{11} + x_{12} = 0,$$

$$x_2 - x_4 - x_5 + x_7 - x_{11} + x_{12} = 0.$$

Similar to the earlier case, we can transform x_t to $z_t = \frac{x_t}{b}$ and our problem becomes to maximize $\psi(z) = \sum_t u_t \; ln(z_t)$ subject to

$$z_t \geq 0, \; t = 1, 2, ..., 12,$$

$$\sum_{t=1}^{12} z_t = 1, \tag{4.3.3}$$

$$z_1 - z_2 + z_5 - z_6 + z_9 - z_{10} = 0,$$

$$z_1 - z_3 - z_6 + z_8 - z_{11} + z_{12} = 0,$$

$$z_2 - z_4 - z_5 + z_7 - z_{11} + z_{12} = 0.$$

We can also write the above as:

$$\underline{z} \in Z = \{\underline{z} : \underline{z} \in \mathbb{R}^{12}, z_t \geq 0, t = 1, 2, ..., 12, C_{\underline{z}} = \underline{a}\}$$

where $C = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & -1 & 1 \\ 0 & 1 & 0 & -1 & -1 & 0 & 1 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}$ and $\underline{a} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$

The vertices of the above convex polygon are:

$$
\begin{aligned}
\underline{v}_1 &= (1/2,\, 0,\, 0,\, 0,\, 0,\, 0,\, 1/2,\, 0,\, 0,\, 0,\, 0,\, 0)^T \\
\underline{v}_2 &= (0,\, 1/2,\, 0,\, 0,\, 0,\, 0,\, 0,\, 1/2,\, 0,\, 0,\, 0,\, 0)^T \\
\underline{v}_3 &= (0,\, 0,\, 1/2,\, 0,\, 0,\, 0,\, 0,\, 0,\, 1/2,\, 0,\, 0,\, 0)^T \\
\underline{v}_4 &= (0,\, 0,\, 0,\, 1/2,\, 0,\, 0,\, 0,\, 0,\, 0,\, 1/2,\, 0,\, 0)^T \\
\underline{v}_5 &= (0,\, 0,\, 0,\, 0,\, 1/2,\, 0,\, 0,\, 0,\, 0,\, 0,\, 1/2,\, 0)^T \\
\underline{v}_6 &= (0,\, 0,\, 0,\, 0,\, 0,\, 1/2,\, 0,\, 0,\, 0,\, 0,\, 0,\, 1/2)^T \\
\underline{v}_7 &= (1/3,\, 0,\, 0,\, 1/3,\, 0,\, 0,\, 0,\, 1/3,\, 0,\, 0,\, 0,\, 0)^T \\
\underline{v}_8 &= (0,\, 1/3,\, 0,\, 0,\, 0,\, 0,\, 1/3,\, 0,\, 0,\, 1/3,\, 0,\, 0)^T \\
\underline{v}_9 &= (0,\, 0,\, 0,\, 1/3,\, 0,\, 1/3,\, 0,\, 0,\, 0,\, 0,\, 1/3,\, 0)^T \\
\underline{v}_{10} &= (0,\, 0,\, 0,\, 0,\, 1/3,\, 0,\, 0,\, 0,\, 0,\, 1/3,\, 0,\, 1/3)^T \\
\underline{v}_{11} &= (0,\, 1/3,\, 0,\, 0,\, 0,\, 1/3,\, 0,\, 0,\, 1/3,\, 0,\, 0,\, 0)^T \\
\underline{v}_{12} &= (0,\, 0,\, 1/3,\, 0,\, 0,\, 0,\, 0,\, 1/3,\, 0,\, 0,\, 0,\, 1/3)^T \\
\underline{v}_{13} &= (1/3,\, 0,\, 0,\, 0,\, 1/3,\, 0,\, 0,\, 0,\, 1/3,\, 0,\, 0,\, 0)^T \\
\underline{v}_{14} &= (0,\, 0,\, 1/3,\, 0,\, 0,\, 0,\, 1/3,\, 0,\, 0,\, 0,\, 1/3,\, 0)^T \\
\underline{v}_{15} &= (1/4,\, 0,\, 0,\, 1/4,\, 0,\, 1/4,\, 0,\, 0,\, 1/4,\, 0,\, 0,\, 0)^T \\
\underline{v}_{16} &= (1/4,\, 0,\, 0,\, 0,\, 1/4,\, 0,\, 0,\, 1/4,\, 0,\, 0,\, 0,\, 1/4)^T \\
\underline{v}_{17} &= (0,\, 1/4,\, 0,\, 0,\, 0,\, 1/4,\, 1/4,\, 0,\, 0,\, 0,\, 1/4,\, 0)^T \\
\underline{v}_{18} &= (0,\, 1/4,\, 0,\, 0,\, 1/4,\, 0,\, 0,\, 1/4,\, 0,\, 0,\, 1/4,\, 0)^T \\
\underline{v}_{19} &= (0,\, 0,\, 1/4,\, 0,\, 0,\, 0,\, 1/4,\, 0,\, 0,\, 1/4,\, 0,\, 1/4)^T \\
\underline{v}_{20} &= (0,\, 0,\, 1/4,\, 1/4,\, 0,\, 0,\, 0,\, 0,\, 1/4,\, 1/4,\, 0,\, 0)^T
\end{aligned}
$$

Now we can answer a version of Problem 1 with $J = 20$. $\underline{z} = \sum_{j=1}^{20} p_j \underline{v}_j$ and $\phi(p) = \sum_{t=1}^{12} u_t \; ln\{\sum_{j=1}^{20} p_j (\underline{v}_j)_t\}$ where $(\underline{v}_j)_t = t^{th}$ element of $\underline{v}_j$.

The partial derivatives $\frac{\partial \phi}{\partial p_j}$'s are given by $\frac{\partial \phi}{\partial p_j} = \underline{v}_j^T \underline{w}$. Thus $\underline{d} = V^T \underline{w}$, where $V = (\underline{v}_1, \underline{v}_2, ..., \underline{v}_{20})$ and $\underline{w} = (\underline{w}_1, \underline{w}_2, ..., \underline{w}_{12})$, $\underline{w}_t = \frac{u_t}{z_t}$.

For example, we still consider the same data [Placket (1974)], but in 4×4 contingency table format. The grading of the unaided distance vision of each eye of 7477 women had the following frequencies, that is:
$(O_{12}, O_{31}, O_{24}, O_{43}, O_{13}, O_{21}, O_{34}, O_{42}, O_{14}, O_{41}, O_{23}, O_{32}) = (266, 124, 66, 432, 78, 205, 234, 117, 36, 362, 82, 179)$, $b = 2181$

For the same choices of $f(.)$ (as in 3×3 case) in algorithm (3.2.1), we record for $n = 1, 2, 3, 4$ the number of iterations needed to achieve $\max_j \{F_j\} \leq 10^{-n}$, for $j = 1, 2, ..., J$ starting from equal initial weights $p_j^{(0)} = 1/J$, $j = 1, 2, ..., J$ where F_j are the vertex directional derivatives. Refer to §4.4 for tables of results computed for various choices of δ.

Table 4.8, shows the numbers of iterations for δ achieving fastest convergence for each $f(.)$. Our results show that the choice of $f(d) = d^\delta$ and $f(d) = exp\{d\delta\}$ are better.

Now we write the following summary of the iteration results.

(i) The expected frequencies in each case were found as:

3×3-case :

$(E_{12}, E_{31}, E_{23}, E_{21}, E_{13}, E_{32}) = (252.022, 173.898, 479.002, 247.740, 169.616, 474.720)$.

4×4-case :

$(E_{12}, E_{31}, E_{24}, E_{43}, E_{13}, E_{21}, E_{34}, E_{42}, E_{14}, E_{41}, E_{23}, E_{32}) =$

$(252.482, 111.843, 56.966, 409.418, 70.585, 195.258, 247.237, 131.269, 42.785, 383.133,$
$91.625, 188.399)$.

(ii) Each choice of $f(d)$ depends on a free parameter δ which is always positive. The value of δ is important for convergence rates. The values of δ recorded in tables 4.4 and 4.8 achieves fastest convergence.

(iii) In both cases, it is clear that d^{δ} and $exp\{d\delta\}$ achieves the fastest and best convergence.

(iv) Remembering that $\sum_j p_j F_j = 0$, since $F_j = d_j - \sum_i p_i d_i$, we might consider replacing d_j by F_j. Results are given in tables 4.9 and 4.10. In the 3×3-case with $f(F) = \Phi(F\delta)$ and $\delta = 2.5$ the number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$ respectively are 2, 3, 4 ,6, whereas taking $f(d) = \Phi(d\delta)$ with same δ takes 56, 142, 247, 428 iterations to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$

respectively.

In the 4×4-case with $f(F) = \Phi(F\delta)$ and $\delta = 3.0$ the number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$ respectively are 3, 6, 10, 12, whereas taking $f(d) = \Phi(d\delta)$ with same δ takes 417, 1102, 2045, 2951 iterations to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$ respectively.

Thus we see that the choice of $f(.)$ and its argument play an important role in the convergence of the algorithm. Also the choice of the free parameter δ is crucial. With the appropriate choices, we set good results as shown in the tables.

4.4 Tables: Iteration Results

3×3 **-case:**

Results for various choices of δ

δ	$n=1$	$n=2$	$n=3$	$n=4$
0.5	10	22	41	62
0.6	9	20	38	57
0.7	9	20	36	55
0.8	9	19	36	54
0.9	9	19	36	54
1.0	9	20	37	56
1.5	13	29	51	80
2.0	24	56	96	160
2.5	56	142	247	428
3.0	138	468	887	1528

Table 4.1: 3×3 case - $f(d) = \Phi(d\delta)$: Number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$.

δ	$n=1$	$n=2$	$n=3$	$n=4$
0.5	5	11	21	31
1.0	3	5	10	14
1.5	2	4	6	9
1.6	2	4	6	8
1.7	2	3	5	7
1.8	2	4	5	7
1.9	2	4	6	8
2.0	3	5	7	9
2.5	5	9	15	19
3.0	9	17	27	35

Table 4.2: 3×3 case - $f(d) = d^{\delta}$: Number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$.

δ	$n=1$	$n=2$	$n=3$	$n=4$
0.5	5	11	21	30
1.0	3	6	10	14
1.5	2	4	6	9
1.6	3	4	5	8
1.7	3	4	5	7
1.8	3	5	6	8
2.0	3	6	9	11
2.5	9	17	25	33

Table 4.3: 3×3 case - $f(d) = exp\{d\delta\}$: Number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$.

$f(.)$	δ	$n=1$	$n=2$	$n=3$	$n=4$
$\Phi(d\delta)$	0.8	9	19	36	54
d^{δ}	1.6	2	4	6	8
$exp\{d\delta\}$	1.7	3	4	5	7

Table 4.4: 3×3 case - Number of iterations for best choices of δ.

4×4 **-case:**

Result for various choices of δ

δ	$n=1$	$n=2$	$n=3$	$n=4$
0.5	15	48	79	111
0.6	14	45	73	103
0.7	14	43	71	99
0.8	14	42	70	98
0.9	15	43	71	100
1.0	15	44	73	103
1.5	24	64	108	153
2.0	48	125	219	312
3.0	417	1102	2045	2951

Table 4.5: 4×4 case - $f(d) = \Phi(d\delta)$: Number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$.

δ	$n=1$	$n=2$	$n=3$	$n=4$
0.5	8	25	41	57
1.0	4	12	20	28
1.5	3	8	13	18
1.8	3	7	11	14
2.0	3	6	9	13
2.2	3	6	9	12
2.3	3	5	8	11
2.5	4	6	10	13
3.0	6	13	35	57

Table 4.6: 4×4 case - $f(d) = d^{\delta}$: Number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$.

δ	$n=1$	$n=2$	$n=3$	$n=4$
0.5	7	24	38	53
1.0	5	12	18	26
1.5	4	8	12	17
1.7	3	7	10	15
1.9	3	5	9	13
2.0	4	6	9	12
2.1	4	6	9	11
2.3	5	8	12	15
2.5	8	13	19	23

Table 4.7: 4×4 case - $f(d) = exp\{d\delta\}$: Number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$.

$f(.)$	δ	$n=1$	$n=2$	$n=3$	$n=4$
$\Phi(d\delta)$	0.8	14	42	70	98
d^{δ}	2.3	3	5	8	11
$exp\{d\delta\}$	2.1	4	6	9	11

Table 4.8: 4×4 case - Number of iterations for best choices of δ.

Results for various choices of δ using F_j in place of d_j

δ	$n=1$	$n=2$	$n=3$	$n=4$
0.5	6	14	26	39
1.0	4	7	13	19
2.0	2	4	6	8
2.3	2	4	5	7
2.5	2	3	4	6
2.7	2	3	5	7
2.8	2	3	6	9
3.0	2	4	7	11

Table 4.9: 3×3 case - $f(F) = \Phi(F\delta)$: Number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$.

δ	$n=1$	$n=2$	$n=3$	$n=4$
0.5	10	30	49	69
1.0	6	16	25	35
1.5	4	11	17	23
2.0	4	9	13	18
2.5	3	7	11	15
2.7	3	7	10	14
2.9	3	6	10	13
3.0	3	6	10	12
3.5	3	6	14	24
4.0	3	88	164	242

Table 4.10: 4×4 case - $f(F) = \Phi(F\delta)$: Number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$.

Chapter 5

Equality of Variances of the Estimates of Two Parametric Functions

5.1 Introduction

In this chapter, we construct approximate optimal designs which optimize a non-standard criterion function. As an example, we take the criterion as the difference of the variances of the estimates of two parametric functions.

In fact, if the parametric functions are $\underline{a}^T\underline{\theta}$ and $\underline{b}^T\underline{\theta}$, then the above optimization problem is equivalent to minimizing covariance between $\underline{c}^T\hat{\underline{\theta}}$ and $\underline{d}^T\hat{\underline{\theta}}$, where $\underline{c} = \frac{\underline{a}+\underline{b}}{2}$ and $\underline{d} = \frac{\underline{a}-\underline{b}}{2}$. In many design problems it is desired to estimate certain parameters or parametric functions independently of others. This can be done by making covariances or correlations between the relevant parameter estimates to zero. Equivalently, this can be done by finding the optimizing distribution with equal variances of the estimates of the parametric functions with the above relationship between $\underline{a}$, $\underline{b}$ and $\underline{c}$, $\underline{d}$.

This is an example of an optimal regression design problem where we need to obtain an optimizing probability distribution. Some recent work in this direction are Torsney and Alahmadi (1995) and Mandal, Torsney and Carriere (2004).

Torsney and Alahmadi (1995) consider constructing designs subject to zero correlations between the estimates of two linear combinations of the parameters. They consider the case of minimal support designs and transform the constrained optimal design problem to a maximization problem with respect to two or three sets of weights. Mandal, Torsney and Carriere (2004) consider constructing optimal designs by maximizing a criterion (D_A- and A-optimality) subject to two constraints. They solve the problem by transforming the constrained optimization problem to one of maximizing three functions of the design weights simultaneously.

Here, we do not consider a constrained optimization problem. Rather we take one of the constraints as our criterion function. Then we optimize that criterion function subject to the basic constraints $p_j \geq 0 \ \forall j$ and $\sum p_j = 1$.

To construct the optimizing distribution, we use the multiplicative algorithms (3.2.1), indexed by a function $f(.)$ which satisfies certain conditions. To improve the convergence rates, we consider some objective choices of the function $f(.)$.

5.2 Formulation of the Optimization Problem

Suppose the two parametric functions we consider are $\underline{a}^T\underline{\theta}$ and $\underline{b}^T\underline{\theta}$, where $\underline{a}$, $\underline{b} \varepsilon \mathbb{R}^K$. We want to find an approximate design (if it exists) such that the variances

of the estimates of the above two parametric functions. Let the estimates of $\underline{a}^T\underline{\theta}$ and $\underline{b}^T\underline{\theta}$ be $\underline{a}^T\hat{\underline{\theta}}$ and $\underline{b}^T\hat{\underline{\theta}}$ respectively.

The variances of each of the above estimates would be:

$$V(\underline{a}^T\hat{\underline{\theta}}) = \underline{a}^T M^{-1}(p)\,\underline{a} \tag{5.2.1}$$

and

$$V(\underline{b}^T\hat{\underline{\theta}}) = \underline{b}^T M^{-1}(p)\,\underline{b} \tag{5.2.2}$$

Let us define the function $g(p)$ as:

$$\begin{aligned} g(p) &= V(\underline{a}^T\hat{\underline{\theta}}) - V(\underline{b}^T\hat{\underline{\theta}}) \\ &= \underline{a}^T M^{-1}(p)\,\underline{a} - \underline{b}^T M^{-1}(p)\,\underline{b} \end{aligned} \tag{5.2.3}$$

One possible motivation for the case

$$g(p) = \underline{a}^T M^{-1}(p)\,\underline{a} - \underline{b}^T M^{-1}(p)\,\underline{b}$$

arises when we take

$$\phi = -tr\{AM^{-1}(p)A^T\}, \;\; A = [\underline{a}, \underline{b}]^T$$

The above choice of $g(p)$ is equivalent to a problem in the case $\phi = -tr\{AM^{-1}A^T\}$.

If both variances have a common value then $tr\{AM^{-1}A^T\}$ is twice this common value and hence is minimized when this common value is minimized.

Now consider the function,

$$\begin{aligned} \phi = G &= -g^2(p) \\ &= -[\underline{a}^T M^{-1}(p)\,\underline{a} - \underline{b}^T M^{-1}(p)\,\underline{b}]^2 \end{aligned} \tag{5.2.4}$$

Note that the above optimization problem is an example of Problem 1. This can be done by maximizing G $(= -g^2(p))$ for appropriate vectors $\underline{a}$ and $\underline{b}$.

Thus, in Problem 1, we maximize $\phi(p) = G(p) = -g^2(p)$ over $P \equiv \{p = (p_1, p_2, ..., p_j) : p_j \geq 0, \sum_{j=1}^{J} p_j = 1\}$.

If a maximum value of zero is attained, we obtain the optimizing distribution with equal variances of $\underline{a}^T\hat{\underline{\theta}}$ and $\underline{b}^T\hat{\underline{\theta}}$.

Now we obtain the partial derivatives:

$$\begin{aligned} d_j^G = \frac{\partial G}{\partial p_j} &= -2g(p) \cdot \frac{\partial g(p)}{\partial p_j} \\ &= 2[\underline{a}^T M^{-1}(p)\,\underline{a} - \underline{b}^T M^{-1}(p)\,\underline{b}][(\underline{a}^T M^{-1}(p)\,\underline{v}_j)^2 - (\underline{b}^T M^{-1}(p)\,\underline{v}_j)^2] \\ &= 2[\underline{a}^T M^{-1}(p)\,\underline{a} - \underline{b}^T M^{-1}(p)\,\underline{b}][(\underline{a}+\underline{b})^T M^{-1}(p)\,\underline{v}_j][(\underline{a}-\underline{b})^T M^{-1}(p)\,\underline{v}_j] \\ &= 2g(p) \cdot [(\underline{a}+\underline{b})^T M^{-1}(p)\,\underline{v}_j][(\underline{a}-\underline{b})^T M^{-1}(p)\,\underline{v}_j] \end{aligned} \tag{5.2.5}$$

Thus, using the definition of directional derivatives, we can obtain the directional derivatives of G as:

$$F_j^G = d_j^G - \sum_{i=1}^{J} p_i d_i^G \tag{5.2.6}$$

5.3 Algorithms

In the case of finding design maximizing the function $G(p)$, that is, finding design with equal variances of the estimates of $\underline{a}^T\underline{\theta}$ and $\underline{b}^T\underline{\theta}$, we maximize $\phi(p) = -G^2(p)$ subject to $p_j \geq 0, \sum p_j = 1$.

For this type of optimization problem, we use algoithm (3.2.1) but with suitable choice of the argument in $f(.)$. That is, we use the algorithm:

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

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

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

$$F_j^{(r)} = d_j^{(r)} - \sum_i^J p_i^{(r)} d_i^{(r)} \text{ [a directional derivative of G at } p = p^{(r)}],$$

We start with the choice of the argument of $f(.)$ by taking the partial derivative of $G = -g^2(p)$, and a suitable choice of the function $f(.)$ which satisfies the required conditions.

5.4 Examples and Results

We consider the following examples. These examples are defined by their design spaces.

Example-1:

$$\begin{aligned} \underline{v}_1 &= (1, -1, -1,)^T \\ \underline{v}_2 &= (1, -1, 1)^T \\ \underline{v}_3 &= (1, 1, -1)^T \\ \underline{v}_4 &= (1, 2, 2)^T \end{aligned}$$

Example-2:

$$\begin{aligned} \underline{v}_1 &= (1, -1, -1,)^T \\ \underline{v}_2 &= (1, -1, 1)^T \\ \underline{v}_3 &= (1, 1, -1)^T \\ \underline{v}_4 &= (1, 2, 3)^T \end{aligned}$$

Example-3:

$$\begin{aligned} \underline{v}_1 &= (1, -1, -2,)^T \\ \underline{v}_2 &= (1, -1, 1)^T \\ \underline{v}_3 &= (1, 1, -1)^T \\ \underline{v}_4 &= (1, 2, 2)^T \end{aligned}$$

Example-4:

$$\begin{aligned} \underline{v}_1 &= (1, 1, -1, -1)^T \\ \underline{v}_2 &= (1, -1, 1, -1)^T \\ \underline{v}_3 &= (1, -1, -1, -1)^T \\ \underline{v}_4 &= (1, 2, 2, -1)^T \\ \underline{v}_5 &= (1, 1, -1, 1)^T \\ \underline{v}_6 &= (1, -1.5, 1, 1)^T \\ \underline{v}_7 &= (1, -1, -1, 2)^T \end{aligned}$$

These examples correspond to linear models with a constant term since the first component of each vertex ($\underline{v}_j$) is always 1.

In fact, we can assume the more realistic regression model:

$$E(y \mid v) = \underline{v}^T \underline{\theta}, \quad \underline{v} \varepsilon \mathcal{V}. \tag{5.4.1}$$

In examples 1-3, the choices of $\underline{a}$ and $\underline{b}$ (in which $g(p) = 0$ is obtained) are:

$$\begin{aligned} \underline{a} &= (1, 0, 1)^T \\ \underline{b} &= (1, 0, -1)^T \end{aligned}$$

In example 4, the choice of $\underline{a}$ and $\underline{b}$ are:

$$\begin{aligned} \underline{a} &= (1, 0, 0, 1)^T \\ \underline{b} &= (1, 0, 0, -1)^T \end{aligned}$$

Other choices of $\underline{a}$ and $\underline{b}$ in example 4 are:

$$\underline{a} = (1, 0, 1, 0)^T$$
$$\underline{b} = (1, 0, -1, 0)^T$$

We report the performance of algorithm (3.2.1) in calculating the optimizing distributions for each of the above examples. We take $f(.)$ such that it satisfies the required conditions.

We first consider the following choices of $f(.)$, taking $x = d$:

$$f(d) = exp\{d\delta\} \tag{5.4.2}$$

$$f(d) = \frac{exp\{d\delta\}}{1 + exp\{d\delta\}} \tag{5.4.3}$$

In tables 5.1, 5.2, 5.4, 5.5, 5.7, 5.8, 5.10, 5.11, 5.13, and 5.14, we recorded for $n = 1, 2, 3, 4$ the number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$, where F_j are the vertex directional derivatives of G. In all the cases, we take the initial design to be $p_j^{(0)} = \frac{1}{j}, j = 1, 2, 3, 4$. We also try to improve convergence rates of the algorithm by using the properties of the directional derivatives of the criterion function under consideration.

The choice of $f(.)$ plays an important role in the convergence of the algorithm.

Note that any criterion has both positive and negative directional derivatives. So the function $f(.)$ needs to be defined for both positive and negative F. Note that $F_j = d_j - \sum p_j d_j$. Thus $\sum p_j d_j = 0$. So, a suitable choice of the function should be one which is centred at zero and changes reasonably quickly about zero. That is,

$$f(F) = \frac{exp\{F\delta\}}{1 + exp\{F\delta\}} \tag{5.4.4}$$

With these in mind, we choose an objective choice of $f(.)$, namely the iteration results are reported in Tables 5.3, 5.6, 5.9, 5.12, 5.15.

5.5 Tables: Iteration Results

As we can see, for each example the numbers of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$, were small, meaning it converged quickly. We also noticed that for each case of $f(d) = \frac{exp\{d\delta\}}{1+exp\{d\delta\}}$ and $f(F) = \frac{exp\{F\delta\}}{1+exp\{F\delta\}}$, the results were almost the same.

Each choice of $f(.)$ depends on a free parameter of δ which is always positive. For example-5, having $\underline{a} = (1, 0, 1, 0)^T$ and $\underline{b} = (1, 0, -1, 0)^T$, all three $f(.)$'s achieves the fastest and best convergence compared to the other examples. Also noted is the values of G for each case were zero.

Below are the iteration results for different functions of $f(.)$. In examples 1-3, the choices of $\underline{a}$ and $\underline{b}$ are $\underline{a} = (1, 0, 1)^T$ and $\underline{b} = (1, 0, -1)^T$. The results for these examples are reported in Tables 5.1-5.9.

δ	$n=1$	$n=2$	$n=3$	$n=4$
0.02	7	11	15	19
0.03	5	7	9	11
0.04	3	4	5	6
0.06	5	7	9	11
0.07	7	11	14	17
0.08	13	21	27	33

Table 5.1: Example 1 - $f(d) = exp\{d\delta\}$: Number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$.

δ	$n=1$	$n=2$	$n=3$	$n=4$
0.001	309	518	727	936
0.04	7	11	15	19
0.05	6	9	12	14
0.06	5	6	9	11
0.07	4	5	7	9
0.08	3	4	5	6
0.09	2	3	4	5
0.1	3	4	5	6

Table 5.2: Example 1 - $f(d) = \frac{exp\{d\delta\}}{1+exp\{d\delta\}}$: Number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$.

δ	$n=1$	$n=2$	$n=3$	$n=4$
0.009	34	56	78	101
0.03	10	16	22	27
0.04	7	11	15	19
0.05	6	9	12	14
0.06	5	7	9	11
0.07	4	5	7	9
0.08	3	4	5	6
0.09	2	3	4	5
0.1	3	4	5	6

Table 5.3: Example 1 - $f(F) = \frac{exp\{F\delta\}}{1+exp\{F\delta\}}$: Number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$.

δ	$n=1$	$n=2$	$n=3$	$n=4$
0.01	26	42	58	74
0.04	6	9	12	15
0.05	4	6	9	11
0.06	3	4	6	7
0.07	3	4	5	6
0.09	4	5	7	8
0.1	5	7	9	11

Table 5.4: Example 2 - $f(d) = exp\{d\delta\}$: Number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$.

δ	$n=1$	$n=2$	$n=3$	$n=4$
0.01	52	85	118	152
0.05	10	16	21	27
0.08	6	9	12	15
0.09	5	8	10	13
0.1	5	7	9	11
0.11	4	6	7	9
0.12	4	5	6	8
0.13	3	4	5	6
0.14	2	3	4	5
0.15	3	4	5	6
0.2	5	7	9	12
0.25	11	17	23	29

Table 5.5: Example 2 - $f(d) = \frac{exp\{d\delta\}}{1+exp\{d\delta\}}$: Number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$.

δ	$n=1$	$n=2$	$n=3$	$n=4$
0.01	52	85	118	152
0.07	7	11	14	18
0.08	6	9	12	15
0.09	5	8	10	13
0.1	4	7	9	11
0.11	4	6	7	9
0.12	3	5	6	8
0.13	2	4	5	6
0.2	5	7	9	12

Table 5.6: Example 2 - $f(F) = \frac{exp\{F\delta\}}{1+exp\{F\delta\}}$: Number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$.

δ	$n=1$	$n=2$	$n=3$	$n=4$
0.005	33	57	80	104
0.01	17	28	39	50
0.03	5	8	11	14
0.04	4	5	7	9
0.05	2	3	4	5
0.06	3	4	5	6
0.07	4	6	8	10
0.08	6	9	12	15
0.09	9	14	19	24
0.1	15	23	32	40

Table 5.7: Example 3 - $f(d) = exp\{d\delta\}$: Number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$.

δ	$n=1$	$n=2$	$n=3$	$n=4$
0.01	33	57	80	104
0.05	6	10	14	17
0.07	4	7	9	11
0.08	4	5	7	9
0.09	3	4	6	7
0.1	3	4	5	6
0.14	4	6	8	10
0.15	5	7	10	12
0.2	19	32	45	58

Table 5.8: Example 3 - $f(d) = \frac{exp\{d\delta\}}{1+exp\{d\delta\}}$: Number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$.

δ	$n=1$	$n=2$	$n=3$	$n=4$
0.01	33	57	80	104
0.05	6	10	14	17
0.08	4	5	7	9
0.09	3	4	6	7
0.1	3	4	5	6
0.13	4	5	7	8
0.15	5	8	10	12
0.2	20	32	44	56

Table 5.9: Example 3 - $f(F) = \frac{exp\{F\delta\}}{1+exp\{F\delta\}}$: Number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$.

δ	$n=1$	$n=2$	$n=3$	$n=4$
0.001	72	158	245	331
0.005	15	31	47	64
0.01	8	15	23	30
0.02	4	7	10	13
0.03	3	4	6	7
0.04	2	3	4	5
0.05	3	5	7	9
0.06	5	9	13	17
0.07	12	24	37	49
0.08	38	53	69	85
0.09	17	19	20	23

Table 5.10: Example 4 with $\underline{a} = (1,\ 0,\ 0,\ 1)^T$ and $\underline{b} = (1,\ 0,\ 0,\ -1)^T$ - $f(d) = exp\{d\delta\}$: Number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$.

δ	$n=1$	$n=2$	$n=3$	$n=4$
0.001	143	317	491	665
0.005	29	63	97	131
0.01	15	31	47	64
0.05	3	5	8	10
0.06	3	4	6	7
0.07	2	3	4	5
0.09	3	4	5	7
0.1	3	5	6	9

Table 5.11: Example 4 with $\underline{a} = (1,\ 0,\ 0,\ 1)^T$ and $\underline{b} = (1,\ 0,\ 0,\ -1)^T$ - $f(d) = \frac{exp\{d\delta\}}{1+exp\{d\delta\}}$: Number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$.

δ	$n=1$	$n=2$	$n=3$	$n=4$
0.001	143	317	491	665
0.005	29	63	97	131
0.01	4	7	10	13
0.04	4	7	10	13
0.05	3	5	8	10
0.06	3	4	6	7
0.07	2	3	4	5
0.09	3	4	5	7
0.1	3	5	7	9

Table 5.12: Example 4 with $\underline{a} = (1, 0, 0, 1)^T$ and $\underline{b} = (1, 0, 0, -1)^T$ - $f(F) = \frac{exp\{F\delta\}}{1+exp\{F\delta\}}$: Number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$.

δ	$n=1$	$n=2$	$n=3$	$n=4$
0.001	1	85	174	264
0.005	1	17	34	51
0.01	1	9	17	25
0.02	1	5	8	11
0.03	1	3	5	6
0.04	1	2	3	4
0.05	1	3	5	7
0.06	1	5	9	12
0.07	1	11	19	29
0.08	1	237	321	405

Table 5.13: Example 4 with $\underline{a} = (1, 0, 1, 0)^T$ and $\underline{b} = (1, 0, -1, 0)^T$ - $f(d) = exp\{d\delta\}$: Number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$.

δ	$n=1$	$n=2$	$n=3$	$n=4$
0.001	1	169	350	530
0.005	1	34	69	105
0.01	1	17	34	51
0.02	1	9	17	25
0.03	1	6	11	16
0.04	1	5	8	11
0.05	1	4	6	8
0.06	1	3	5	6
0.07	1	2	4	5
0.08	1	2	3	4
0.09	1	3	4	5
0.1	1	3	5	7

Table 5.14: Example 4 with $\underline{a} = (1, 0, 1, 0)^T$ and $\underline{b} = (1, 0, -1, 0)^T$ - $f(d) = \frac{exp\{d\delta\}}{1+exp\{d\delta\}}$: Number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$.

δ	$n=1$	$n=2$	$n=3$	$n=4$
0.001	1	169	350	530
0.005	1	34	69	105
0.01	1	17	33	51
0.02	1	9	17	25
0.03	1	6	11	16
0.04	1	5	8	11
0.05	1	4	6	8
0.06	1	3	5	6
0.07	1	2	4	5
0.08	1	2	3	4
0.09	1	3	4	5
0.1	1	3	5	7

Table 5.15: Example 4 with $\underline{a} = (1, 0, 1, 0)^T$ and $\underline{b} = (1, 0, -1, 0)^T$ - $f(F) = \frac{exp\{F\delta\}}{1+exp\{F\delta\}}$: Number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$.

Chapter 6

Conclusions

6.1 Summary

The application of optimal design theory has become an increasing interest in the field of statistics.

We considered constructing optimizing distributions with applications in estimation by exploring a class of multiplicative algorithms, indexed by a function $f(.)$ is positive and strictly increasing. The function may depend on a free positive parameter δ.

First we provided some basic introduction to linear design theory. We also provided some standard design criteria and discussed their properties.

We discussed optimality conditions. These are based on directional derivatives. We also discussed the properties of these derivatives and the General Equivalence Theorem.

We considered a class of multiplicative algorithms:

$$p_j^{(r+1)} \propto p_j^{(r)} f(x_j^{(r)})$$

where $x_j^{(r)} = d_j^{(r)} or F_j^{(r)}$

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

$$F_j^{(r)} = d_j^{(r)} - \sum_i^J p_i^{(r)} d_i^{(r)}.$$

Properties of the algorithms were also discussed.

Then we considered some estimation problems and their properties. For finding optimizing distributions, we considered the problem of determining maximum likelihood estimates under a hypothesis of marginal homogeneity for data in a square contingency tables. We considered two cases: namely 3×3 case and 4×4 case. We also discussed how we improved convergence rates.

We also considered another estimation/design problem of constructing optimizing distributions with equality of variances of the estimates of two parametric functions of interest. Here also, we discussed how we improved convergence rates of the algorithm by objectively choosing the function $f(.)$, its argument and the free parameter δ.

6.2 Future Work

Many design selections for optimal criteria are highly dependent on the approximation of a response surface model. The model is usually proposed before we collect data. The optimal design generated by a computer algorithm is only optimal for that specific proposed model.

However, in many situations, the regression model is not known at the beginning of the designing stage. In this case, we need to implement a design that is not only efficient for a model but rather for two or more models that might fit the experiment to discriminate between them. By selecting the best model, we can proceed with the optimization techniques. Once the model is selected, it is possible to obtain the optimal design of the chosen model. We would like to work in this direction, that is, on model selection.

As we have discussed in Chapter 1, there are many design criteria in the field of optimal design. The most popular and widely used criteria in computer generated design experiments is $D-$optimality. The $D-$optimality criterion, or determinant criterion, claims that the best set of points in the experiment maximizes the determinant $|X^TX|$. From a statistical point of view, a $D-$optimal design leads to response surface models for which the maximum variance of the predicted responses is minimized. In other words, the points of the experiment will minimize the error in the estimated coefficients of the response model. The advantages of this criterion are the use of irregular shapes and the possibility to include extra design points while the quantitative factors do not depend on the scale of variables. Not only does this criterion use all

relevant information, it is also invariant under linear transformation of the parameter.

In general, an optimal design will rely on the assumed model with its parameters and on the chosen optimization criterion. In the beginning, the focus in optimal design research is on linear models, and later on the developments involved more around nonlinear models. This is an interesting area to work further. We would like to focus on working on optimal regression design problems in the future.

6.3 Further Readings

For further study in optimal design theory, with literature on optimality being vast, a widely popular text is A.C. Atkinson and A.N. Donev, *Optimum Exsperimental Designs*, Oxford University Press, 1992. Other texts include F. Pukelsheim, *Optimal Design of Experiments*, New York, Wiley, 1993, V. V. Fedorov, *Theory of Optimal Experiments*, Academic Press, New York and London, 1972 and S. D. Silvey, *Optimal Design*, Chapman and Hall, London, 1980.

Bibliography

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

[2] Atwood, C. L. (1969) Optimal and Efficient Designs of Experiments. *Annals of Mathematical Statistics* **40**, 1570-1602.

[3] Atwood, C. L. (1976) Convergent Design Sequences, for Sufficiently Regular Optimality Criteria. *Annals of Statistics* **4**, 1124-1138.

[4] Atwood, C. L. (1980) Convergent Design Sequences, for Sufficiently Regular Optimality Criteria, II: Singular Case. *Annals of Statistics* **8**, 894-912.

[5] Bradley, R. A. (1965) Another Interpretation of a Model for Paired Comparisons. *Psychometrika* **30**, 315-318.

[6] Bradley, R. A. and Terry, M. E. (1952) Rank Analysis of Incomplete Block Designs I. The Method of Paired Comparisons. *Biometrika* **39**, 324-345.

[7] Chernoff, H. (1953) Locally Optimal Designs for Estimating Parameters. *Annals of Mathematical Statistics* **24**, 586-602.

[8] Davidson, R. R. (1969) On a Relationship Between Two Representations of a Model for Paired Comparisons. *Biometrics* **25**, 597-599.

[9] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977) Maximum Likelihood from Incomplete Data Via the EM Algorithm (with discussion). *J. Roy. Statist. Soc. Series B* **39**, 1-38.

[10] Elfving, G. (1952) Optimum Allocation in Linear Regression Theory. *Annals of Mathematical Statistics* **23**, 255-262.

[11] Farrel, R. H., Kiefer, J., and Walbran, A. (1967) Optimum Multivariate Designs. *Proc. 5th Berkeley Symp* **Vol. 1**, 113-138 University of California Press, Berkeley.

[12] Fedorov, V. V. (1972) *Theory of Optimal Experiments.* Academic Press, New York and London.

[13] Karlin, S. and Studden, W. J. (1966) Optimal Experimental Designs. *Annals of Mathematical Statistics*, **37**, 783-815.

[14] Kiefer, J. (1959) Optimum Experimental Designs (with discussion). *J. Roy. Statist.Soc. Series B* **2**, 849-879.

[15] Kiefer, J. and Wolfowitz, J. (1960) The Equivalence of Two Extremum Problems. *Canadian J. Math.* **12**, 363-366.

[16] Mandal, S. (2000) *Construction of Optimizing Distributions with Applications in Estimation and Optimal Design* Ph.D Thesis, University of Glasgow, U.K.

[17] Mandal, S. and Torsney, B. (2000) Algorithms for the Construction of Optimizing Distributions. *Communications in Statistics - Theory and Methods* **29**, 1219-1231.

[18] Mandal, S. and Torsney, B. (2004) Construction of Optimal Designs Using a Clustering Approach. *Journal of Statistical Planning and Inference* To Appear.

[19] Mandal, S., Torsney, B. and Carriere, K. C. (2004) Constructing Optimal Designs with Constraints. *Journal of Statistical Planning and Inference* To Appear.

[20] Pazman, A. (1986) *Foundations of Optimum Experimental Design.* Reidel, Dordrecht.

[21] Placket, R. L. (1974) *The Analysis of Categorical Data.* Griffin, London.

[22] Pukelsheim, F. (1993) *Optimal Design of Experiments.* Wiley, New York.

[23] Rhode, C. A. (1965) Generalized Inverses of Partitioned Matrices. *J. Soc. Indust. Appl. Math.* **13**, 1033-1053.

[24] Shah, K. R. and Sinha, B. K. (1989) *Theory of Optimal Designs. Lecture Notes in Statistics.* **Vol. 54**, Springer-Verlag.

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

[26] Silvey, S. D. and Titterington, D. M. (1973) A Geometric Approach to Optimal Design Theory. *Biometrika* **60**, 21-32.

[27] Silvey, S. D., Titterington, D. M., and Torsney, B. (1978) An Algorithm for Optimal Designs on a Finite Design Space. *Communications in Statistics A 7*, 1379-1389.

[28] Smith, A. F. M. and Makov, U. E. (1978) A Quasi-Bayes Sequential Procedure for Mixtures. *J. Roy. Statist. Soc. Series B* **40**, 106-112.

[29] Smith, K. (1918) On the Standard Deviations of Adjusted and Interpolated Values of an Observed Polynomial function and its Constraints and the Guidance they Give Towards a Proper Choice of the Distribution of Observations. *Biometrika* **12**, 1-85.

[30] Smith, A. F. M. and Makov, U. E. (1978) A Quasi-Bayes Sequential Procedure for Mixtures. *J. Roy. Statist. Soc. Series B* **40**, 106-112.

[31] Titterington, D. M. (1976) Algorithms for Computing D-optimal Designs on a Finite Design Space. *Proc. 1976 Conf. on Information Sciences and Systems* Dept. of Elect. Eng., John Hopkins Univ., Baltimore, MD, 213-216.

[32] Titterington, D. M., Smith, A. F. M., and Markov, U. E. (1985) *Statistical Analysis of Finite Mixture Distributions.* Wiley, Chichester.

[33] Torsney, B (1977) Contribution to Discussion of "Maximum Likelihood from Incomplete Data Via the EM Algorithm" by Dempster et al. *J. Roy. Statist. Soc. Series B* **39**, 26-27.

[34] Torsney, B. (1981) *Algorithms for Constrained Optimization Problem with Applications in Statistics and Optimum Design.* Ph.D Thesis, University of Glasgow, Glasgow.

[35] Torsney, B (1983) A Moment Inequality and Monotonicity of an Algorithm. *Proc. Internat. Symp. on Semi-Infinite Programming and Applications (Edited*

by Kortanek, K. O. and Fiacco, A. V.). Lecture Notes in Economics and Mathematical Systems **Vol. 215**, 249-260, University of Texas, Austin.

[36] Torsney, B. (1988) Computing Optimizing Distributions with Applications in Design, Estimation and Image Processing. *Optimal Design and Analysis of Experiments (Edited by Dodge, Y., Fedorov, V. V., and Wynn, H. P.)* 361-370 Elsevier Science Publishers B. V., North Holland.

[37] Torsney, B. and Alahmadi, A. M. (1992) Further Development of Algorithms for Constructing Optimizing Distributions. *Model Oriented Data Analysis. Proc. 2nd IIASA Workshop in St. Kyrik, Bulgaria (Edited by Fedorov, V. V., Müller, W. G., and Vuchkov, I. N.)* 121-129, Physica-Verlag.

[38] Torsney, B. and Alahmadi, A. M. (1995) Designing for Minimally Dependent Observations. *Statistica Sinica* **5** 499-514.

[39] Torsney, B. and Mandal, S. (2001) Construction of Constrained Optimal Designs. *Optimum Designs 2000* 141-152, Kluwer Academic Publishers.

[40] Whittle, P. (1973) Some General Points in Theory of Optimal Experimental Designs. *J. Roy. Statist. Soc. Series B* **35**, 123-130.

[41] Wu, C. F. J. (1978) Some Iterative Procedures for Generating Nonsingular Optimal Designs. *Communications in Statistics A* **7**, 1399-1412.

[42] Wynn, H. P. (1972) Results in the Theory and Construction of D-optimum experimental Designs (with discussion). *J. Roy. Statist. Soc. Series B* **34**, 133-147, 170-186.