Construction of Optimising Distributions with Applications in Estimation and Optimal Design

Saumendranath Mandal

A Dissertation Submitted to the
University of Glasgow
Faculty of Science
for the degree of
Doctor of Philosophy

UNIVERSITY
of
GLASGOW

Department of Statistics
September 2000

To My Mother

Abstract

This thesis considers constructing optimising distributions with applications in estimation and optimal design by exploring a class of multiplicative algorithms.

Chapter 1 opens with an introduction to the area of linear design theory. It begins with an outline of a linear regression design problem including properties of the information matrix of the design. The second half of this chapter focuses on several design criteria and their properties. This part consists of two cases : when interest is in inference about all of the parameters of the model and when interest is in some of these parameters. The criteria include D-, A-, G-, E-, D_A-, L- (linear) and E_A-optimality.

Chapter 2 considers classes of optimisation problems. These include problems [labelled (P1), (P2)] in which the aim is to find an optimising distribution p. In examples of problem (P2) p is seen to define a distribution on a design space. Optimality conditions are determined for such optimisation problems. The emphasis is on a differential calculus approach in contrast to a Lagrangian one. An important tool is the directional derivative $F_\phi\{p, q\}$ of a criterion function $\phi(.)$ at p in the direction of q. The properties of $F_\phi\{p, q\}$ are studied, differentiability is expressed in terms of it, and further properties are considered when differentiability is defined. The chapter ends with providing some optimality theorems based on the results of the previous sections.

Chapter 3 proposes a class of multiplicative algorithms for these problems. Iterations are of the form: $p_j^{(r+1)} \propto p_j^{(r)} f(x_j^{(r)})$, where $x_j^{(r)} = d_j^{(r)}$ or $F_j^{(r)}$ and $d_j^{(r)} = \partial\phi/\partial p_j$ while $F_j^{(r)} = F_\phi\{p^{(r)}, e_j\} = d_j^{(r)} - \sum_i p_i^{(r)} d_i^{(r)}$ (a vertex directional derivative) at $p = p^{(r)}$ and $f(.)$ satisfies some suitable properties (positive and strictly increasing) and may depend on one or more free parameters. We refer to this as algorithm (3.1) [the label it is assigned]. These iterations neatly satisfy

the constraints of problems (P1), (P2). Some properties of this algorithm are demonstrated.

Chapter 4 focuses on an estimation problem which in the first instance is a seeming generalisation of problem (P1). It is an example of an optimisation problem [labelled (P3) in chapter 2] with respect to variables which should be nonnegative and satisfy several linear constraints. However, it can be transformed to an example of problem (P2). The problem is that of determining maximum likelihood estimates under a hypothesis of marginal homogeneity for data in a square contingency table. The case of a 3×3 and of a 4×4 contingency table are considered.

Chapter 5 investigates the performance of the above algorithm in constructing optimal designs by exploring a variety of choices of $f(.)$ including a class of functions based on a distribution function. These investigations also explore various choices of the argument of $f(.)$. Convergence of the above algorithm are compared for these choices of $f(.)$ and it's argument. Convergence rates can also be controlled through judicious choice of free parameters.
The work for this chapter along with the work in chapter 4 has appeared in Mandal and Torsney (2000a).

Chapter 6 explores more objective choices of $f(.)$. It mainly considers two approaches - approach I and approach II to improve convergence. In the first $f(.)$ is based on a function $h(.)$ which can have both positive and negative arguments. This approach is appropriate when taking x_j in $f(x_j)$ to be F_j, since these vertex directional derivatives being 'centred' on zero, take both positive and negative values. The second bases $f(.)$ on a function $g(.)$ defined only for positive arguments. This is appropriate when taking x_j to be d_j if these partial derivatives are positive as in the case with design criteria. These enjoy improved convergence rates.

Chapter 7 is devoted to a more powerful improvement - a 'clustering approach'. This idea emerges while running algorithm (3.1) in a design space which is a

discretisation of a continuous space. It can be observed that 'clusters' start forming in early iterations of the above algorithm. Each cluster centres on a support point of the optimal design on the continuous space. The idea is that, at an appropriate iterate $p^{(r)}$, the single distribution $p^{(r)}$ should be replaced by conditional distributions within clusters and a marginal distribution across the clusters. This approach is formulated for a general regression problem and, then is explored through several regression models, namely, trigonometric, quadratic, cubic, quartic and a second-order model in two design variables. Improvements in convergence are seen considerably for each of these examples.

Chapter 8 deals with the problem of finding an 'approximate' design maximising a criterion under a linear model subject to an equality constraint. The constraint is the equality of variances of the estimates of two linear functions ($\underline{a}^T\underline{\theta}$ and $\underline{b}^T\underline{\theta}$) of the parameters of interest. The criteria considered are D-, D_A- and A-optimality, where $A = [\underline{a}, \underline{b}]^T$. Initially the Lagrangian is formulated but the Lagrange parameter is removed through a substitution, using linear equation theory, in an approach which transforms the constrained optimisation problem to a problem of maximising two functions (Q and G) of the design weights simultaneously. They have a common maximum of zero which is simultaneously attained at the constrained optimal design weights. This means that established algorithms for finding optimising distributions can be considered.
The work for this chapter has appeared in Torsney and Mandal (2000).

Chapter 9 concludes with a brief review of the main findings of the thesis and a discussion of potential future work on three topics: estimation problems, optimisation with respect to several distributions and constrained optimisation problems.

Acknowledgements

I am very much indebted to my supervisor, Dr. Ben Torsney, for suggesting the problems we studied and for all his advice, support, patience and encouragement during the course of this research. Ben's commitment, extending far beyond the realms of academia has been an inspiration and I hope my future path will reflect his excellence.

I acknowledge the financial support of a University of Glasgow Scholarship and a UK CVCP Overseas Research Student (ORS) Award (96017016).

I would like to thank all the staff members and postgraduate students of the Department of Statistics, in particular Professor D. Michael Titterington, Professor Adrian Bowman, Dr. Jim Kay and Miss Mary Nisbet.

My gratitude also goes to Jill, for her invaluable moral support, and David, Athanase and Paul for their inspiration and encouragement.

My thanks are also due to Mou, Aurnab, Mayurika, Gautam, Debjani, Bappa, Sreenath and Anita for their support and encouragement.

Finally, acknowledgements are due to my mother Kalpana, my father Asoke, my sister Tripti, my uncles Ajit and Shibkumar and my cousins for their understanding, patience, support and faith in me.

Contents

1 Linear Design Theory **1**

1.1 Introduction 1

1.2 Discretising the Design Space 6

1.2.1 Definition 1.2.1. Design Measure 9

1.2.2 Definition 1.2.2. Support of a Design Measure 9

1.2.3 Definition 1.2.3. Standardised Variance of the Predicted Response 10

1.3 Properties of the Matrix $M(p)$ 11

1.4 Choice of Design Criteria and Their Properties 14

1.4.1 Definition: Criteria of a Design 14

1.4.2 Case-I : D-, A-, G-, E-optimality 14

1.4.3 Case-II : D_A-, L-, E_A-optimality 25

2 Optimality Conditions **30**

2.1 Introduction 30

2.2 A Hierarchy of Problems 31

2.2.1 Problem (P1) 31

2.2.2 Problem (P2) 31

2.2.3 Problem (P3) 32

2.3 Directional Derivatives 34

2.3.1 Definition 2.3.1. 34

2.3.2 Definition 2.3.2. 35

2.4 Properties of $F_\phi\{p, q\}$ 36

2.5 Further Properties of $F_\phi\{p, q\}$ when ϕ is differentiable 40

2.5.1 Definition 2.5.1 Differentiability (of a concave function) . . 42

2.6 Optimality Theorems 48

3 Algorithms **55**
3.1 Introduction 55
3.2 A Class of Algorithms 56
3.3 Properties of the Iteration (3.1) 58
3.3.1 General Properties 58
3.3.2 Properties of some specific cases 61

4 An Estimation Problem **67**
4.1 Introduction 67
4.1.1 Example (E1) 67
4.1.2 Example (E2) 68
4.1.3 Example (E3) 68
4.2 Maximum Likelihood Estimation - 3×3 case 73
4.3 Maximum Likelihood Estimation - 4×4 case 80
4.4 Comments 87

5 Construction of Optimal Designs **89**
5.1 Introduction 89
5.2 Examples 90
5.3 A variation in algorithm (3.1) : Replacing d_j by F_j 101
5.4 A further variation 105
5.5 A Class of Functions based on a Distribution Function 109
5.6 Optimal Weights 113

6 Objective Choices of $f(.)$ **114**
6.1 Introduction 114
6.2 Approach I 115
6.3 Approach II 121

7 Construction of Optimal Designs using a Clustering Approach **128**
7.1 Introduction 128
7.2 Trigonometric Regression 129
7.3 Polynomial Regression in One Variable 134
7.4 Second-Order Model in Two Variables 140

7.4.1 Derivation of the exact solution of the weights in the second-order model in two variables . . . 142
7.5 Clustering Approach . . . 146
7.5.1 General Problem . . . 146
7.5.2 Algorithms . . . 150
7.5.3 Applications . . . 152
7.5.3.1 Trigonometric Regression . . . 152
7.5.3.2 Polynomial Regression in One Variable . . . 158
7.5.3.3 Second-Order Model in Two Variables . . . 171

8 Construction of Constrained Optimal Designs 176
8.1 Introduction . . . 176
8.2 Constrained Optimising Distributions . . . 178
8.2.1 General Problem (GP) . . . 178
8.2.2 1st order Lagrangian conditions transformed . . . 181
8.3 Example . . . 191
8.4 Algorithms . . . 194
8.5 Empirical Results . . . 200
8.5.1 Quadratic Model . . . 200
8.5.2 Further Examples . . . 202
8.5.3 Optimal Solution and Discussion . . . 206

9 Discussions and Future Work 218
9.1 Discussions - A Brief Review . . . 218
9.2 Future Work . . . 221
9.2.1 Estimation Problems . . . 221
9.2.2 Optimisation w.r.t. Several Distributions . . . 222
9.2.3 Constrained Optimisation Problems . . . 225

A Interation Results Summarised in Chapter 7 226

B Derivation of $\frac{\partial \phi}{\partial p_j}$ and $\frac{\partial^2 \phi}{\partial p_i \partial p_j}$ 234

References 240

List of Tables

4.1 3×3 - case : $f(d) = d^{\delta}$: Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3$. 77
4.2 3×3 - case : $f(d) = exp(\delta d)$: Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3$. 77
4.3 3×3 - case : $f(d) = \Phi(\delta d)$: Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3$. 78
4.4 3×3 - case : $f(d) = ln(e + \delta d)$: Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3$. 78
4.5 3×3 - case : $f(d) = \frac{exp(\delta d)}{1+exp(\delta d)}$: Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3$. 79
4.6 3×3 - case : Number of iterations for best choices of δ 79
4.7 4×4 - case : $f(d) = d^{\delta}$: Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3$. 84
4.8 4×4 - case : $f(d) = exp(\delta d)$: Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3$. 84
4.9 4×4 - case : $f(d) = \Phi(\delta d)$: Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3$. 85
4.10 4×4 - case : $f(d) = ln(e + \delta d)$: Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3$. 85
4.11 4×4 - case : $f(d) = \frac{exp(\delta d)}{1+exp(\delta d)}$: Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3$. 86
4.12 4×4 - case : Number of iterations for best choices of δ 86

5.1 Example-1:Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$. 96
5.2 Example-2:Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$. 97
5.3 Example-3:Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$. 98
5.4 Example-4:Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$. 99

5.5 Example-5:Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$.100
5.6 Example-1 : Iteration counts : Replacing d_j by F_j. 102
5.7 Example-2 : Iteration counts : Replacing d_j by F_j. 103
5.8 Example-3 : Iteration counts : Replacing d_j by F_j. 103
5.9 Example-4 : Iteration counts : Replacing d_j by F_j. 104
5.10 Example-5 : Iteration counts : Replacing d_j by F_j. 104
5.11 Iteration counts : $f(x) = ln(e + \delta x), x_j = d_j - c$, examples 1-3. . . 107
5.12 Iteration counts : $f(x) = ln(e + \delta x), x_j = d_j - c$, examples 4 & 5. 108
5.13 Iteration counts : $f(x) = ln(e + \delta x), x_j = (d_j - c)^{\beta}$, examples 1-5. 108
5.14 Iteration counts for $H_{\beta}(F)$, $G(\delta F) = \Phi(\delta F)$, examples 1 & 2. . . . 111
5.15 Iteration counts for $H_{\beta}(F)$, $G(\delta F) = \Phi(\delta F)$, examples 3-5. 112

6.1 Approach I : $h(x) = exp(\delta x), x = F$, examples 1-5 118
6.2 Approach I : $h(x) = 1/(1 - \delta x), x = F$, examples 1-5 119
6.3 Approach I : $h(x) = (1 - x)^{-\delta}, x = F$, examples 1-5 120
6.4 Approach II : $g(x) = x^{\delta}, x = d/M_t(d)$, examples 1-5 125
6.5 Approach II : $g(x) = e^{\delta(x-1)}, x = d/M_t(d)$, examples 1-5 126
6.6 Approach II : $g(x) = ln(e + \delta(x - 1)), x = d/M_t(d)$, examples 1-5 127

7.1 Trigonometric Regression : Best choice of δ : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$. 133
7.2 Quadratic Regression : Best choice of δ : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$ 139
7.3 Cubic Regression : Best choice of δ : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$ 139
7.4 Quartic Regression : Best choice of δ : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$ 139
7.5 Second-Order Model : Best choice of δ : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$ 141
7.6 Clustering Approach : Trigonometric Regression : Best choice of δ 156
7.7 Clustering Approach : Quadratic Regression : Best choice of δ . . 161
7.8 Clustering Approach : Cubic Regression : Best choice of δ 165
7.9 Clustering Approach : Quartic Regression : Best choice of δ . . . 169
7.10 Clustering Approach : Second-Order Model : Best choice of δ . . 173

A.1 Trigonometric Regression : Iteration Results : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$. 226
A.2 Quadratic Regression : Iteration Results : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$ . . . 227
A.3 Cubic Regression : Iteration Results : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$ 227
A.4 Quartic Regression : Iteration Results : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$ 228
A.5 Second-Order Model : Iteration Results : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$ 228
A.6 Clustering Approach:Trigonometric Regression:Iteration Results . 229

A.7 Clustering Approach:Quadratic Regression:Iteration Results . . . 230
A.8 Clustering Approach:Cubic Regression:Iteration Results 231
A.9 Clustering Approach:Quartic Regression:Iteration Results 232
A.10 Clustering Approach:Second-Order Model:Iteration Results 233

List of Figures

2.1 Vector diagram showing the rate of change in $\phi(.)$ at p in directions which remain the same for all $c > 0$. 37
2.2 Geometrical Representation of Optimality Conditions. 54

5.1 $H_\beta(F)$ versus F, where F is directional derivative. $H_\beta(F)$ is given in (5.11), where $G(\delta F) = \Phi(\delta F)$,the Normal cdf; δ=2.0 and β=1.0 110

7.1 Showing the division of p_i's to total and within cluster weights . . 148
7.2 Trigonometric Regression : Weights versus Design Points after 5 iterations of (3.1), $f(x) = x^\delta$, $x = d$ and $\delta = 1.7$. 156
7.3 Trigonometric Regression:Variance function $d(x, p)$ after 5 iterations of (3.1),$f(x) = x^\delta$,$x = d$,δ=1.7. $d(x, p)$ is given in (7.28) . . 157
7.4 Trigonometric Regression : Variance function $d(x, p^*)$ at the optimum design (7.33). $d(x, p^*)$ is given in (7.34) 157
7.5 Quadratic Regression : Weights versus Design Points after 5 iterations of (3.1), $f(x) = x^\delta$, $x = d$ and $\delta = 1.9$. 161
7.6 Quadratic Regression : Variance function $d(x, p)$ after 5 iterations of (3.1), $f(x) = x^\delta$, $x = d$, $\delta = 1.9$. $d(x, p)$ is given in (7.36). . . . 162
7.7 Quadratic Regression : Variance function $d(x, p^*)$ at the optimum design (7.38). $d(x, p^*)$ is given in (7.39) 162
7.8 Cubic Regression : Weights versus Design Points after 10 iterations of (3.1), $f(x) = x^\delta$, $x = d$ and $\delta = 0.05$. 165
7.9 Cubic Regression : Variance function $d(x, p)$ after 10 iterations of (3.1), $f(x) = x^\delta$, $x = d$, $\delta = 0.05$. $d(x, p)$ is given in (7.40). 166
7.10 Cubic Regression : Variance function $d(x, p^*)$ at the optimum design (7.41). $d(x, p^*)$ is given in (7.42) 166

7.11 Quartic Regression : Weights versus Design Points after 5 iterations of (3.1), $f(x) = x^{\delta}$, $x = d$ and $\delta = 0.05$. 169
7.12 Quartic Regression : Variance function $d(x, p)$ after 5 iterations of (3.1), $f(x) = x^{\delta}$, $x = d$, $\delta = 0.05$. $d(x, p)$ is given in (7.43). 170
7.13 Quartic Regression : Variance function $d(x, p^*)$ at the optimum design (7.44). $d(x, p^*)$ is given in (7.45) 170
7.14 Second-Order Model : Weights versus Design Points after 10 iterations of (3.1), $f(x) = x^{\delta}$, $x = d$, $\delta = 2.3$ 174
7.15 Second-Order Model : Weights versus Design Points at the optimum174
7.16 Second-Order Model : Variance function $d(x, p)$ after 10 iterations of (3.1), $f(x) = x^{\delta}$, $x = d$, $\delta = 2.3$. $d(x, p)$ is given in (7.46) . . . 175
7.17 Second-Order Model : Variance function $d(x, p^*)$ at the optimum design (7.50). $d(x, p^*)$ is given in (7.51) 175

8.1 An example of a *minimax* problem 180
8.2 D-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = -\left[ln(1 + h^T P^2 h)\right]$. 208
8.3 D-optimality: $\phi(p_2)$ $[= \phi(p)]$ versus p_2. 208
8.4 A-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = -[ln(1 + h^T P^2 h)]$. . 209
8.5 A-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = -\left[h(E)\right]^T \left[h(E)\right]$. . 209
8.6 A-optimality: $\phi(p_2)$ $[= \phi(p)]$ versus p_2. 210
8.7 A-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = [R_3 - 1]$. 210
8.8 D-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = -[h^T h]$. 211
8.9 D-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) =$ -$[\log(1 + h^T h)]^2$. . . 211
8.10 D-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) =$ -$[\log\{1+h(E)^T h(E)\}]^2$.212
8.11 D-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = -\left[h(E)\right]^T \left[h(E)\right]$. . 212
8.12 D-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = -[1 - R_1]^2$. 213
8.13 D-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = -[1 - R_2]^2$. 213
8.14 D-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = -[1 - R_3]^2$. 214
8.15 D-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = Q_2 = [R_1 - 1]$ . . . 214
8.16 D-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = [R_2 - 1]$. 215
8.17 D-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = [R_3 - 1]$. 215
8.18 D-optimality: $Q(p_3)$ versus p_3, where $Q(p_3) = -[ln(1 + h^T P^2 h)]$. . 216
8.19 D-optimality: $\phi(p_3)$ versus p_3. 216
8.20 A-optimality: $Q(p_3)$ versus p_3, where $Q(p_3) = -[ln(1 + h^T P^2 h)]$. . 217
8.21 A-optimality: $\phi(p_3)$ versus p_3. 217

Chapter 1

Linear Design Theory

1.1 Introduction

An experiment can be designed to answer a variety of questions of interest. Values or levels of inputs must be chosen before running an experiment and observing a measurement on some (or several) variable(s) of interest. There will be a set of or several combinations of the inputs allowed. We must decide how many observations to take at each combination of inputs. This defines a design.

In regression experiments the inputs are often numerical and the objective is good estimation of the parameters of the regression model. There are a variety of criteria defining good estimation. We choose a design to optimise a chosen criterion.

The general theory of optimal design was originally developed for the linear models. The aim of this chapter is to give a general description of optimal design theory for linear models. We give some fundamental concepts of optimal design theory, such as the definition of a design, variance function, information matrix, various criterion functions and their properties.

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

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

where y is the response variable. In a particular experimental condition, y is considered as a sum of a real-valued response function evaluated at $\underline{x}$, and a random error.

$\underline{x} = (x_1, x_2, \ldots, x_m)^T$ are design variables. These can be chosen by the experimenter, their values being restricted to a space $\mathcal{X}$, i.e. $\underline{x} \in \mathcal{X} \subseteq \mathbb{R}^m$. Thus $\mathcal{X}$ is the set of experimental conditions. $\mathcal{X}$ is called the design space. Typically it will be continuous but can be discrete.

$\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)^T$ is a k-dimensional vector of unknown parameters. The true value of $\underline{\theta}$ is known to belong to a set $\Theta \in \mathbb{R}^k$.

σ is a nuisance parameter; this also is fixed and unknown but is not of primary interest.

$p(.)$ is a probability model.

The experimenter can freely choose the experimental conditions from the given experimental domain $\mathcal{X}$. In most applications, $\mathcal{X}$ is taken to be compact. For each $x \in \mathcal{X}$, an experiment can be performed whose outcome is a random variable $y = y(\underline{x})$, where $var(y(\underline{x})) = \sigma^2$. We generally suppose that σ does not depend on the experimental condition $\underline{x}$.

In linear regression design the model is linear in the unknown parameters $\underline{\theta}$ but not necessarily linear in $\underline{x}$. So in linear models $y(\underline{x})$ has an expected value of the

explicit form:

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

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

In order to obtain an observation on y, a value for $\underline{x}$ must first be selected from $\mathcal{X}$. It is assumed that $\underline{x}$ can be set to any chosen value in $\mathcal{X}$.

Given this control over the selection of $\underline{x}$, a natural question to consider is at what values of $\underline{x}$ should observations, say n, on y be taken in order to obtain a 'best' inference or as reliable an inference as possible for all or some of the parameters $\underline{\theta}$.

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

The mode of inference must first be decided upon. For the moment let us suppose that it is point estimation. It will be seen that the solution proposed for this case will hold good for other modes of inference too.

It is desired then to choose n values $(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n)$ to yield 'best' point estimates $\hat{\underline{\theta}}$ of some or all of the parameters $\underline{\theta}$.

Suppose by some method of point estimation the estimator $\hat{\underline{\theta}}$ of $\underline{\theta}$ is obtained. Let $\hat{\underline{\theta}}$ be unbiased for $\underline{\theta}$. Typically the components $\hat{\theta}_j$ will be correlated. Arguably then the $k \times k$ matrix $D(\hat{\underline{\theta}}) = E([\hat{\underline{\theta}} - \underline{\theta}][\hat{\underline{\theta}} - \underline{\theta}]^T)$, the dispersion matrix of $\hat{\underline{\theta}}$ about

$\underline{\theta}$, contains information about the accuracy of $\hat{\underline{\theta}}$ not only in its diagonal elements, which of course measure the mean squared deviation of $\hat{\theta}_j$ from θ_j, but also in its off-diagonal cross product deviation terms. Generally speaking the "smaller" is $D(\hat{\underline{\theta}})$ the better is the accuracy of $\hat{\underline{\theta}}$.

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

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

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

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

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

Least squares estimators are a conventional choice for this model having the optimality of being best linear unbiased estimators (BLUE). They are solutions of:

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

The $k \times k$ matrix $(X^T X)$ is the information matrix for $\underline{\theta}$. The larger $(X^T X)$, the greater is the information in the experiment. If all the parameters $\underline{\theta}$ are of

interest, then the selection of $\underline{x}$ must at least ensure that the matrix (X^TX) is non-singular, in which case the unique solution for (1.5) is given by:

$$\hat{\underline{\theta}} = (X^TX)^{-1}X^TY \tag{1.6}$$

with

$$\begin{aligned} E(\hat{\underline{\theta}}) &= \underline{\theta} \\ D(\hat{\underline{\theta}}) &= \sigma^2(X^TX)^{-1}. \end{aligned}$$

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

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

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

Clearly the dispersion matrix of $\hat{\underline{\theta}}$ does not depend on $\underline{\theta}$ and only depends proportionally on the parameter σ^2. We have to select $\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n\}$ to make the matrix $D(\hat{\underline{\theta}})$ as small as possible, namely a $\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n\}$ which makes the matrix (X^TX) large in some sense.

1.2 Discretising the Design Space

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

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

where

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

Clearly choosing a vector $\underline{x}$ in the design space $\mathcal{X}$ is equivalent to choosing a k-vector $\underline{v}$ in the closed bounded k-dimensional space $\mathcal{V} = \underline{f}(\mathcal{X})$, where $\underline{f}$ is the vector valued function $(f_1, f_2, \ldots, f_k)^T$. That is, $\mathcal{V}$ is the image under f of $\mathcal{X}$. So, $\mathcal{V}$ is an induced design space. Typically this design space is continuous but we can assume that $\mathcal{V}$ is discrete. A 'justification' for this will be given later on in this section.

Suppose that the discrete design space $\mathcal{V}$ consists of J distinct vectors $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_J$. In order to obtain an observation on y, a value for $\underline{v}$ must first be chosen from the J elements of $\mathcal{V}$ to be the point at which to take this observation. That $\mathcal{V}$ is taken to be discrete suggests that this can be done without error.

The design problem can now be expressed more precisely. At which of the points $\underline{v}_j$ should observations be taken and, if n observations in total are allowed, how many observations should be taken at these points in order to obtain 'best' least squares estimators of $\underline{\theta}$?

Given n observations, we must decide how many of these, say n_j, to take at $\underline{v}_j$,

$\sum_{j=1}^{J} n_j = n$. Given these choices the matrix (X^TX) can be expressed in the form:

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

where

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

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

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

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

However,

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

where

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

$$= VPV^T \quad (1.11)$$

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

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

This is a simpler or more flexible problem to solve and yet one that is not much visibly different from the original.

Naturally an approximate solution that would be preferred to the original exact design problem would be np^*, rounded to a 'nearest' exact design. Hopefully this would be a near fully optimum exact design.

Note that we can view p as defining a probability distribution on $\mathcal{V}$ to yield

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

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

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

Definition 1.2.1. Design Measure

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

$$p = \begin{Bmatrix} \underline{x}_1 & \underline{x}_2 & \cdots & \underline{x}_J \\ p_1 & p_2 & \cdots & p_J \end{Bmatrix} \tag{1.13}$$

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

A more conventional notation is

$$\xi = \begin{Bmatrix} \underline{x}_1 & \underline{x}_2 & \cdots & \underline{x}_J \\ p_1 & p_2 & \cdots & p_J \end{Bmatrix} \tag{1.14}$$

with ξ defined to be the design measure. However, we use the same symbol (p usually) to represent both the vector of weights and the resultant measures on both $\mathcal{X}$ and $\mathcal{V}$. We also do not distinguish notationally between a design and a design measure. This does not cause confusion. If p is a design measure or distribution, it is by definition defined on the design space. We will always be clear about what the design space is and about which design point receives which weight.

Definition 1.2.2. Support of a Design Measure

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

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

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

Definition 1.2.3. Standardised Variance of the Predicted Response

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

$$d(\underline{x}, p) = \underline{f}^T(\underline{x})\, M^{-1}(p)\, \underline{f}(\underline{x}), \tag{1.15}$$

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

If a design has N trials and $\hat{Y}(\underline{x})$ is the predicted response at $\underline{x}$, then $d(\underline{x}, p)$ can be written as

$$d(\underline{x}, p) = N \frac{var\{\hat{Y}(\underline{x})\}}{\sigma^2}. \tag{1.16}$$

1.3 Properties of the Matrix $M(p)$

(i) The information matrix $M(p)$ is symmetric nonnegative definite. The symmetry of this matrix follows from its definition (1.9), and, the nonnegativeness of the appropriate quadratic form is easy to verify.

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

If a design has less that k (number of parameters) support points then the determinant of its information matrix is zero.

(ii) Let $\mathcal{M} = \{M(p) : p \text{ is any probability measure on } \mathcal{V}\}$. The set $\mathcal{M}$ is the convex hull of the set $\{\underline{v}\,\underline{v}^T : \underline{v} \in \mathcal{V}\}$. Note that if p_v is the probability measure that puts unit weight at the point $\underline{v} \in \mathcal{V}$, then $M(p_v) = \underline{v}\,\underline{v}^T$, see Silvey (1980).

Theorem 1.3.1. (Caratheodory's Theorem) Each element M of the convex hull $\mathcal{M}$ of any subset $\mathcal{S}$ of n-dimensional space can be expressed as a convex combination of at most $n+1$ elements of $\mathcal{S}$:

$$M = \sum_{i=1}^{n+1} \alpha_i s_i, \quad \sum_{i=1}^{n+1} \alpha_i = 1, \ \alpha_i \geq 0, \ s_i \in S, \ i = 1, 2, \ldots, n+1 \qquad (1.17)$$

If M is a boundary point of the set $\mathcal{M}$ then α_{n+1} can be set equal to zero.

Such representations are not unique. [See Fedorov (1972) and Silvey (1980).]

Now applying the problem to the case

$$S = \{\underline{v}\,\underline{v}^T : \underline{v} \in \mathcal{V}\},$$

we have that each $M \in \mathcal{M}$ has at least one representation of the form

$$M = \sum_{l=1}^{L} p_l \, \underline{v}_l \, \underline{v}_l^T,$$

where $\underline{v}_l \in \mathcal{V}$, $l = 1, 2, \ldots, L$ and $L \leq \left[\frac{k(k+1)}{2} + 1\right]$. Also by the same theorem if M is a boundary point of $\mathcal{M}$, the inequality involving L can be strengthened to $L \leq \frac{k(k+1)}{2}$.

Some Remarks

From a practical point of view this property is extremely important. It says that, for any experimental design $\tilde{p}$ with support exceeding $\left[\frac{k(k+1)}{2} + 1\right]$ points, and information matrix $M(\tilde{p})$ it is always possible to find a design p with a support of at most $\left[\frac{k(k+1)}{2} + 1\right]$ points, which for this allocation (general number of distinct observations) will have the information matrix $M(p) = M(\tilde{p})$. Thus we have that any continuous measure and in particular any continuous optimal design measure can be replaced by at least one finite discrete probability distribution, and so we have an explanation for having initially assumed $\mathcal{V}$ discrete, for such an optimal design will have a discrete optimal support. This optimal support we could regard as the discrete

space. However, typically, this optimal support will not be known except in some instances when it, or some finite discrete subset of $\mathcal{V}$ containing it, can be identified using intuition or, geometrical or symmetry arguments.

Typically, the optimal support must in a sense be computed, possibly only approximately, as a prelude to determining the optimal weights p^* and this is essentially done by some of the algorithms which we will consider. Any programmed numerical technique must discretise a continuous space; if solutions are not discrete, numerical techniques will only produce discrete approximations. Effectively design algorithms will work with discrete $\mathcal{V}$'s, say $\mathcal{V}_D$'s, containing approximations, as indicated below, to the optimal support $Supp(p^*)$.

An ideal discretisation would seem intuitively to be some form of "uniform grid" on a continuous $\mathcal{V}$, but typically this is difficult to determine when $\mathcal{V}$ is an image under some $\underline{f}$ of some $\mathcal{X}$. In practice the discretisation that is used is the image under $\underline{f}$ of a uniform grid on $\mathcal{X}$.

1.4 Choice of Design Criteria and Their Properties

1.4.1 Definition: Criteria of a Design

It may be possible to obtain a best inference for all or some of the unknown parameters $\underline{\theta} \in \Theta$ by making the matrix $M(p)$ large in some sense. So we consider various ways in which to make the matrix $M(p)$ large, namely by maximising some real valued function $\phi(p) = \psi\{M(p)\}$. Note that the function ϕ is called the criterion function, and in turn, the criterion defined by the function ϕ is usually called ϕ-optimality. A design maximising $\phi(p)$ is called a ϕ-optimal design.

Now we consider different design criteria of interest and their properties.

1.4.2 Case-I

First we consider the case when interest is in inference about all of the parameters $\underline{\theta}$ of the linear model (1.7). The information matrix $M(p)$ must therefore be non-singular and hence positive definite. Possible criteria in this case include D-optimality, A-optimality, G-optimality and E-optimality.

(I) D-optimality

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

$$\phi_D(p) = \psi_D\{M(p)\} = log\, det\{M(p)\} = -log\, det\{M^{-1}(p)\}. \quad (1.18)$$

That is, a measure p^* is called D-optimal if

$$det\{M(p^*)\} = \sup_p det\{M(p)\}. \quad (1.19)$$

Kiefer and Wolfowitz (1960) show that this is equivalent to

$$\inf_p \sup_x d(x,p) = \sup_x d(x,p^*), \quad (1.20)$$

and also to

$$\sup_x d(x,p^*) = k, \quad (1.21)$$

where $d(x,p)$ is the standardised variance of the predicted response

$$d(x,p) = \underline{f}^T(x)\, M^{-1}(p)\, \underline{f}(x) \quad (1.22)$$

and k, as before, is the number of parameters.

Various motivations for D-optimality exist. These extend beyond the idea of point estimation and all fall into the realm of explicit joint inference. There is an interesting statistical interpretation of D-optimal design. If we assume normality of the errors in the linear model (1.7), then the general form of the joint confidence region for the vector of unknown parameters $\underline{\theta} \in \Theta$ is described by an ellipsoid

of the form:

$$\{\underline{\theta} : (\underline{\theta} - \underline{\hat{\theta}})^T M(p) (\underline{\theta} - \underline{\hat{\theta}}) \leq c\}, \text{ for some critical value } c \qquad (1.23)$$

where $\underline{\hat{\theta}}$ is the least squares estimate or the maximum likelihood estimate of $\underline{\theta}$. The D-optimal criterion chooses $M(p)$ to make the volume of the above ellipsoid as small as possible because it is the case that this volume is proportional to $[det\{M(p)\}]^{-\frac{1}{2}}$. The value of $[log\, det\{M(p)\}]$ is finite if and only if $M(p)$ is non-singular i.e. when all the unknown parameters are estimable. This is the most extensively studied of all design criteria; see Kiefer (1959), Farrell et al (1967), Fedorov (1972), Silvey (1980), Pazman (1986), Shah and Sinha (1989), Atkinson and Donev (1992), Pukelsheim (1993).

Other motivations for D-optimality lie in hypothesis testing under a normal linear model, though these would be equivalent to taking the ellipsoid above to be a classical confidence ellipsoid.

Properties of $\phi_D(p) = \psi_D\{M(p)\}$

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

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

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

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

follows from theorem 1.4.2.

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

$$\frac{\partial \phi_D}{\partial p_j} = \underline{v}_j^T M^{-1}(p) \underline{v}_j. \tag{1.24}$$

(iv) ϕ_D is invariant under a non-singular linear transformation of $\mathcal{V}$.

This property can be easily seen to follow from formula (1.9) for $M(p)$. Suppose $\mathcal{V} = [\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_J]$ is transformed to $\mathcal{W} = [\underline{\omega}_1, \underline{\omega}_2, \ldots, \underline{\omega}_J]$ under the linear transformation $\underline{\omega}_j = A\underline{v}_j$, where A is a $k \times k$ matrix. Then a design assigning weight p_j to $\underline{\omega}_j$ has information matrix:

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

Then

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

Theorem 1.4.1. The weighted sum of the variances of the estimates of the predicted response $d(\underline{x}, p)$, taken over all points of the design p, is equal to the number of parameters k.

i.e.,

$$\sum_{j=1}^{J} p_j \, d(\underline{x}_j, p) \;=\; k. \tag{1.25}$$

In the case of a continuous design

$$\int_{\mathcal{X}} d(\underline{x}, p) \, dp(\underline{x}) \;=\; k. \tag{1.26}$$

Proof. From (1.15) we can write $d(\underline{x}_j, p)$ as

$$d(\underline{x}_j, p) \;=\; \underline{f}^T(\underline{x}_j) \, M^{-1}(p) \, \underline{f}(\underline{x}_j). \tag{1.27}$$

So from (1.27) above

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

Theorem 1.4.2. The function $[log\, det\{M(p)\}]$, where $M(p)$ is the information matrix, is a concave function.

Proof. We know that the set of matrices $\mathcal{M} = \{M(p) : p$ is any probability measure on $\mathcal{V}\}$ is a convex set. Therefore to prove the theorem it is sufficient to show that

$$logdet\{M\} > (1-\alpha)\, logdet\{M_1\} + \alpha\, logdet\{M_2\} \tag{1.28}$$

where $M_1, M_2 \in \mathcal{M}$, $M_1 \neq M_2$ and $M = (1-\alpha)\, M_1 + \alpha\, M_2$, $0 < \alpha < 1$. Now (1.28) immediately follows from the inequality

$$[det\{M\}] > [det\{M_1\}]^{(1-\alpha)}\ [det\{M_2\}]^{\alpha}. \tag{1.29}$$

Equality will hold when $M_1 = M_2$.
See Fedorov (1972, p.20).

Theorem 1.4.3. Let ξ_1 and ξ_2 be two designs with distinct information matrices M_1 and M_2, $M_i = M(\xi_i)$, $i = 1, 2$ for which $det\{M_1\} = det\{M_2\}$. Let

$$\xi = (1-\alpha)\,\xi_1 + \alpha\xi_2, \quad 0 < \alpha < 1. \tag{1.30}$$

Then the information matrix $M(\xi)$ of the design ξ has the determinant

$$det\{M(\xi)\} > det\{M_1\}. \tag{1.31}$$

Proof. The proof of this theorem follows immediately from the strict concavity of $log\, det\{M\}$

(II) A-optimality

A-optimality is defined by the following criterion function:

$$\phi_A(p) = \psi_A\{M(p)\} = -Trace\{M^{-1}(p)\} \tag{1.32}$$

Thus an A-optimum design seeks to minimise the sum of the variances of the parameter estimates or their average variance, but does not take correlations between these estimates into account. This criterion was considered by Elfving (1952) and Chernoff (1953).

From the point of view of computational complexity, the criterion $\phi_A(p)$ is particularly simple to evaluate since it only requires the computation of the k diagonal entries of the matrix $M^{-1}(p)$.

Properties of $\phi_A(p) = \psi_A\{M(p)\}$

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

(ii) ψ_A is concave on $\mathbb{M}$.

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

$$\frac{\partial \phi_A}{\partial p_j} = \underline{v}_j^T M^{-2}(p)\, \underline{v}_j. \tag{1.33}$$

(III) G-optimality

G-optimality is defined by the criterion function:

$$\phi_G(p) \;=\; \psi_G\{M(p)\} \;=\; -\underset{\underline{v}\in\mathcal{V}}{Max}\; \underline{v}^T M^{-1}(p)\,\underline{v} \tag{1.34}$$

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

Properties of $\phi_G(p) = \psi_G\{M(p)\}$

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

(ii) ψ_G is concave on $\mathbb{M}$.

(iii) ϕ_G is invariant under a non-singular linear transformation of $\mathcal{V}$. To see this consider the same linear transformation in this section for D-optimality. Then

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

(iv) Suppose that uniquely $\underline{v}_i^T M^{-1}(p)\,\underline{v}_i = \underset{t}{Max}\ \underline{v}_t^T M^{-1}(p)\,\underline{v}_t$, then ϕ_G has unique partial derivatives corresponding to positive weights, namely

$$\frac{\partial \phi_G}{\partial p_j} = [\underline{v}_j^T M^{-1}(p)\,\underline{v}_j]^2, \tag{1.35}$$

otherwise ϕ_G is not differentiable.

(IV) E-optimality

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

$$\phi_E(p) = \psi_E\{M(p)\} = -\lambda_{max}[M^{-1}(p)] = -\lambda_{max}^{-1} \tag{1.36}$$

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

Properties of $\phi_E(p) = \psi_E\{M(p)\}$

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

(ii) ψ_E is concave on $\mathbb{M}$.

(iii) Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ denote the eigenvalues of $M(p)$. If λ_{max} is unique then ϕ_E has unique partial derivatives corresponding to positive weights. Otherwise ϕ_E is not differentiable.

Some General Discussion about A-, D- and E-optimality

All these three criteria are special cases of

$$\psi_t\{M\} = -[(1/k)\, tr(M^{-t})]^{1/t}. \tag{1.37}$$

That D- and E-optimality emerge as particular cases is due to the respective facts that for positive definite M

(i) $\lim_{t\to 0} \psi_t\{M\} = -[det\{M^{-1}\}]^{1/k} = -[det\{M\}]^{-1/k}$

and

(ii) $\lim_{t\to\infty} \psi_t\{M\} = -\lambda_{max}\{M^{-1}\},$

while A-optimality is clearly equivalent to the case $t = 1$.

Results (i) and (ii) can be seen directly in a number of ways. In particular they can be established by a proof analogous to that which would prove the two corresponding moment results, below, of which, interestingly, (i) and (ii) are particular cases.

Suppose x is a discrete positive valued random variable with probability distribution given by

$$P(x = x_i) = q_i,\ i = 1, 2, \ldots, k,$$

where $x_i > 0$, $q_i > 0$ and $\sum q_i = 1$.

Let $f(t) = [E(x^t)]^{1/t}$. $f(t)$ is increasing in t.

Then

(i) $\lim_{t\to 0} f(t) = \prod_{i=1}^{k} x_i^{q_i}$

and

(ii) $\lim_{t\to\infty} f(t) = \max_{1\le i\le k} \{x_i\}$.

For a proof see Beckenback and Bellman (1961, p.16).

Since the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$ of M are positive, M being positive definite, and

$$tr(M^{-t}) = \sum_{i=1}^{k} \lambda_i^{-t},$$

the above matrix results are corollaries which arise in the case

$$q_i = 1/k, \; x_i = \lambda_i^{-1}.$$

It was Kiefer (1974) who observed the above generalization which of course has the advantage of making possible a unified treatment of D-, A- and E-optimality.

1.4.3 Case-II

Sometimes 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 (1.2). Suppose we are interested in s linear combinations of the parameters $\theta_1, \theta_2, \ldots, \theta_k$, namely those s linear combinations which are elements of the vector $\underline{\alpha} = A\underline{\theta}$, where A is an $s \times k$ matrix of rank $s \leq k$. In particular when $A = [I_s : O]$ where I_s is the $s \times s$ identity matrix and O is the $s \times (k - s)$ zero matrix, then in this case we are interested only in estimating the first s parameters $\theta_1, \theta_2, \ldots, \theta_s$ of $\underline{\theta} \in \Theta$.

Now if $M(p)$ is non-singular, then the variance matrix of the least squares estimator of $A\underline{\theta}$ is proportional to the matrix $AM^{-1}(p)A^T$. But 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 (column space) of $M(p)$ which results in the invariance of the matrix $AM^{-}(p)A^T$ to the choice of generalised inverse $M^{-}(p)$ of $M(p)$ [see Graybill (1969), theorem 6.6.9].

Note that a generalised inverse of a matrix M is defined as any matrix M^{-} satisfying the condition $MM^{-}M = M$. This generalised inverse exists for each matrix M, but it is not unique except when M is a square non-singular matrix, in which case $M^{-} = M^{-1}$ uniquely. A particular example is when $M^{-} = M^{+}$, where M^{+} is the Moore-Penrose generalised inverse [some authors call it the pseudo inverse of the p-inverse; see Seber (1977), p.76] which not only satisfies $MM^{+}M = M$, but also $M^{+}MM^{+} = M^{+}$ and symmetry of $M^{+}M$ and MM^{+}.

So a good design will be one which makes the matrix $AM^{-}(p)A^T$ as small as possible. Specific criteria which have been proposed include D_A-, linear and E_A-optimality.

(I) D_A-optimality

The criterion function for this optimality is defined by

$$\phi_{D_A}(p) = \psi_{D_A}\{M(p)\} = -\log\det\{AM^{-}(p)A^{T}\}. \tag{1.38}$$

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

Properties of $\phi_{D_A}(p) = \psi_{D_A}\{M(p)\}$

(i) ψ_{D_A} is an increasing function over the set of positive definite symmetric matrices.

(ii) ψ_{D_A} is concave on $\mathbb{M}$.

(iii) ϕ_{D_A} has unique partial derivatives corresponding to positive weights, namely

$$\frac{\partial\phi_{D_A}}{\partial p_j} = \underline{v}_j^{T} M^{-}(p)\, A^{T}[AM^{-}(p)A^{T}]^{-1}AM^{-}(p)\underline{v}_j\,,\ p_j > 0. \tag{1.39}$$

These derivatives are invariant for any generalised inverse $M^{-}(p)$ of $M(p)$ if $\underline{v}_j$'s and A are in the column space of $M(p)$ [see Graybill (1969), theorem 6.6.9 and corollaries 6.6.9.1, 6.6.9.2].

We now consider an important special case of D_A-optimality.

Note that if $A = [I_s : O]$ and we partition the matrix $M(p)$ as follows:

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

then the matrix $(AM^{-}(p)A^{T})^{-1}$ can be expressed as $(M_{11} - M_{12}\,M_{22}^{-}\,M_{12}^{T})$ [see Rhode (1965) and Torsney (1981)] and our design criterion becomes that of choosing p to maximise the determinant of this matrix. So maximising ϕ_{D_A} in this particular case is equivalent to maximising

$$\phi_{D_s}(p) \;=\; log\,det\{M_{11} - M_{12}\,M_{22}^{-}\,M_{12}^{T}\} \tag{1.40}$$

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

(II) Linear Optimality

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

$$\phi_L(p) \;=\; \psi_L\{M(p)\} \;=\; -tr\{M^{-}(p)L\} \tag{1.41}$$

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

If L is of rank $s \leq k$ it can be expressed in the form $L = A^{T}A$ where A is a $s \times k$ matrix of rank s. Then the criterion function (1.41) can be expressed as

$$\phi_L(p) \;=\; -tr\{M^{-}(p)L\} = -tr\{M^{-}(p)A^{T}A\} = -tr\{AM^{-}(p)A^{T}\}. \tag{1.42}$$

This form stresses the relationship with the D_A-optimum design of (1.38), where

the determinant, rather than the trace, of $\{-AM^{-}(p)A^{T}\}$ was maximised.

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

Properties of $\phi_L(p) = \psi_L\{M(p)\}$

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

(ii) ψ_L is concave on $\mathbb{M}$.

(iii) ϕ_L has unique partial derivatives corresponding to positive weights, namely

$$\frac{\partial \phi_L}{\partial p_j} = \underline{v}_j^T M^{-}(p)\, A^T A M^{-}(p) \underline{v}_j\,,\ p_j > 0. \tag{1.43}$$

Note that the case $A = \underline{c}^T$, where $\underline{c}$ is a $k \times 1$ vector, corresponds to another standard criterion known in the literature as the c-optimality criterion; see Elfving (1952). The criterion function is of the form:

$$\phi_c(p) = -\underline{c}^T M^{-}(p) \underline{c}. \tag{1.44}$$

(III) E_A-optimality

E_A-optimality is defined by the following criterion function.

$$\phi_{E_A}(p) = \psi_{E_A}\{M(p)\} = -\lambda_{max}[AM^{-}(p)A^{T}] \tag{1.45}$$

where λ_{max} denotes the largest eigenvalue of the matrix $AM^{-}(p)A^{T}$; see Pazman (1986).

Properties of $\phi_{E_A}(p) = \psi_{E_A}\{M(p)\}$

(i) ψ_{E_A} is an increasing function over the set of positive definite symmetric matrices.

(ii) ψ_{E_A} is concave on $\mathbb{M}$.

(iii) The differentiability properties of this criterion are similar to those of E-optimality.

Chapter 2

Optimality Conditions

2.1 Introduction

In this chapter we determine conditions for optimality, in which p^* will be optimal for an optimisation problem. The emphasis is on a differential calculus approach in contrast to a Lagrangian one. An important tool is the directional derivative $F_\phi\{p, q\}$ of a criterion function $\phi(.)$ at p in the direction of q, and also a normalised directional derivative. We discuss the properties of $F_\phi\{p, q\}$. We also consider further properties of it when $\phi(.)$ is differentiable. This plays an important simplifying role in the calculus of optimisation. At the end of the chapter we consider some optimality theorems.

We first consider various classes of optimisation problems in which we wish to find an optimising distribution (or which are generalisations of this). Optimal regression designs are a particular example. Other examples include maximum likelihood estimation, stratified sampling, image processing.

2.2 A Hierarchy of Problems

Consider the following problems.

Problem (P1)

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

The equality constraint $\sum_j p_j = 1$ renders the problem a nondegenerate constrained optimisation problem, the full constraint region being a closed bounded convex set.

Problem (P2)

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

$$\mathcal{CH}\{\mathcal{G}(\mathcal{V})\} \;=\; \left\{ x = x(p) = \sum_{j=1}^{J} p_j \, G(\underline{v}_j) \, : \, p = (p_1, p_2, \ldots, p_J) \in \mathcal{P} \right\} \quad (2.1)$$

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

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

That is, solve (P1) for

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

Problem (P3)

Maximise $\Phi(\theta)$ over $\Theta = \left\{\theta = (\theta_1, \theta_2, \ldots, \theta_t) : \theta_j \geq 0,\ C\theta = a\right\}$

where C is an $s \times t$ matrix of rank s, and a is in the range space of C.

Problem (P3) is clearly a generalisation of problem (P1). Of intriguing interest is that problem (P3) can generate an example of problem (P2) and hence of problem (P1).

One such occurence of problem (P3) arises when testing linear hypothesis about the parameters in multinomial models for categorical data. These parameters are of course 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 state that various subsets of the components of θ should sum to unity. We will consider an example of such a linear hypothesis in chapter 4.

Clearly an example of problem (P2) is a general optimal linear regression design problem. Note that, as with the design problem, a generalisation of problem (P2) would be to seek a probability measure defined on a continuous and possibly unbounded space $\mathcal{V}$ to maximise a function $\phi(.)$ as in problem (P2). However, Caratheodory's theorem guarantees that at least one optimising distribution is discrete. Problem (P2) may differ in a number of ways from other examples of problem (P1):

- One may only be interested in an optimising x^* as opposed to an optimising p^*, $x^* = x(p^*)$.

- While there may be a unique optimising x^* there could be many optimising p^*'s.

- An optimising p^* may put $p_j^* = 0$, i.e., the optimum lies on the boundary of $\mathcal{P}$.

In contrast there is almost certainly a unique optimising p^* in the case of examples considered in chapter 4, otherwise the parameters would be inestimable. Also p^* does not lie on the boundary of $\mathcal{P}$.

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

2.3 Directional Derivatives

We define optimality conditions in terms of point to point directional derivatives. There are two derivatives of interest. We define these in terms of a function $\phi(p)$, but this could be any function with no constraints on p.

2.3.1 Definition 2.3.1.

Let

$$g(p,q,\varepsilon) = \phi\{(1-\varepsilon)p+\varepsilon q\} \tag{2.3}$$

$$F_\phi\{p,\, q\} = \lim_{\varepsilon\downarrow 0}\frac{g(p,q,\varepsilon)-\phi(p)}{\varepsilon} = \left.\frac{dg(p,q,\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0+} \tag{2.4}$$

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

The choice of notation is due to a convention which regards this derivative as specifically a Fréchet derivative. In the later chapters generally we use the notation $F\{p,q\}$ instead of $F_\phi\{p,q\}$ except when we need to emphasise which function is under consideration. In another context concerning influence curves the term $F\{p,q\}$ has been referred to by Andrews et al (1972, p.30), as a Von Mises derivative. They refer to Von Mises (1947). See also Hampel (1968, 1971), Eplett (1980).

2.3.2 Definition 2.3.2.

Let

$$h(p, m, \varepsilon) = \phi\{p + \varepsilon m\} \tag{2.5}$$

$$G_\phi\{p, m\} = \lim_{\varepsilon \downarrow 0} \frac{h(p, m, \varepsilon) - \phi(p)}{\varepsilon} = \left. \frac{dh(p, m, \varepsilon)}{d\varepsilon} \right|_{\varepsilon=0^+} \tag{2.6}$$

$G_\phi\{p, m\}$ is called Gâteaux derivative of $\phi(.)$ at p in the direction of m. Clearly $F_\phi\{p, q\} = G_\phi\{p, m\}$ where $m = q - p$, while $G_\phi\{p, m\} = F_\phi\{p, p + m\}$. We note that differentiability of $\phi(.)$ at p implies that G_ϕ is linear in its second argument (see Rockafellar (1970), p.241).

Whittle (1971) uses this alternative but equivalent definition of 2.3.1. Kiefer (1974) also used the concept of Gâteaux derivative in his design theory though he did not call it a directional derivative. Certainly it does not benefit from concavity of $\phi(.)$. However, this representation of $F_\phi\{p, q\}$ in terms of $G\{., .\}$ proves useful.

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

We will see that definition 2.3.1, which allows the direction of interest, to be determined by a point q as above, is more useful and indeed leads to a generalisation of some standard calculus. The derivative $F_\phi\{p, q\}$ will serve our purpose better than $G_\phi\{p, q\}$.

Now we discuss some general properties (GP[1]) of $F_\phi\{p,q\}$.

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

(GP1) If $p,q \in S$, where S is a convex set, then so does $\{(1-\varepsilon)p+\varepsilon q\}$, which is clearly an advantage if one wishes $F_\phi\{p,q\}$ only for $p,q \in S$. In contrast, $G_\phi(p,q)$ does not particularly benefit from such convexity.

(GP2) $F_\phi\{p,q\} \geq \phi(q)$ - $\phi(p)$ if $\phi(.)$ is concave.

Proof:

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

(GP3) $F_\phi\{p,p\} = 0$, a desirable property since no change is effected in $\phi(.)$ if one does not move from p. In contrast $G_\phi\{p,p\} = F_\phi\{p,2p\} \neq 0$.

(GP4) Intuitively $F_\phi\{p,q\}$ in some sense measures the rate of change in $\phi(.)$ at p in the direction of q. However, it does so in units of measurement which depend on the distance between p and q. $F_\phi\{p,q\}$ depends on this distance as well as on the said rate of change.

Note that to move from p in the direction of q is to move from p in the direction

[1]The notation 'GP' is used for 'General Property'

of the vector $m = q - p$ and hence in the direction of the vector cm, $c > 0$. If we pass along the full length of the vector cm from p, we will arrive (according to the theory of vectors) at $\{p + c(q-p)\}$. This phenomenon is shown in figure 2.1.

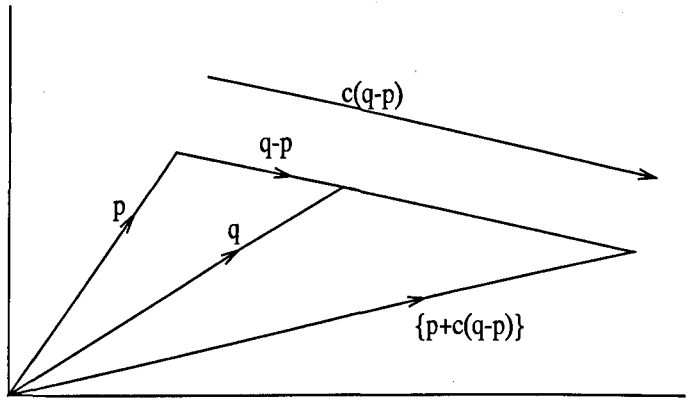


Figure 2.1. Vector diagram showing the rate of change in $\phi(.)$ at p in directions which remain the same for all $c > 0$.

So $F_\phi\{p, p + c(q-p)\}$ measures the rate of change in $\phi(.)$ at p in directions which remain the same for all positive c.

We prove this in the following:

$$\begin{aligned} F_\phi\{p,\, p + c(q-p)\} &= \lim_{\varepsilon \downarrow 0} [\phi\{(1-\varepsilon)p + \varepsilon[p + c(q-p)]\} - \phi(p)]/\varepsilon \\ &= \lim_{\varepsilon \downarrow 0} [\phi\{p + c\varepsilon(q-p)\} - \phi(p)]/\varepsilon \\ &= \lim_{\varepsilon \downarrow 0} c\,[\phi\{p + \delta(q-p)\} - \phi(p)]/\delta, \quad \delta = c\varepsilon \\ &= \lim_{\delta \downarrow 0} c\,[\phi\{(1-\delta)p + \delta q\} - \phi(p)]/\delta \end{aligned}$$

Hence

$$F_\phi\{p,\, p + c(q-p)\} = c\, F_\phi\{p,\, q\}. \tag{2.7}$$

We can write $F_\phi\{p,q\} = f'_+(0)$ where $f(\varepsilon) = \phi\{(1-\varepsilon)p+\varepsilon q\}$. Since $f'_+(0)$ is the amount of change induced in the linear approximation to $f(.)$ at 0 by a unit increase in ε, it follows that $F_\phi\{p,q\}$ defines the amount of change induced in a corresponding linear approximation to $\phi(.)$ at p by a step towards q, the magnitude of which is the distance between q and p, namely, $||q-p|| = \sqrt{(q-p)^T(q-p)}$.

Thus it suggests that we should calculate $F_\phi\{p,q\}$ only for a q which is a unit distance from p. However, the problem is that we will be presented with a q of interest which will not typically be a unit distant from p. Such a q must be scaled up or down appropriately. Clearly the solution is to choose the constant above such that $c(q-p)$ has unit length, namely c such that $c^{-1} = ||r||$, $r = q-p$.

Thus it creates the normalised directional derivative

$$F^I_\phi\{p,\, q\} \;=\; F_\phi\{p,q\} \,\Big/\, \sqrt{r^T r}\,. \tag{2.8}$$

This uses only one particular norm. A more general normalised directional derivative would be

$$F^A_\phi\{p,q\} \;=\; F_\phi\{p,q\} \,/\, \sqrt{r^T A\, r}, \tag{2.9}$$

where A is symmetric nonnegative definite.

We conclude this section by commenting on some relatives or generalisations of $F_\phi\{p,q\}$.

(i) A converse concept is the directional derivatives of $\phi(.)$ at p as p is approached from the direction of q, namely,

$$\bar{F}_\phi\{p,q\} = \lim_{\delta\uparrow 0} [\phi\{(1+\delta)p - \delta q\} - \phi(p)]/\delta \tag{2.10}$$

However,

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

a result which is to be expected.

Thus $\bar{F}_\phi\{p,q\}$ enjoy properties analogous to that of $F_\phi\{p,q\}$. In particular,

$$\bar{F}_h\{p,p-1\} = h'_-(p),$$

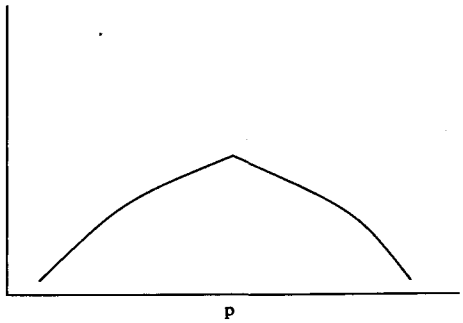
the left hand derivative at p of a function $h(.)$ of one variable.

(ii) An offspring of $F_\phi\{p,q\}$ defines higher order directional derivatives of $\phi(.)$ at p in the direction of q, namely,

$$F_\phi^{(n)}\{p,q\} = \left.\frac{d^n f(\varepsilon)}{d\varepsilon^n}\right|_{\varepsilon=0+},\ f(\varepsilon) = \phi\{(1-\varepsilon)p + \varepsilon q\}. \tag{2.11}$$

2.5 Further Properties of $F_\phi\{p, q\}$ when ϕ is differentiable

We have not so far made any assumptions about differentiability of the criterion function ϕ. A function need not be differentiable at a point p in order that it should have well defined directional derivatives in all directions. Whittle (1971, p.61) quotes the following figure in support of this.



The figure demonstrates the following, as stated by Whittle: "that a function could have a discontinuity in slope at a point p although sloping away from this point in a perfectly smooth fashion in any given direction".

However, when the criterion function ϕ is differentiable, it plays an important simplifying role in the optimisation of ϕ. Now we proceed to redefine the concept in terms of $F_\phi\{p, q\}$.

The idea is that, at p, $\phi(.)$ should be smoothly changing in all directions. A more precise definition is that, at 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. Then 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 coincide in such a manner, they must have some common characteristics at the point of contact p. Apart from sharing a common value they must be changing at the same rate, otherwise, they will not be in parallel and will definitely cross. They must have common first derivatives, partial, directional or Gâteaux, and hence whatever properties are enjoyed by the derivatives of one function at p must be enjoyed by those of the other function.

Then consider the form of the directional derivative of a linear function $L(p) = a^T p + b$

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

Similarly

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

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

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

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

or

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

The condition on $G_\phi\{p,q\}$ is a familiar definition of differentiability. If we were to accept this as such a definition an equivalent and as we shall see a more useful form would be the following definition applied to any function $\phi(.)$.

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

$$F_\phi\{p,e_j\} \;=\; d_j \;-\; \sum_{i=1}^{J} p_i\, d_i \tag{2.12}$$

Definition 2.5.1 Differentiability (of a concave function)

A concave function $\phi(p)$ is differentiable at p if

$$F_\phi\Big\{p,\sum_r c_r q_r\Big\} \;=\; \sum c_r F_\phi\{p,q_r\} + \Big[\sum_r c_r - 1\Big]F_\phi\{p,2p\} \tag{2.13}$$

or

$$G_\phi\Big\{p,\sum_r c_r q_r\Big\} \;=\; \sum_r c_r G_\phi\{p,q_r\}. \tag{2.14}$$

These two conditions are equivalent as we shall see later. The latter, clearly the simpler, states that $G_\phi\{p,\, q\}$ must be linear in the second argument, while in general this need not be the case with $F_\phi\{p,\, q\}$.

Below is a list of properties[2] which follow from this definition.

(DP1) $G_\phi\{p, q\} = q^T d$ since $q = (q_1, q_2, \ldots, q_J)^T = \sum_{i=1}^{J} q_i e_i$ and $d_i = \frac{\partial \phi}{\partial p_i} = G_\phi\{p, e_i\}$.

Conversely suppose $G_\phi\{p, q\} = q^T d$ for all q.

Then

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

Hence condition (2.14) is equivalent to requiring that $G_\phi\{p, q\} = q^T d$.

Of interest is that, according to Rockafellar (1970, p.244, theorem 25.2), a sufficient condition for (2.14) to hold in the case of concave functions, is that the two-sided partial derivatives exist at p and are finite.

(DP2) $F_\phi\{p, q\} = G_\phi\{p, q - p\} = G_\phi\{p, q\} - G_\phi\{p, p\} = (q - p)^T d$

As in the design context, if we regard the argument of the function $\psi(.)$ as a symmetric $k \times k$ matrix A, then this result can be re-expressed in the form

$$F_\psi\{A, B\} = tr\{(B - A)^T \Delta\psi(A)\},$$

where $\Delta\psi(A)$ is the $k \times k$ matrix whose $(i, j)^{\text{th}}$ element is $(\frac{\partial \psi}{\partial a_{ij}})(1 + \delta_{ij})/2$.

[2]The Proeprties are denoted by 'DP' to mean that these are after assuming Differentiability

(DP3) A proof that (2.14) implies (2.13) is

$$\begin{aligned} F_\phi\Big\{p, \sum_r c_r q_r\Big\} &= G_\phi\Big\{p, \sum_r c_r q_r\Big\} - G_\phi\{p,p\} \\ &= \sum_r c_r G_\phi\{p, q_r\} - G_\phi\{p,p\} \\ &= \sum_r c_r\Big[G_\phi\{p,q_r\} - G_\phi\{p,p\}\Big] + \Big[\sum_r c_r - 1\Big] G_\phi\{p,p\} \\ &= \sum_r c_r F_\phi\{p,q_r\} + \Big[\sum_r c_r - 1\Big] F_\phi\{p,2p\} \qquad \text{(by GP3)} \end{aligned}$$

A proof that (2.13) implies (2.14) is

$$\begin{aligned} G_\phi\Big\{p, \sum_r c_r q_r\Big\} &= F_\phi\Big\{p, p + \sum_r c_r q_r\Big\} \\ &= F_\phi\{p,p\} + \sum_r c_r F_\phi\{p,q_r\} + \Big[1 + \sum_r c_r - 1\Big] F_\phi\{p,2p\} \\ &= \sum_r c_r F_\phi\{p,q_r\} + \Big[\sum_r c_r\Big] F_\phi\{p,2p\} \qquad \text{(by GP3)} \\ &= \sum_r c_r\Big[F_\phi\{p,q_r\} + F_\phi\{p,2p\}\Big] \\ &= \sum_r c_r\Big[F_\phi\{p,p\} + F_\phi\{p,q_r\} + F_\phi\{p,2p\}\Big] \\ &= \sum_r c_r F_\phi\{p, p+q_r\} \qquad \text{(by (2.13))} \\ &= \sum_r c_r G_\phi\{p,q_r\}. \end{aligned}$$

(DP4) $F_\phi\{p, 2p-q\} = 2F_\phi\{p,p\} - F_\phi\{p,q\} + [(2-1)-1]F_\phi\{p,2p\}$

i.e. $F_\phi\{p,2p-q\} = -F_\phi\{p,q\}$ (by GP3)

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

This is the distinguishing feature of differentiability from nondifferentiability, that, as we pass through p in the direction of q, the rate of change in $\phi(.)$ should be the same on the approach to and the departure from p. In the case of a function $h(x)$ of a one-dimensional variable x, a consequence is that there is no need to distinguish between right and left hand derivatives, for

$$\lim_{\varepsilon \uparrow 0} \left[\{h(x+\varepsilon) - h(x)\}/\varepsilon\right] = \lim_{\varepsilon \downarrow 0} \left[\{h(x+\varepsilon) - h(x)\}/\varepsilon\right]$$

i.e.,

$$\bar{F}_\phi\{p, p-1\} = F_\phi\{p, p+1\}.$$

(DP5) If $\sum_r c_r = 1$ then clearly $F_\phi\left\{p, \sum_r c_r q_r\right\} = \sum_r c_r F_\phi\{p, q_r\}$.

This proves to be a very useful result for us, when the criterion $\phi(.)$ or $\psi(.)$ is defined on a convex set S. For example, $S = \mathcal{P}$ in the case of problem (P1) and $S = \mathcal{CH}\{\mathcal{G}(\mathcal{V})\}$ as in the case of problem (P2). For example, if $y \in \mathcal{CH}\{\mathcal{G}(\mathcal{V})\}$ then $y = x(q) = \sum_j q_j G(\underline{v}_j)$, $\sum_j q_j = 1$, $q_j \geq 0$.

Hence

$$F_\psi\{x, y\} = \sum_j q_j F_\psi\{x, G(\underline{v}_j)\}. \tag{2.15}$$

In problem (P1) $\phi(.)$ is a function of $p \in \mathcal{P}$. Then for $q \in S$,

$$F_\phi\{p, q\} = \sum_j q_j F_\phi\{p, e_j\}. \tag{2.16}$$

Note that if $\phi(p) = \psi\{x(p)\}$, $x(p) \in \mathcal{CH}\{\mathcal{G}(\mathcal{V})\}$,

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

If we consider $S = \mathcal{M}$ and ψ is defined on $\mathcal{M}$

$$F_\psi\{M(p), M(q)\} = \sum_j q_j F_\psi\{M(p), \underline{v}_j \underline{v}_j^T\}. \tag{2.17}$$

In the remaining properties the set S is crucial.

(DP6) For $S = \mathcal{CH}\{\mathcal{G}(\mathcal{V})\}$, $\sum_j p_j F_\psi\{x(p), G(\underline{v}_j)\} = 0$.

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

In particular $\sum_j p_j F_\psi\{M(p), \underline{v}_j \underline{v}_j^T\} = 0$.

Note one particular consequence of this result.

For at least one $\underline{v}_r \in Supp(p)$, $F_\psi\{x(p), G(\underline{v}_r)\} > 0$ while for at least one $\underline{v}_s \in Supp(p)$, $F_\psi\{x(p), G(\underline{v}_s)\} < 0$, unless $F_\psi\{x(p), G(\underline{v}_j)\} = 0$ for all $\underline{v}_j \in Supp(p)$.

(DP7) If $S = \mathcal{P}$ then

$$\begin{aligned}\max_{q \in S} F_\phi\{p, q\} &= \max_{1 \le j \le J} F_\phi\{p, e_j\} \\ \min_{q \in S} F_\phi\{p, q\} &= \min_{1 \le j \le J} F_\phi\{p, e_j\}\end{aligned}$$

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

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

and hence

$$\Big[\sum_j q_j\Big] \min_{1 \le t \le J} F_\phi\{p, e_t\} \le F_\phi\{p, q\} \le \Big[\sum_j q_j\Big] \max_{1 \le s \le J} F_\phi\{p, e_s\}$$

The result follows since $\sum_j q_j = 1$.

(DP8) For $S = \mathcal{CH}\{\mathcal{G}(\mathcal{V})\}$,

$\max_{y \in S} F_\psi\{x, y\} \ge 0$, $\min_{y \in S} F_\psi\{x, y\} \le 0$. [see problem (P2)]

For $S = \mathcal{P}$,

$\max_{q \in S} F_\phi\{p, q\} \ge 0$, $\min_{q \in S} F_\phi\{p, q\} \le 0$. [see problem (P1)]

These follow from (DP6) and (DP7).

2.6 Optimality Theorems

We are now in a position to state optimality theorems. Their truth is, in the main, self evident in the light of the results of the previous section.

We have already mentioned that differentiability plays an important simplyfying role in optimisation. This is so because differentiability of a point x^* demands in the case of unconstrained optimisation that x^* can be a stationary value if it is to be an optimum. It can also prove to be the case in a constrained optimisation problem that a differentiable point x^* requires to be what is called a constrained stationary value. We will see later in this section that this leads to simple optimality conditions.

Theorem 2.6.1. Let $S = \mathcal{CH}\{\mathcal{G}(\mathcal{V})\}$ and assume that $\psi(x)$ is concave on S, then $x(p^*)$, $p^* \in \mathcal{P}$, maximises $\psi(.)$ on S iff

$$\begin{aligned} & F_\psi\{x(p^*), x(q)\} \leq 0 \qquad \forall q \in \mathcal{P} \\ \text{i.e.} \quad & \max_{q \in \mathcal{P}} F_\psi\{x(p^*), x(q)\} \leq 0 \qquad \forall x(q) \in S. \end{aligned}$$

An alternative but equivalent condition is that

$$\begin{aligned} & F_\psi\{x(q), x(p^*)\} \geq 0 \qquad \forall q \in \mathcal{P} \\ \text{i.e.} \quad & \min_{q \in \mathcal{P}} F_\psi\{x(q), x(p^*)\} \geq 0 \qquad \forall x(q) \in S. \end{aligned}$$

Proof.

Necessity: $\psi(.)$ is concave and $x(p^*)$, $p^* \in \mathcal{P}$, maximises $\psi(.)$ on S imply

$$\psi\{(1-\varepsilon)\,x(p^*) + \varepsilon\, x(q)\} - \psi\{x(p^*)\} \leq 0 \tag{2.18}$$

for all $\varepsilon \in [0,1]$ and all $q \in \mathcal{P}$.

Now from the definition of directional derivative $F_\psi\{x(p^*, x(q)\}$,
(2.18) implies that

$$F_\psi\{x(p^*), x(q)\} \leq 0 \qquad \forall q \in \mathcal{P}.$$

Sufficiency: From the definition of $F_\psi\{x(p^*, x(q)\}$ and concavity of $\psi(.)$,

$$\begin{aligned} F_\psi\{x(p^*, x(q)\} &= \lim_{\varepsilon \downarrow 0} [\psi\{(1-\varepsilon)x(p^*) + \varepsilon x(q)\} - \psi(x(p^*)]/\varepsilon \\ &\geq \lim_{\varepsilon \downarrow 0} [(1-\varepsilon)\psi(x(p^*) + \varepsilon\psi(x(q)) - \psi(x(p^*)]/\varepsilon \\ &= \psi(x(q)) - \psi(x(p^*)). \end{aligned}$$

This is in fact ψ-version of property (GP2).
Hence

$$F_\psi\{x(p^*), x(q)\} \leq 0 \qquad \forall q \in \mathcal{P}$$

implies

$$\psi(x(q)) - \psi(x(p^*)) \leq 0 \qquad \forall q \in \mathcal{P}.$$

i.e., $x(p^*)$ maximises $\psi(.)$.

Note that the theorem is saying that derivatives at $x(p^*)$ should in all directions

be nonpositive.

The theorem has little practical value. It only will be of practical value if it suggests an explicit solution for $x(p^*)$, and that is unlikely if $x(p^*)$ lies on the boundary of $\mathcal{CH}\{\mathcal{G}(\mathcal{V})\}$. The optimality conditions it defines are infinite. If we employ numerical techniques to compute the optimum and the solution it suggests is in fact the correct solution, it may not be easy to verify this.

The above criticism can not be made of the next theorem.

Theorem 2.6.2. Vertex Direction Optimality Theorem

Let $S = \mathcal{CH}\{\mathcal{G}(\mathcal{V})\}$ and assume that $\psi(x)$ is concave on S and $x(p^*)$ is a differentiable point of $\psi(.)$, then $x(p^*)$ maximises $\psi(.)$ on S iff

$$F_\psi\{x(p^*), G(\underline{v}_j)\} = 0 \quad \text{when } p_j^* > 0 \tag{2.19}$$

$$F_\psi\{x(p^*), G(\underline{v}_j)\} \leq 0 \quad \text{when } p_j^* = 0. \tag{2.20}$$

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

A key result in proving this is that

$$F_\psi\{x(p^*), x(q)\} = \sum_j q_j F_\psi\{x(p^*), G(\underline{v}_j)\} \tag{2.21}$$

where $q_j \geq 0$ and $\sum_j q_j = 1$. [Property (DP5)].

Clearly this is nonnegative if (2.19) and (2.20) are satisfied. Proving the opposite is more involved. Suppose all $p_j > 0$ so that $x(p^*)$ is not on the boundary of $\mathcal{CH}\{\mathcal{G}(\mathcal{V})\}$. Then the need for the zero derivatives in (2.19) is that if $F_\psi\{x(p^*), G(\underline{v}_j)\}$ were negative then the derivative in the opposite direction would be positive. Hence theorem 2.6.1 would be violated.

Theorem 2.6.2 plays an important role in constructing optimal designs, specifying a finite set of optimality conditions. It should be easy to check whether or not these are satisfied by a postulated solution obtained by numerical techniques. Differentiability though is an essential requirement.

We illustrate the above in figure 2.2. If the optimum lies inside the basic feasible region (case-1 in figure 2.2) then under theorem 2.6.2 the directional derivatives at the optimum towards each vertex would be zero.

If the optimum lies on the boundary of the feasible region, for example, on the line joining vertices $\underline{v}_1$ and $\underline{v}_5$ (as shown in case-2 of figure 2.2), then the directional derivatives at the optimum towards these two verices should be zero, but should be negative (or not positive) towards the other vertices.
i.e., $F_1 = F_5 = 0$, F_2, F_3, $F_4 \leq 0$.

Corollary (i)

If $S = \mathcal{M}$, $\psi(M)$ is concave on $\mathcal{M}$ and $M(p^*)$ is differentiable point of $\psi(.)$, then $M(p^*)$ maximises $\psi(.)$ on $\mathcal{M}$ iff

$$F_\psi\{M(p^*), \underline{v}_j\underline{v}_j^T\} = 0 \quad \text{when } p_j^* > 0 \tag{2.22}$$

$$F_\psi\{M(p^*), \underline{v}_j\underline{v}_j^T\} \leq 0 \quad \text{when } p_j^* = 0 \tag{2.23}$$

Corollary (ii)

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

$$\frac{\partial \phi}{\partial p_j^*} = \sum_{i=1}^{J} p_i^* \frac{\partial \phi}{\partial p_i^*} \qquad \text{when } p_j^* > 0 \tag{2.24}$$

$$\frac{\partial \phi}{\partial p_j^*} \leq \sum_{i=1}^{J} p_i^* \frac{\partial \phi}{\partial p_i^*} \qquad \text{when } p_j^* = 0 \tag{2.25}$$

These follow from theorem 2.6.2 and equation (2.12).

Corollary (iii)

p^* solves $\min_{p \in \mathcal{P}} \max_{y \in \mathcal{CH}\{\mathcal{G}(\mathcal{V})\}} [F_\psi\{x(p), y\}]$.

Proof: From (DP7) $\max_{y \in \mathcal{CH}\{\mathcal{G}(\mathcal{V})\}} [F_\psi\{x, y\}] = \max_{1 \leq j \leq J} [F_\psi\{x, G(\underline{v}_j)\}]$.

Clearly $\max_{1 \leq j \leq J} [F_\psi\{x(p^*), G(\underline{v}_j)\}] = 0$.

Consequently $\max_{y \in \mathcal{CH}\{\mathcal{G}(\mathcal{V})\}} [F_\psi\{x(p^*), y\}] = 0$.

From (DP8) $\max_{y \in \mathcal{CH}\{\mathcal{G}(\mathcal{V})\}} [F_\psi\{x(p), y\}] \geq 0 \quad \forall p$.

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

Theorem 2.6.2 was in fact derived by Whittle (1973) but with only a general optimum design problem in mind. So also did Kiefer (1974) though using the Gâteaux

derivative. Wu (1976) derived it by appealing to the Kuhn-Tucker theorem in a more general setting than the design problem. The latter is admittedly a standard result in constrained optimisation, but it is not one that is conventionally stated in terms of directional derivatives.

In the D-optimum version of $\psi(.)$, Corollary (iii) establishes the equivalence of D-optimality and G-optimality. This follows since,

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

Hence corollary (iii) implies that p^* solves $\min_{p\in\mathcal{P}} \max_{1\le j\le J}\{\underline{v}_j^T M^{-1}(p)\underline{v}_j\}$, which is the G-optimal criterion mentioned in section 1.4. Kiefer and Wolfowitz (1960) derived this result directly thereby providing theorem 2.6.2 for D-optimality as well, and this was the first appearance of the theorem.

Other authors too have derived the general theorem but using Lagrangian theory and duality. Sibson (1974) and Silvey and Titterington (1974) have respectively established dual problems and corresponding duality theorems for D_A-optimality and for a general design criterion.

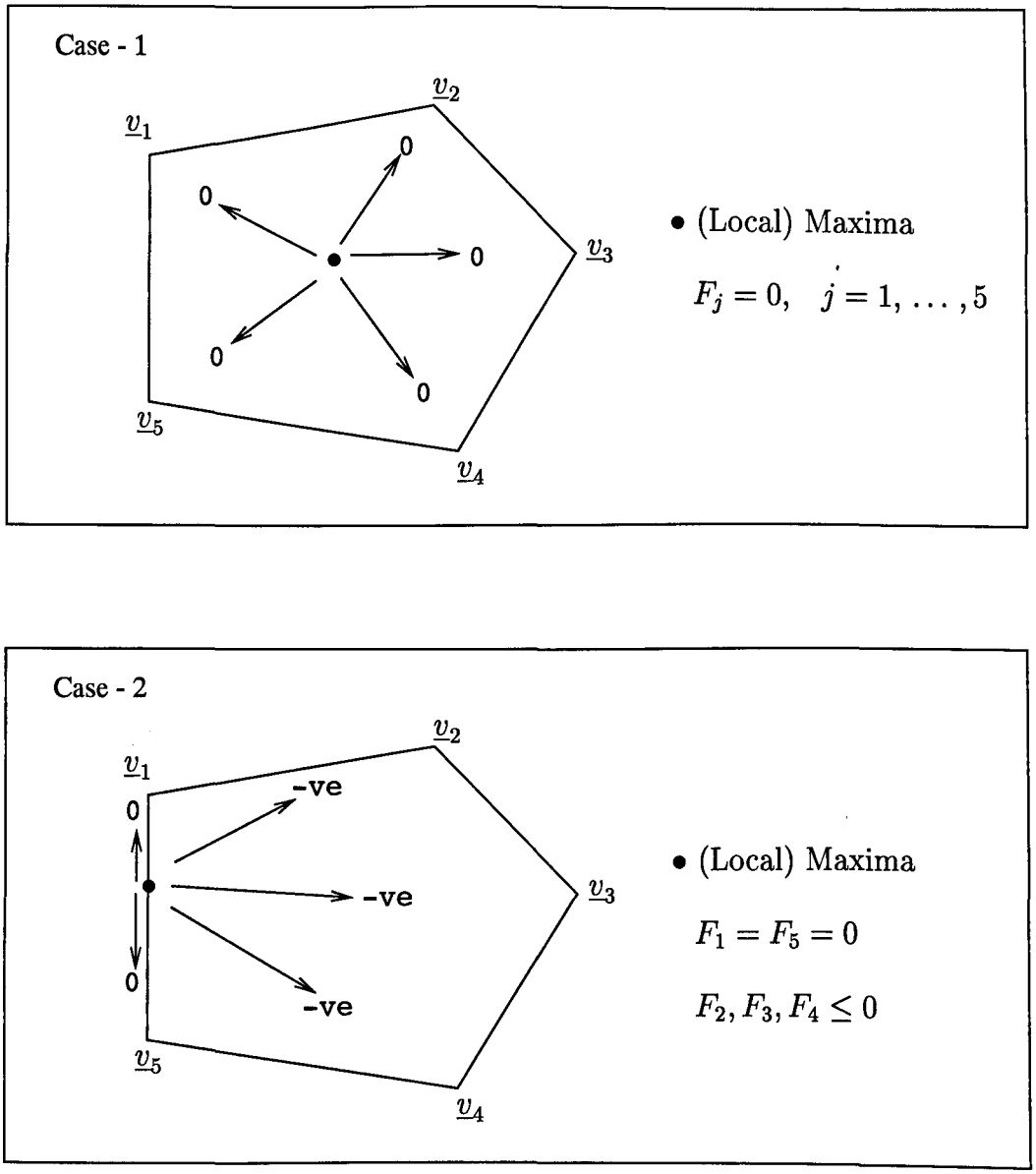


Figure 2.2. Geometrical Representation of Optimality Conditions.

Chapter 3

Algorithms

3.1 Introduction

An analytic solution of the problem of constructing optimal designs is possible only in simple cases. It is typically not possible to evaluate an explicit solution p^* to problems (P1) and (P2) or in particular to derive an optimal regression design explicitly. Iterative techniques must be employed and so special algorithms have been devised for a constrained optimisation problem (particularly for the design problem) which requires the calculation of an optimising probability distribution.

In chapter 1 we have introduced why there is a need for special algorithms. Now we know that there always exists an optimal measure with finite support (section 1.3). We wish to identify an optimising p^*. Typically p^* will be on the boundary of $\mathcal{CH}\{\mathcal{G}(\mathcal{V})\}$. Certainly this will be the case if $\mathcal{V}$ is a discritisation 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.

3.2 A Class of Algorithms

Problems (P1) and (P2) have a distinctive set of constraints, namely the variables $p_1, p_2, \ldots, p_J$ must be nonnegative and sum to 1. An iteration which neatly submits to these and has some suitable properties is the multiplicative algorithm:

$$p_j^{(r+1)} = p_j^{(r)} f(x_j^{(r)}) \Big/ \sum_{i=1}^{J} p_i^{(r)} f(x_i^{(r)}) \tag{3.1}$$

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

$$\begin{aligned} d_j^{(r)} &= \left.\frac{\partial \phi}{\partial p_j}\right|_{p=p^{(r)}} \\ F_j^{(r)} &= d_j^{(r)} - \sum_{i}^{J} p_i^{(r)} d_i^{(r)} \quad \text{[a directional derivative at } p = p^{(r)}\text{]}, \end{aligned}$$

while $f(x)$ 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.

Thus in view of the conditions for (local) optimality, a solution to problem (P1) is a fixed point of the iteration but so also are the solutions to problem (P2) for

any subset of $\mathcal{V}$.

This type of iteration was first proposed by Torsney (1977), taking $x = d$, $f(d) = d^{\delta}$, with $\delta > 0$. This, of course, requires derivatives to be positive. Subsequent 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 $f(d) = e^{\delta d}$ in a variety of applications, for which one criterion $\phi(.)$ could have negative derivatives; Torsney and Alahmadi (1992) consider other choices of $f(.)$; Mandal and Torsney (2000a) explore systematic choices of $f(.)$; see also chapters 4 and 5.

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 and also for estimating maximum likelihood estimates in an estimation problem.

3.3 Properties of the Iteration (3.1)

Under the conditions imposed on $f(.)$, iterations under (3.1) possess the following properties considered by Torsney (1988), Torsney and Alahmadi (1992) and Mandal and Torsney (2000a).

3.3.1 General Properties

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

(b) $F_\phi\{p^{(r)}, p^{(r+1)}\} \geq 0$ 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

$$F_\phi\{p^{(r)}, p^{(r+1)}\} = Cov\Big[D, f(D)\Big] \Big/ E\Big[f(D)\Big]. \tag{3.2}$$

Proof:

$$
\begin{aligned}
F_\phi\{p^{(r)}, p^{(r+1)}\} &= \left[p^{(r+1)} - p^{(r)}\right]^T \underline{d} \\
&= \sum_{i=1}^{J} \left[p_i^{(r+1)} - p_i^{(r)}\right] d_i \\
&= \sum_{i=1}^{J} p_i^{(r+1)} d_i - \sum_{i=1}^{J} p_i^{(r)} d_i \\
&= \frac{\sum_{i=1}^{J} p_i f(d_i)\, d_i}{\sum_{i=1}^{J} p_i f(d_i)} - \sum_{i=1}^{J} p_i\, d_i \qquad (3.3) \\
&= \frac{\left[\sum_{i=1}^{J} p_i f(d_i)\, d_i\right] - \left[\sum_{i=1}^{J} p_i\, d_i\right]\left[\sum_{i=1}^{J} p_i f(d_i)\right]}{\sum_{i=1}^{J} p_i f(d_i)} \qquad (3.4) \\
&= \frac{Cov\left[D, f(D)\right]}{E\left[f(D)\right]}
\end{aligned}
$$

The argument then is that 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 $p^{(r+1)}$.

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

(d) An iterate $p^{(r)}$ is a fixed point of the iteration if the derivatives $\partial\phi/\partial p_j^{(r)}$ corresponding to nonzero $p_j^{(r)}$ are all equal. 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 (P1) or (P2). Thus in view

of the conditions for (local) optimality, a solution to problem (P1) is a fixed point of the iteration but so also are the solutions to problem (P2) for any subset of $\mathcal{V}$, see Torsney (1988).

(e) We mentioned earlier that $f(.)$ may depend on one or more free parameters. Torsney and Alahmadi (1992) explore the following idea.

Suppose $f(.)$ is indexed by a free parameter δ, e.g. $f(x) = exp(\delta x)$ or x^{δ} $(x > 0)$, where $\delta > 0$.

Consider the case $x = d$ and let $h(\delta) = F_{\phi}\{p^{(r)}, p^{(r+1)}\}$. Then from (3.3)

$$h(\delta) = \frac{\sum_{i=1}^{J} p_i f(d_i)\, d_i}{\sum_{i=1}^{J} p_i f(d_i)} - \sum_{i=1}^{J} p_i\, d_i$$

and then

$$h'(\delta) = \frac{\left[\sum_{i=1}^{J} p_i f_i\right]\left[\sum_{i=1}^{J} p_i f_i' d_i\right] - \left[\sum_{i=1}^{J} p_i f_i d_i\right]\left[\sum_{i=1}^{J} p_i f_i'\right]}{\left[\sum_{i=1}^{J} p_i f_i\right]^2}$$

where $f_i = f_i(d)$ and $f_i' = \dfrac{\partial f_i(d)}{\partial \delta}$

$$= \frac{\left[\sum_{i=1}^{J} p_i f_i\right]\left[\sum_{i=1}^{J} p_i f_i d_i \frac{f_i'}{f_i}\right] - \left[\sum_{i=1}^{J} p_i f_i d_i\right]\left[\sum_{i=1}^{J} p_i f_i \frac{f_i'}{f_i}\right]}{\left[\sum_{i=1}^{J} p_i f_i\right]^2}$$

$$= \frac{\left[\sum_{i=1}^{J} p_i f_i d_i \frac{f_i'}{f_i}\right]}{\left[\sum_{i=1}^{J} p_i f_i\right]} - \frac{\left[\sum_{i=1}^{J} p_i f_i d_i\right]\left[\sum_{i=1}^{J} p_i f_i \frac{f_i'}{f_i}\right]}{\left[\sum_{i=1}^{J} p_i f_i\right]^2}$$

$$= \left[\sum_{i=1}^{J} q_i d_i \frac{f_i'}{f_i}\right] - \left[\sum_{i=1}^{J} q_i d_i\right]\left[\sum_{i=1}^{J} q_i \frac{f_i'}{f_i}\right]$$

where

$$q_i = \frac{p_i f(d_i)}{\sum_{j=1}^{J} p_j f(d_j)}.$$

Hence $h'(\delta) = Cov(D, G)$

where

$$G = \left[\frac{\partial f(D)}{\partial \delta}\right] \Big/ f(D)$$

$$= \frac{\partial ln[f(D)]}{\partial \delta} \tag{3.5}$$

and D is a random variable taking the value d_i with probability q_i.

3.3.2 Properties of some specific cases

Case I

We consider two choices of the function $f(d)$, namely $f(d) = d^{\delta}$ and $f(d) = exp(\delta d)$. These share the follwoing two properties.

(i) If there is a unique maximum derivative at $p^{(r)}$, say $d_m = \left.\frac{\partial \phi}{\partial p_m}\right|_{p=p^{(r)}}$ then for the function $f(d) = exp(\delta d)$:

$$p^{(r+1)} \longrightarrow e_m \quad as \quad \delta \longrightarrow \infty \tag{3.6}$$

where e_m is the m^{th} unit vector.

Proof:

$$\lim_{\delta\to\infty} p_j^{(r+1)} = \lim_{\delta\to\infty} \frac{p_j^{(r)} exp(\delta d_j)}{\sum_{i=1}^{J} p_i^{(r)} exp(\delta d_i)}$$

$$= \lim_{\delta\to\infty} \frac{p_j^{(r)} exp(\delta d_j)}{exp(\delta d_m) \sum_{i=1}^{J} p_i^{(r)} \left[\frac{exp(\delta d_i)}{exp(\delta d_m)}\right]}$$

$$= \lim_{\delta\to\infty} \frac{p_j^{(r)} \left[\frac{exp(\delta d_j)}{exp(\delta d_m)}\right]}{\sum_{i=1}^{J} p_i^{(r)} \left[\frac{exp(\delta d_i)}{exp(\delta d_m)}\right]}$$

$$= \lim_{\delta\to\infty} \frac{p_j^{(r)} \left[\frac{exp(d_j)}{exp(d_m)}\right]^{\delta}}{\sum_{i=1}^{J} p_i^{(r)} \left[\frac{exp(d_i)}{exp(d_m)}\right]^{\delta}} = \begin{cases} 1 & \text{if } m = j \\ 0 & \text{if } m \neq j \end{cases}$$

Hence $p^{(r+1)} \longrightarrow e_m$ as $\delta \longrightarrow \infty$.

Similarly the same is true for the function $f(d) = d^{\delta}$.

(ii) Recall the function $h(\delta) = F_{\phi}\{p^{(r)}, p^{(r+1)}\}$ in section (3.3.1 e). $h(\delta)$ is nondecreasing in δ. Note that the function G is given by $G(D) = ln(D)$ and $G(D) = D$ for the two above functions respectively. Both are increasing functions and therefore

$$h'(\delta) = Cov(D, G) \geq 0.$$

In the optimal design context the vector e_m corresponds to a single point design. For some optimal design criteria $\phi(e_m) = -\infty$. The implication is that for such

criteria iteration (3.1) in unlikely to be monotonic and possibly not convergent if δ is quite large. In fact, non-convergence occurs under the following cases.

$$\phi(p) = \prod_{i=1}^{J} p_i, \quad f(d) = d^{\delta}, \ \delta = 2;$$

$$\phi(p) = \sum_{i=1}^{J} p^{-t}, \quad f(d) = d^{\delta}, \ \delta = \tfrac{2}{t+1};$$

$$\phi(p) = \sum_{i=1}^{J} p_i \, ln(p_i), \quad f(d) = exp(\delta d), \ \delta = 2.$$

In each of the above combinations iterations oscillate between two values unless the initial value is the optimising p^*, which is $p^* = 1/J$ for each $\phi(p)$. We have given a proof of the first case in chapter 7.

In contrast the optimum is attained in one step from any initial $p^{(0)}$ if $\delta = 1$, $\frac{1}{t+1}$, 1 respectively in the above three cases. An implication seems to be that iteration (3.1) would be convergent if not monotonic at least for $\delta \leq 1$, $\delta \leq \frac{1}{t+1}$, $\delta \leq 1$ in the three cases repectively. We recall that, for large δ, property (b) in section (3.3.1) only guarantees an increase in the criterion if we take a small enough step from $p^{(r)}$ in the direction of $p^{(r+1)}$. This means a different formula from (3.1) for the next iterate. If we adopt such a method, property (3.6) suggests taking $\delta = \infty$. Then the revised iterative rule will be a vertex direction one but not a steepest ascent method since $F_{\phi}\{p, q\}$ depends on the distance between p and q. Constrained steepest ascent techniques choose directions which maximise normalised directional derivatives.

Case II

Here we consider two other choices of $f(d)$, namely $f(d) = ln(e + \delta d)$ and $f(d) = F(\delta d)$ where $F(x)$ is increasing in x and bounded above so that it must have asymptote as $x \longrightarrow \infty$. Examples include cumulative distribution functions (c.d.f.).

In these two choices:

(i) $p^{(r+1)} \longrightarrow p^{(r)}$ as $\delta \longrightarrow \infty$;

(ii) $h(\delta)$ $\left[= F_\phi\{p^{(r)}, p^{(r+1)}\}\right]$ is maximised by some finite δ, say δ^*.

The first one is trivial. Since $F_\phi\{p, p\} = 0$, it implies that $h(\infty) = h(0) = 0$.

Given that $h(\delta) \geq 0$ from (3.3.1 b), property (ii) follows.

It will be a possibility then that convergence, if not monotonicity are obtained for any δ. An optimal choice might be the δ^* of (ii). In general, there is no explicit formula for δ^* in terms of $p^{(r)}$ and $d^{(r)}$ $[= \frac{\partial \phi}{\partial p}|_{p=p^{(r)}}]$, but one can suggest an approximation to it in the case of $f(d) = F(\delta d)$. Note that $h'(\delta)$ in section (3.3.1 e) is a covariance between random variables D and G where $G = \frac{\partial ln[f(D)]}{\partial \delta}$. Therefore, if δ is such that G has a turning point in the range of $d_1, d_2, \ldots, d_J$, $h'(\delta)$ is likely to be zero.

The partial derivative of the random variable G with respect to d is

$$\frac{\partial G(d)}{\partial d} = \frac{\partial[\{\partial f(d)/\partial \delta\}/f(d)]}{\partial d} = \frac{\partial^2 ln[f(d)]}{\partial d\, \partial \delta}. \tag{3.7}$$

For $f(d) = F(\delta d)$,

$$\frac{\partial f(d)}{\partial \delta} = d\,F'(\delta d) \text{ and } \frac{\partial f(d)}{\partial \delta} \Big/ f(d) = dF'(\delta d)/f(d). \tag{3.8}$$

Thus from (3.8) the derivatives (3.7) will be of the form:

$$\frac{\partial G(d)}{\partial d} = \frac{F(\delta d)\,[\delta d\,F''(\delta d)] - [\delta d\,\{F'(\delta d)\}^2]}{[F(\delta d)]^2} \tag{3.9}$$

$$= \frac{F'(\delta d)}{F(\delta d)} + \frac{\delta d\,F''(\delta d)}{F(\delta d)} - \frac{\delta d\,[F'(\delta d)]^2}{[F(\delta d)]^2} \tag{3.10}$$

Let $x = \delta d$ and $H(x) = \frac{\partial G(d)}{\partial d}$. Thus (3.10) would be

$$H(x) = \frac{F'(x)}{F(x)} + \frac{xF''(x)}{F(x)} - \frac{x[F'(x)]^2}{[F(x)]^2}. \tag{3.11}$$

Let $H(x^*) = 0$. A simplistic suggestion is to approximate δ^* by $\delta^* = \frac{x^*}{\sum_{i=1}^{J} p_i d_i}$ or by corresponding terms based on other moments of the d_i's.

Other iterations for problems like (P2) have been proposed. Vertex direction algorithms which perturb one p_j and change the others proportionately were first proposed by Fedorov (1972) and Wynn (1972). These are useful when many of the p_j are zero at the optimum as happens in design problems. At the other extreme, when all p_j are positive at the optimum or when it has been established

which are positive, constrained steepest ascent or Newton type iterations may be appropriate. See Wu (1978) and Atwood (1976, 1980) on these respectively. Molchanov and Zuyev (2000) consider steepest descent algorithms based on the gradient function. Torsney (1983) suggested that iteration (3.1) might be useful in a context intermediate to these, when only a few optimal weights might be zero.

Certainly some modification would be needed if there are many zero optimal weights. We explore one such modification based on a 'clustering approach' in chapter 7. This is related to the fact that the support points of a discretised design space can be viewed as consisting of some 'clusters' of points. These clusters begin to emerge in early iteartions of algorithm (3.1). At this point the current set of weights are transformed to weights within clusters and total cluster weights. Optimal values of these are then sought using a modified version of algorithm (3.1). We explore this idea through several regression models and enjoy improved convergence.

Chapter 4

An Estimation Problem

4.1 Introduction

As we mentioned earlier there are many problems in statistics which demand the calculation of one or more optimising distributions or measures and hence are examples of the general problems (P1, P2 and P3) considered in chapter 2.

We cite some examples of problem (P1).

(E1) Possibly the simplest example is that of finding the maximum likelihood estimators of the probability parameters of a multinomial likelihood:

$$\phi(p) \;=\; c(x)\, p_1^{x_1} p_2^{x_2} \ldots p_J^{x_J} \tag{4.1}$$

The optimum choice of p_j is $p_j^* = x_j/x$, $x = \sum_j x_j$.

(E2) A second example is that of estimating the mixing parameters (probabilities) of a mixture distribution given data $\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_n$. The simplest example of this will arise when the component probability models $f_j(\underline{y})$ of the mixture are themselves free of any unknown parameters, and then the likelihood function is

$$\phi(p) = \prod_{i=1}^{n} \left\{ \sum_{j=1}^{J} p_j f_j(\underline{y}_i) \right\}. \tag{4.2}$$

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

(E3) A third example arises in the field of paired comparisons. Suppose J treatments $T_1, T_2, \ldots, T_J$ are compared on a pairwise basis, n_{ij} comparisons being made on treatments T_i and T_j, $i < j$. Assume that in any single comparison of T_i and T_j there is a probability π_{ij} that T_i will be preferred to T_j $(i \neq j)$, the same for all such pairwise comparisons, with

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

Let O_{ij} denote the observed number of times that T_i is preferred to T_j $(i \neq j)$ and assume that there are no ties so that, for $i < j$,

$$O_{ij} + O_{ji} = n_{ij}.$$

Assuming also independence between each pairwise comparison, the likelihood for the data is given by

$$L_O(\pi) = \prod_{i<j} (\pi_{ij})^{O_{ij}} (\pi_{ji})^{O_{ji}}. \tag{4.3}$$

Many models suggest that π_{ij} is of the form

$$\pi_{ij} = \frac{p_i}{(p_i + p_j)}, \quad p_i > 0. \tag{4.4}$$

See Bradley and Terry (1952), Davidson (1969), Bradley (1965).

However, this relationship only defines the p_i's relative to each other, for it will follow that

$$\pi_{ij} = \frac{cp_i}{(cp_i + cp_j)}. \tag{4.5}$$

In order to find a particular set of p_i's corresponding to the maximum likelihood estimator of π_{ij}, a restriction must be imposed on $\sum p_i$, and $\sum p_i = 1$ is a natural choice. Finding the corresponding estimates of the p_i's requires solution of problem (P1) with

$$\phi(p) = \frac{\prod_{i=1}^{J} p_i^{O_i}}{\prod_{i<j} (p_i + p_j)^{n_{ij}}}, \quad O_i = \sum_{\substack{j=1 \\ j \neq i}}^{J} O_{ij}. \tag{4.6}$$

There is almost certainly a unique optimising p^* in the case of examples (E1) to (E3) [certainly in the case of example (E1)], otherwise the parameters will be inestimable. Furthermore p^* certainly does not lie on the boundary of $\mathcal{P}$ of problem (P1) in the case of examples (E1) and (E2) and it is unlikely to do so in example (E3) assuming $n > J$, where $n = \sum_i \sum_j n_{ij}$.

In such a case we effectively have a simpler constrained optimisation problem, a problem having one active constraint, the simple linear equality $\sum_{i=1}^{J} p_i = 1$.

We consider now some of the properties of examples (E1) to (E3) that will be seen to be relevant later.

(i) All three functions are homogeneous. This is not all that surprising since independence is a common assumption in the formulation of probability models.

Note that the equality $\sum_i p_i = 1$ is an informative constraint to impose on a function satisfying the homogeneity condition that $\phi(cp) = c^t\phi(p)$. Study $\phi(p)$ subject to $\sum_i p_i = 1$ and one has an informed picture of the general behaviour of $\phi(.)$ on the positive quadrant at least.

(ii) With the exception of example (E3) the functions have positive derivatives as is evident from the following respective expressions for $\frac{\partial\phi}{\partial p_j}$:

$$(E1) \ : \ \phi(p)\left[\frac{x_j}{p_j}\right]$$

$$(E2) \ : \ \phi(p)\left[\sum_i \frac{f_j(\underline{y}_i)}{\sum_s p_s f_s(\underline{y}_i)}\right]$$

$$(E3) \ : \ \phi(p)\left[\frac{O_j}{p_j} - \sum_{s\neq j} \frac{n_{js}}{(p_j+p_s)}\right]$$

In the case of example (E3) there will typically be both positive and negative derivatives when p is in the positive quadrant, because

$$\sum_j p_j \frac{\partial \phi}{\partial p_j} = 0,$$

a consequence of the fact that $\phi(p)$ is a homogeneous function of degree zero; in fact then $F_j = d_j$.

(iii) In some instances the functions are concave.

The latter property is nice in the fact that it guarantees the existence of a unique maximum while the first two properties, not important in themselves, prove basic ingredients in the formulation of an algorithm.

It is possibly not surprising that problem (P1) crops up in various forms in the statistical literature given that probabilities are not infrequently parameters of probability models. This is particularly so in the case of likelihoods for categorical data.

In addtion to the above, we now consider an estimation problem which in the first instance is an example of problem (P3) a seeming generalisation of problem (P1). 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, first considered by Torsney (1988). This estimation problem could be transformed to an example of problem (P2).

Given observed frequencies O_{ij}, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, n$, and assuming a single multinomial distribution conditional on $N = \sum_{i=1}^{n} \sum_{j=1}^{n} O_{ij}$, with cell

probabilities θ_{ij}, we wish to solve the following version of problem (P2).

Maximise $\psi(\theta) = \sum_{i=1}^{n}\sum_{j=1}^{n} O_{ij}\, ln(\theta_{ij})$ subject to

$$\begin{aligned} &\theta_{ij} \geq 0, \;\; i = 1, 2, \ldots, n, \; j = 1, 2, \ldots, n, \\ &\sum_{i=1}^{n}\sum_{j=1}^{n} \theta_{ij} = 1, \\ &\sum_{j=1}^{n} \theta_{rj} = \sum_{j=1}^{n} \theta_{jr} \;\text{ for }\; r = 1, 2, \ldots, n. \end{aligned} \tag{4.7}$$

Some simplification of the problem is possible in view of the fact that at the solution

$$\theta_{ii} \;=\; \frac{O_{ii}}{N}, \;\; i = 1, 2, \ldots, n,$$

and also that one of the linear constraints, e.g. that corresponding to $r = n$, can be removed since they are linearly dependent.

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

4.2 Maximum Likelihood Estimation - 3×3 case

It is convenient to introduce the following notation at this point. Let
$(y_1, y_2, y_3, y_4, y_5, y_6) = (O_{12}, O_{31}, O_{23}, O_{21}, O_{13}, O_{32})$ 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 $E_{ij} = N\theta_{ij}$,
$i = 1, 2, 3,\ j = 1, 2, 3$ and hence are expected frequencies.

Thus our simplified problem in terms of x_i's and y_i's is now

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

$$\begin{aligned} & x_t \geq 0,\ \ t = 1, 2, \ldots, 6 \\ & \sum_{t=1}^{6} x_t = b = (N - \sum_{i=1}^{3} O_{ii}), \\ & x_1 - x_2 - x_4 + x_5 = 0, \\ & -x_1 + x_3 + x_4 - x_6 = 0. \end{aligned} \tag{4.8}$$

A standardised version of this problem is given by the transformation
$z_t = x_t / b$. So

$$\psi \ = \ \psi(z) \ = \ \sum_t y_t \, ln(z_t) \ + \ \sum_t y_t \, ln(b), \tag{4.9}$$

where $\sum_t z_t = 1$.

So our problem is: Maximise $\psi(z) = \sum_t y_t \, ln(z_t)$ subject to

$$\begin{aligned} & z_t \geq 0,\ \ t = 1, 2, \ldots, 6, \\ & \sum_{t=1}^{6} z_t = 1, \\ & z_1 - z_2 - z_4 + z_5 = 0, \\ & -z_1 + z_3 + z_4 - z_6 = 0. \end{aligned} \tag{4.10}$$

Clearly,

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

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, whose vertices are, in linear programming terminology, basic feasible solutions. In fact, vertices are given by:

$$\begin{aligned} \underline{v}_1 &= (1/2, 0, 0, 1/2, 0, 0)^T \\ \underline{v}_2 &= (0, 1/2, 0, 0, 1/2, 0)^T \\ \underline{v}_3 &= (0, 0, 1/2, 0, 0, 1/2)^T \\ \underline{v}_4 &= (1/3, 1/3, 1/3, 0, 0, 0)^T \\ \underline{v}_5 &= (0, 0, 0, 1/3, 1/3, 1/3)^T \end{aligned}$$

Thus we wish to solve a version of problem (P2) for which $J = 5$.
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
$\phi(p) = \sum_{t=1}^{6} y_t \, ln\{\sum_{j=1}^{5} p_j (\underline{v}_j)_t\}$ where $(\underline{v}_j)_t = t^{th}$ element of $\underline{v}_j$.

Now we derive the partial derivatives $\underline{d} = \frac{\partial \phi}{\partial p}$. Let $V = (\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_5)$.

$$\begin{aligned}
\frac{\partial \phi}{\partial p_s} &= \sum_{t=1}^{6} \frac{y_t}{z_t} \frac{\partial z_t}{\partial p_s} \\
&= \sum_{t=1}^{6} \frac{y_t}{z_t} (\underline{v}_s)_t \\
&= \sum_{t=1}^{6} u_t V_{ts} \qquad \text{where } V_{ts} = (t,s)^{\text{th}} \text{ element of } V \text{ and } u_t = \frac{y_t}{z_t} \\
&= \underline{v}_s^T \underline{u} \qquad \text{where } \underline{u} = (u_1, u_2, \ldots, u_6)^T.
\end{aligned}$$

Thus

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

The derivatives d_j's are positive and the criterion $\phi(p)$ is concave.

The vertex directional derivatives are given by

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

Thus

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

A particular example of data for which the hypothesis of marginal homogeneity is of interest is given in Plackett (1974), p.77-79. A grading of the unaided

distance vision of each eye of 3242 men resulted in the following frequencies, namely $(O_{12}, O_{31}, O_{23}, O_{21}, O_{13}, O_{32}) = (230, 77, 87, 223, 62, 106), b = 785$.

Taking $x = d$ in various choices of $f(x)$ in algorithm (3.1), we record for n=1, 2, 3 the number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$, for $j = 1, 2, \ldots, J$ starting from $p_j^{(0)} = 1/J$, $j = 1, 2, \ldots, J$, where F_j are the vertex directional derivatives. The results are given in tables 4.1-4.5 for various choices of δ.

In table 4.6 the numbers of iterations are given for best choices of δ (i.e. achieving fastest convergence) for each of $f(.)$.

From the results it is clear that numbers of iterations depends on the choice of $f(.)$. Clearly the choices $f(d) = d^\delta$ and $f(d) = exp(\delta d)$ are better than the other three. We make some general comments about this at the end of this chapter.

δ	n = 1	n = 2	n = 3
0.5	5	10	24
1.0	2	4	11
1.5	2	4	7
1.6	2	4	6
1.7	2	4	6
2.0	3	5	7
2.5	5	9	15
3.0	6	14	269

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

δ	n = 1	n = 2	n = 3
0.5	5	11	24
1.0	3	6	11
1.4	2	4	7
1.5	2	3	6
1.6	3	4	6
2.0	3	7	9
2.5	17	27	39

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

δ	n = 1	n = 2	n = 3
0.1	32	76	167
0.5	9	21	48
0.8	9	20	43
0.9	8	19	43
1.0	9	19	44
1.5	13	28	65
2.0	24	57	135

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

δ	n = 1	n = 2	n = 3
0.5	18	42	92
1.0	11	27	60
1.5	9	22	49
2.0	9	20	45
2.5	8	19	42
3.0	8	18	41
4.0	8	17	39
5.0	7	17	39
6.0	7	17	39
7.0	7	17	39
8.0	7	17	39
9.0	7	17	39
9.1	8	17	39
9.2	8	17	40
10.0	8	18	40

Table 4.4. 3×3 - case : $f(d) = ln(e + \delta d)$: Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3$.

δ	n = 1	n = 2	n = 3
0.5	12	29	65
1.0	9	20	46
1.1	9	20	45
1.2	9	20	45
1.3	9	20	45
1.4	9	20	45
1.5	9	20	46
2.0	11	24	54
2.5	14	31	70
3.0	18	42	97

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

$f(d)$	δ	n = 1	n = 2	n = 3
$exp(\delta d)$	1.5	2	3	6
d^{δ}	1.6	2	4	6
$\Phi(\delta d)$	0.9	8	19	43
$ln(e + \delta d)$	9.0	8	17	39
$\frac{exp(\delta d)}{1+exp(\delta d)}$	1.2	9	20	45

Table 4.6. 3×3 - case : Number of iterations for best choices of δ

4.3 Maximum Likelihood Estimation - 4×4 case

As in the previous case considering only the independent constraints of the hypothesis of marginal homogenity, our problem is:

Maximise $\psi(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{n} O_{ij}\, ln(\theta_{ij})$ subject to

$$\begin{aligned} &\theta_{ij} \geq 0, \;\; i = 1, 2, \ldots, 4, \; j = 1, 2, \ldots, 4, \\ &\sum_{i=1}^{4} \sum_{j=1}^{4} \theta_{ij} = 1, \\ &\theta_{12} + \theta_{13} + \theta_{14} - \theta_{21} - \theta_{31} - \theta_{41} = 0, \\ &-\theta_{12} + \theta_{21} + \theta_{23} + \theta_{24} - \theta_{32} - \theta_{42} = 0, \\ &-\theta_{13} - \theta_{23} + \theta_{31} + \theta_{32} + \theta_{34} - \theta_{43} = 0. \end{aligned} \tag{4.11}$$

Let

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

and

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

where $E_{ij} = N\theta_{ij}$ are expected frequencies.

In view of the fact that at the solution $\theta_{ii} = O_{ii}/N$, $i = 1, 2, \ldots, 4$ and in terms of x_i's and y_i's, our simplified problem is

Maximise $\psi(x) = \sum_{t=1}^{12} y_t\, ln(x_t)$ subject to

$$\begin{aligned} &x_t \geq 0, \;\; t = 1, 2, \ldots, 12 \\ &\sum_{t=1}^{12} x_t = b = (N - \sum_{i=1}^{4} O_{ii}), \\ &x_1 - x_2 + x_5 - x_6 + x_9 - x_{10} = 0, \\ &x_1 - x_3 - x_6 + x_8 - x_{11} + x_{12} = 0, \\ &x_2 - x_4 - x_5 + x_7 - x_{11} + x_{12} = 0. \end{aligned} \tag{4.12}$$

If we transform x_t to z_t as $z_t = x_t/b$, $t = 1, 2, \ldots, 12$ then the simplified version of the problem is

Maximise $\psi(z) = \sum_t y_t \, ln(z_t)$ subject to

$$\begin{aligned} & z_t \geq 0, \quad t = 1, 2, \ldots, 12 \\ & \sum_{t=1}^{12} z_t = 1 \\ & z_1 - z_2 + z_5 - z_6 + z_9 - z_{10} = 0, \\ & z_1 - z_3 - z_6 + z_8 - z_{11} + z_{12} = 0, \\ & z_2 - z_4 - z_5 + z_7 - z_{11} + z_{12} = 0. \end{aligned} \tag{4.13}$$

Clearly,

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

where

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

This is a convex polygon, where vertices (i.e. the basic feasible solutions) are given by

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

Now we can solve a version of problem (P2) for which $J = 20$.
$\underline{z} = \sum_{j=1}^{20} p_j \underline{v}_j$ and $\phi(p) = \sum_{t=1}^{12} y_t \, ln\{\sum_{j=1}^{20} p_j (\underline{v}_j)_t\}$ 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 $\frac{\partial \phi}{\partial p_j} = \underline{v}_j^T \underline{u}$.
Thus $\underline{d} = V^T \underline{u}$, where $V = (\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_{20})$ and $\underline{u} = (u_1, u_2, \ldots, u_{12})$, $u_t = \frac{y_t}{z_t}$.

The criterion $\phi(p)$ is concave with positive derivatives d_j's.

We consider the same example of data as in 3×3 case, but as a 4×4 contingency table [Plackett (1974)]. The grading of the unaided distance vision of each eye of 3242 men resulted in the following frequencies, namely

$(O_{12}, O_{31}, O_{24}, O_{43}, O_{13}, O_{21}, O_{34}, O_{42}, O_{14}, O_{41}, O_{23}, O_{32}) = (112, 85, 35, 145, 27, 87, 116, 72, 43, 151, 34, 106$
$b = 1013.$

For the same choices of $f(d)$ (as in 3×3 case) in algorithm (3.1), we record for n=1, 2, 3 the number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$, for $j = 1, 2, \ldots, J$ starting from $p_j^{(0)} = 1/J$, $j = 1, 2, \ldots, J$, where F_j are the vertex directional derivatives. For various choices of δ, results are given in tables 4.7-4.11.

Similar to table 4.6 the numbers of iterations for δ achieving fastest convergence for each of $f(.)$ are given in table 4.12.

δ	n = 1	n = 2	n = 3
0.5	6	23	38
1.0	3	11	18
1.5	2	7	11
2.0	2	5	8
2.2	2	4	7
2.3	2	4	6
2.4	2	4	6
2.5	2	4	8

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

δ	n = 1	n = 2	n = 3
0.5	7	23	37
1.0	4	11	18
1.5	4	7	11
2.0	3	5	8
2.1	3	5	7
2.2	3	5	7
2.3	3	6	9
2.5	4	7	13
3.0	13	41	67

Table 4.8. 4×4 - case : $f(d) = exp(\delta d)$: Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3$.

δ	n = 1	n = 2	n = 3
0.5	13	45	75
0.7	11	40	67
0.8	11	39	66
0.9	11	40	67
1.0	12	41	69
1.5	18	59	101
2.0	36	116	201

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

δ	n = 1	n = 2	n = 3
0.5	24	86	144
1.0	16	56	94
1.5	13	46	78
2.0	12	42	70
2.5	11	40	66
3.0	11	38	64
4.0	10	37	62
5.0	10	36	61
6.0	10	36	61
7.0	10	37	62

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

δ	n = 1	n = 2	n = 3
0.5	17	60	101
1.0	12	43	72
1.1	12	42	70
1.2	12	42	70
1.3	12	42	70
1.4	12	42	71
1.5	12	43	72
2.0	15	50	85
2.5	19	64	110
3.0	26	87	150

Table 4.11. 4×4 - case : $f(d) = \frac{exp(\delta d)}{1+exp(\delta d)}$: Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3$.

$f(d)$	δ	n = 1	n = 2	n = 3
$exp(\delta d)$	2.1	3	5	7
d^{δ}	2.3	2	4	6
$\Phi(\delta d)$	0.8	11	39	66
$ln(e + \delta d)$	5.0	10	36	61
$\frac{exp(\delta d)}{1+exp(\delta d)}$	1.2	12	42	70

Table 4.12. 4×4 - case : Number of iterations for best choices of δ

4.4 Comments

(i) For all choices of $f(d)$ and δ, convergence was to the same solution, namely

3×3 - case :

$(E_{12}, E_{31}, E_{23}, E_{21}, E_{13}, E_{32}) =$
(227.33, 70.26, 97.43, 225.67, 68.56, 95.74).

4×4 - case :

$(E_{12}, E_{31}, E_{24}, E_{43}, E_{13}, E_{21}, E_{34}, E_{42}, E_{14}, E_{41}, E_{23}, E_{32}) =$
$(110.02, 83.10, 38.28, 144.30, 30.13, 97.60, 118.13, 73.68, 39.60, 151.73, 30.80, 95.61)$.

(ii) Each choice of $f(d)$ depends on a free parameter δ which should be positive. Clearly the value of δ is crucial in such a choice. The values of δ reported in tables 4.6 and 4.12 achieve fastest convergence.

(iii) Here d_j's are positive and two of these choices of $f(.)$ are defined for only positive d (d^δ, $ln(1+\delta d)$). The other three are defined for both positive and negative d.

(iv) Clearly $exp(\delta d)$ and d^δ are best in both cases. While $\Phi(\delta d)$, $ln(1+\delta d)$ and $\frac{exp(\delta d)}{1+exp(\delta d)}$ take many more iterations than $exp(\delta d)$ and d^δ.

(v) Here d_j's are positive and $\Phi(\delta d)$ and $\frac{exp(\delta d)}{1+exp(\delta d)}$ change slowly for high values, say values above 1, whereas they change more quickly at zero. Noting 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. In the 3×3-case with $f(F) = \Phi(\delta F)$ and $\delta = 0.9$ the number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3$ respectively are 3,

6, 14, whereas taking $f(d) = \Phi(\delta d)$ with same δ takes 8, 19, 43 iterations (from table 4.6) to achieve $\max_j\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3$ respectively.

Note that iterations under $f(d) = exp(\delta d)$ and under $f(F) = exp(\delta F)$ will be identical.

Chapter 5

Construction of Optimal Designs

5.1 Introduction

In the previous chapter we have seen that many examples of problems (P1) and (P2) can arise in the area of estimation. In this chapter we consider examples in constructing D-optimal designs.

These are constrained optimisation problems having the single linear equality constraint $\sum_j p_j = 1$ as well as $p_j \geq 0$, $j = 1, 2, \ldots, J$, the full constraint region being a closed bounded convex set.
Therefore we have an example of problem (P1) with

(i) $\phi(p) = log\{det(M(p))\}$;

(ii) $M(p) = \sum_{j=1}^{J} p_j \underline{v}_j \underline{v}_j^T$, the information matrix.

It is also an example of problem (P2) where

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

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

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

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

5.2 Examples

We consider D-optimal designs for five examples considered by Silvey et al (1978), Wu (1978). The first example originated in Wynn (1970). The examples are defined by their design spaces.

Example - 1: $\mathcal{V} = \mathcal{V}_1 = \left\{(1,-1,-1)^T, (1,-1,1)^T, (1,1,-1)^T, (1,2,2)^T\right\}$

Example - 2: $\mathcal{V} = \mathcal{V}_2 = \left\{(1,-1,-1)^T, (1,-1,1)^T, (1,1,-1)^T, (1,2,3)^T\right\}$

Example - 3: $\mathcal{V} = \mathcal{V}_3 = \left\{(1,-1,-2)^T, (1,-1,1)^T, (1,1,-1)^T, (1,2,2)^T\right\}$

Example - 4: $\mathcal{V} = \mathcal{V}_4 = \Big\{(1,1,-1,-1)^T, (1,-1,1,-1)^T, (1,-1,-1,-1)^T,$
$(1,2,2,-1)^T, (1,1,-1,1)^T, (1,-1.5,1,1)^T,$
$(1,-1,-1,2)^T\Big\}$

Example - 5: $\mathcal{V} = \mathcal{V}_5 = \Big\{(1,1,-1,-1)^T, (1,-1,1,-1)^T, (1,-1,-1,-1)^T,$
$(1,2,2,-1)^T, (1,1,-1,1)^T, (1,-1.5,1,1)^T,$
$(1,-1,-1,2)^T, (1,1,1.5,1)^T\Big\}$

These correspond to linear models with a constant term since the first component of each vertex ($\underline{v}_j$) is always 1. Thus in examples 1-3 if $\underline{v}_j \in \mathcal{V}$ then $\underline{v}_j$ is of the form $(1, x_2, x_3)$.

In fact, we can assume the more realistic regression model

$$E(y) = \underline{v}^T\underline{\theta}, \qquad \underline{v} \in \mathcal{V} \tag{5.1}$$

where

$$\mathcal{V} = \left\{\underline{x} = (x_1, x_2, x_3)^T : x_1 = 1, (x_2, x_3) \in \mathcal{Q}\right\} \tag{5.2}$$

where $\mathcal{Q}$ is the quadrilateral with vertices

$$\begin{array}{ll} & (-1,-1), (-1,1), (1,-1), (2,2) \quad \text{for example-1,} \\ & (-1,-1), (-1,1), (1,-1), (2,3) \quad \text{for example-2,} \\ \text{and} & (-1,-2), (-1,1), (1,-1), (2,2) \quad \text{for example-3.} \end{array}$$

These could be viewed as the design space, but it is well established that observations should only be taken at the vertices.

Examples 4 and 5 have similar characterisations. Also

$$\mathcal{V}_5 = \mathcal{V}_4 \cup \{(1, 1, 1.5, 1)^T\}.$$

We now derive the optimal weights (p^*) for example 1. This derivation is simplified by employing symmetry arguments to justify $p_2^* = p_3^*$.

Let $p_1^* = q$, $p_2^* = p_3^* = r$. Then $p_4^* = 1 - q - 2r$.

Let

$$\begin{aligned} P &= diag(p_1^*, p_2^*, p_3^*, p_4^*) \\ &= diag(q, r, r, 1 - q - 2r) \end{aligned} \tag{5.3}$$

and

$$\begin{aligned} V &= [\underline{v}_1, \underline{v}_2, \underline{v}_3, \underline{v}_4] \\ &= \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 2 \\ -1 & 1 & -1 & 2 \end{bmatrix} \end{aligned} \tag{5.4}$$

Then the information matrix would be

$$\begin{aligned} M &= VPV^T \\ &= \begin{bmatrix} 1 & -3q - 4r + 2 & -3q - 4r + 2 \\ -3q - 4r + 2 & -3q - 6r + 4 & -3q - 10r + 4 \\ -3q - 4r + 2 & -3q - 10r + 4 & -3q - 6r + 4 \end{bmatrix} \end{aligned}$$

and its determinant is given by

$$det(M) = 72qr + 64r^2 - 72q^2r - 192qr^2 - 128r^3. \tag{5.5}$$

The above expression has two unknowns q and r. The maximum of the determinant is found by differentiation, giving the pair of derivatives

$$\frac{\partial[det\{M\}]}{\partial q} = 72r - 144qr - 192r^2$$

$$\frac{\partial[det\{M\}]}{\partial r} = 72q + 128r - 72q^2 - 384qr - 384r^2.$$

At the maximum both of the above derivatives are zero. Thus solution of the resulting equations yields

$$q = 0.125$$
$$r = 0.28125.$$

Hence the optimum solution p^* would be

$$\begin{aligned} p_1^* &= 0.125 \\ p_2^* = p_3^* &= 0.28125 \\ p_4^* &= 0.3125. \end{aligned} \tag{5.6}$$

Note that in chapter 1 while considering the properties of different design criteria we mentioned that D-optimality is a special case of the standardised criteria

$$\psi_t\{M(p)\} = -\left[\frac{1}{k} tr\left(M^{-t}(p)\right)\right]^{1/t}.$$

This is the case in view of the fact that

$$\lim_{t\to 0} \psi_t\{M(p)\} = -[det\{M^{-1}(p)\}]^{1/k} = -[det\{M(p)\}]^{-1/k}.$$

Thus we consider the standardised criterion

$$\phi(p) = \frac{1}{k} log\{det(M(p))\}, \quad k\text{=number of parameters.} \tag{5.7}$$

The criterion $[log\, det\{M(p)\}]$ is concave on the set of symmetric positive definite matrices and is invariant under non-singular linear transformation of $\mathcal{V}$.

We need to calculate the the partial derivatives $\frac{\partial \phi}{\partial p_j}$. They can be derived as

$$\begin{aligned} d_j = \frac{\partial \phi}{\partial p_j} &= \frac{1}{k} G\left\{M(p), \underline{v}_j \underline{v}_j^T\right\} \\ &= \frac{1}{k} tr\left\{\underline{v}_j \underline{v}_j^T M^{-1}(p)\right\} \\ &= \frac{1}{k} tr\left\{\underline{v}_j^T M^{-1}(p) \underline{v}_j\right\} \\ &= \frac{1}{k} \underline{v}_j^T M^{-1}(p) \underline{v}_j. \end{aligned}$$

where $G\left\{M(p), \underline{v}_j \underline{v}_j^T\right\}$ is the Gâteaux derivative of $[log\, det\{M(p)\}]$.

Note that $\sum_{j=1}^{J} p_j d_j = 1$.

We report the performance of algorithm (3.1) in calculating D-optimal designs when $f(.)$ satisfies the conditions of section (3.2).

We first consider three choices of $f(.)$, taking $x = d$: namely, $f(d) = ln(e + \delta d)$; $f(d) = \frac{exp(\delta d)}{1+exp(\delta d)}$ and $f(d) = a - exp(-\delta d)$, $a > 1$. Note that for a close to 1 the last choice of $f(d)$ is close to an exponential cumulative distribution function. These choices were also considered by Torsney and Alahmadi (1992). In contrast to these authors we consider taking the standardised criterion function (5.7).

In tables 5.1-5.5 we record for n=1, 2, 3, 4 the number of iterations needed to achieve $\max_{1\leq j\leq J}\{F_j\} \leq 10^{-n}$, where F_j are the vertex directional derivatives. In all the cases we take the initial design to be $p_j^{(0)} = 1/J$, $j = 1, 2, \ldots, J$.

Convergence of the algorithm can be slow as the results in tables 5.1-5.5 demonstrate. However arguably it is fast to begin with. If we compare the results for the five examples (corresponding to the best choices of δ, given in bold numbers[1]) for the three choices of the functions $f(.)$, $f(d) = \frac{exp(\delta d)}{1+exp(\delta d)}$ is not good whereas $f(d) = ln(e + \delta d)$ and $f(d) = a - exp(-\delta d)$ are not bad.

One marginally positive result is that convergence is faster under the case $f(d) = a - exp(-\delta d)$ with $a = 1.0001$. Convergence was slower for large values of a. Interestingly if $a \approx 1$ and δ is small then $f(d) \approx \delta d$. This suggests that $f(d) = d$ is an efficient choice for D-optimality criterion. Certainly it is known to be monotonic for this criterion.

[1]In all tables iteration counts for best choices of δ are given in bold font

Example - 1

$f(d) = ln(e + \delta d)$				
δ	n=1	n=2	n=3	n=4
2.0	3	17	44	74
3.0	3	15	40	67
4.0	2	15	38	64
5.0	**2**	**14**	**38**	**63**
6.0	2	14	38	63
7.0	2	14	38	63
7.8	2	14	38	63
7.9	2	14	38	64
8.0	2	14	38	64
$f(d) = \frac{exp(\delta d)}{1+exp(\delta d)}$				
δ	n=1	n=2	n=3	n=4
1.0	3	17	45	75
1.2	3	16	43	73
1.3	**3**	**16**	**43**	**72**
1.4	3	16	43	73
1.5	3	16	44	73
1.6	3	17	45	75
2.0	3	18	50	84
$f(d) = a - exp(-\delta d), a = 1.0001$				
δ	n=1	n=2	n=3	n=4
0.005	1	4	11	19
0.01	**1**	**4**	**11**	**18**
0.05	1	4	11	19
0.1	1	5	11	19
0.5	1	6	14	24
0.8	1	7	17	29
0.9	1	7	18	31
1.0	1	8	20	33
1.5	2	10	27	45

Table 5.1. Example-1:Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$.

Example - 2

$f(d) = ln(e + \delta d)$				
δ	n=1	n=2	n=3	n=4
3.0	3	26	69	123
4.0	3	25	66	118
4.5	3	25	65	117
5.0	**3**	**24**	**65**	**116**
6.0	3	24	65	116
7.0	3	24	65	116
7.5	3	24	65	116
8.0	3	25	65	117
$f(d) = \frac{exp(\delta d)}{1+exp(\delta d)}$				
δ	n=1	n=2	n=3	n=4
1.0	3	29	78	138
1.1	3	29	76	135
1.3	**3**	**28**	**74**	**132**
1.4	3	28	75	133
1.5	3	28	75	134
2.0	4	32	86	153
$f(d) = a - exp(-\delta d), a = 1.0001$				
δ	n=1	n=2	n=3	n=4
0.005	1	8	20	36
0.009	1	7	20	35
0.01	**1**	**7**	**20**	**35**
0.05	1	8	20	36
0.5	1	10	26	46
1.0	2	15	44	89
2.0	3	23	64	115

Table 5.2. Example-2:Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$.

Example - 3

$f(d) = ln(e + \delta d)$				
δ	n=1	n=2	n=3	n=4
4.0	2	10	36	57
4.5	2	10	36	57
4.8	2	10	35	56
5.0	**2**	**10**	**35**	**56**
6.0	2	10	35	56
7.0	2	10	35	56
7.5	2	10	35	56
7.7	2	10	35	56
8.0	2	10	36	57
$f(d) = \frac{exp(\delta d)}{1+exp(\delta d)}$				
δ	n=1	n=2	n=3	n=4
1.0	2	12	42	67
1.2	2	12	41	65
1.3	**2**	**12**	**40**	**64**
1.4	2	12	41	65
1.5	2	12	41	65
2.0	3	13	47	75
$f(d) = a - exp(-\delta d), a = 1.0001$				
δ	n=1	n=2	n=3	n=4
0.001	1	5	11	18
0.005	1	5	10	16
0.01	**1**	**4**	**10**	**16**
0.05	1	5	10	16
0.5	1	5	13	21
2.0	2	10	35	56

Table 5.3. Example-3:Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$.

Example - 4

$f(d) = ln(e + \delta d)$				
δ	n=1	n=2	n=3	n=4
4.0	5	54	192	481
5.0	5	54	190	474
6.0	**5**	**53**	**189**	**473**
6.5	5	54	190	474
7.0	5	54	190	474
$f(d) = \frac{exp(\delta d)}{1+exp(\delta d)}$				
δ	n=1	n=2	n=3	n=4
1.1	6	61	217	543
1.2	6	61	217	542
1.3	**6**	**61**	**216**	**540**
1.5	6	62	219	549
2.0	7	71	251	629
$f(d) = a - exp(-\delta d), a = 1.0001$				
δ	n=1	n=2	n=3	n=4
0.001	1	17	61	153
0.005	1	17	61	153
0.01	**1**	**17**	**60**	**151**
0.05	1	17	61	153
2.0	6	53	190	477

Table 5.4. Example-4:Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$.

Example - 5

$f(d) = ln(e + \delta d)$				
δ	n=1	n=2	n=3	n=4
4.0	5	75	299	645
5.0	5	74	295	636
6.0	**5**	**74**	**294**	**634**
7.0	5	74	295	636
8.0	5	75	297	640
$f(d) = \frac{exp(\delta d)}{1+exp(\delta d)}$				
δ	n=1	n=2	n=3	n=4
1.2	6	85	337	726
1.3	**6**	**84**	**336**	**724**
1.5	6	86	341	735
2.0	7	99	391	842
$f(d) = a - exp(-\delta d), a = 1.0001$				
δ	n=1	n=2	n=3	n=4
0.005	2	24	95	205
0.007	2	24	95	204
0.009	2	24	95	204
0.01	**2**	**24**	**95**	**204**
0.05	2	24	96	206
2.0	6	75	296	639

Table 5.5. Example-5:Number of iterations needed to achieve $max\{F_j\} \leq 10^{-n}$.

5.3 A variation in algorithm (3.1) : Replacing d_j by F_j

We now attempt to improve convergence by considering some choices of $f(.)$ for which we replace d_j by F_j. Some of the above choices of $f(.)$ could be bad because D-optimal derivatives are positive and 'centred' on 1 $\left[\sum_j p_j d_j = 1\right]$; namely the logistic cumulative distribution choice of $f(.)$ which can accept negative arguments. Similarly we noted in the previous chapter the choice of $f(d) = \Phi(\delta d)$ is bad because the d_j's are positive and 'centred' on 1. $\Phi(\delta d)$ slowly changes at 1 whereas it changes more quickly at zero. Torsney (1988) first considered such choices of $f(.)$ for criterion with negative derivatives d_j. In its conception algorithm (3.1) was originally evolved for standard optimal design criteria which have positive derivatives and $f(d) = d^\delta$ proved to be a natural choice for particular values of δ; in particular $\delta = 1$ for D-optimality and $\delta = 1/2$ for c-optimality yield monotonic iterations; see Titterington (1976) and Torsney (1983).

However, any criterion has both positive and negative vertex directional derivatives F_j $\left[\sum_j p_j F_j = 0\right]$. This motivates the idea of replacing d_j by F_j in algorithm (3.1) to yield

$$p_j^{(r+1)} = p_j^{(r)} f(F_j^{(r)}) \Big/ \sum_{i=1}^{J} p_i^{(r)} f(F_i^{(r)}) \tag{5.8}$$

where $F_j^{(r)}$ are directional derivatives at r^{th} iteration, and $f(.)$ satisfies the conditions of section 3.2.

Here we consider two choices of $f(x)$: $f(x) = \Phi(x)$, and $f(x) = \frac{exp(x)}{1+exp(x)}$, $x = \delta F$. Note that the function $f(x) = ln(e + \delta x)$ (considered in the previous section) cannot be considered since x can be negative.

In tables 5.6 - 5.10 we record for n=1, 2, 3, 4 the number of iterations needed to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$ for these two choices. The initial design is $p_j^{(0)} = 1/J$, $j = 1, 2, \ldots, J$.

Results in tables 5.6 - 5.10 clearly illustrate that convergence is improved considerably (for both of the choices $f(x) = \Phi(x)$ and $f(x) = \frac{exp(x)}{1+exp(x)}$) over all the choices of $f(.)$ in the previous section. For example, in example-2 with $f(x) = \frac{exp(x)}{1+exp(x)}$, $x = \delta F$ and $\delta = 3.0$ the number of iterations needed to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$ for $n = 1, 2, 3, 4$ respectively are 2, 4, 12, 22 (table 5.7) whereas with the same function $f(x)$ but taking $x = \delta d$ with $\delta = 1.3$ takes 3, 28, 74, 132 iterations (table 5.2).

Example - 1

$f(x) = \Phi(x)$, $x = \delta F$

δ	n=1	n=2	n=3	n=4
1.0	1	6	14	24
1.5	1	4	9	15
2.0	**1**	**3**	**7**	**10**
2.5	1	9	21	33

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

δ	n=1	n=2	n=3	n=4
2.0	1	4	11	18
2.5	1	3	8	14
3.0	**1**	**3**	**6**	**11**
3.5	1	5	9	13
4.0	1	9	23	35

Table 5.6. Example - 1 : Iteration counts : Replacing d_j by F_j.

Example - 2

$f(x) = \Phi(x),\ x = \delta F$				
δ	n=1	n=2	n=3	n=4
1.0	1	9	25	45
1.5	1	6	16	29
2.0	**2**	**4**	**12**	**21**
2.5	3	21	49	77
$f(x) = \frac{exp(x)}{1+exp(x)},\ x = \delta F$				
δ	n=1	n=2	n=3	n=4
2.0	1	7	20	35
2.5	1	6	15	27
3.0	**2**	**4**	**12**	**22**
4.0	3	21	55	81

Table 5.7. Example - 2 : Iteration counts : Replacing d_j by F_j.

Example - 3

$f(x) = \Phi(x),\ x = \delta F$				
δ	n=1	n=2	n=3	n=4
1.0	1	5	13	21
1.5	1	4	8	13
2.0	**1**	**3**	**6**	**8**
2.5	2	7	17	25
$f(x) = \frac{exp(x)}{1+exp(x)},\ x = \delta F$				
δ	n=1	n=2	n=3	n=4
2.0	1	4	10	16
2.5	1	3	8	12
3.0	**1**	**2**	**6**	**9**
4.0	2	7	17	27

Table 5.8. Example - 3 : Iteration counts : Replacing d_j by F_j.

Example - 4

$f(x) = \Phi(x),\ x = \delta F$				
δ	n=1	n=2	n=3	n=4
1.0	2	21	75	188
1.5	1	14	49	124
2.0	1	11	37	93
2.5	**1**	**10**	**29**	**74**
$f(x) = \frac{exp(x)}{1+exp(x)},\ x = \delta F$				
δ	n=1	n=2	n=3	n=4
2.0	1	17	60	149
3.0	1	11	39	99
3.5	1	10	33	84
4.0	**1**	**10**	**29**	**73**

Table 5.9. Example - 4 : Iteration counts : Replacing d_j by F_j.

Example - 5

$f(x) = \Phi(x),\ x = \delta F$				
δ	n=1	n=2	n=3	n=4
1.0	2	30	117	252
1.5	1	20	78	167
2.0	1	15	58	125
2.5	**1**	**12**	**46**	**100**
$f(x) = \frac{exp(x)}{1+exp(x)},\ x = \delta F$				
δ	n=1	n=2	n=3	n=4
2.0	2	24	93	201
3.0	1	16	62	133
3.5	1	14	53	114
4.0	**1**	**12**	**46**	**99**

Table 5.10. Example - 5 : Iteration counts : Replacing d_j by F_j.

5.4 A further variation

Consider functions $f(d)$ defined for only positive d. We cannot replace d_j by F_j since at least one F_j must be negative. However, noting that $F_j = d_j - c$ where $c = \sum_{i=1}^{J} p_i d_i$, we might improve convergence by replacing d_j by $x_j = (d_j - c)$ and choosing c to be a value in the range $(0, min\{d_j\})$.

One possible choice of c would be a lower limit on the d_j if such were known. One example of this arises in the case of a linear model with a constant term. Then the standardised D-optimal derivatives have a lower limit of $1/k$. In fact, the following is true.

$$d_j^{(D)} = d_j^{(D_s)} + 1 \tag{5.9}$$

where the $d_j^{(D)}$ are the raw partial derivatives of $log\{det(M(p))\}$ and the $d_j^{(D_s)}$ are the (nonnegative) partial derivatives of the D_s-criteria when interest is in all parameters excepting the constant term. See Silverman and Titterington (1980) and Torsney (1981).

Since $d_j^{(D_s)} \geq 0$, $d_j^{(D)} > 1$, and of course,

$$d_j^{(D_s)} = d_j^{(D)} - c, \quad c = 1. \tag{5.10}$$

As an alternative to this choice of c we consider $c = min\{d_j\}/2$ with $f(x) = ln(e + \delta x)$.

Iteration counts in tables 5.11 and 5.12 show that convergence is improved by replacing d_j by $(d_j - c)$, $c = min\{d_j\}/2$. For example, in example-1 with

$f(d) = (d - c)$, $c = min\{d_j\}/2$ and $\delta = 8.0$ the number of iterations needed to achieve $\max_{1 \le j \le J}\{F_j\} \le 10^{-n}$ for $n = 1, 2, 3, 4$ respectively are 2, 8, 19, 31 (table 5.11) whereas taking $f(d) = ln(e + \delta d)$ with $\delta = 5.0$ takes 2, 14, 38, 63 iterations (table 5.1). Similar results are true for the other examples as well.

Further improvement is possible through taking $x_j = (d_j - c)^\beta$ with a suitable choice of β. Iteration counts are reported in table 5.13.

Table 5.13 clearly illustrates that convergence is further improved by taking x_j as $(d_j - c)^\beta$ than by taking simply $(d_j - c)$, $c = min\{d_j\}/2$. In example-5 with $f(d) = (d - c)^\beta$, $c = min\{d_j\}/2$, $\delta = 10.0$, $\beta = 2.0$ the number of iterations needed to achieve $\max_{1 \le j \le J}\{F_j\} \le 10^{-n}$ for $n = 1, 2, 3, 4$ respectively are 1, 21, 82, 176 whereas taking $f(d) = (d - c)$, $c = min\{d_j\}/2$, $\delta = 10.0$ takes 3, 39, 152, 327 iterations. Similar results are true for the other examples as well (see table 5.13).

$f(x) = ln(e + \delta x)$, $x_j = d_j - c$, $c = min\{d_j\}/2$

Example - 1

δ	n=1	n=2	n=3	n=4
1.0	3	18	47	77
2.0	2	12	30	49
5.0	2	8	21	34
6.0	2	8	20	33
8.0	**2**	**8**	**19**	**31**

Example - 2

δ	n=1	n=2	n=3	n=4
2.0	2	20	52	91
3.0	2	17	42	75
4.0	2	15	38	67
5.0	2	14	36	63
6.0	2	14	35	61
7.0	2	13	34	60
8.0	2	13	33	59
9.0	**2**	**13**	**33**	**58**

Example - 3

δ	n=1	n=2	n=3	n=4
5.0	1	6	19	30
6.0	1	5	18	29
7.0	1	5	18	28
8.0	1	5	18	28
9.0	**1**	**5**	**17**	**27**
10.0	1	5	17	27
12.0	1	5	17	27
20.0	1	5	18	28

Table 5.11. Iteration counts for $f(x) = ln(e + \delta x)$, $x_j = d_j - c$, $c = min\{d_j\}/2$, examples 1-3.

$f(x) = ln(e + \delta x)$, $x_j = d_j - c$, $c = min\{d_j\}/2$

Example - 4				
δ	n=1	n=2	n=3	n=4
2.0	4	43	149	370
4.0	3	32	112	277
6.0	3	29	102	252
10.0	3	28	96	239
12.0	**3**	**28**	**96**	**238**
15.0	3	28	97	240
20.0	3	29	99	245
Example - 5				
δ	n=1	n=2	n=3	n=4
2.0	4	59	232	501
6.0	3	41	159	343
8.0	3	40	154	331
10.0	**3**	**39**	**152**	**327**
12.0	3	39	152	327
15.0	3	39	153	329

Table 5.12. Iteration counts for $f(x) = ln(e + \delta x)$, $x_j = d_j - c$, $c = min\{d_j\}/2$, examples 4 & 5.

$f(x) = ln(e + \delta x)$, $x_j = (d_j - c)^\beta$, $c = min\{d_j\}/2$

EXAMPLES	δ	β	n=1	n=2	n=3	n=4
Example-1	8.0	3.0	1	4	9	14
Example-2	9.0	2.0	1	7	18	31
Example-3	9.0	2.0	1	4	9	14
Example-4	12.0	3.0	1	11	40	101
Example-5	10.0	2.0	1	21	82	176

Table 5.13. Iteration counts for $f(x) = ln(e+\delta x), x_j = (d_j - c)^\beta, c = min\{d_j\}/2$, examples 1-5.

5.5 A Class of Functions based on a Distribution Function

We further attempt to improve convergence considering a class of functions based on a distribution function. Consider functions $f(.)$ defined for negative F. Since at the optimum

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

a desirable choice of $f(.)$ should be one which is changing at a reasonable rate at zero. Below is a class of such functions based on a distribution function $G(.)$.

$$f(F) = H_\beta(F) \begin{cases} = [1 + [G(\delta F)]^\beta - 2(1/2)^\beta]/2[1 - (1/2)^\beta] & \text{for } F > 0 \\ = [1 - [G(-\delta F)]^\beta]/2[1 - (1/2)^\beta] & \text{for } F < 0 \end{cases} \tag{5.11}$$

More generally,

$$f(F) = H_\beta(F) \begin{cases} = [1 + [G(\delta F)]^\beta - 2[G(0)]^\beta]/2[1 - [G(0)]^\beta] & \text{for } F > 0 \\ = [1 - [G(-\delta F)]^\beta]/2[1 - [G(0)]^\beta] & \text{for } F < 0 \end{cases} \tag{5.12}$$

Clearly $H_\beta(F)$ satisfies the conditions stated in section 3.2.

Figure 5.1 shows a plot $H_\beta(F)$ versus F where $G(\delta F) = \Phi(\delta F)$, the Normal cdf; $\delta = 2.0$ and $\beta = 1.0$. The plot shows that $H_\beta(F)$ changes quickly at zero.

We use the algorithm

$$p_j^{(r+1)} = p_j^{(r)} H_\beta(F_j^{(r)}) \Big/ \sum_{i=1}^{J} p_i^{(r)} H_\beta(F_i^{(r)}) \tag{5.13}$$

where $H_\beta(F)$ is given in (5.11).

Iteration counts using algorithm (5.13) similar to that of the previous tables are contained in tables 5.14 (for examples 1 and 2) and 5.15 (for examples 3, 4 and 5).

The results demonstrate that, convergence is improved by considering this class of functions $[H_\beta(F)]$ than that of considering only d_j (tables 5.1-5.5). For example, in example-5 with $f(F) = H_\beta(F)$, $G(\delta F) = \Phi(\delta F)$ and $\delta = 4.0$, $\beta = 3.0$ the number of iterations needed to achieve $\underset{1\leq j\leq J}{max}\{F_j\} \leq 10^{-n}$ for $n = 1, 2, 3, 4$ respectively are 1, 16, 64, 140 (table 5.15) whereas taking $f(d) = a - exp(-\delta d), a = 1.0001$ with δ =0.01 takes 2, 24, 95, 204 iterations (table 5.5).

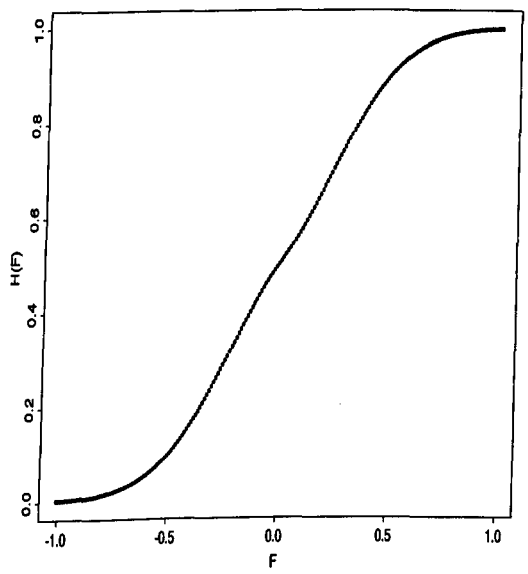


Figure 5.1. $H_\beta(F)$ versus F, where F is directional derivative. $H_\beta(F)$ is given in (5.11), where $G(\delta F) = \Phi(\delta F)$, the Normal cdf; $\delta = 2.0$ and $\beta = 1.0$.

$f(F) = H_\beta(F)$, $G(\delta F) = \Phi(\delta F)$

Example - 1

δ	β	n=1	n=2	n=3	n=4
3.0	2.0	1	5	7	12
2.0	2.0	1	4	10	16
2.0	3.0	1	6	15	26
3.0	**2.5**	**1**	**3**	**7**	**12**
3.0	3.0	1	3	9	16
2.5	2.5	1	3	9	16

Example - 2

δ	β	n=1	n=2	n=3	n=4
1.0	1.0	1	9	25	45
1.0	2.0	2	14	38	68
1.0	0.5	1	8	21	37
1.0	0.1	1	7	18	31
1.0	0.01	1	7	17	30
2.0	0.5	2	9	23	37
2.0	1.0	2	4	12	21
2.0	2.0	1	6	17	32
2.0	3.0	1	10	27	50
3.0	**2.0**	**2**	**6**	**10**	**20**
3.0	3.0	1	5	17	31
3.0	1.5	5	31	53	73

Table 5.14. Iteration counts for $H_\beta(F)$, $H_\beta(F)$ is given in (5.11), where $G(\delta F) = \Phi(\delta F)$, the Normal cdf; examples 1 & 2.

$f(F) = H_\beta(F)$, $G(\delta F) = \Phi(\delta F)$

Example - 3

δ	β	n=1	n=2	n=3	n=4
1.0	1.0	1	5	13	21
1.0	2.0	1	7	20	32
1.0	0.5	1	5	11	17
2.0	2.0	1	4	9	14
2.0	0.5	1	5	11	19
2.0	**1.0**	**1**	**3**	**6**	**8**
3.0	2.0	1	4	7	9
3.0	3.0	1	3	8	14

Example - 4

δ	β	n=1	n=2	n=3	n=4
1.0	1.0	2	21	75	188
2.0	2.0	2	17	68	147
2.5	**1.0**	**1**	**10**	**29**	**74**
2.5	2.0	1	12	43	110

Example - 5

δ	β	n=1	n=2	n=3	n=4
2.0	2.0	1	22	86	187
3.0	3.0	1	22	87	189
4.0	**3.0**	**1**	**16**	**64**	**140**
4.0	4.0	1	24	101	221

Table 5.15. Iteration counts for $H_\beta(F)$, $H_\beta(F)$ is given in (5.11), where $G(\delta F) = \Phi(\delta F)$, the Normal cdf; examples 3 - 5.

5.6 Optimal Weights

We now report, for the five examples, the values of the weights to which convergence was obtained for the choices of $f(.)$ and δ considered.

These are the optimal designs.

Example - 1: {0.125, 0.281, 0.281, 0.313}

Example - 2: {0.074, 0.291, 0.311, 0.324}

Example - 3: {0.243, 0.305, 0.161, 0.291}

Example - 4: {0.030, 0.012, 0.231, 0.234, 0.183, 0.208, 0.102}

Example - 5: {0.030, 0.012, 0.231, 0.234, 0.183, 0.208, 0.102, 0.000}

Note that the design space for example 5 added a design point to that of example 4, i.e., $\mathcal{V}_5 = \mathcal{V}_4 \cup \{(1, 1, 1.5, 1)^T\}$. The optimal weight (p_8) for this vertex $(\underline{v}_8)$ is zero and the other optimal weights are therefore the same as in example 4.

Also note that the optimal weights for example 1 above match with the solution (5.6) derived in section 5.2.

Chapter 6

Objective Choices of $f(.)$

6.1 Introduction

In the previous chapter we considered a variety of choices of the function $f(.)$ in iteration (3.1). These were to some extent arbitrary. Convergence rates varied although judicious choice of a free parameter could yield improvements. We also considered changing the argument of $f(.)$ from d_j to F_j. This yielded further improvements especially through the class of iteration (5.13).

We now consider more objective choices of $f(x)$ for both choices of x, $x = d_j$ and $x = F_j$. We mainly consider two approaches - approach I and approach II to improve convergence. In approach I, the criterion function $\phi(p)$ can have both positive and negative partial derivatives d_j, whereas in approach II it can have only positive d_j. We consider appropriate choices of the class of functions $f(x)$ for the two approaches.

6.2 Approach I

In this approach we replace d_j by F_j. Since there are always positive and negative F_j ($\sum_j p_j F_j = 0$) this requires a function $f(F)$ which is defined for positive and negative F, where we take F to represent a directional derivative. We note that properties (a) to (d) of (3.2) are still satisfied provided $f(F)$ is increasing in F. A suitable choice of $f(.)$ should be governed by the fact that at the optimum

$$F_j^* \begin{cases} = 0 & \text{for} \quad p_j^* > 0 \\ \leq 0 & \text{for} \quad p_j^* = 0 \end{cases} \tag{6.1}$$

This suggests that a suitable function is one that is 'centred' on zero and changes reasonably quickly about $F = 0$. It should also be desirable to treat positive and negative F_j's symmetrically, at least when all p_j^*'s are positive.

A class of functions $f(x)$ with the potential to satisfy these requirements has the following structure.

Suppose a function $h(x)$ is defined on $\mathbb{R}$ such that

$$h(0) = 1, \quad h(x) > 0 \quad \text{and} \quad h'(x) > 0. \tag{6.2}$$

Then take

$$f(x) = \begin{cases} h(x) & \text{for} \quad x < 0 \\ 2 - h(-x) & \text{for} \quad x > 0 \end{cases} \tag{6.3}$$

i.e. $f(x) = (1+s) - s\,h(-sx)$, $s = sign(x)$.

Clearly $f(x)$ is increasing, while for $y > 0$, $(y, f(y))$ and $(-y, f(-y))$ are reflections of each other in the point $(0,1) = (0, f(0))$, i.e., $f(-y) = 2 - f(y)$. Equivalently $f'(y)$ is symmetric about zero.

Note that $0 < f(x) < 2$, so that $f(x)$ is bounded. Also $f(0) = 1$.

Possible choices of $h(x)$ with $x = F$ and δ being a positive free parameter are

$$h(x) = \begin{cases} exp(\delta x) \\ 1/(1-\delta x) \\ (1-x)^{-\delta} \\ 2H(\delta x) \end{cases} \tag{6.4}$$

where $H(.)$ is a c.d.f. such that $H(0) = 1/2$.

Note that if $h(x)$ is bounded on $\mathbb{R}^+$ then a converse choice would be possible. In that case we need $h(x)$ to be

$$h^*(x) = 1 + \frac{h(x) - h(0)}{h(\infty) - h(0)}$$

so that $h^*(0) = 1$ and $h^*(\infty) = 2$.

Then

$$f(x) = \begin{cases} h^*(x) & \text{for } x > 0 \\ 2 - h^*(-x) & \text{for } x < 0. \end{cases}$$

Using the functions $h(x) = exp(\delta x)$, $h(x) = 1/(1-\delta x)$ and $h(x) = (1-x)^{-\delta}$ in (6.4) and considering $f(x)$ in (6.3) we calculate D-optimal designs for the 5 examples given in section (5.2).

We use the algorithms

$$p_j^{(r+1)} = \begin{cases} p_j^{(r)} h(F_j^{(r)}) \Big/ D & \text{for } F_j^{(r)} < 0 \\ p_j^{(r)} \left[2 - h(-F_j^{(r)})\right] \Big/ D & \text{for } F_j^{(r)} > 0. \end{cases} \tag{6.5}$$

where

$$D = \sum_{i=1}^{J} p_i^{(r)} f(F_i^{(r)}),$$

with $f(.)$ as in (6.3).

We record for $n = 1, 2, 3, 4,$ the number of iterations needed to achieve $\max_j \{F_j\} \leq 10^{-n}$, for $j = 1, 2, \ldots, J$. In all the cases we take the initial design to be $p_j^{(0)} = 1/J$, $j = 1, 2, \ldots, J$.

Comparing the results in tables 6.1, 6.2 and 6.3 to those in tables 5.1-5.5 shows that convergence (to the same solution as in section 5.6) is improved by considering approach I. The choices of $h(.)$ give almost same number of iterations within each eaxmple. Overall for all choices of $h(.)$ the best values of δ (written in bold font) seem to be somewhere between 1.5 and 2.0. Note values higher than those presented in the tables were investigated.

$h(x) = exp(\delta x), x = F$				
Example - 1				
δ	n=1	n=2	n=3	n=4
1.0	1	5	11	18
1.5	1	3	7	11
1.6	**1**	**3**	**6**	**10**
1.7	1	3	7	11
2.0	1	7	17	29
Example - 2				
δ	n=1	n=2	n=3	n=4
1.0	1	8	20	36
1.5	1	5	13	23
1.7	2	4	12	20
1.8	**2**	**5**	**12**	**20**
1.9	2	7	17	29
2.0	2	9	35	65
Example - 3				
δ	n=1	n=2	n=3	n=4
0.5	1	6	22	35
1.0	1	5	10	16
1.4	1	3	7	11
1.5	**1**	**3**	**6**	**10**
2.0	1	5	13	23
Example - 4				
δ	n=1	n=2	n=3	n=4
1.0	2	17	60	150
1.5	1	12	40	100
1.8	1	10	33	83
2.0	**1**	**8**	**30**	**74**
2.1	1	10	30	92
Example - 5				
δ	n=1	n=2	n=3	n=4
1.0	2	24	94	202
1.5	1	16	63	135
2.0	1	12	47	101
2.1	**1**	**12**	**45**	**96**

Table 6.1. Approach I : Iteration counts : $f(x) = (1+s)-sh(-sx)$, $s = sign(x)$, $h(x) = exp(\delta x), x = F$, examples 1-5

$h(x) = 1/(1-\delta x), x = F$				
Example - 1				
δ	n=1	n=2	n=3	n=4
1.0	1	5	12	19
1.5	1	3	7	12
1.6	1	3	7	11
1.7	**1**	**3**	**6**	**10**
2.0	1	5	13	25
Example - 2				
δ	n=1	n=2	n=3	n=4
1.0	1	8	21	36
1.5	1	6	14	23
1.6	2	5	13	22
1.7	**1**	**5**	**12**	**20**
2.0	2	7	25	53
Example - 3				
δ	n=1	n=2	n=3	n=4
1.0	1	5	11	16
1.5	1	3	7	10
1.6	1	3	6	9
1.7	**1**	**3**	**6**	**8**
1.8	1	2	6	11
2.0	1	4	11	21
Example - 4				
δ	n=1	n=2	n=3	n=4
1.0	2	18	61	151
1.5	1	12	41	100
2.0	**1**	**9**	**30**	**75**
2.1	1	9	29	73
Example - 5				
δ	n=1	n=2	n=3	n=4
1.0	2	25	95	204
1.5	1	17	64	136
2.0	1	13	48	102
2.1	**1**	**12**	**46**	**97**
2.2	1	12	44	101

Table 6.2. Approach I : Iteration counts : $f(x) = (1+s) - sh(-sx)$, $s = sign(x)$, $h(x) = 1/(1-\delta x), x = F$, examples 1-5

$h(x) = (1-x)^{-\delta}, x = F$				
Example - 1				
δ	n=1	n=2	n=3	n=4
1.0	1	5	12	19
1.5	1	3	7	12
1.6	1	3	6	11
1.7	**1**	**3**	**7**	**10**
1.6	1	3	8	13
2.0	1	5	15	27
Example - 2				
δ	n=1	n=2	n=3	n=4
1.0	1	8	21	36
1.5	1	5	13	23
1.7	1	4	12	20
1.8	**2**	**4**	**12**	**18**
1.9	2	6	15	27
2.0	2	7	31	59
Example - 3				
δ	n=1	n=2	n=3	n=4
1.0	1	5	11	16
1.5	1	3	6	10
1.7	**1**	**2**	**6**	**9**
1.8	1	3	7	11
2.0	1	5	13	21
Example - 4				
δ	n=1	n=2	n=3	n=4
1.0	2	18	61	151
1.5	1	12	40	100
2.0	**1**	**9**	**30**	**75**
2.1	1	8	28	76
Example - 5				
δ	n=1	n=2	n=3	n=4
1.0	2	25	95	204
1.5	1	17	63	136
2.0	1	13	48	102
2.1	**1**	**12**	**45**	**97**

Table 6.3. Approach I : Iteration counts : $f(x) = (1+s) - sh(-sx)$, $s = sign(x)$, $h(x) = (1-x)^{-\delta}, x = F$, examples 1-5

6.3 Approach II

This approach can only apply to criterion $\phi(p)$ with positive derivatives d_j. This is true for standard optimal design criteria. In choosing $f(d)$ we are guided by the fact that at the optimum

$$d_j^* \begin{cases} = \sum_i p_i^* d_i^* & \text{if} \quad p_j^* > 0 \\ \leq \sum_i p_i^* d_i^* & \text{if} \quad p_j^* = 0 \end{cases} \tag{6.6}$$

So at the optimum the derivatives corresponding to positive p_j share a common value, a value which equals the 'mean' $\sum p_i^* d_i^*$.

More generally, letting

$$M_t(\underline{d}^*) = \left[\sum_i p_i^* (d_i^*)^t\right]^{\frac{1}{t}}, \tag{6.7}$$

we must have, for $t \geq 0$,

$$d_j^* \begin{cases} = M_t(\underline{d}^*) & \text{if} \quad p_j^* > 0 \\ \leq M_t(\underline{d}^*) & \text{if} \quad p_j^* = 0 \end{cases} \tag{6.8}$$

or, letting

$$x_j^* = \left[\frac{d_j^*}{M_t(\underline{d}^*)}\right], \tag{6.9}$$

$$x_j^* \begin{cases} = 1 & \text{if} \quad p_j^* > 0 \\ \leq 1 & \text{if} \quad p_j^* = 0. \end{cases} \tag{6.10}$$

Now we consider choices of $f(d)$ in which we replace d_j by x_j. Again properties (a) to (d) of (3.2) are satisfied. A suitable choice of $f(d)$ is now one which changes reasonably quickly about $x = 1$ and which treats values of x equidistant from 1 equally.

A class of functions $f(x)$ with the potential to satisfy these requirements has the following structure.

Suppose a function $g(x)$ is defined on $\mathbb{R}^+ = \{x : x > 0\}$ such that

$$g(1) = 1, \quad g(x) > 0 \text{ and } g'(x) > 0. \tag{6.11}$$

Then take

$$f(x) = \begin{cases} g(x) & \text{for } 0 < x < 1 \\ 1/g(1/x) & \text{for } x > 1 \end{cases} \tag{6.12}$$

i.e. $f(x) = \{g(x^{-t})\}^{-t}$, $t = sign(x-1)$.

Clearly $f(x)$ is increasing, while for $0 < y < 1$, $f(1/y) = 1/f(y)$ or $f(y)f(1/y) = 1$.

Note that $g(0) \leq f(x) \leq 1/g(0)$, so that $f(x)$ is bounded if $g(0) > 0$. Also $f(1) = 1$.

Possible choices of x are

$$x = x_j = \begin{cases} d_j \\ \frac{d_j}{\{\sum_i p_i (d_i)^t\}^{1/t}} \\ \frac{d_j}{\sum_i p_i d_i} \\ \frac{d_j}{\prod_i d_i^{p_i}}. \end{cases} \tag{6.13}$$

With (6.11) and (6.12) in mind possible choices of $g(x)$ are

$$g(x) = \begin{cases} x^{\delta} \\ exp(\delta(x-1)) \\ ln(e+\delta(x-1)) \\ \frac{1+\delta}{1+\delta/x} \\ \frac{\Phi(\delta x)}{\Phi(\delta)} \end{cases} \tag{6.14}$$

δ being a free positive parameter.

Note that if $g(x) = x^{\delta}$ then $f(x) = x^{\delta}$ $\forall x$ and then if $x_j = \left[\frac{d_j}{M_t(\underline{d})}\right]$,

$$p_j^{(r+1)} = p_j^{(r)}(d_j^{(r)})^{\delta} \Big/ \sum p_i^{(r)}(d_i^{(r)})^{\delta}.$$

The other choices of $g(.)$ are attempts to identify other choices of $f(.)$ which mimic the choice $f(d) = d^{\delta}$.

Note that if $g(1) \neq 1$, a generalisation of (6.12) will be

$$f(x) = \begin{cases} g(x)/g(1) & \text{for } 0 < x < 1 \\ g(1)/g(1/x) & \text{for } x > 1. \end{cases} \tag{6.15}$$

Here also $f(x)$ is bounded as

$$\left[\frac{g(0)}{g(1)}\right] \leq f(x) \leq \left[\frac{g(1)}{g(0)}\right]$$

and $f(1) = 1$.

In applications, we consider

$$x = x_j = \left[\frac{d_j}{M_0(d)}\right], \text{ where } M_0(d) = \left[\prod_i d_i^{p_i}\right]. \tag{6.16}$$

Using the functions $g(x) = x^\delta$, $g(x) = exp(\delta(x-1))$ and $g(x) = ln(e + \delta(x-1))$ in (6.14) and considering $f(x)$ in (6.12) we calculate D-optimal designs for the 5 examples given in section (5.2).

We use the algorithms

$$p_j^{(r+1)} = \begin{cases} p_j^{(r)} g(x_j^{(r)}) \Big/ D & \text{for } 0 < x_j^{(r)} < 1 \\ \\ p_j^{(r)} \left[1/g(1/x_j^{(r)})\right] \Big/ D & \text{for } x_j^{(r)} > 1. \end{cases} \tag{6.17}$$

where

$$D = \sum_{i=1}^{J} p_i^{(r)} f(x_i^{(r)}),$$

with $f(.)$ as in (6.12).

Similar to the tables in the previous section (approach I) we record for $n = 1, 2, 3, 4$, the number of iterations in tables 6.4, 6.5 and 6.6 for the above three choices of $g(x)$ respectively. We take the initial weights to be equal, i.e., $p_j^{(0)} = 1/J$, $j = 1, 2, \ldots, J$.

Here also comparing the results in tables 6.4, 6.5 and 6.6 to those in tables 5.1-5.5 shows that convergence (to the same solution as in section 5.6) is improved by considering approach II. The choices of $g(.)$ give almost same number of iterations within each eaxmple. Note that values higher those presented in the tables were investigated.

$g(x) = x^{\delta}, x = d/M_t(d)$				
Example - 1				
δ	n=1	n=2	n=3	n=4
1.0	1	3	9	17
1.4	1	2	6	10
1.5	**1**	**2**	**5**	**9**
1.6	1	3	5	9
2.0	1	9	21	35
Example - 2				
δ	n=1	n=2	n=3	n=4
1.0	1	6	17	32
1.5	1	4	10	19
1.6	2	4	8	17
1.7	**2**	**5**	**10**	**16**
1.8	2	7	13	21
2.0	4	21	51	81
Example - 3				
δ	n=1	n=2	n=3	n=4
1.0	1	2	9	15
1.5	1	2	4	8
1.6	**1**	**2**	**5**	**7**
1.7	1	3	5	9
2.0	2	7	17	25
Example - 4				
δ	n=1	n=2	n=3	n=4
1.0	3	16	57	146
1.5	2	11	38	97
2.0	**1**	**7**	**28**	**72**
2.1	1	7	29	93
Example - 5				
δ	n=1	n=2	n=3	n=4
1.0	2	24	91	196
1.5	2	16	61	131
2.0	1	12	45	98
2.1	**1**	**11**	**43**	**93**

Table 6.4. Approach II : Iteration counts : $f(x) = \{g(x^{-t})\}^{-t}$, $t = sign(x-1)$, $g(x) = x^{\delta}, x = d/M_t(d)$, examples 1-5

$g(x) = exp(\delta(x-1)), x = d/M_t(d)$				
Example - 1				
δ	n=1	n=2	n=3	n=4
1.0	1	5	11	18
1.5	1	4	7	11
1.6	**1**	**4**	**7**	**10**
1.7	1	5	9	13
2.0	1	11	23	35
Example - 2				
δ	n=1	n=2	n=3	n=4
1.0	1	8	20	36
1.5	2	5	13	23
1.6	2	6	12	21
1.7	**2**	**6**	**12**	**20**
1.8	2	9	15	23
2.0	5	27	55	85
Example - 3				
δ	n=1	n=2	n=3	n=4
1.0	1	5	10	16
1.5	1	3	6	10
1.6	**1**	**3**	**6**	**9**
1.7	1	4	7	11
2.0	2	7	17	27
Example - 4				
δ	n=1	n=2	n=3	n=4
1.0	1	17	60	150
1.5	1	11	40	99
2.0	**3**	**13**	**30**	**74**
2.1	3	23	90	268
Example - 5				
δ	n=1	n=2	n=3	n=4
1.0	1	24	94	202
1.5	1	16	62	134
2.0	1	12	47	101
2.1	**1**	**12**	**44**	**95**

Table 6.5. Approach II : Iteration counts : $f(x) = \{g(x^{-t})\}^{-t}$, $t = sign(x-1)$, $g(x) = exp(\delta(x-1)), x = d/M_t(d)$, examples 1-5

$g(x) = ln(e + \delta(x-1)), x = d/M_t(d)$				
Example - 1				
δ	n=1	n=2	n=3	n=4
4.0	1	3	6	10
4.2	1	3	6	10
4.3	**1**	**3**	**6**	**10**
4.4	1	4	7	11
5.0	1	7	13	19
Example - 2				
δ	n=1	n=2	n=3	n=4
4.0	2	4	11	21
4.4	2	6	10	18
4.5	**3**	**7**	**11**	**16**
4.6	3	7	13	18
4.7	3	9	13	19
Example - 3				
δ	n=1	n=2	n=3	n=4
4.0	1	3	5	9
4.1	**1**	**3**	**5**	**8**
4.2	1	3	5	8
4.3	1	3	5	8
4.5	1	3	7	9
Example - 4				
δ	n=1	n=2	n=3	n=4
4.0	3	11	39	100
4.1	3	11	38	97
4.2	3	11	37	95
4.3	**7**	**10**	**37**	**93**
Example - 5				
δ	n=1	n=2	n=3	n=4
4.0	1	16	62	134
5.0	1	13	49	107
5.6	1	12	44	95
5.7	**1**	**12**	**43**	**93**

Table 6.6. Approach II : Iteration counts : $f(x) = \{g(x^{-t})\}^{-t}$, $t = sign(x-1)$, $g(x) = ln(e + \delta(x-1)), x = d/M_t(d)$, examples 1-5

Chapter 7

Construction of Optimal Designs using a Clustering Approach

7.1 Introduction

In the previous chapter we considered two approaches (approach I and approach II) for improving the convergence of algorithm (3.1). In this chapter we consider a more powerful improvement - a 'clustering approach'. We shall see that the support points of an optimum design on a design space $\mathcal{V}$ which is a discretisation of a continuous space consist of clusters, each cluster 'centred' on a support point of the design on the continuous space. We shall see also that corresponding clusters begin to emerge in early iterations of algorithm (3.1). Note that the finer is the discretisation the larger are the clusters.

This suggests the idea that at an appropriate iterate $p^{(r)}$, the single distribution $p^{(r)}$ should be 'replaced' by conditional distributions within clusters and a marginal distribution across clusters.

We first consider constructing D-optimal designs for some examples which we will later use to study the performance of this approach. These are more substantial problems than those so far considered. We consider five problems namely, trigonometric regression, quadratic regression, cubic regression, quartic regression and a second order model in two design variables.

In sections 7.2 to 7.4 we first report results of using the raw form of algorithm (3.1) with $x = d$, $f(x) = x^{\delta}$ but with one modification to aid convergence to the optimal design on the discretised design space, which will have many zero weights: namely, we ran the algorithm until $\underset{j}{max}\{F_j\} \leq 10^{-3}$ was achieved; thereafter we set weight $p_j^{(r)}$ to zero if $p_j^{(r)} < \varepsilon_1$ and $F_j^{(r)} < -\varepsilon_2$, where $\varepsilon_1 = 0.001$ and $\varepsilon_2 = 0.001$. Silvey, Titterington and Torsney (1978) used a similar idea.

7.2 Trigonometric Regression

First we consider the trigonometric regression model

$$E(y|\underline{v}_x) = \underline{v}_x^T\underline{\theta}, \tag{7.1}$$

where $\underline{v}_x = (x, x^2, sin2\pi x, cos2\pi x)^T$, $x \in [0,1]$ and $\underline{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)^T$.

$\underline{v}_x \in \mathcal{V} = \{\underline{v}_x : \underline{v}_x = (x, x^2, sin2\pi x, cos2\pi x)^T,\ 0 \leq x \leq 1\}$, the design space.

Hoel (1958), Wynn (1969), Torsney (1983) consider this model. We consider the design space approximated by a grid of 101 points equally spaced at intervals of 0.01 between 0 and 1.

So we are considering a design of the form

$$p = \left\{ \begin{array}{cccc} x_1 & x_2 & \dots & x_{101} \\ p_1 & p_2 & \dots & p_{101} \end{array} \right\}. \tag{7.2}$$

Consider first the use of algorithm (3.1) with $f(x) = x^\delta$, $x = d$. Results are reported in tables (7.1) and (A.1[1]). We record for $n = 1, 2, 3, 4$ the number of iterations needed to achieve $\max_j\{F_j\} \leq 10^{-n}$, for $j = 1, 2, \dots, J$ where F_j are the vertex directional derivatives. We take the initial design to be $p_j^{(0)} = 1/J$, $j = 1, 2, \dots, J$ (J=101).

As we approximated the design space at intervals of 0.01, the solution converges to

$$p = \left\{ \begin{array}{cccccccc} 0.08 & 0.09 & 0.37 & 0.38 & 0.39 & 0.73 & 0.74 & 1.00 \\ 0.21853 & 0.03147 & 0.00639 & 0.24117 & 0.00244 & 0.16808 & 0.08192 & 0.25 \end{array} \right\} \tag{7.3}$$

We note that in (7.3) the support points can be viewed as consisting of four 'clusters' of points - either a pair of neighbouring points or a single point. But this is the solution for the discretised design space. It suggests that the solution for the continuous design space is a 4-point design, with the 4 support points contained 'within' the clusters, and each point having the total design weight of its cluster. Atwood (1976) suggested that approximations to these support points are convex combinations of the relevant cluster members, convex weights being proportional to design weights. This yields the solution

[1]Because of the large volume of this chapter the detailed results for the iteration counts are given in appendix A

$$p^* = \left\{ \begin{array}{cccc} 0.081 & 0.380 & 0.733 & 1.00 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{array} \right\}. \tag{7.4}$$

From table (A.1) it is clear that convergence of algorithm (3.1) is very slow, especially after achieving $\max_{1 \le j \le J} \{F_j\} \le 10^{-n}$ for $n = 3$. Convergence is much slower for small values of δ compared to that of higher values. For example, for $\delta = 0.5$, the number of iterations needed to achieve $n = 4$ is 8793, whereas for $\delta = 1.9$, this number is 2314. This value of δ (1.9) turns out as the best value and the iteration numbers are written in bold font. Iterations are slow to attain the above conditions for δ higher than 1.9. For $\delta = 2.0$, oscillatory features develop and the above condition is satisfied for $n = 1$ at iteration 27, but there is no attainment for $n \ge 3$ in the first 4000 iterations.

This is consistent with results of Fellman (1989) for c-optimality. It is also consistent with the following phenomenon, noted in Torsney (1981, 1983).

In a design p with vertices $\underline{v}_j$ and the associated weights p_j, $j = 1, 2, \ldots, J$, the information matrix $M(p)$ is

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

where $V = [\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_J]$ and $P = \text{diag}(p_1, p_2, \ldots, p_J)$.

If $J = k$, the number of parameters, then V is a square matrix and the D-optimal

criterion is equivalent to

$$\begin{aligned}\phi(p) &= det\{M(p)\} &= [det\{V\}]^2[det\{P\}] \\ &&= [det\{V\}]^2 p_1 p_2 \ldots p_k.\end{aligned}$$

So the partial derivatives of $\phi(p)$ are

$$\frac{\partial \phi}{\partial p_t} = \frac{\phi}{p_t} = d_t. \tag{7.5}$$

Let

$$\begin{aligned}q_j &= \frac{p_j d_j^\delta}{\sum_i p_i d_i^\delta} \\ &= \frac{p_j^{1-\delta}}{\sum_i p_i^{1-\delta}} \qquad \text{[from (7.5)]}.\end{aligned}$$

so that $q_j = p_j^{(r+1)}$ if $p_j = p_j^{(r)}$ under iteration (3.1) with $f(x) = x^\delta$, $x = d$.

Thus for $\delta = 1$,

$$q_j = \frac{1}{k},$$

So this choice of iteration (3.1) attains the optimum in one step.

For $\delta = 2$, at the first iteration

$$q_j = \frac{[1/p_j]}{\sum_i [1/p_i]}.$$

At the second iteration let the weights be s_j. Then

$$\begin{aligned} s_j &= \frac{[1/q_j]}{\sum_i [1/q_i]} \\ &= p_j. \end{aligned}$$

This implies oscillatory iterations.

This suggests that we might usually want to keep $\delta \leq 2$.

Trigonometric Regression				
	Number of iterations needed to achieve $\max_{1 \leq j \leq J}\{F_j\} \leq 10^{-n}$ for			
δ	n=1	n=2	n=3	n=4
1.9	8	31	261	2314

Table 7.1. Iteration Results for the best choice of δ: $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$

7.3 Polynomial Regression in One Variable

In this section we find the D-optimal solution to some polynomial regression (in one variable) problems for which explicit solutions can be obtained.

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

$$E(y|\underline{v}_x) = \underline{v}_x^T\underline{\theta} \tag{7.6}$$

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

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

Here we have a standardised continuous design space. Fedorov (1972) reports that the discrete D-optimal design is unique, having a minimal support of k points which are the k roots of the polynomials $(1-x^2)P'_{k-1}(x)$ where $P_k(x)$ is the k^{th} Legendre polynomial

$$P_k(x) = \sum_{n=0}^{N} \left[\frac{(-1)^n\,(2k-2n)!\,x^{k-2n}}{2^k\,n!\,(k-n)!\,(k-2n)!}\right] \tag{7.7}$$

where

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

Since $Supp(p^*)$ contains k points the D-optimal design on it assigns weight $(1/k)$ to each of these.

In the cases of $k = 3$ (quadratic Regression), $k = 4$ (cubic regression) and $k = 5$ (quartic regression) (7.7) simplifies to

$$\begin{aligned} \text{Quadratic Regression}: \quad & (1-x^2)\,P_2'(x) &&= 3x\,(1-x^2), \\ \text{Cubic Regression}: \quad & (1-x^2)\,P_3'(x) &&= \tfrac{(15x^2-3)\,(1-x^2)}{2}, \\ \text{Quartic Regression}: \quad & (1-x^2)\,P_4'(x) &&= \tfrac{5x(7x^2-3)\,(1-x^2)}{2}, \end{aligned} \tag{7.8}$$

so the support points of p^* are given by

$$\begin{aligned} \text{Quadratic Regression}: \quad & x &&= 0,\ \pm 1, \\ \text{Cubic Regression}: \quad & x &&= \pm 1,\ \pm 1/\sqrt{5} = \pm 0.447, \\ \text{Quartic Regression}: \quad & x &&= 0,\ \pm 1,\ \pm\sqrt{3/7} = \pm 0.655. \end{aligned} \tag{7.9}$$

It will be of interest to compare numerically constructed designs with these analytic solutions.

Note that the set of orthogonal polynomials in (7.7) can also be obtained by the recurrence relation

$$(k+1)P_{k+1}(x) = (2k+1)xP_k(x) - kP_{k-1}(x) \tag{7.10}$$

with $P_0(x) = 1$ and $P_1(x) = x$.

Fedorov also reports unique solutions to the D-optimal design problems for polynomial regression when the constant variance assumption is replaced by $Var(y) = \sigma^2\lambda(x)$ where $\lambda(x)$ is of known form. Smith (1918) also considers designs for $\lambda(x)$ of the asymmetrical form $(1+ax)^2$ $(0 \leq a < 1)$ and of the

symmetrical form $(1+ax^2)^2$ $(a > -1)$.

However, we continue to find designs for constant variance. We first consider constructing D-optimal designs for quadratic regression, cubic regression and quartic regression using algorithm (3.1) i.e.

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

with $f(x) = x^\delta$, $x = d$ for various values of δ.

In each case we approximate the design interval by a grid of 201 points equally spaced at intervals of 0.01 between -1 and 1.

i.e., a design of the form:

$$p = \left\{ \begin{array}{cccc} x_1 & x_2 & \cdots & x_{201} \\ p_1 & p_2 & \cdots & p_{201} \end{array} \right\}. \tag{7.11}$$

Numerical information about the performance of the algorithm similar to that of tables 7.1 and A.1 is contained in tables A.2, A.3 and A.4 (summerised in tables 7.2, 7.3 and 7.4) for quadratic, cubic and quartic regressions respectively. Clearly convergence of algorithm (3.1) is slow for each of the examples. Convergence is slower for small values of δ compared to that of higher values. Here also oscillatory features occur for $\delta = 2.0$.

For the various choices of δ in each example the algorithm converged to the same design. These were as follows:

Quadratic Regression

$$p = \left\{ \begin{array}{ccccc} -1.0 & -0.01 & 0.00 & 0.01 & 1.00 \\ 0.33333 & 0.05922 & 0.21489 & 0.05922 & 0.33333 \end{array} \right\}$$

Cubic Regression

$$p = \left\{ \begin{array}{cccccccc} -1.00 & -0.46 & -0.45 & -0.44 & 0.44 & 0.45 & 0.46 & 1.00 \\ 0.25 & 0.007278 & 0.172138 & 0.070584 & 0.070584 & 0.172138 & 0.007278 & 0.25 \end{array} \right\}$$

Quartic Regression

$$p = \left\{ \begin{array}{ccccccccc} -1.00 & -0.66 & -0.65 & -0.01 & 0.00 & 0.01 & 0.65 & 0.66 & 1.00 \\ 0.20 & 0.09692 & 0.10305 & 0.01608 & 0.16790 & 0.01608 & 0.10305 & 0.09692 & 0.20 \end{array} \right\}$$

In each case the support points can be viewed as consisting of 'clusters' of points. In the case of quadratic regression there is one cluster centred on zero with three points while the others are the two end points. This suggests that the solution for the continuous design space is a 3-point design, with the 3 support points contained 'within' the clusters, and each point having the total design weight of its cluster. Similarly in the cases of cubic and quartic regressions the support points are viewed as consisting of four and five clusters of points respectively.

Taking convex combination of the relevant cluster members (convex weights being proportional to design weights) yields the solutions

Quadratic Regression

$$p^* = \left\{ \begin{array}{ccc} -1.0 & 0.0 & 1.0 \\ 0.33333 & 0.33333 & 0.33333 \end{array} \right\} \tag{7.12}$$

Cubic Regression

$$p^* = \left\{ \begin{array}{cccc} -1.00 & -0.447 & 0.447 & 1.00 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{array} \right\} \tag{7.13}$$

Quartic Regression

$$p^* = \left\{ \begin{array}{ccccc} -1.00 & -0.655 & 0.00 & 0.655 & 1.00 \\ 0.20 & 0.20 & 0.20 & 0.20 & 0.20 \end{array} \right\} \tag{7.14}$$

Clearly these are the optimal design solutions and match those of (7.9) obtained by using the Legendre polynomials of (7.7).

Quadratic Regression				
	Number of iterations needed to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$ for			
δ	n=1	n=2	n=3	n=4
1.9	7	27	262	2630

Table 7.2. Iteration Results for the best choice of δ: $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$

Cubic Regression				
	Number of iterations needed to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$ for			
δ	n=1	n=2	n=3	n=4
1.9	9	27	256	2127

Table 7.3. Iteration Results for the best choice of δ : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$

Quartic Regression				
	Number of iterations needed to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$ for			
δ	n=1	n=2	n=3	n=4
1.9	11	29	261	2571

Table 7.4. Iteration Results for the best choice of δ: $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$

7.4 Second-Order Model in Two Variables

Now we consider a regression model with more than one design variable.

The second-order polynomial in m design variables is

$$E(Y|\underline{x}) = \theta_o + \sum_{j=1}^{m} \theta_j x_j + \sum_{j=1}^{m-1} \sum_{k=j+1}^{m} \theta_{jk} x_j x_k + \sum_{j=1}^{m} \theta_{jj} x_j^2. \tag{7.15}$$

This is a $\frac{(m+1)(m+2)}{2}$ parameter model.

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

Thus the model is, with a revised parameterisation,

$$\begin{aligned} E(Y|x_1, x_2) &= \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 + \theta_4 x_1^2 + \theta_5 x_2^2 \\ &= \underline{v}_x^T \underline{\theta} \end{aligned} \tag{7.16}$$

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

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

We consider the discretised design space consisting of all pairs (x_1, x_2) arising when the values for each x_i, $i = 1, 2$ are those between -1 and +1 taken at steps of 0.1. That is, the space consists of $(21)^2 = 441$ pairs (x_1, x_2).

We write the design with arbitrary weights as:

$$p = \left\{ \begin{array}{cccc} (x_{11}, x_{21}) & (x_{12}, x_{22}) & \cdots & (x_{1\,441}, x_{2\,441}) \\ p_1 & p_2 & \cdots & p_{441} \end{array} \right\}$$

We consider the use of algorithm (3.1) with $f(x) = x^\delta$, $x = d$. Results are reported in table A.5 (summarised in table 7.5). We record the number of iterations needed to achieve $\max_{1 \le j \le J} \{F_j\} \le 10^{-n}$, for $n = 1, 2, 3, 4$. The initial design is $p_j^{(0)} = 1/J$, $j = 1, 2, \ldots, J$ with $J = 441$ as we approximated the design space consisting of 441 pairs of (x_1, x_2). Iteration counts show that convergence is slow especially for smaller values of δ. For example, for $\delta = 0.5$, the number of iterations needed to achieve the above condition at $n = 4$ is 1310.

The design converged to is

$$p^* = \left\{ \begin{array}{ccccccccc} -1.0 & -1.0 & -1.0 & 0 & 0 & 0 & 1.0 & 1.0 & 1.0 \\ -1.0 & 0 & 1.0 & -1.0 & 0 & 1.0 & -1.0 & 0 & 1.0 \\ 0.14579 & 0.08016 & 0.14579 & 0.08016 & 0.09620 & 0.08016 & 0.14579 & 0.08016 & 0.14579 \end{array} \right\} \quad (7.17)$$

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

Second-Order Model				
	Number of iterations needed to achieve $\max_{1 \le j \le J} \{F_j\} \le 10^{-n}$ for			
δ	n=1	n=2	n=3	n=4
2.3	6	44	171	286

Table 7.5. Iteration Results for the best choice of δ : $f(d) = d^\delta$, $d = \frac{\partial \phi}{\partial p}$

7.4.1 Derivation of the exact solution of the weights in the second-order model in two variables

We note the following analytic considerations about this optimal design.

Because of the symmetry of the above support it is intuitive that the weights at the four mid-points [(-1, 0), (0, 1), (1, 0), (0, -1)] of the sides should be equal and that weights at the four corners [(-1, -1), (-1, 1), (1, 1), (1, -1)] should be equal.

Let q, r be these common weights respectively and let s be the weight at the centre (0, 0).

Then obviously

$$4q + 4r + s = 1, \qquad q, r, s > 0. \tag{7.18}$$

The distribution defined on x_1 and x_2 is then

$x_1 \backslash x_2$	-1	0	1
-1	r	q	r
0	q	s	q
1	r	q	r

Reordering the vertex $\underline{v}_x$ in (7.16) to be

$$\underline{v}_x = (1,\ x_1^2,\ x_2^2,\ x_1,\ x_2,\ x_1x_2)^T,$$

the information matrix M is

$$M = \begin{bmatrix} 1 & E(x_1^2) & E(x_2^2) & E(x_1) & E(x_2) & E(x_1x_2) \\ E(x_1^2) & E(x_1^4) & E(x_1^2x_2^2) & E(x_1^3) & E(x_1^2x_2) & E(x_1^3x_2) \\ E(x_2^2) & E(x_1^2x_2^2) & E(x_2^4) & E(x_1x_2^2) & E(x_2^3) & E(x_1x_2^3) \\ E(x_1) & E(x_1^3) & E(x_1x_2^2) & E(x_1^2) & E(x_1x_2) & E(x_1^2x_2) \\ E(x_2) & E(x_1^2x_2) & E(x_2^3) & E(x_1x_2) & E(x_2^2) & E(x_1x_2^2) \\ E(x_1x_2) & E(x_1^3x_2) & E(x_1x_2^3) & E(x_1^2x_2) & E(x_1x_2^2) & E(x_1^2x_2^2) \end{bmatrix}$$

where expectations are with respect to the above distribution.

Because of its symmetry the following expectations are zero:

$$E(x_1) = E(x_2) = E(x_1x_2) = E(x_1^2x_2) = E(x_1x_2^2) = 0.$$

Also since the only nonzero values for each x_i are ± 1, we have

$$E(x_1^2) = E(x_2^2) = E(x_1^4) = E(x_2^4) = 4r + 2q.$$

Finally

$$E(x_1^2x_2^2) = 4r.$$

With these values the information matrix becomes

$$M = \begin{bmatrix} 1 & 4r+2q & 4r+2q & 0 & 0 & 0 \\ 4r+2q & 4r+2q & 4r & 0 & 0 & 0 \\ 4r+2q & 4r & 4r+2q & 0 & 0 & 0 \\ 0 & 0 & 0 & 4r+2q & 0 & 0 \\ 0 & 0 & 0 & 0 & 4r+2q & 0 \\ 0 & 0 & 0 & 0 & 0 & 4r \end{bmatrix}$$

The above matrix is of block-diagonal structure and its determinant is given by

$$det\{M\} = -16qr(q+2r)^2(4q^2+16r^2-q-4r+16qr).$$

The above expression has two unknowns q and r, which satisfy

$$(q+r) \leq \frac{1}{4}, \quad \text{since } s = 1-4q-4r \geq 0.$$

The maximum of the determinant of the information matrix M is found by differentiation. The Maple package yields the following derivatives

$$\frac{\partial[det\{M\}]}{\partial q} = -64[2qr(q+2r)+r(q+2r)^2](4q^2+16r^2-q-4r+16qr) \\ -64qr(q+2r)^2(8q+16r-1)$$

$$\frac{\partial[det\{M\}]}{\partial r} = -64[4qr(q+2r)+q(q+2r)^2](4q^2+16r^2-q-4r+16qr) \\ -256qr(q+2r)^2(4q+8r-1).$$

At the maximum both derivatives are zero. Solution of the resulting equations using Maple yields

$$\begin{aligned} q &= 0.08016 \\ r &= 0.14579. \end{aligned}$$

So the value of s is given by

$$s = 1 - 4q - 4r = 0.09620.$$

Hence the optimal weights with the values of x_1 and x_2 are given by

x_1	x_2	p
-1	-1	0.14579
-1	1	0.14579
1	1	0.14579
1	-1	0.14579
0	0	0.09620
-1	0	0.08016
0	1	0.08016
1	0	0.08016
1	-1	0.08016

Note that the above solution exactly matches with the solution in (7.17).

7.5 Clustering Approach

Inevitably convergence of the above type of algorithm is slow when there are many non-support points with zero weights. This happens in regression models with continuous design spaces as can be seen in examples of sections 7.2, 7.3 and 7.4. We have noted that the optimal support points can be viewed as consisting of a number of 'clusters' when we consider a discretised design space. In an attempt to improve convergence we consider a modification of the algorithm based on a clustering approach. This is motivated by the 'clusters' noted in the above examples. We introduce the idea first considering a general regression problem and then through the above examples.

7.5.1 General Problem

Consider a regression model with one continuous design variable x. Suppose that the design space is approximated by a regular grid of J points in the space and that the following represents an arbitrary design on this discretisation.

$$p = \left\{ \begin{array}{cccc} x_1 & x_2 & \dots & x_J \\ p_1 & p_2 & \dots & p_J \end{array} \right\}$$

With the design above consider running algorithm (3.1), i.e.

$$p_j^{(r+1)} = p_j^{(r)} f(d_j^{(r)}) \Big/ \sum_{i=1}^{J} p_i^{(r)} f(d_i^{(r)})$$

After running the above algorithm for a small number of iterations (with a suitable choice of $f(.)$) and then plotting weights or the variance function versus the

design points we argue that these plots depict a curve with a number of maximal and minimal turning points. For example, in trigonometric regression, after running algorithm (3.1) 5 times (with $f(x) = x^{\delta}$, $x = d$, $\delta = 1.7$) consider the plots of weights and variance function in figures 7.2 and 7.3 respectively. The plots clearly depict curves with three maximal, three minimal turning points and a final peak on the boundary at $x = 1$.

The minimal turning points clearly 'correspond' to regions of zero weight. The weights on the points in these regions are converging to zero. Assuming such zero weights we would be left with a number of disjoint clusters of points which must contain the optimal support points.

This clearly ties in with the clustering feature of optimal designs on discretised spaces noted in sections 7.2, 7.3. The design after 5 or 10 iterations has the same number of clusters as the final solution.

It is also of interest to note that the Wynn (1970) algorithm has a related feature. This algorithm 'adds' weight to 1 point subject to normalisation. For the trigonometric regression problem Torsney (1983) ran this algorithm from over 100 different initial supports each of 4 points selected from $[x : x = 0, 0.01, 0.02, \ldots, 0.99, 1]$. He observed that after on average 8 (=$2k$) iterations the point selected to receive added weight came from the following set.

$$\{V_{(x)} : x = 0.07, 0.08, 0.09, 0.37, 0.38, 0.39, 0.72, 0.73, 0.74, 0.75, 0.99, 1\}$$

This consists of 4 clusters each cluster being a cluster of the support of the optimal design expanded to include nearest neighbours.

At this point suppose there are m clusters and let the clusters of weights be $C_1, C_2, \ldots, C_m$. Let there be n_j weights in the j^{th} cluster (i.e. in cluster C_j). So

n_j is the j^{th} cluster size, $j = 1, 2, \ldots, m$.

Now consider transforming from the weights p_i to weights within clusters and total cluster weights. In fact, we first divide the p_i's into m clusters and then within each cluster we assign within cluster weights. This is shown in figure 7.1.

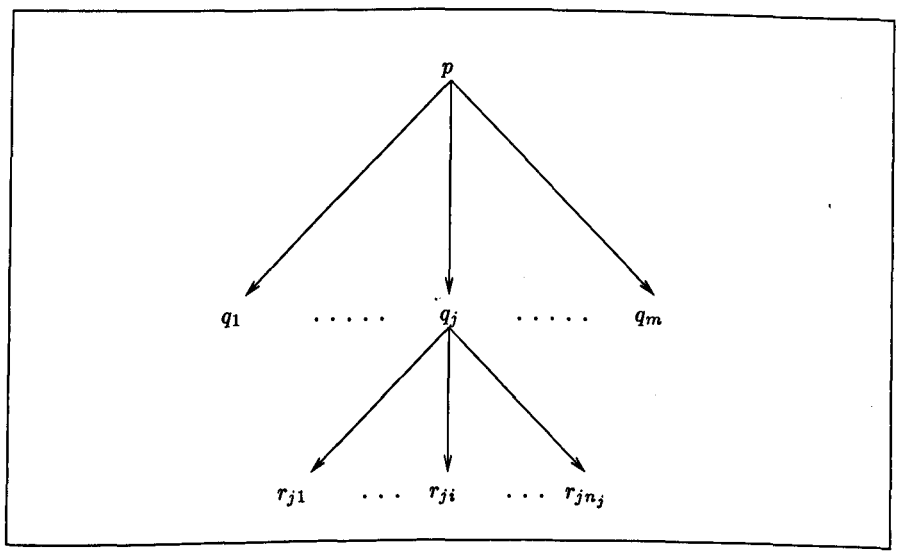


Figure 7.1. Showing the division of p_i's to total and within cluster weights

Let

$$q_j = \sum_{p_i \in C_j} p_i, \quad j = 1, 2, \ldots, m \tag{7.19}$$

$$r_{ji} = p_i/q_j, \quad p_i \in C_j, \quad j = 1, 2, \ldots, m, \quad i = 1, 2, \ldots, n_j. \tag{7.20}$$

Thus

q_j = j^{th} total cluster weight, $j = 1, 2, \ldots, m$

r_{ji} = i^{th} within cluster weight for C_j, $j = 1, 2, \ldots, m$, $i = 1, 2, \ldots, n_j$.

Hence

$$\begin{aligned} &\sum_{j=1}^{m} q_j = 1, \quad q_j \geq 0, \quad j = 1, 2, \ldots, m \\ &\sum_{i=1}^{n_j} r_{ji} = 1, \quad r_{ji} \geq 0, \quad i = 1, 2, \ldots, n_j, \; j = 1, 2, \ldots, m. \end{aligned} \tag{7.21}$$

So we have transformed the original weights p_i to q_j's, the total cluster weights and to r_{ji}'s, the within cluster weights.

Let

$$\begin{aligned} \underline{q} &= (q_1, q_2, \ldots, q_m)^T \\ \underline{r}_j &= (r_{j1}, r_{j2}, \ldots, r_{jn_j})^T, \; j = 1, 2, \ldots, m. \end{aligned} \tag{7.22}$$

Clearly $\underline{r}_j$ contains the within cluster weights for cluster C_j.

Thus the criterion $\phi(p)$ becomes a function of the total cluster weights and the within cluster weights. We write $\phi(p)$ as

$$\phi(p) = \phi_m(\underline{q}, \underline{r}_1, \underline{r}_2, \ldots, \underline{r}_m). \tag{7.23}$$

We need to choose $\underline{q}, \underline{r}_1, \underline{r}_2, \ldots, \underline{r}_m$ optimally.

Thus our problem can be written as
Maximise $\phi_m(\underline{q}, \underline{r}_1, \underline{r}_2, \ldots, \underline{r}_m)$ subject to

$$\begin{aligned} &\sum_{j=1}^{m} q_j = 1, \quad q_j \geq 0, \quad j = 1, 2, \ldots, m \\ &\sum_{i=1}^{n_j} r_{ji} = 1, \quad r_{ji} \geq 0, \quad i = 1, 2, \ldots, n_j, \; j = 1, 2, \ldots, m. \end{aligned} \tag{7.24}$$

Now we consider using appropriate algorithms for finding the optimal weights.

7.5.2 Algorithms

A natural extension of algorithm (3.1) is

$$\begin{aligned} \underline{q}^{(n+1)} &= \underline{q}^{(n)} f_q(d^{q^{(n)}}) \Big/ \sum_{j=1}^{m} q_j^{(n)} f_q(d_j^{q^{(n)}}) \\ \underline{r}_j^{(n+1)} &= \underline{r}_j^{(n)} f_{r_j}(d^{r_j^{(n)}}) \Big/ \sum_{i=1}^{n_j} r_{ji}^{(n)} f_{r_j}(d_i^{r_j^{(n)}}) \end{aligned} \tag{7.25}$$

where

$\underline{q}^{(n)}$ = total cluster weights at n^{th} iteration,
$\underline{r}_j^{(n)}$ = within cluster weights for C_j at n^{th} iteration, $j = 1, 2, \ldots, m$

$$\begin{aligned} d_j^q &= \frac{\partial \phi_m}{\partial q_j} \\ &= \sum_{i=1}^{J} \frac{\partial \phi_m}{\partial p_i} \frac{\partial p_i}{\partial q_j} \\ &= \sum_{i : p_i \in C_j} r_{ji} \frac{\partial \phi_m}{\partial p_i} \end{aligned}$$

$$\begin{aligned} d_i^{r_j} &= \frac{\partial \phi_m}{\partial r_{ji}} \\ &= \sum_{i=1}^{J} \frac{\partial \phi_m}{\partial p_t} \frac{\partial p_t}{\partial r_{ji}} \\ &= q_j \frac{\partial \phi_m}{\partial p_i} \end{aligned}$$

$f_q(.)$ and $f_{r_j}(.)$ are positive and strictly increasing and may depend on some free positive parameters δ_q and δ_{r_j} respectively.

Properties of algorithm (7.25)

Under the conditions imposed on the functions $f_l(.)$, $l = q, r_1, r_2, \ldots, r_m$, algorithm (7.25) has similar properties to algorithm (3.1); in particular the following:

(a) $\underline{q}^{(n)}, \underline{r}_1^{(n)}, \ldots, \underline{r}_m^{(n)}$ are always feasible.

(b) $F_{\phi_m}\{\underline{q}^{(n)}, \underline{r}_1^{(n)}, \ldots, \underline{r}_m^{(n)} \; ; \; \underline{q}^{(n+1)}, \underline{r}_1^{(n+1)}, \ldots, \underline{r}_m^{(n+1)}\} \geq 0.$

Proof.

$$F_{\phi_m}\{\underline{q}^{(n)}, \underline{r}_1^{(n)}, \ldots, \underline{r}_m^{(n)} \; ; \; \underline{q}^{(n+1)}, \underline{r}_1^{(n+1)}, \ldots, \underline{r}_m^{(n+1)}\}$$

$$= \left[(\underline{q}^{(n+1)} - \underline{q}^{(n)})^T : (\underline{r}_1^{(n+1)} - \underline{r}_1^{(n)})^T : \cdots : (\underline{r}_m^{(n+1)} - \underline{r}_m^{(n)})^T\right] \left[\underline{d}^{q^{(n)}} : \underline{d}^{r_1^{(n)}} : \cdots : \underline{d}^{r_m^{(n)}}\right]$$

$$= \left[(\underline{q}^{(n+1)} - \underline{q}^{(n)})^T \underline{d}^{q^{(n)}}\right] + \left[(\underline{r}_1^{(n+1)} - \underline{r}_1^{(n)})^T \underline{d}^{r_1^{(n)}}\right] + \cdots + \left[(\underline{r}_m^{(n+1)} - \underline{r}_m^{(n)})^T \underline{d}^{r_m^{(n)}}\right]$$

$$= \left[F_{\phi_m}^{(q)}\{\underline{q}^{(n)}, \underline{q}^{(n+1)}\}\right] + \left[F_{\phi_m}^{(r_1)}\{\underline{r}_1^{(n)}, \underline{r}_1^{(n+1)}\}\right] + \cdots + \left[F_{\phi_m}^{(r_m)}\{\underline{r}_m^{(n)}, \underline{r}_m^{(n+1)}\}\right]$$

where

$$F_{\phi_m}^{(q)}\{\underline{q}^{(n)}, \underline{q}^{(n+1)}\} \; = \; F_{\phi_m}\{\underline{q}^{(n)}, \underline{r}_1^{(n)}, \ldots, \underline{r}_m^{(n)} \; ; \; \underline{q}^{(n+1)}, \underline{r}_1^{(n)}, \ldots, \underline{r}_m^{(n)}\}.$$

and $F_{\phi_m}^{(r_j)}\{\underline{r}_j^{(n)}, \underline{r}_j^{(n+1)}\}$ is similarly defined.

These component directional derivatives would be the directional derivatives if we changed only one of the distributions $\underline{q}$ or $\underline{r}_1$ or ... or $\underline{r}_m$ keeping the others fixed using algorithm (3.1).

From property (3.3.1 b), such directional derivatives are nonnegative. Hence the result.

Equality will hold when the partial derivatives corresponding to nonzero weights of each distribution have a common value.

(c) An iterate $\{\underline{q}^{(n)}, \underline{r}_1^{(n)}, \ldots, \underline{r}_m^{(n)}\}$ is a fixed point of the iteration if partial derivatives corresponding to nonzero weights of each distribution are all equal; equivalently if the corresponding vertex directional derivatives are zero.

7.5.3 Applications

We apply this clustering approach to our earlier regression problems.

7.5.3.1 Trigonometric Regression

Consider the trigonometric regression problem. The model is given in (7.1). Consider starting algorithm (3.1) with equal weights of 1/101 on the grid of 101 points equally spaced at intervals of 0.01 between 0 and 1. i.e. the discretised design space is given by

$$\mathcal{V} = \{\underline{v}_x : \underline{v}_x = (x, x^2, sin2\pi x, cos2\pi x)^T, \ x = 0 : 1/0.01\} \tag{7.26}$$

and the design is of the form

$$p = \begin{Bmatrix} x_1 & x_2 & \ldots & x_{101} \\ p_1 & p_2 & \ldots & p_{101} \end{Bmatrix}. \tag{7.27}$$

Figure 7.2 is a plot of the weights versus the design points after running algorithm (3.1) 5 times (with $f(x) = x^\delta$, $x = d$, $\delta = 1.7$). The plot depicts a curve with three maximal, three minimal turning points and a final peak on the boundary at $x = 1$. The minimal turning points clearly 'correspond' to regions of zero weight. The weights on the points in these regions are converging to zero. Given such zero

weights we would be left with four disjoint clusters of points which must contain the optimal support points. We have no zero weights in the above picture but we can still perceive or define four clusters by defining boundaries between clusters to be at or near the three minimal turning points.

Figure 7.3 is a plot of the variance function $d(x, p)$ versus the design points at the same number of iterations above. $d(x, p)$ is given by the expression $\underline{v}^T(x)\, M^{-1}(p)\, \underline{v}(x)$.
Here the variance function[2] $d(x, p)$ is given by

$$\begin{aligned} d(x,p) &= 179.696x^4 - 268.028x^3 + (31.524\sin 2\pi x - 40.452\cos 2\pi x + 103.198)x^2 \\ &\quad -(21.276\sin 2\pi x - 28.766\cos 2\pi x)x - 2.004\sin 4\pi x + 4.034\sin^2 2\pi x \\ &\quad + 4.217\cos^2 2\pi x. \end{aligned} \tag{7.28}$$

This plot has a similar shape with same number of maximal and minimal points.

Now consider transforming from the weights p_i to weights within clusters and total cluster weights as considered in (7.19) and (7.20).

In this case the total and within cluster weights are given by

$$\begin{aligned} \underline{q} &= (q_1, q_2, q_3, q_4)^T \\ \underline{r}_j &= (r_{j1}, r_{j2}, \ldots, r_{jn_j})^T, \quad j = 1, 2, 3, 4. \end{aligned} \tag{7.29}$$

Hence

$$\begin{aligned} &\sum_{j=1}^{4} q_j = 1, \quad q_j \geq 0, \quad j = 1, 2, 3, 4 \\ &\sum_{i=1}^{n_j} r_{ji} = 1, \quad r_{ji} \geq 0, \quad i = 1, 2, \ldots, n_j, \; j = 1, 2, 3, 4. \end{aligned} \tag{7.30}$$

[2]The variance functions were obtained by using the package 'Maple'

So the criterion becomes

$$\phi(p) = \phi_m(\underline{q},\ \underline{r}_1,\ \underline{r}_2,\ \underline{r}_3,\ \underline{r}_4),\ \text{a function of } \underline{q}, \underline{r}_1, \underline{r}_2, \underline{r}_3 \text{ and } \underline{r}_4. \qquad (7.31)$$

We need to choose $\underline{q}$, $\underline{r}_1$, $\underline{r}_2$, $\underline{r}_3$, $\underline{r}_4$ optimally. To find these optimal values we maximise $\phi_m(\underline{q},\ \underline{r}_1,\ \underline{r}_2,\ \underline{r}_3,\ \underline{r}_4)$ subject to (7.30).

We use algorithm (7.25). Results are reported in table A.6. In the left half of this table we record for n=1, 2, 3 and 4 the number of iterations needed to achieve

$$\max_{1\le j\le m+J}\{F_j^{\underline{q},\underline{r}_1,\ldots,\underline{r}_m}\} \le 10^{-n},$$

where the term $\max_{1\le j\le m+J}\{F_j^{\underline{q},\underline{r}_1,\ldots,\underline{r}_m}\}$ denotes the maximum among all directional derivatives for total and within cluster weights. As we have m clusters and the number of within cluster weigths is same as the number of p_i (i.e. J), we will have $m + J$ such directional derivatives in total.

In the right half of the table we record for n=1, 2, 3 and 4 the number of iterations needed to achieve

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

where $\max_{1\le j\le J}\{F_j^{p}\} \le 10^{-n}$ denotes the maximum among all directional derivatives for p_j.

As we started after 5 iterations of algorithm (3.1), the actual numbers of iterations should be 5 plus the numbers given in table A.6.

In table 7.6 iteration counts are recorded for the best choice of δ. This proved to be $\delta = 60$. We investigated the values of δ higher than those presented in table A.6. Trial and error revealed that this gave fastest convergence.

Iterations converge to the design

$$p = \left\{ \begin{array}{cccccccc} 0.08 & 0.09 & 0.37 & 0.38 & 0.39 & 0.73 & 0.74 & 1.00 \\ 0.21857 & 0.03143 & 0.00618 & 0.24131 & 0.00251 & 0.16810 & 0.08190 & 0.25 \end{array} \right\} \quad (7.32)$$

Taking the convex combination of the relevant cluster members (convex weights being proportional to design weights) in (7.32) yields the solution

$$p^* = \left\{ \begin{array}{cccc} 0.081 & 0.380 & 0.733 & 1.00 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{array} \right\}, \quad (7.33)$$

which is exactly the same design we obtained in section 7.2.

Figure 7.4 is a plot showing the variance function $d(x, p^*)$ for the above D-optimal design, where $d(x, p^*)$ is given by

$$\begin{aligned} d(x,p) &= 116.427x^4 - 184.664x^3 + (13.150\sin 2\pi x - 28.032\cos 2\pi x + 76.181)x^2 \\ &\quad -(8.894\sin 2\pi x - 20.374\cos 2\pi x)x - 1.066\sin 4\pi x + 2.931\sin^2 2\pi x \\ &\quad + 3.715\cos^2 2\pi x. \end{aligned} \quad (7.34)$$

Clearly the plot has maximum value of 4 occuring at the 4 support points.

Tables 7.6 and A.6 clearly demonstrate that the convergence of algorithm (7.25) is much faster than that of algorithm (3.1). For example, in table 7.1, for $\delta = 1.9$ and $n = 4$, number of iterations needed is 2314, whereas using the clustering approach, for $\delta = 60.0$ and $n = 4$ this number reduces to 76 (=71+5). Thus the convergence of the algorithm using the clustering approach is much faster than that by using algorithm (3.1).

Thus the clustering approach has improved convergence considerably.

Trigonometric Regression								
	Number of iterations needed to achieve $\max_{1\le j\le m+J}\{F_j^{q,r_1,\ldots,r_m}\} \le 10^{-n}$ for				Number of iterations needed to achieve $\max_{1\le j\le J}\{F_j^{p}\} \le 10^{-n}$ for			
δ	n=1	n=2	n=3	n=4	n=1	n=2	n=3	n=4
60.0	2	6	7	22	2	7	9	71

Table 7.6. Clustering Approach : Iteration Results for the best choice of δ : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$

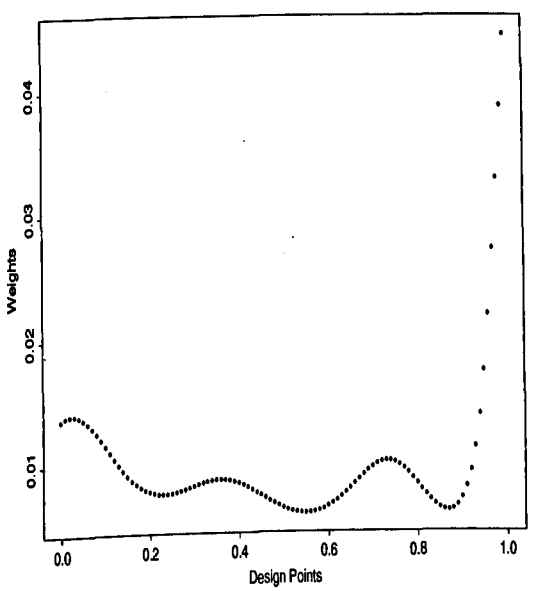


Figure 7.2. Trigonometric Regression : Weights versus Design Points after 5 iterations of (3.1), $f(x) = x^{\delta}$, $x = d$ and $\delta = 1.7$.

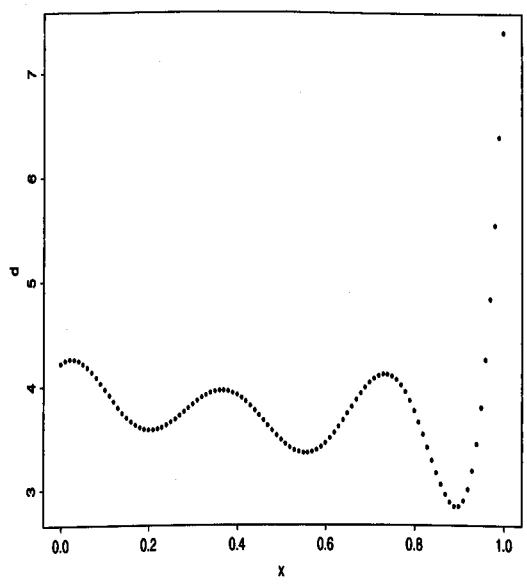


Figure 7.3. Trigonometric Regression : Variance function $d(x, p)$ after 5 iterations of (3.1), $f(x) = x^{\delta}$, $x = d$, $\delta = 1.7$. $d(x, p)$ is given in (7.28).

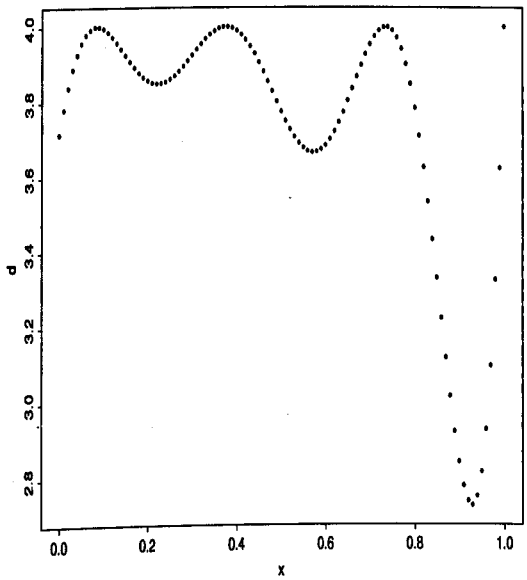


Figure 7.4. Trigonometric Regression : Variance function $d(x, p^*)$ at the optimum design (7.33). $d(x, p^*)$ is given in (7.34). The maximum value of 4 occurs at the support points.

7.5.3.2 Polynomial Regression in One Variable

In polynomial regression in one variable of order $k-1$ the model is given in 7.6. In this section we apply the clustering approach to our polynomial regression problems, namely, quadratic, cubic and quartic regressions. In each case we approximate the design interval by a grid of 201 points equally spaced at intervals of 0.01 between -1 and 1. i.e., we consider a design of the form

$$p = \begin{Bmatrix} x_1 & x_2 & \cdots & x_{201} \\ p_1 & p_2 & \cdots & p_{201} \end{Bmatrix}. \tag{7.35}$$

With the design above we start with the algorithm (3.1), i.e.,

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

with $f(x) = x^\delta$, $x = d$ for various values of δ.

After running this for 5 to 10 iterations the design points start forming clusters. We take each of the three regressions in turn.

Quardratic Regression

Figure 7.5 is a plot of the weights versus the design points after running algorithm (3.1) 5 times (with $f(x) = x^\delta$, $x = d$, $\delta = 1.9$). The plot shows a curve with one maximal, two minimal turning points and two peaks on the boundaries at $x = -1$ and $x = 1$. Assuming that the minimal turning points 'correspond' to zero weights, we would be left with three disjoint clusters of points which must contain the optimal support points.

Figure 7.3 is a plot of the variance function $d(x, p)$ at the same number of iterations. Here $d(x, p)$ is given by

$$d(x,p) = 6.20179x^4 - 6.63447x^2 + 3.68109. \tag{7.36}$$

This plot has a similar shape with one maximal, two minimal turning points and two peaks on the boundaries.

At this point we consider transforming from the weights p_i to weights within clusters and total cluster weights, with the boundaries between clusters at or near the minimal turning points.

Based on 3 clusters we now use algorithm (7.25).

Numerical information about the performance of algorithm (7.25) similar to that of tables A.6 and 7.6 is contained in tables A.7 and 7.7. As we started after 5 iterations of the algorithm (3.1), the actual numbers of iterations should be 5 plus the numbers given in tables A.7 and 7.7. If we compare these numbers to those in tables A.2 and 7.2 we see that the convergence is quite improved by considering the clustering approach. As the iteration results in table A.7 demonstrate, the value of $\delta = 78$ gave the fastest convergence.

Iterations converge to the design

$$p = \left\{ \begin{array}{ccccc} -1.0 & -0.01 & 0.00 & 0.01 & 1.00 \\ 0.33333 & 0.05922 & 0.21490 & 0.05922 & 0.33333 \end{array} \right\} \tag{7.37}$$

Taking convex combinations of the relevant cluster members as mentioned earlier,

the approximate design is given by

$$p^* = \left\{ \begin{array}{ccc} -1.0 & 0.0 & 1.0 \\ 0.33333 & 0.33333 & 0.33333 \end{array} \right\}. \tag{7.38}$$

Figure 7.7 is a plot showing the variance function $d(x, p^*)$ for the above D-optimal design. $d(x, p^*)$ is given by

$$d(x, p^*) = 4.5x^4 - 4.5x^2 + 3.0. \tag{7.39}$$

The maximum value of 3 occurs at the 3 support points -1, 0 and 1, confirming the D-optimality of the above quadratic regression.

Quadratic Regression								
	Number of iterations needed to achieve $\max_{1\le j\le m+J}\{F_j^{\underline{q},\underline{r}_1,\ldots,\underline{r}_m}\} \le 10^{-n}$ for				Number of iterations needed to achieve $\max_{1\le j\le J}\{F_j^{p}\} \le 10^{-n}$ for			
δ	n=1	n=2	n=3	n=4	n=1	n=2	n=3	n=4
78.0	1	3	4	23	2	3	7	65

Table 7.7. Clustering Approach : Iteration Results for the best choice of δ : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$

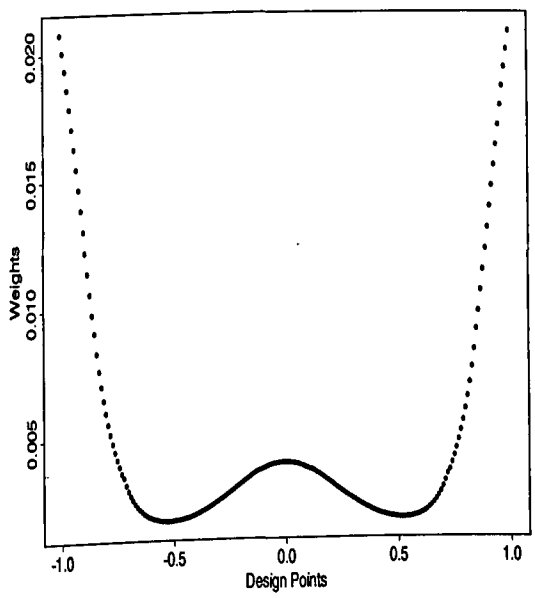


Figure 7.5. Quadratic Regression : Weights versus Design Points after 5 iterations of (3.1), $f(x) = x^{\delta}$, $x = d$ and $\delta = 1.9$.

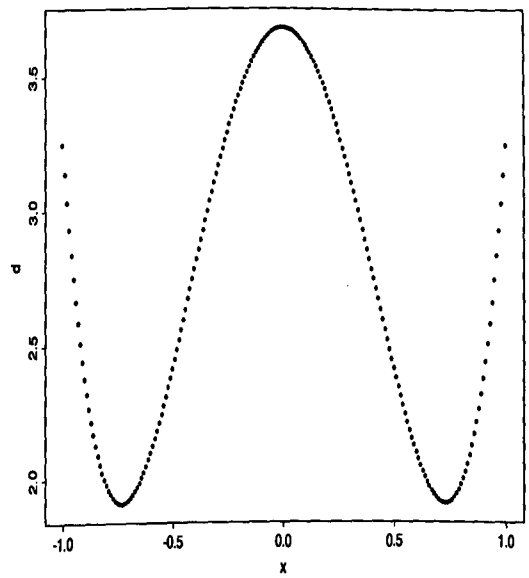


Figure 7.6. Quadratic Regression : Variance function $d(x, p)$ after 5 iterations of (3.1), $f(x) = x^{\delta}$, $x = d$, $\delta = 1.9$. $d(x, p)$ is given in (7.36).

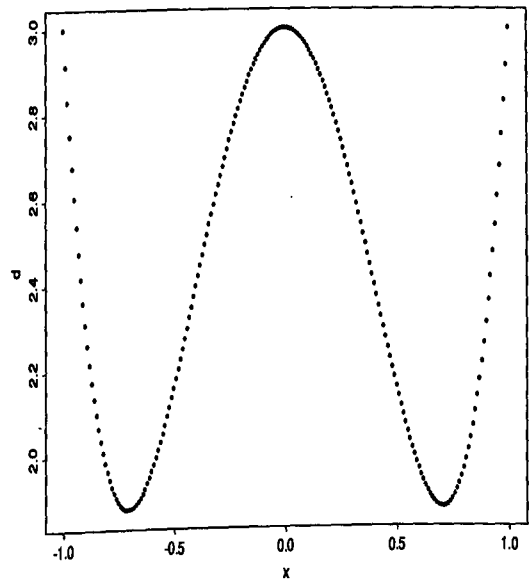


Figure 7.7. Quadratic Regression : Variance function $d(x, p^*)$ at the optimum design (7.38). $d(x, p^*)$ is given in (7.39). The maximum value of 3 occurs at the support points.

Cubic Regression

Figure 7.8 is a plot of the weights versus the design points after running algorithm (3.1) 10 times (with $f(x) = x^{\delta}$, $x = d$, $\delta = 0.05$). The plot shows a curve with two maximal, three minimal turning points and two peaks on the boundaries at $x = -1$ and $x = 1$. Assuming that the minimal turning points 'correspond' to zero weights, we would be left with four disjoint clusters of points which must contain the optimal support points.

Figure 7.9 is a plot of the variance function $d(x, p)$ at the same iteration, where $d(x, p)$ is given by

$$d(x,p) = 40.801x^6 - 39.903x^4 + 11.105x^2 + 2.289. \qquad (7.40)$$

This plot has a similar shape with with same number of maximal and minimal points.

Now consider transforming from the weights p_i to weights within clusters and total cluster weights.

Based on 4 clusters we now use algorithm (7.25). Iteration counts under algorithm (7.25) is contained in tables 7.8 and A.8. Here again the iteration numbers are much smaller that those in tables 7.3 and A.3. Values of δ higher those given in table A.8 were also investigated. The value of $\delta = 61$ gave the fastest convergence.

Iterations converge to the design

$$p = \left\{ \begin{array}{cccccccc} -1.00 & -0.46 & -0.45 & -0.44 & 0.44 & 0.45 & 0.46 & 1.00 \\ 0.25 & 0.007279 & 0.172137 & 0.070584 & 0.070584 & 0.172137 & 0.007279 & 0.25 \end{array} \right\}.$$

Taking convex combinations of the relevant cluster members, the approximate design is given by

$$p^* = \left\{ \begin{array}{cccc} -1.00 & -0.447 & 0.447 & 1.00 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{array} \right\}, \tag{7.41}$$

which is the same design we obtained in (7.13) of section 7.3.

Figure 7.10 is a plot of the variance function $d(x, p^*)$ for the above design. The maximum value of 4 occurs at the 4 support points ± 1, ± 0.447. This confirms the D-optimality of the above design. $d(x, p^*)$ is given by

$$d(x, p^*) = 4.5x^4 - 4.5x^2 + 3.0. \tag{7.42}$$

Cubic Regression								
	Number of iterations needed to achieve $\max_{1\le j\le m+J}\{F_j^{\underline{q},\underline{r}_1,\ldots,\underline{r}_m}\} \le 10^{-n}$ for				Number of iterations needed to achieve $\max_{1\le j\le J}\{F_j^{p}\} \le 10^{-n}$ for			
δ	n=1	n=2	n=3	n=4	n=1	n=2	n=3	n=4
61.0	2	3	4	21	2	3	9	67

Table 7.8. Clustering Approach : Iteration Results for the best choice of δ : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$

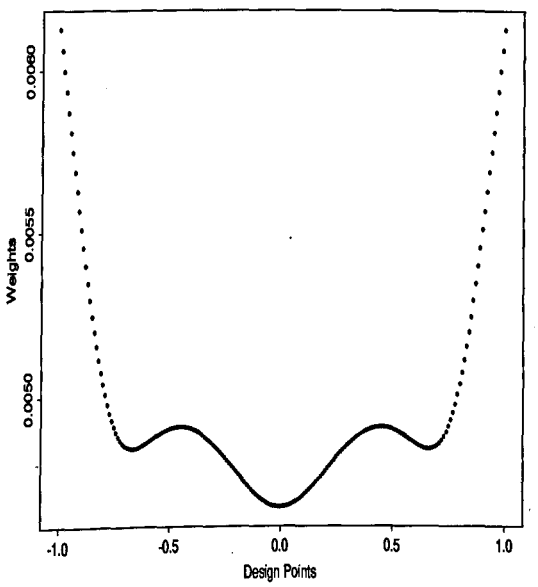


Figure 7.8. Cubic Regression : Weights versus Design Points after 10 iterations of (3.1), $f(x) = x^{\delta}$, $x = d$ and $\delta = 0.05$.

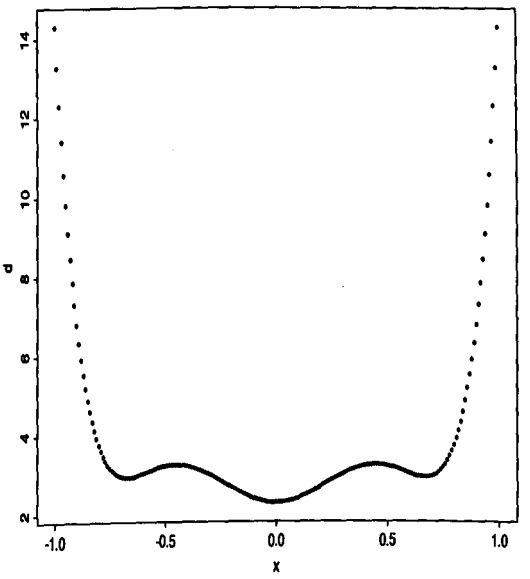


Figure 7.9. Cubic Regression : Variance function $d(x, p)$ after 10 iterations of (3.1), $f(x) = x^{\delta}$, $x = d$, $\delta = 0.05$. $d(x, p)$ is given in (7.40).

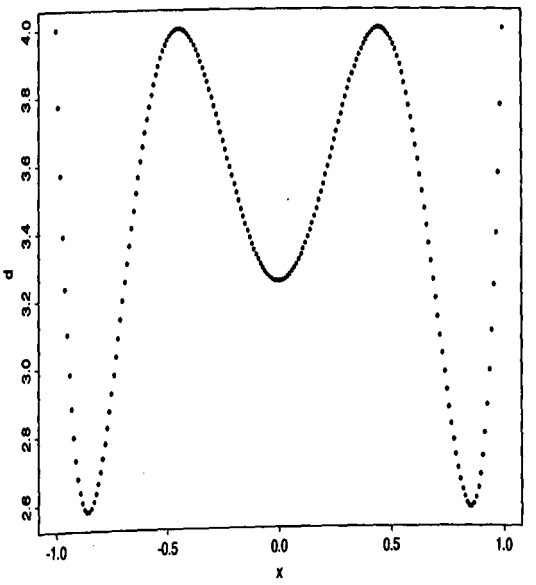


Figure 7.10. Cubic Regression : Variance function $d(x, p^*)$ at the optimum design (7.41). $d(x, p^*)$ is given in (7.42). The maximum value of 4 occurs at the support points.

Quartic Regression

Figure 7.11 is a plot of the weights versus the design points after running algorithm (3.1) 5 times (with $f(x) = x^{\delta}$, $x = d$, $\delta = 0.05$). The plot shows a curve with three maximal, four minimal turning points and two peaks on the boundaries at $x = -1$ and $x = 1$. Assuming that the minimal turning points 'correspond' to zero weights, we would be left with five disjoint clusters of points which must contain the optimal support points.

Figure 7.12 is a plot of variance function $d(x, p)$ at the same point above, where $d(x, p)$ is given by

$$d(x,p) = 153.623x^8 - 234.354x^6 + 111.263x^4 - 14.554x^2 + 3.675 \quad (7.43)$$

This plot has a similar shape with same number of maximal and minimal turning points. As the design points start forming clusters, we consider transforming from the weights p_i to weights within clusters and total cluster weights.

Based on 5 clusters we now use algorithm (7.25). Numerical information about the performance of algorithm (7.25) is contained in tables 7.9 and A.9. Values of δ higher those presented in table A.9 were investigated. The value of $\delta = 53$ provided the fastest convergence. Here again convergence is improved a lot if we compare the numbers in these tables to those in tables 7.4 and A.4. In this example, for $\delta = 1.9$ and $n = 4$, number of iterations needed is 2571, whereas using clustering approach, for $\delta = 53$ this number reduces to 97 (=92+5). This clearly indicates improvement in convergence.

Iterations converge to the design

$$p = \left\{ \begin{array}{ccccccccc} -1.00 & -0.66 & -0.65 & -0.01 & 0.00 & 0.01 & 0.65 & 0.66 & 1.00 \\ 0.20 & 0.09695 & 0.10305 & 0.01605 & 0.16790 & 0.01605 & 0.10305 & 0.09695 & 0.20 \end{array} \right\}.$$

Taking convex combinations of the relevant cluster members, the approximate design is given by

$$p^* = \left\{ \begin{array}{ccccc} -1.00 & -0.655 & 0.00 & 0.655 & 1.00 \\ 0.20 & 0.20 & 0.20 & 0.20 & 0.20 \end{array} \right\}, \tag{7.44}$$

which is in turn the same design we obtained in (7.14).

Figure 7.13 is a plot of the variance function $d(x, p^*)$ for the above design. $d(x, p^*)$ is given by

$$d(x, p^*) = 76.495x^8 - 141.999x^6 + 79.527x^4 - 14.023x^2 + 5.0. \tag{7.45}$$

The maximum value of 5 occurs at the 5 support points ± 1, ± 0.655 and 0 confirming the D-optimality of the above design.

Quartic Regression								
	Number of iterations needed to achieve $\max_{1\le j\le m+J}\{F_j^{\underline{q},\underline{r}_1,\ldots,\underline{r}_m}\} \le 10^{-n}$ for				Number of iterations needed to achieve $\max_{1\le j\le J}\{F_j^{\underline{p}}\} \le 10^{-n}$ for			
δ	n=1	n=2	n=3	n=4	n=1	n=2	n=3	n=4
53.0	1	3	4	20	2	3	10	92

Table 7.9. Clustering Approach : Iteration Results for the best choice of δ : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$

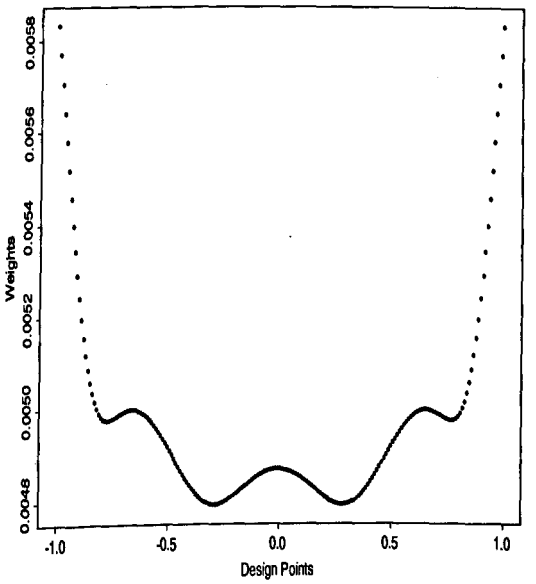


Figure 7.11. Quartic Regression : Weights versus Design Points after 5 iterations of (3.1), $f(x) = x^{\delta}$, $x = d$ and $\delta = 0.05$.

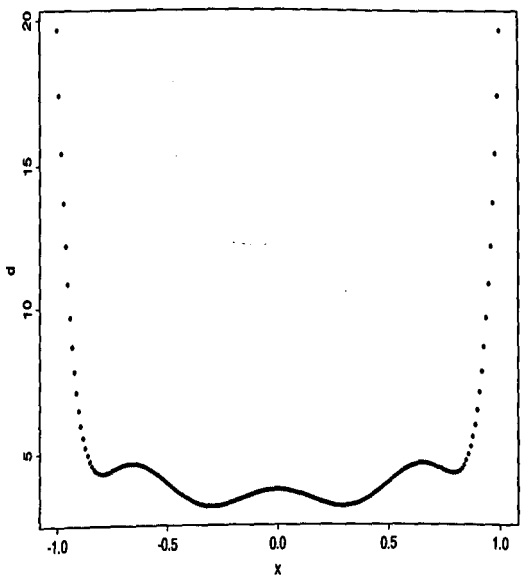


Figure 7.12. Quartic Regression : Variance function $d(x, p)$ after 5 iterations of (3.1), $f(x) = x^{\delta}$, $x = d$, $\delta = 0.05$. $d(x, p)$ is given in (7.43).

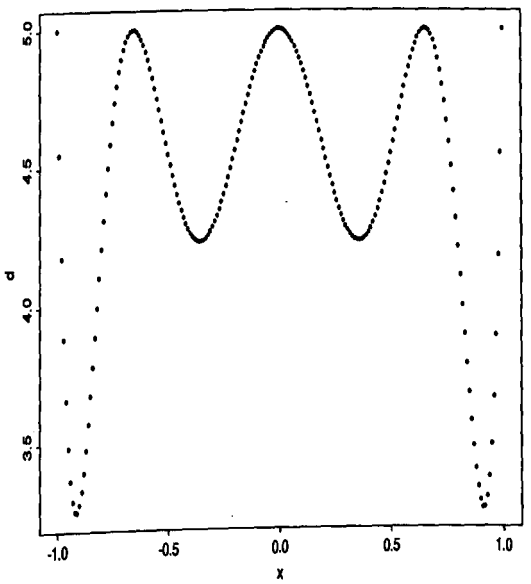


Figure 7.13. Quartic Regression : Variance function $d(x, p^*)$ at the optimum design (7.44). $d(x, p^*)$ is given in (7.45). The maximum value of 5 occurs at the support points.

7.5.3.3 Second-Order Model in Two Variables

As a last example we consider the second-order model in two variables. The model is given in (7.16).

We consider the discretised design space consisting of all pairs (x_1, x_2) arising when the values for each x_i, $i = 1, 2$ are those between -1 and +1 taken at intervals of 0.1. That is, the space consists of $(21)^2 = 441$ pairs (x_1, x_2). i.e. the discretised design space is given by

$$\mathcal{V} = \{\underline{v}_x : \underline{v}_x = (1,\ x_1,\ x_2,\ x_1x_2,\ x_1^2,\ x_2^2)^T,\ x_i = -1 : 1/0.1,\ i = 1, 2\}$$

Consider starting algorithm (3.1) with equal weights of $1/441$ on each pair (x_1, x_2).

We write the design with arbitrary weights as:

$$p = \left\{ \begin{array}{cccc} (x_{11}, x_{21}) & (x_{12}, x_{22}) & \ldots & (x_{1\,441}, x_{2\,441}) \\ p_1 & p_2 & \ldots & p_{441} \end{array} \right\}$$

Figure 7.14 is a plot of the weights p_i versus the design points (x_1, x_2) after running algorithm (3.1) 10 times (with $f(x) = x^\delta$, $x = d$, $\delta = 2.3$). Clearly the plot shows 9 clusters - 4 in the corners, 4 in the middle of the sides and 1 in the centre.

Figure 7.16 shows the variance function $d(\underline{x}, p)$ versus (x_1, x_2) at the above iteartion. This plot also shows the 9 clusters as above. $d(\underline{x}, p)$ is given by

$$\begin{aligned} d(\underline{x}, p) &= \underline{v}^T(\underline{x})\, M^{-1}(p)\, \underline{v}(\underline{x}) \\ &= 6.88179(x_1^4 + x_2^4) - 6.92214(x_1^2 + x_2^2) \\ &\quad -0.20284x_1^2x_2^2 + 6.93500. \end{aligned} \tag{7.46}$$

As the design points start forming clusters we consider transforming from weights p_i to weights within clusters and total cluster weights. Based on 9 clusters the total and within cluster weights are given by

$$\begin{aligned} \underline{q} &= (q_1, q_2, \ldots, q_9)^T \\ \underline{r}_j &= (r_{j1}, r_{j2}, \ldots, r_{jn_j})^T, \quad j = 1, 2, \ldots, 9. \end{aligned} \tag{7.47}$$

The above weights satisfy the following:

$$\begin{aligned} &\sum_{j=1}^{9} q_j = 1, \quad q_j \geq 0, \quad j = 1, 2, \ldots, 9 \\ &\sum_{i=1}^{n_j} r_{ji} = 1, \quad r_{ji} \geq 0, \quad i = 1, 2, \ldots, n_j, \; j = 1, 2, \ldots, 9. \end{aligned} \tag{7.48}$$

After transforming the weights the criterion is given by

$$\phi(p) = \phi_m(\underline{q}, \underline{r}_1, \underline{r}_2, \ldots, \underline{r}_9), \text{ a function of } \underline{q}, \underline{r}_1, \ldots, \underline{r}_9. \tag{7.49}$$

We need to choose $\underline{q}, \underline{r}_1, \underline{r}_2, \ldots, \underline{r}_9$ optimally. So the problem is to maximise $\phi_m(\underline{q}, \underline{r}_1, \underline{r}_2, \ldots, \underline{r}_9)$ subject to (7.48).

The optimal solution is obtained by using algorithm (7.25). Iteration counts are recorded in table A.10 and summarised in table 7.10 (for the best choice of δ). Note values of δ higher those presented in table A.10 were also investigated. This proved to be $\delta = 33$. Iteration results, as in the other regression models of previous sections, confirm that using the clustering approach improves the convergence considerably.

Iterations converge to the design

$$p^* = \left\{ \begin{array}{ccccccccc} -1.0 & -1.0 & -1.0 & 0 & 0 & 0 & 1.0 & 1.0 & 1.0 \\ -1.0 & 0 & 1.0 & -1.0 & 0 & 1.0 & -1.0 & 0 & 1.0 \\ 0.14579 & 0.08016 & 0.14579 & 0.08016 & 0.09620 & 0.08016 & 0.14579 & 0.08016 & 0.14579 \end{array} \right\} \tag{7.50}$$

which is the optimal solution obtained analytically in section 7.4.1.

Figure 7.15 shows a plot of weights versus design points (x_1, x_2) at the above optimum design. Figure 7.17 is a plot showing the variance function $d(\underline{x}, p^*)$ for the above D-optimal design. The maximum value of 6 occurs at the 9 design points (-1, -1), (-1, 0), (-1, 1), (0, -1), (0, 0), (0, 1), (1, -1), (1, 0) and (1, 1). $d(\underline{x}, p^*)$ is given by the expression

$$d(\underline{x}, p^*) = 5.38007(x_1^4 + x_2^4) - 5.38007(x_1^2 + x_2^2) + 6.0. \quad (7.51)$$

Second-Order Model								
	Number of iterations needed to achieve $\max_{1\le j\le m+J}\{F_j^{\underline{q},\underline{r}_1,\ldots,\underline{r}_m}\} \le 10^{-n}$ for				Number of iterations needed to achieve $\max_{1\le j\le J}\{F_j^{p}\} \le 10^{-n}$ for			
δ	n=1	n=2	n=3	n=4	n=1	n=2	n=3	n=4
33.0	1	2	5	13	1	2	12	21

Table 7.10. Clustering Approach : Iteration Results for the best choice of δ : $f(d) = d^\delta$, $d = \frac{\partial \phi}{\partial p}$

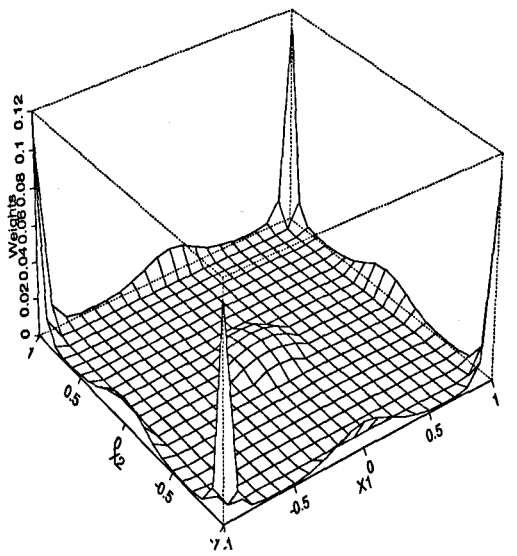


Figure 7.14. Second-Order Model : Weights versus Design Points after 10 iterations of (3.1), $f(x) = x^{\delta}$, $x = d$, $\delta = 2.3$

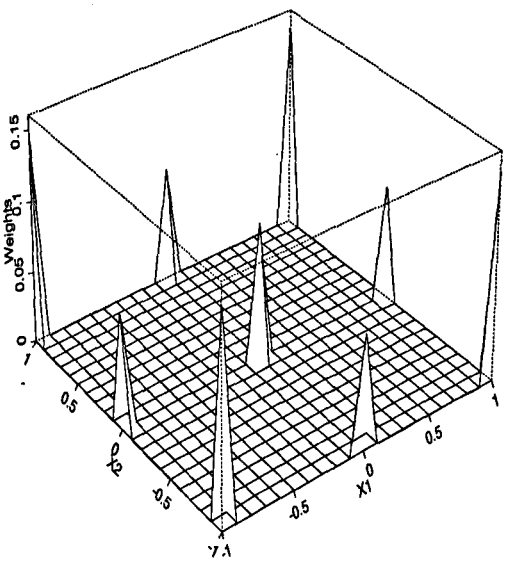


Figure 7.15. Second-Order Model : Weights versus Design Points at the optimum

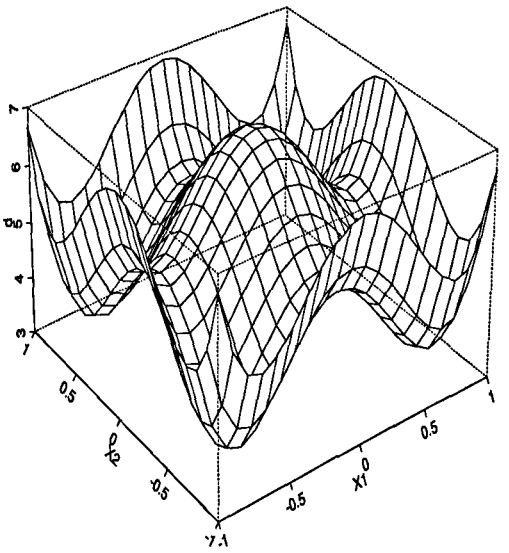


Figure 7.16. Second-Order Model : Variance function $d(x, p)$ after 10 iterations of (3.1), $f(x) = x^{\delta}$, $x = d$, $\delta = 2.3$. $d(x, p)$ is given in (7.46)

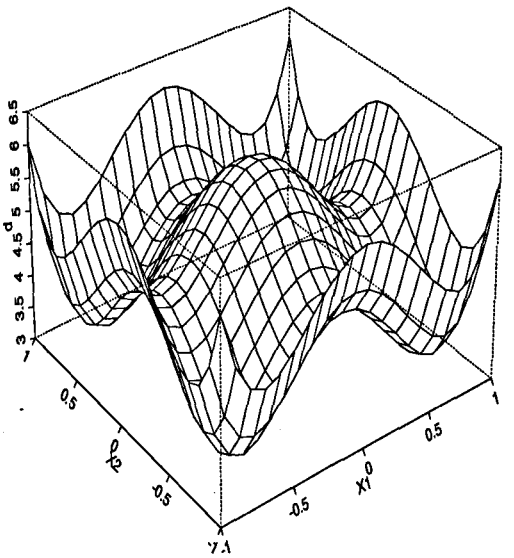


Figure 7.17. Second-Order Model : Variance function $d(x, p^*)$ at the optimum design (7.50). $d(x, p^*)$ is given in (7.51). The maximum value of 6 occurs at the support points.

Chapter 8

Construction of Constrained Optimal Designs

8.1 Introduction

We now consider the problem of computing designs which optimise standard regression design criteria subject to an equality constraint. Examples of such problems include determining designs subject to a given efficiency. See Pukelsheim and Rosenberger (1993). Cook and Fedorov (1995) is an invited discussion paper on constrained optimal design.

Torsney and Alahmadi (1995) consider the case of finding designs subject to zero correlations between the estimates of two linear combinations of the underlying parameters. In particular, they consider the case of minimal support designs and transform the constrained optimal design problem to one of maximising a criterion with respect to two or three sets of weights or distributions. They employed extensions of the multiplicative algorithms (3.1).

This approach does not extend to other equality constraints or to constrained optimal designs subject to zero correlation when the support exceeds minimum size. It is our purpose to consider the general case and, in fact, to extend the work of Alahmadi (1993).

We consider problems of maximising the D_A- and A-optimal criteria under a linear model subject to a constraint. The criteria are the -(log)determinant (D_A-optimality) and trace (A-optimality) of the covariance matrix of two linear functions of the parameters of interest. These two linear functions define the matrix A. The constraint is the equality of the variances of the estimates of two of these linear functions.

Initially the Lagrangian is formulated but the Lagrange parameter is removed through a substitution, using linear equation theory, in an approach which transforms the constrained optimisation problem to a problem of maximising two functions of the design weights simultaneously. They have a common maximum of zero which is simultaneously attained at the constrained optimal design weights. This means that established algorithms for finding optimising distributions can be considered. The approach can easily be extended to the case of several constraints.

8.2 Constrained Optimising Distributions

8.2.1 General Problem (GP)

The general problem which we will consider is the following:
Maximise a criterion $\phi(p)$ subject to the constraint $g(p) = 0$ as well as

$$p_j \geq 0, \quad \sum_j p_j = 1.$$

$\phi(p)$ could be one of the criteria A-, c-, D- or D_A-optimality.

The function $g(p)$ could be one of the following:

$g(p) = \underline{a}^T M^{-1}(p)\,\underline{a} - \underline{b}^T M^{-1}(p)\,\underline{b}$ if we want $var(\underline{a}^T\hat{\underline{\theta}}) = var(\underline{b}^T\hat{\underline{\theta}})$,
or
$g(p) = \underline{r}^T M^{-1}(p)\underline{s}$ if we want $Cov(\underline{r}^T\hat{\underline{\theta}}, \underline{s}^T\hat{\underline{\theta}}) = 0$,
or
$g(p) = \phi(p)/\phi(p^*_U) - e$ if we want an efficiency of e $(0 < e < 1)$, where p^*_U is the unconstrained optimiser of $\phi(p)$.

In the latter case the criterion to be optimised is $\phi(p)$.

In the first case an appropriate criterion might be

$$\phi(p) = -logdet\{AM^{-1}(p)A^T\}$$
or
$$\phi(p) = -tr\{AM^{-1}(p)A^T\},$$

where $A = [\underline{a}, \underline{b}]^T$.

The second case is, in fact, equivalent to the first case with $\underline{r} = \frac{\underline{a}+\underline{b}}{2}$ and $\underline{s} = \frac{\underline{a}-\underline{b}}{2}$.

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

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

We illustrate this phenomenon in figure 8.1.

Suppose we wish to miminise the maximum of two functions, and let these two functions be $f_1(x)$ and $f_2(x)$. Let $h(x)$ be the maximum of these two functions. i.e., $h(x) = max\{f_1(x), f_2(x)\}$.

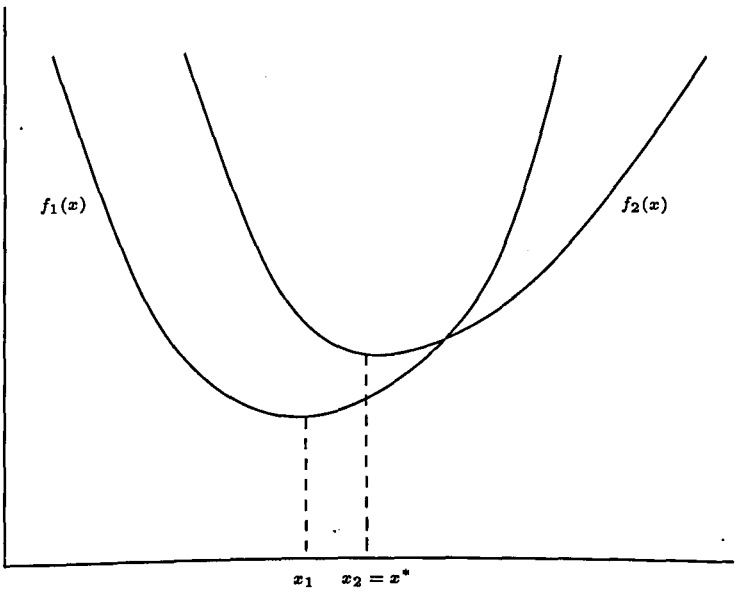
Now two cases may arise. In the first case (Case-1 in figure 8.1) the minimum of $h(x)$ occurs at the point where one of the functions is minimised, e.g. $f_2(x)$.

In the second case (Case-2 in figure 8.1) the minimum of $h(x)$ occurs at a point where both functions are equal. If this point is to be the optimal solution, it must minimise both functions subject to them being equal. The above choice of $g(p)$ is equivalent to such a problem in the case $\phi = -tr\{AM^{-1}(p)A^T\}$.

If both variances have a common value then $tr\{AM^{-1}(p)A^T\}$ is twice this common value and is therefore minimised when this common value is minimised.

Note we have to be sure that the constraint $g(p) = 0$ can be satisfied. One check on this would be to solve problem (P1) with $\phi(p) = -[g(p)]^2$. In fact consideration of $g(p) = \underline{r}^T M^{-1}(p)\underline{s}$ evolved from a problem considered by Torsney (1988), in which a numerical covariance had to be minimised.

Case - 1



Case - 2

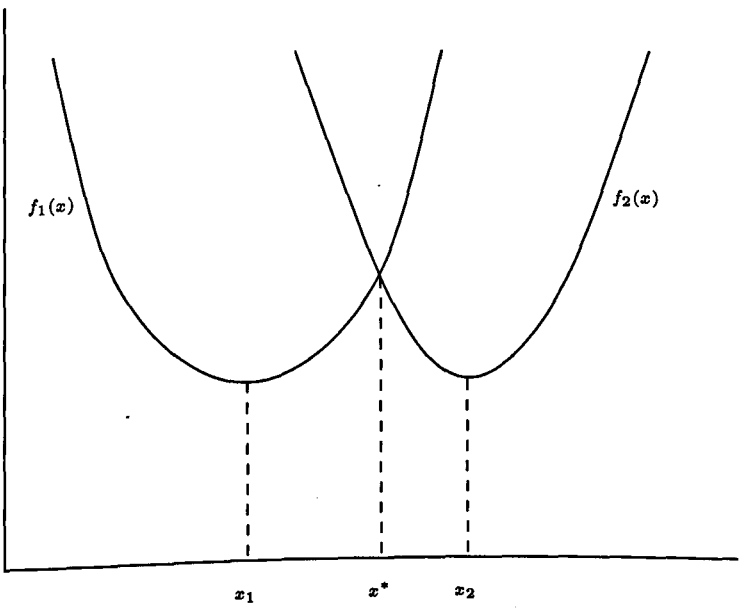


Figure 8.1. An example of a *minimax* problem

8.2.2 1st order Lagrangian conditions transformed

First we establish first order conditions for a constrained local maximum using Lagrangian theory applied to the equality constraints $g(p) = 0$ and $\sum_j p_j = 1$.

Let

$$L(\phi, p, \lambda) = \phi(p) + \lambda g(p) + \mu\left(\sum_j p_j - 1\right). \tag{8.1}$$

Then for $i = 1, 2, \ldots, J$

$$\begin{aligned} d_i^L &= \frac{\partial L}{\partial p_i} \\ &= \frac{\partial \phi}{\partial p_i} + \lambda \frac{\partial g}{\partial p_i} + \mu \\ &= d_i^\phi + \lambda d_i^g + \mu. \end{aligned} \tag{8.2}$$

Suppose $p_j > 0$, $j = 1, 2, \ldots, J$. Then we must have

$$\begin{aligned} d_i^L &= 0 \\ \text{i.e. } \sum_{i=1}^{J} p_i d_i^L &= 0 \\ \text{i.e. } \mu &= -\sum_{i=1}^{J} p_i (d_i^\phi + \lambda d_i^g). \end{aligned}$$

Thus

$$d_i^\phi + \lambda d_i^g = -\mu = \sum_{j=1}^{J} p_j (d_j^\phi + \lambda d_j^g). \tag{8.3}$$

Hence the vertex directional derivatives of L are

$$\begin{aligned} F_i^L &= d_i^L - \sum_{j=1}^{J} p_j d_j^L \\ &= (d_i^\phi + \lambda d_i^g) - \sum_{j=1}^{J} p_j (d_j^\phi + \lambda d_j^g) \\ &= 0 \qquad [\text{from (8.3)}] \end{aligned} \tag{8.4}$$

Note we would arrive at this point by applying this Lagrangian approach only to the constraint $g(p) = 0$ and then invoking the equivalence theorem to ensure $\sum_j p_j = 1$.

Also

$$\begin{aligned} F_i^L &= d_i^L - \sum_{j=1}^{J} p_j d_j^L \\ &= d_i^\phi - \sum_{j=1}^{J} p_j d_j^\phi + \lambda (d_i^g - \sum_{j=1}^{J} p_j d_j^g) \\ &= F_i^\phi + \lambda F_i^g. \end{aligned} \tag{8.5}$$

where $F_i^\phi = d_i^\phi - \sum_{j=1}^{J} p_j d_j^\phi$ and $F_i^g = d_i^g - \sum_{j=1}^{J} p_j d_j^g$.

Equivalently

$$\underline{F}^L = \underline{F}^\phi + \lambda \underline{F}^g. \tag{8.6}$$

Now from (8.4)

$$\begin{aligned} F_j^L &= 0 \quad \text{given } p_j > 0,\ j = 1, 2, \ldots, J \\ \text{i.e.,} \quad \underline{F}^L &= \underline{0} \\ \text{i.e.,} \quad \underline{F}^g \lambda &= -\underline{F}^\phi \end{aligned} \tag{8.7}$$

Suppose $A = \underline{F}^g$, $\underline{b} = -\underline{F}^\phi$ and $\underline{\lambda} = \lambda$.

Then $\underline{\lambda}$ must satisfy

$$A\underline{\lambda} = \underline{b} \tag{8.8}$$

We now deal with λ using the approach of Alahmadi (1993).

The set of solutions to the system of equations (8.8), if solutions exists, is given by

$$\underline{\lambda} = A^- \underline{b} + (I - A^- A)\underline{z} \quad \text{for any } \underline{z}, \tag{8.9}$$

where A^- is any generalised inverse of A.

Assuming $A^T A$ is non-singular one choice of A^- is

$$\begin{aligned} A^- &= (A^T A)^{-1} A^T \\ &= \left[(\underline{F}^g)^T \underline{F}^g\right]^{-1} (\underline{F}^g)^T. \end{aligned} \tag{8.10}$$

Then

$$\begin{aligned} A^{-}\underline{b} &= \frac{-(\underline{F}^{g})^{T}\underline{F}^{\phi}}{(\underline{F}^{g})^{T}\underline{F}^{g}} = \hat{\lambda}, \\ AA^{-}\underline{b} &= \hat{\lambda}A = \hat{\lambda}\underline{F}^{g}. \end{aligned}$$

Thus p^* should be such that

$$\hat{\lambda}\underline{F}^{g} = -\underline{F}^{\phi} \quad \text{i.e.,} \quad \hat{\lambda}\underline{F}^{g} + \underline{F}^{\phi} = \underline{0} \tag{8.11}$$

and then substituting by $\hat{\lambda}$ in (8.11), we get

$$-\frac{(\underline{F}^{g})^{T}\underline{F}^{\phi}}{(\underline{F}^{g})^{T}\underline{F}^{g}}\underline{F}^{g} + \underline{F}^{\phi} = \underline{0} \tag{8.12}$$

$$\text{or,} \quad \underline{h} = \left[(\underline{F}^{g})^{T}\underline{F}^{g}\right]\underline{F}^{\phi} - \left[(\underline{F}^{g})^{T}\underline{F}^{\phi}\right]\underline{F}^{g} = \underline{0} \tag{8.13}$$

$$\text{i.e.,} \quad \underline{h}^{T}\underline{h} = 0. \tag{8.14}$$

where

$$\begin{aligned} h_i &= \left[(\underline{F}^{g})^{T}\underline{F}^{g}\right]F_i^{\phi} - \left[(\underline{F}^{g})^{T}\underline{F}^{\phi}\right]F_i^{g} \\ &= \left[\sum_{j=1}^{J}(F_j^{g})^2\right]F_i^{\phi} - \left[\sum_{j=1}^{J}(F_j^{g}F_j^{\phi})\right]F_i^{g}. \end{aligned} \tag{8.15}$$

So since $\underline{h}^{T}\underline{h} \geq 0$, p^* should minimise $\underline{h}^{T}\underline{h}$ or maximise $Q = Q_1 = [-\underline{h}^{T}\underline{h}]$ with a maximum value of zero.

Alternatively

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

$$= \left[(\underline{F}^g)^T\underline{F}^g\right]^2\left[(\underline{F}^\phi)^T\underline{F}^\phi\right] + \left[(\underline{F}^g)^T\underline{F}^\phi\right]^2\left[(\underline{F}^g)^T\underline{F}^g\right] - \left[(\underline{F}^g)^T\underline{F}^\phi\right]^2\left[(\underline{F}^g)^T\underline{F}^g\right] - \left[(\underline{F}^g)^T\underline{F}^g\right]\left[(\underline{F}^g)^T\underline{F}^\phi\right]^2$$

$$= \left[(\underline{F}^g)^T\underline{F}^g\right]^2\left[(\underline{F}^\phi)^T\underline{F}^\phi\right] - \left[(\underline{F}^g)^T\underline{F}^\phi\right]^2\left[(\underline{F}^g)^T\underline{F}^g\right]$$

$$= \left[(\underline{F}^g)^T\underline{F}^g\right]^2\left[(\underline{F}^\phi)^T\underline{F}^\phi\right]\left[1 - \frac{\left[(\underline{F}^g)^T\underline{F}^\phi\right]^2}{\left[(\underline{F}^\phi)^T\underline{F}^\phi\right]\left[(\underline{F}^g)^T\underline{F}^g\right]}\right]$$

$$= \left[(\underline{F}^g)^T\underline{F}^g\right]^2\left[(\underline{F}^\phi)^T\underline{F}^\phi\right]\left[1 - R_1\right] \tag{8.16}$$

where

$$R_1 = \frac{\left[(\underline{F}^g)^T\underline{F}^\phi\right]^2}{\left[(\underline{F}^\phi)^T\underline{F}^\phi\right]\left[(\underline{F}^g)^T\underline{F}^g\right]} = \frac{\xi_1^2}{\xi_2\,\xi_3}, \tag{8.17}$$

$$\xi_1 = (\underline{F}^g)^T\underline{F}^\phi,\ \xi_2 = (\underline{F}^\phi)^T\underline{F}^\phi,\ \xi_3 = (\underline{F}^g)^T\underline{F}^g. \tag{8.18}$$

So $\underline{h}^T\underline{h}$ is given by (in terms of ξ_1, ξ_2 and ξ_3)

$$\underline{h}^T\underline{h} = \xi_2\,\xi_3^2 - \xi_3\,\xi_1^2. \tag{8.19}$$

Clearly $0 \leq R_1 \leq 1$. So p^* should maximise $Q = Q_2 = [R_1 - 1]$ yielding a maximum value of zero.

Now consider the possibility of zero p^* which was not considered by Alahmadi (1993). Then by the equivalence theorem

$$F_j^L \begin{cases} = 0 & \text{if } p_j^* > 0 \\ \leq 0 & \text{if } p_j^* = 0. \end{cases} \tag{8.20}$$

There are two ways in which we can adapt the above approach to this case. One is to argue that

$$\begin{aligned} l_i &= p_i h_i = 0 \quad \text{for } i = 1, 2, \ldots, J \\ \text{i.e.} \quad \underline{l} &= P\underline{h} = \underline{0} \end{aligned} \tag{8.21}$$

where $P = \text{diag}(p_1, p_2, \ldots, p_J)$. So we replace $\underline{h}$ in the argument above by $\underline{l} = P\underline{h}$ to conclude that p^* should maximise

$$Q = Q_3 = -\underline{h}^T P^2 \underline{h}. \tag{8.22}$$

We try to simplify $\underline{h}^T P^2 \underline{h}$ as

$$\begin{aligned} \underline{h}^T P^2 \underline{h} &= \left[\left((\underline{F}^g)^T \underline{F}^g\right) \underline{F}^\phi - \left((\underline{F}^g)^T \underline{F}^\phi\right)\right]^T P^T P \left[\left((\underline{F}^g)^T \underline{F}^g\right) \underline{F}^\phi - \left((\underline{F}^g)^T \underline{F}^\phi\right)\right] \\ &= \left[(\underline{F}^g)^T \underline{F}^g\right]^2 \left[(\underline{F}^\phi)^T P^T P \underline{F}^\phi\right] + \left[(\underline{F}^g)^T \underline{F}^\phi\right]^2 \left[(\underline{F}^g)^T P^T P \underline{F}^g\right] - \\ &\quad (\underline{F}^g)^T \left[(\underline{F}^\phi)^T \underline{F}^g\right] P^T P \left[(\underline{F}^g)^T \underline{F}^g\right] (\underline{F}^\phi) - \\ &\quad (\underline{F}^\phi)^T \left[(\underline{F}^g)^T \underline{F}^g\right] P^T P \left[(\underline{F}^g)^T \underline{F}^\phi\right] (\underline{F}^g) \\ &= \left[(\underline{F}^g)^T \underline{F}^g\right]^2 \left[(\underline{F}^\phi)^T P^T P \underline{F}^\phi\right] + \left[(\underline{F}^g)^T \underline{F}^\phi\right]^2 \left[(\underline{F}^g)^T P^T P \underline{F}^g\right] - \\ &\quad 2\left[(\underline{F}^g)^T \underline{F}^g\right] \left[(\underline{F}^g)^T \underline{F}^\phi\right] \left[(\underline{F}^g)^T P^T P \underline{F}^\phi\right] \\ &= \left\{\left[(\underline{F}^g)^T \underline{F}^g\right]^2 \left[(\underline{F}^\phi)^T P^T P \underline{F}^\phi\right] + \left[(\underline{F}^g)^T \underline{F}^\phi\right]^2 \left[(\underline{F}^g)^T P^T P \underline{F}^g\right]\right\} [1 - R_2] \end{aligned} \tag{8.23}$$

where

$$R_2 = \frac{2\left[(\underline{F}^g)^T\underline{F}^g\right]\left[(\underline{F}^g)^T\underline{F}^\phi\right]\left[(\underline{F}^g)^TP^TP\underline{F}^\phi\right]}{\left[(\underline{F}^g)^T\underline{F}^g\right]^2\left[(\underline{F}^\phi)^TP^TP\underline{F}^\phi\right]+\left[(\underline{F}^g)^T\underline{F}^\phi\right]^2\left[(\underline{F}^g)^TP^TP\underline{F}^g\right]}$$

$$= \frac{2\,\xi_1\,\xi_3\,\xi_4}{\xi_3^2\,\xi_5+\xi_1^2\,\xi_6}, \tag{8.24}$$

$$\xi_4 = \left[(\underline{F}^g)^TP^2\underline{F}^\phi\right],\ \xi_5 = (\underline{F}^\phi)^TP^2\underline{F}^\phi,\ \xi_6 = (\underline{F}^g)^TP^2\underline{F}^g, \tag{8.25}$$

ξ_1, ξ_2, ξ_3 are same as in (8.18). In terms of ξ_i's $\underline{h}^TP^2\underline{h}$ is given by

$$\underline{h}^TP^2\underline{h} = \xi_3^2\,\xi_5+\xi_1^2\,\xi_6-2\,\xi_1\,\xi_3\,\xi_4. \tag{8.26}$$

Result: $R_2 \le 1$.

Proof:

For notational convenience let

$$\alpha_1 = \left[(\underline{F}^g)^T\underline{F}^g\right], \alpha_2 = \left[(\underline{F}^\phi)^T\underline{F}^\phi\right],$$

$$\underline{c} = P\underline{F}^\phi, \underline{d} = P\underline{F}^g.$$

Then R_2 can be written as

$$R_2 = \frac{2\,\alpha_1\,\alpha_2\,(\underline{c}^T\underline{d})}{\alpha_1^2\,(\underline{c}^T\underline{c})+\alpha_2^2\,(\underline{d}^T\underline{d})}$$

$$= \frac{2\,(\underline{u}^T\underline{v})}{\underline{u}^T\underline{u}+\underline{v}^T\underline{v}} \qquad [\underline{u}=\alpha_1\underline{c} \text{ and } \underline{v}=\alpha_2\underline{d}]$$

$$= \frac{2\sum_{i=1}^{J}u_iv_i}{\sum_{i=1}^{J}u_i^2+\sum_{i=1}^{J}v_i^2} \tag{8.27}$$

where $\underline{u} = (u_1, u_2, \ldots, u_J)^T$ and $\underline{v} = (v_1, v_2, \ldots, v_J)^T$.

Now, as u_i, v_i are real numbers, from Cauchy-Schwarz inequality

$$u_i v_i \leq \frac{u_i^2 + v_i^2}{2},$$

$$\text{i.e., } \sum_{i=1}^{J} u_i v_i \leq \frac{\sum_{i=1}^{J} u_i^2 + \sum_{i=1}^{J} v_i^2}{2}.$$

Hence, from (8.27) $R_2 \leq 1$.

Note that the above can be directly proved from (8.23) as in that expression $[1 - R_2]$ is one of the two factors of a factorisation of $[\underline{h}^T P^2 \underline{h}]$ which is nonnegative. Since the other factor is nonnegative so is $[1 - R_2]$.

So p^* should maximise $Q = Q_4 = [R_2 - 1]$ yielding a maximum value of zero.

Another option is to argue that

$$\begin{aligned} E_j^L &= p_j F_j^L = 0 \quad \text{for } j = 1, 2, \ldots, J \\ \text{or, } \underline{E}^L &= P\underline{F}^L = \underline{0}. \end{aligned} \tag{8.28}$$

So we replace $\underline{F}^L$ in the above argument by $\underline{E}^L$ concluding that we should have at p^*

$$\underline{h} = \underline{h}(\underline{E}) = \underline{0}. \tag{8.29}$$

As $[\underline{h}(\underline{E})]^T [\underline{h}(\underline{E})] \geq 0$, p^* should maximise

$$Q = Q_5 = -[\underline{h}(\underline{E})]^T [\underline{h}(\underline{E})] \tag{8.30}$$

yielding a maximum of zero.

Here also we can write $[\underline{h}(\underline{E})]^T [\underline{h}(\underline{E})]$ as

$$[\underline{h}(\underline{E})]^T [\underline{h}(\underline{E})] = \left[(\underline{E}^g)^T \underline{E}^g\right]^2 \left[(\underline{E}^\phi)^T \underline{E}^\phi\right] [1 - R_3], \tag{8.31}$$

where

$$R_3 = \frac{\left[(\underline{E}^g)^T \underline{E}^\phi\right]^2}{\left[(\underline{E}^\phi)^T \underline{E}^\phi\right] \left[(\underline{E}^g)^T \underline{E}^g\right]} = \frac{\xi_4^2}{\xi_5 \xi_6}, \tag{8.32}$$

ξ_4, ξ_5 and ξ_6 are given in (8.25).

Clearly from (8.32) $0 \leq R_3 \leq 1$. Thus from (8.30) it is clear that $\left\{-[\underline{h}(\underline{E})]^T [\underline{h}(\underline{E})]\right\}$ is maximised or $\left\{[\underline{h}(\underline{E})]^T [\underline{h}(\underline{E})]\right\}$ is minimised when R_3 is maximised. i.e.,

$$[\underline{h}(\underline{E})]^T [\underline{h}(\underline{E})] = 0 \Leftrightarrow R_3 = 1. \tag{8.33}$$

So p^* should maximise $Q = Q_6 = [R_3 - 1]$ yielding a maximum value of zero.

So we have dealt with the Lagrange multiplier by 'substitution' in terms of p and transformed attainment of the first order conditions in the Lagrangian to an example of problem (P1), where $\phi_L(p) = Q$ with several possible choices of Q $(Q_1, Q_2, \ldots, Q_6)$ and the optimal value is zero. Alternatively p^* must maximise any increasing function of Q. We consider some such transformations choosing

them such that the maximum of the revised criterion is still zero e.g. $(e^Q - 1)$ or $[-log(1-Q)]$.

The reason for this is that we still must ensure the constraint $g(p^*) = 0$. We deal with this by transformation to an optimisation problem too; namely p^* should maximise $G(p) = -[g(p)]^2$ subject to $p_j \geq 0$, $\sum p_j = 1$. So $G(p^*) = 0$. We have already noted in section 8.2.1 that we might solve problem (P1) with $\phi(p) = G(p)$ to check that the constraint can be satisfied.

Thus p^* should simultaneously maximise $G(p)$ and $Q(p)$ with a common maximum of zero if $Q(p^*) = 0$ subject to $p_j \geq 0$, $\sum p_j = 1$. The sharing of a common optimal value of zero means that we have transformed the original constrained optimisation problem to an example of (P1) with various choices of $\phi(p)$ namely

$$\begin{aligned}
\phi_L(p) &= \phi_1(p) = G(p) + Q(p) \\
\phi_L(p) &= \phi_2(p) = \alpha G(p) + (1-\alpha) Q(p), \quad 0 < \alpha < 1 \\
\phi_L(p) &= \phi_3(p) = \min\{G(p), Q(p)\}
\end{aligned}$$

Hence we can consider applying methods for solving problem (P1).

8.3 Example

First we illustrate the above theory by simple examples in the context of quadratic regression on three points : -1, 0, 2. Thus the design corresponds to the distribution

x	-1	0	2
$p(x)$	p_1	p_2	p_3

and the information matrix is

$$M(p) = \begin{bmatrix} 1 & E(X) & E(X^2) \\ E(X) & E(X^2) & E(X^3) \\ E(X^2) & E(X^3) & E(X^4) \end{bmatrix}$$

$$= \begin{bmatrix} 1 & -p_1 + 2p_3 & p_1 + 4p_3 \\ -p_1 + 2p_3 & p_1 + 4p_3 & -p_1 + 8p_3 \\ p_1 + 4p_3 & -p_1 + 8p_3 & p_1 + 16p_3 \end{bmatrix}$$

We focus on constraints concerning the element $(M^{-1}(p))_{23}$ of $M^{-1}(p)$. This is given by

$$\begin{aligned} C &= [(8p_3 - p_1) - (2p_3 - p_1)(4p_3 + p_1)]/det(M(p)) \\ &= [(p_2 - 1/2)^2 - 9(p_3 - 1/2)^2 + 2]/det(M(p)). \end{aligned} \tag{8.34}$$

Of course the numerator is quadratic in the weights. We consider first the constraint $C = 0$ i.e. taking $g(p) = C$ so that we want $Cov(\hat{\theta}_2, \hat{\theta}_3) = 0$. To produce plots we deal with this constraint and the '$\sum_j p_j = 1$' constraint by substitution to

transform the constrained optimisation problem to an unconstrained optimisation with respect to one weight, thereby ensuring $G(p) = 0$.

One possible pair of substitutions is

$$\begin{aligned} p_1 &= (\frac{1}{2} - p_2) \mp \frac{1}{3}\sqrt{(p_2 - \frac{1}{2})^2 + 2}, \\ p_3 &= \frac{1}{2} \pm \frac{1}{3}\sqrt{(p_2 - \frac{1}{2})^2 + 2}. \end{aligned}$$

Note that there can be two possible formulae given the quadratic nature of C. However, the nonnegativity constraints on p_1, p_2 and p_3 can eliminate one of these and can also limit the range of p_2. In fact only the following formulae are valid and there is no restriction on p_2 (i.e. $0 < p_2 < 1$).

$$\begin{aligned} p_1 &= (\frac{1}{2} - p_2) + \frac{1}{3}\sqrt{(p_2 - \frac{1}{2})^2 + 2}, \\ p_3 &= \frac{1}{2} - \frac{1}{3}\sqrt{(p_2 - \frac{1}{2})^2 + 2}. \end{aligned}$$

We now show plots of $\phi(p_2)$, $Q(p_2)$ for $\phi(p) = \text{logdet}(M(p))$, -tr$(M^{-1}(p))$, and for various choices of Q. First see figures 8.2, 8.3 and 8.4. Clearly these functions are unimodal and all are maximised at $p_2^* = 0.4925325$ with $Q(p_2^*) = 0$. Also in figures 8.7, 8.9, 8.14, 8.15, 8.16 and 8.17, all functions are unimodal and all are maximised at the same point above. There is clearly difficulty in attaining the first order conditions at least in respect of Q as figures 8.2, 8.4 and 8.9 show. The functions seem virtually non-differentiable on either side.

In figures 8.3, 8.8, 8.12, 8.13 and 8.15 it is clear that the functions are pretty flat near the optimum indicating that the convergence of iteration is very slow in that region.

In figures 8.5, 8.10 and 8.11 the functions have more than one local maximum including at the points $p_2 = 0$ and 1.

In figures 8.6 and 8.8 the functions are very flat over the whole region because of the numerically high values of Q at $p_2 = 0$ and 1. It seems that if we draw those plots over the region $0.2 < p_2 < 0.8$, we might have a better picture of the curvature of the functions.

Functions like $Q(p_2) = [R_2 - 1]$ and $Q(p_2) = [R_3 - 1]$ in figures 8.16 and 8.17 respectively change reasonably quickly near the optimum and hence are better choices than the others.

We also consider the constraint $\det(M(p))\,C = 2$ since this leads to linear constraints on the weights. It is simpler to substitute for p_1, p_2 in terms of p_3. Relevant formulae are

$$\begin{aligned} p_1 &= -2(2p_3 - 1), \quad p_2 = 3p_3 - 1 \quad \text{for} \quad \tfrac{1}{3} \le p_3 \le \tfrac{1}{2} \\ p_1 &= 2p_3 - 1, \qquad\quad p_2 = 2 - 3p_3 \quad \text{for} \quad \tfrac{1}{2} \le p_3 \le \tfrac{2}{3}. \end{aligned}$$

Plots of $\phi(p_3)$ and $Q(p_3)$ [figures 8.18, 8.19, 8.20 and 8.21] reveal that they have two maximal turning points, one in each of the regions $p_3 < 0.5$ and $p_3 > 0.5$. Further $Q(p_3)$ has a maximum of zero at both turning points, while $\phi(p_3)$ is maximised at the lower value of p_3.

8.4 Algorithms

We have transformed the original constrained optimisation problem (GP) to a problem of maximising two functions (G and Q) of the design weights simultaneously. They have a common maximum of zero which is simultaneously attained at the constrained optimal design weights. This means that established algorithms for finding optimising distributions can be considered. We consider the following algorithm.

$$p_j^{(r+1)} = p_j^{(r)} f(x_j^{(r)}) \Big/ \sum_{i=1}^{J} p_i^{(r)} f(x_i^{(r)}) \tag{8.35}$$

where $f(x)$ is a positive and strictly increasing function in x, and may depend on a free parameter δ.

$x = F^{\phi_L}$, where

$$\begin{aligned} \phi_L(p) &= G(p) + Q(p) && (8.36)\\ \text{or } & \alpha G(p) + (1-\alpha) Q(p), \quad 0 < \alpha < 1 && (8.37)\\ \text{or } & \min\{G(p), Q(p)\}. && (8.38) \end{aligned}$$

F^{ϕ_L} are directional derivatives for $\phi_L(p)$. In the latter case F^{ϕ_L} will be either F^G or F^Q depending on which function (between G and Q) is minimum.
F^G and F^Q are given by

$$\begin{aligned} F_j^G &= d_j^G - \sum_{i=1}^{J} p_i d_i^G \\ F_j^Q &= d_j^Q - \sum_{i=1}^{J} p_i d_i^Q \end{aligned}$$

where d_j^G and d_j^Q are first partial derivatives for the two functions G and Q respectively.

d_j^G are given by

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

d_j^Q depends on the choice of Q. We have explored various choices of Q. For $Q = -\underline{h}^T P^2 \underline{h}$,

$$\begin{aligned} d_j^Q &= \frac{\partial Q}{\partial p_j} \\ &= 2\left[\xi_3\xi_4 - \xi_1\xi_6\right]\frac{\partial \xi_1}{\partial p_j} + 2\left[\xi_1\xi_4 - \xi_3\xi_5\right]\frac{\partial \xi_3}{\partial p_j} + 2\xi_1\xi_3\frac{\partial \xi_4}{\partial p_j} - \xi_3^2\frac{\partial \xi_5}{\partial p_j} - \xi_1^2\frac{\partial \xi_6}{\partial p_j}. \end{aligned}$$

For $Q = -\underline{h}^T \underline{h}$,

$$\begin{aligned} d_j^Q &= \frac{\partial Q}{\partial p_j} \\ &= 2\xi_1\xi_3\frac{\partial \xi_1}{\partial p_j} - \xi_3^2\frac{\partial \xi_2}{\partial p_j} - \left[\xi_1^2 + \xi_2\xi_3\right]\frac{\partial \xi_3}{\partial p_j} \end{aligned}$$

where $\xi_1, \xi_2, \ldots, \xi_6$ are given in (8.18) and (8.25), and

$\frac{\partial \xi_i}{\partial p_j}$'s are given by

$$\frac{\partial \xi_1}{\partial p_j} = \sum_{i=1}^{J} \left[F_i^{\phi} \frac{\partial F_i^g}{\partial p_i} + F_i^g \frac{\partial F_i^{\phi}}{\partial p_i} \right]$$

$$\frac{\partial \xi_2}{\partial p_j} = 2 \sum_{i=1}^{J} \left[F_i^{\phi} \frac{\partial F_i^{\phi}}{\partial p_i} \right]$$

$$\frac{\partial \xi_3}{\partial p_j} = 2 \sum_{i=1}^{J} \left[F_i^g \frac{\partial F_i^g}{\partial p_i} \right]$$

$$\frac{\partial \xi_4}{\partial p_j} = 2 p_j F_j^g F_j^{\phi} + \sum_{i=1}^{J} p_i^2 \left[F_i^{\phi} \frac{\partial F_i^g}{\partial p_i} + F_i^g \frac{\partial F_i^{\phi}}{\partial p_i} \right]$$

$$\frac{\partial \xi_5}{\partial p_j} = 2 p_j (F_j^{\phi})^2 + 2 \sum_{i=1}^{J} \left[p_i^2 F_i^{\phi} \frac{\partial F_i^{\phi}}{\partial p_i} \right]$$

$$\frac{\partial \xi_6}{\partial p_j} = 2 p_j (F_j^g)^2 + 2 \sum_{i=1}^{J} \left[p_i^2 F_i^g \frac{\partial F_i^g}{\partial p_i} \right],$$

$$\frac{\partial F_j^{\phi}}{\partial p_i} = \frac{\partial^2 \phi}{\partial p_i \partial p_j} - \left[\frac{\partial \phi}{\partial p_i} + \sum_{t=1}^{J} p_t \frac{\partial^2 \phi}{\partial p_t p_i} \right]$$

and

$$\frac{\partial F_j^g}{\partial p_i} = \frac{\partial^2 g}{\partial p_i \partial p_j} - \left[\frac{\partial g}{\partial p_i} + \sum_{t=1}^{J} p_t \frac{\partial^2 g}{\partial p_t p_i} \right].$$

$\frac{\partial^2 g}{\partial p_i \partial p_j}$ are given by

$$\frac{\partial^2 g}{\partial p_i \partial p_j} = 2 \underline{v}_j^T M^{-1}(p)(\underline{a} + \underline{b})(\underline{a} - \underline{b})^T M^{-1}(p) \underline{v}_i \underline{v}_i^T M^{-1}(p) \underline{v}_j.$$

Derivatives[1] $\frac{\partial \phi}{\partial p_j}$ and $\frac{\partial^2 \phi}{\partial p_i \partial p_j}$ depend on the choice of the criterion ϕ.

For A-optimality,

$$\frac{\partial \phi}{\partial p_j} = \underline{v}_j^T M^{-1} A^T A M^{-1} \underline{v}_j$$

$$\frac{\partial^2 \phi}{\partial p_i \partial p_j} = -2 \left[\underline{v}_j^T M^{-1}(p) A^T A M^{-1}(p) \underline{v}_i\right] \left[\underline{v}_j^T M^{-1}(p) \underline{v}_i\right],$$

and for D_A-optimality,

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

$$\frac{\partial^2 \phi}{\partial p_i \partial p_j} = \left[\underline{v}_j^T M^{-1}(p) A^T (A M^{-1}(p) A^T)^{-1} A M^{-1}(p) \underline{v}_i\right]^2 - 2 \left[\underline{v}_j^T M^{-1}(p) A^T (A M^{-1}(p) A^T)^{-1} A M^{-1}(p) \underline{v}_i\right] \left[\underline{v}_j^T M^{-1}(p) \underline{v}_i\right].$$

$\frac{\partial g}{\partial p_i}$, $\frac{\partial \phi}{\partial p_i}$, $\frac{\partial^2 g}{\partial p_j p_i}$ and $\frac{\partial^2 \phi}{\partial p_j p_i}$ are the first and second order partial derivatives for the function $g(p)$ and the design criterion $(\phi(p))$.

In the next section we report some empirical results, but before that we summarise the choices of the various terms involved.

$\phi(p)$:

The choices of $\phi(p)$ (the criterion to be maximised) are

$$\phi(p) = log\, det\{M(p)\} \qquad [D\text{-optimality}]$$

$$\phi(p) = -log\, det\{A M^{-1}(p) A^T\} \qquad [D_A\text{-optimality}]$$

$$\phi(p) = -tr\{A M^{-1}(p) A^T\} \qquad [A\text{-optimality}]$$

[1] These derivatives are derived in appendix B

$g(p)$:

We consider only:

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

$\phi_L(p)$:

The preferred choice of $\phi_L(p)$ is $min\,\{G(p), Q(p)\}$. A danger with the others, such as $\phi_L(p) = G(p) + Q(p)$, is that it might be possible to have $F_j^G + F_j^Q = 0\ \forall j$ without achieving the necessary conditions of zero values for all F_j^G and for all F_j^Q when all weights are positive.

$G(p)$:

We consider $G(p) = \text{-}g^2(p)$.

Note that we could consider $G(p) = [-g(p)]^{2k}$, k being a positive interger.

$Q(p)$:

We have considered a variety of choices of $Q(p)$. Here is a list of the choices considered in the next section where we report empirical results.

$$\begin{aligned}
Q = Q_1 &= -\left[\underline{h}^T \underline{h}\right] \\
Q = Q_2 &= [R_1 - 1] \\
Q = Q_3 &= -\left[\underline{h}^T P^2 \underline{h}\right] \\
Q = Q_4 &= [R_2 - 1] \\
Q = Q_5 &= -\left[[\underline{h}(\underline{E})]^T\,[\underline{h}(\underline{E})]\right] \\
Q = Q_6 &= [R_3 - 1] \\
Q = Q_7 &= -[1 - R_1]^2 \\
Q = Q_8 &= -[1 - R_2]^2 \\
Q = Q_9 &= -[1 - R_3]^2 \\
Q = Q_{10} &= -[ln(1 - Q_5)]^2 \\
Q = Q_{11} &= -[ln(1 - Q_3)]^2 \\
Q = Q_{12} &= -[1 - exp(Q_5)]^2
\end{aligned}$$

where

$$
\begin{aligned}
R_1 &= \frac{\xi_1^2}{\xi_2 \xi_3} \\
R_2 &= \frac{2\,\xi_1 \xi_3 \xi_4}{\xi_3^2 \xi_5 + \xi_1^2 \xi_6} \\
R_3 &= \frac{\xi_4^2}{\xi_5 \xi_6}
\end{aligned}
$$

and

$$
\begin{aligned}
\xi_1 &= \left[(\underline{F}^g)^T \underline{F}^\phi\right] \\
\xi_2 &= \left[(\underline{F}^\phi)^T \underline{F}^\phi\right] \\
\xi_3 &= \left[(\underline{F}^g)^T \underline{F}^g\right] \\
\xi_4 &= \left[(\underline{F}^g)^T P^2 \underline{F}^\phi\right] \\
\xi_5 &= \left[(\underline{F}^\phi)^T P^2 \underline{F}^\phi\right] \\
\xi_6 &= \left[(\underline{F}^g)^T P^2 \underline{F}^g\right].
\end{aligned}
$$

Note the maximum of all choices of Q is zero. The reason for using various choices of Q is that the convergence of algorithm (8.35) is likely to depend on curvature of the function Q which we saw to vary from one choice to another in plots of Q.

$f(.)$:

In algorithm (8.35) we consider $x{=}F^{\phi_L}$, where $\phi_L(p) = \min\{G(p), Q(p)\}$. As F^G and F^Q can have negative values, we have to consider choices of $f(x)$ which are defined for both positive and negative values. To serve this purpose we first consider two choices of $f(x)$, namely, $f(x) = exp(\delta x)$ and $f(x) = (1 + sx)^{s\delta}$, $s = \text{sign}(x)$.

We also consider

$$f(x) = \begin{cases} h(x) & \text{for } x < 0 \\ 2 - h(-x) & \text{for } x > 0. \end{cases}$$

This type of function was considered in section 6.2 of chapter 6. We take $h(x) = (1-x)^{\delta}$.

8.5 Empirical Results

8.5.1 Quadratic Model

First for the above quadratic model we consider deriving the D-optimal design subject to the constraint $g(p) = \underline{a}^T M^{-1}(p)\,\underline{a} - \underline{b}^T M^{-1}(p)\,\underline{b}$, where $\underline{a} = (0,1,1)^T$ and $\underline{b} = (0,1,-1)^T$. Note that this constraint is equivalent to $Cov(\hat{\theta}_2, \hat{\theta}_3) = 0$.

(i) At equal initial weights with $Q = Q_7 = -[1 - R_1]^2$,

$$\begin{aligned} (G,Q) &= (-28.4444, -1.0) \\ -67.5556 &\leq F_i^G \leq 39.1111 \\ -83508 &\leq F_i^Q \leq 144240. \end{aligned}$$

After 200 iterations starting from this design with $f(x) = \exp(\delta x)$, $x = F$ and $\delta = 0.001$, we converge to the design above i.e. $p_2^* = 0.4925325$ and the new values are

$$\begin{aligned} (G,Q) &= (-2.30\,10^{-8}, -9.48\,10^{-7}) \\ -0.0012280 &\leq F_i^G \leq 0.0415182 \\ -0.0640491 &\leq F_i^Q \leq 0.0027849. \end{aligned}$$

These derivatives could be closer to zero. This illustrates that while G and

various choices of Q can be approximately zero, directional derivatives can be distinctly nonzero.

(ii) At initial weights $p^{(0)} = (0.4, 0.5, 0.1)$ with $Q = Q_6 = [R_3 - 1]$,

$$\begin{aligned} (G, Q) &= (-9.67901, -0.193743) \\ -88.4938 &\leq F_i^G \leq 15.2099 \\ -3.62321 &\leq F_i^Q \leq 4.92886. \end{aligned}$$

After 700 iterations starting from this design with $f(x) = \exp(\delta x)$, $x = F$ and $\delta = 0.001$,

$$\begin{aligned} (G, Q) &= (0, 0) \\ -0.0093914 &\leq F_i^G \leq 0.0002848 \\ -0.0002144 &\leq F_i^Q \leq 0.0002176. \end{aligned}$$

(iii) At initial weights $p^{(0)} = (0.3, 0.3, 0.4)$ with $Q = Q_6 = [R_3 - 1]$,

$$\begin{aligned} (G, Q) &= (-36.2226, -0.0063) \\ -80.8042 &\leq F_i^G \leq 61.2997 \\ -0.823343 &\leq F_i^Q \leq 0.803477. \end{aligned}$$

After 900 iterations starting from this design with $f(x) = \exp(\delta x)$, $x = F$ and $\delta = 0.001$,

$$\begin{aligned} (G, Q) &= (0, 0) \\ -0.0000047 &\leq F_i^G \leq 0.0000014 \\ -0.0000051 &\leq F_i^Q \leq 0.0000052. \end{aligned}$$

Clearly the latter results are better than the previous two. We almost attain F_i^G and F_i^Q to be zero with also zero values of the functions G and Q.

8.5.2 Further Examples

We further consider three examples considered by Silvey et al (1978), Wu (1978) and Alahmadi (1993). These are examples 1, 2 and 3 considered in chapters 5 and 6. The examples are defined by their design spaces.

Example - 1: $\mathcal{V} = \mathcal{V}_1 = \{(1,-1,-1)^T, (1,-1,1)^T, (1,1,-1)^T, (1,2,2)^T\}$

Example - 2: $\mathcal{V} = \mathcal{V}_2 = \{(1,-1,-1)^T, (1,-1,1)^T, (1,1,-1)^T, (1,2,3)^T\}$

Example - 3: $\mathcal{V} = \mathcal{V}_3 = \{(1,-1,-2)^T, (1,-1,1)^T, (1,1,-1)^T, (1,2,2)^T\}$

A-optimality

First we consider the design criterion $\phi(p) = -tr(AM^{-1}(p)A^T)$ where $A = [\underline{a}, \underline{b}]^T$, while the constraint is $g(p) = \underline{a}^T M^{-1}(p)\,\underline{a} - \underline{b}^T M^{-1}(p)\,\underline{b}$, $\underline{a} = (1,0,1)^T$ and $\underline{b} = (1,0,-1)^T$. Here also note that this constraint is equivalent to $Cov(\hat{\theta}_1, \hat{\theta}_3) = 0$. Alahmadi (1993) took $\phi_L(p)$=$\phi_1(p)$. He found convergence to be slow too with $Q = Q_7 = -[1 - R_1]^2$, $f(x)$=$(1 + sx)^{s\delta}$, $x = d$, s=sign(x), $\delta = 0.3$, 0.08, 0.2 for examples 1, 2 and 3 respectively. Approximate solutions seemed to be (0.237,0.270,0.330,0.163), (0.259,0.230,0.359,0.152) and (0.255,0.355,0.215,0.175) for examples 1, 2 and 2 respectively.

We have endeavoured to improve on this with a view to accommodating zero optimal weights. We took $\phi_L(p)$=$\phi_3(p)$ since there is a danger of converging to zero values for $\underline{F}^L = \underline{F}^G + \underline{F}^Q$ but not for $\underline{F}^G$, $\underline{F}^Q$.

In example-1 starting from equal initial weights with $Q = Q_{12} = -[1 - exp(Q_5)]^2$, $f(x) = (1+sx)^{s\delta}$, s=sign(x) and $x = F$, the pair (G, Q) change from (-0.1773,-0.1527)

to (-0.0402, -0.0512) after 700 iterations. Yet at this point

$$\begin{aligned} -1.39321 &\leq F_i^G \leq 2.04547 \\ -1.88422 &\leq F_i^Q \leq 2.01673. \end{aligned}$$

To explore this issue further, in the remaining illustrations, we take $p^{(0)}$ to be the approximate optimal solution of Alahmadi (1993). We denote this by $p^{(A)}$.

(i) In example-1, at initial weights $p^{(A)} = (0.237,\ 0.270,\ 0.330,\ 0.163)$ with $Q = Q_{11} = -[ln(1 - Q_3)]^2$,

$$\begin{aligned} (G, Q) &= (-5.2\,10^{-8},\ -8.9\,10^{-4}) \\ -0.00170 &\leq F_i^G \leq 0.00139 \\ -38.7634 &\leq F_i^Q \leq 81.6632. \end{aligned}$$

After 450 iterations starting from this design with $h(x) = (1-x)^{-\delta}$, $x = F$, and $\delta = 0.001$,

$$\begin{aligned} (G, Q) &= (-0.000016,\ -0.000007) \\ -0.00254 &\leq F_i^G \leq 0.00198 \\ -0.04262 &\leq F_i^Q \leq 0.00019. \end{aligned}$$

(ii) In example-1 again, at a variation of $p^{(A)}$, when $Q = Q_{12} = -[1 - exp(Q_5)]^2$,

$$\begin{aligned} (G, Q) &= (-0.000087,\ 0) \\ -0.05678 &\leq F_i^G \leq 0.06889 \\ F_i^Q &= 0. \end{aligned}$$

After 500 iterations with $f(x) = exp(\delta x)$, $x = F$, and $\delta = 0.001$,

$$\begin{aligned} (G, Q) &= (0,\ 0) \\ -0.000026 &\leq F_i^G \leq 0.000031 \\ F_i^Q &= 0. \end{aligned}$$

(iii) In example-2, at $p^{(0)} = p^{(A)}$ with $Q = Q_7 = -[1 - R_1]^2$,

$$\begin{aligned} (G, Q) &= (-0.0000102, -0.0000001) \\ -0.01621 &\leq F_i^G \leq 0.02209 \\ -0.08275 &\leq F_i^Q \leq 0.17188. \end{aligned}$$

After 700 iterations starting from this design with $f(x) = exp(\delta x)$, $x = F$, and $\delta = 0.001$,

$$\begin{aligned} (G, Q) &= (0, 0) \\ -0.00013 &\leq F_i^G \leq 0.00054 \\ -0.00409 &\leq F_i^Q \leq 0.00084. \end{aligned}$$

(iv) In example-3, at $p^{(0)} = p^{(A)}$ with $Q = Q_{10} = -[ln(1 - Q_5)]^2$,

$$\begin{aligned} (G, Q) &= (-0.000005, 0) \\ -0.0134 &\leq F_i^G \leq 0.0121 \\ -0.000003 &\leq F_i^Q \leq 0.000001. \end{aligned}$$

After 400 iterations starting from this design with $f(x) = exp(\delta x)$, $x = F$, and $\delta = 0.001$,

$$\begin{aligned} (G, Q) &= (0, 0) \\ -0.00016 &\leq F_i^G \leq 0.00014 \\ F_i^Q &= 0. \end{aligned}$$

From the above results it is clear that the F_i^Q attain zero values quicker than the F_i^G, especially in examples 1 and 3. So convergence is slow. However, in all cases the functions G and Q almost attain the maximum value of zero.

D_A-optimality

Now we report some results on D_A-optimality, i.e., the design criterion is $\phi(p) = -log\, det\{AM^{-1}(p)A^T\}$, where $A = [\underline{a}, \underline{b}]^T$, and the constraint is same as above, i.e., $g(p) = \underline{a}^T M^{-1}(p)\,\underline{a} - \underline{b}^T M^{-1}(p)\,\underline{b}$, $\underline{a} = (1, 0, 1)^T$ and $\underline{b} = (1, 0, -1)^T$.

(i) In example-1 starting from equal initial weights with $Q = Q_6 = [R_3 - 1]$,

$$\begin{aligned} (G, Q) &= (-0.177285, -0.975964) \\ -3.52704 \leq\ & F_i^G \leq 3.56437 \\ -1.70462 \leq\ & F_i^Q \leq 1.47282. \end{aligned}$$

After 1200 iterations starting from this design with $f(x) = exp(\delta x)$, $x = F$, and $\delta = 0.001$,

$$\begin{aligned} (G, Q) &= (0, 0) \\ -0.0000165 \leq\ & F_i^G \leq 0.0000139 \\ -0.0000134 \leq\ & F_i^Q \leq 0.0000366. \end{aligned}$$

(ii) In example-3, at equal initial weights, with $Q = Q_6 = [R_3 - 1]$,

$$\begin{aligned} (G, Q) &= (-0.096522, -0.901059) \\ -1.97729 \leq\ & F_i^G \leq 2.80194 \\ -1.89622 \leq\ & F_i^Q \leq 2.09638. \end{aligned}$$

After 1300 iterations starting from this design with $f(x) = exp(\delta x)$, $x = F$, and $\delta = 0.01$,

$$\begin{aligned} (G, Q) &= (0, 0) \\ -0.00072 \leq\ & F_i^G \leq 0.00005 \\ -0.00052 \leq\ & F_i^Q \leq 0.00028. \end{aligned}$$

In the above cases as we started with equal initial weights the iteration is

naturally slow. Some of the values of F_i^G and F_i^Q are zero to only 3 decimal places. However, the functions G and Q attain the maximum value zero.

8.5.3 Optimal Solution and Discussion

We now report, for the quadratic regression model on the design space {-1, 0, 2} and for examples 1, 2 and 3, the values of the weights to which convergence was obtained for the choices of $f(.)$ and δ considered.

Quadratic Regression : D-optimality : {0.478878, 0.492533, 0.028589}

A-optimality:

Example - 1: {0.237469, 0.270394, 0.329606, 0.162531}

Example - 2: {0.258470, 0.230047, 0.358880, 0.152603}

Example - 3: {0.254785, 0.355270, 0.214827, 0.175118}

D_A-optimality:

Example - 1: {0.0, 0.192822, 0.578465, 0.228713}

Example - 3: {0.514191, 0.257095, 0.0, 0.228714}

Note that we have a zero optimal weight in these two D_A-optimal designs, a scenario for which the methods of this chapter were designed to cater.

Discussion:

There is clearly difficulty in attaining the first order conditions at least in respect of Q. This is not surprising in view of the plots of Q in figures 8.2, 8.4 and 8.9. The functions seem virtually non-differentiable at the optimum while being convex with steep derivatives on either side.

Convergence was slow in all examples although G and Q can often be both close to zero despite in some cases unsatisfactory attainment of first order conditions. The directional derivatives could be closer to zero. This indicates that these derivatives could be nonzero while G and Q could be approximately zero.

However, in the case of the quadratic model with $Q = [R_3 - 1]$, $f(x) = \exp(\delta x)$ and $\delta = 0.001$ for D-optimality; in the case of example-1 with $Q = -[1 - exp(Q_5)]^2$, $f(x) = \exp(\delta x)$ and $\delta = 0.001$ for A-optimality; and in the case of example-1 with $Q = [R_3 - 1]$, $f(x) = exp(\delta x)$ and $\delta = 0.001$ we almost attain zero values for F_i^G and F_i^Q as well as for the functions G and Q, although in the latter case it takes 1200 iterations compared to 900 and 500 iterations in the first two cases respectively to achieve optimality conditions. Among the different choices of Q, it seems that $Q = Q_6 = [R_3 - 1]$ is better than the others. This is consistent with the curvature of this function as figures 8.7, 8.17 illustrate. The function changes reasonably quickly near its optimum.

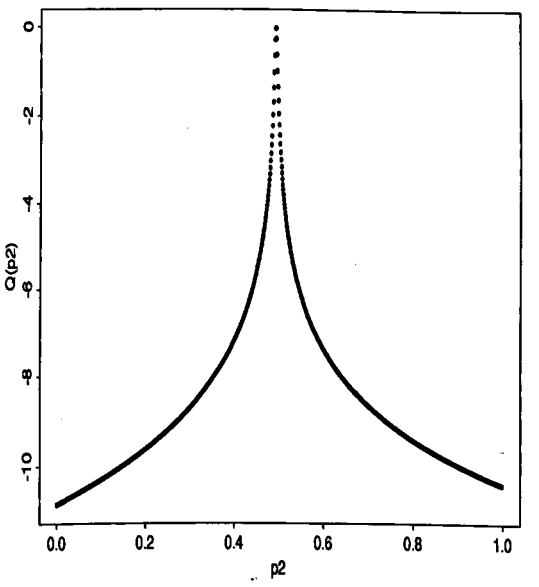


Figure 8.2. *D*-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = -\left[ln(1 + h^T P^2 h)\right]$.

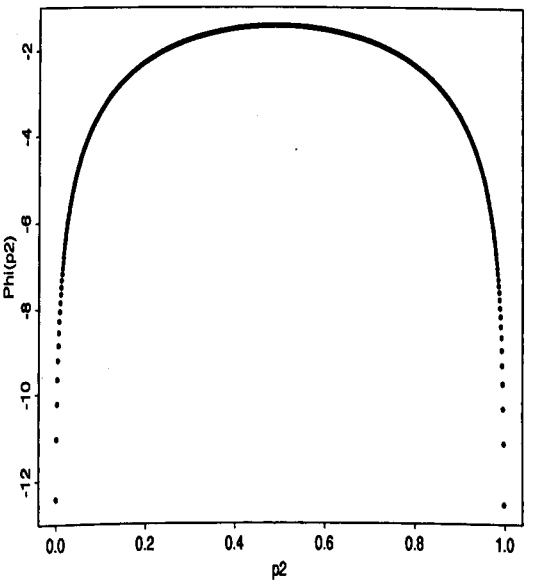


Figure 8.3. *D*-optimality: $\phi(p_2)$ $[= \phi(p)]$ versus p_2.

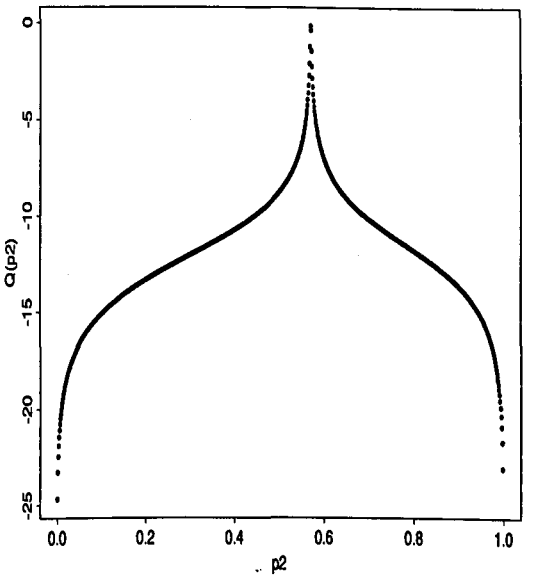


Figure 8.4. *A*-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = -[ln(1 + h^T P^2 h)]$.

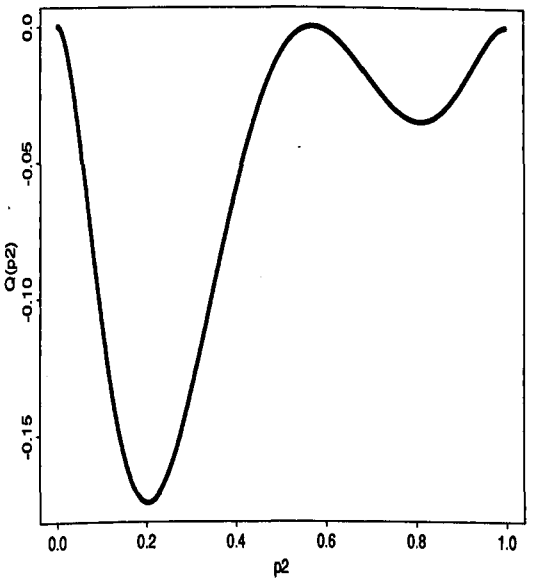


Figure 8.5. *A*-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = -[h(E)]^T [h(E)]$.

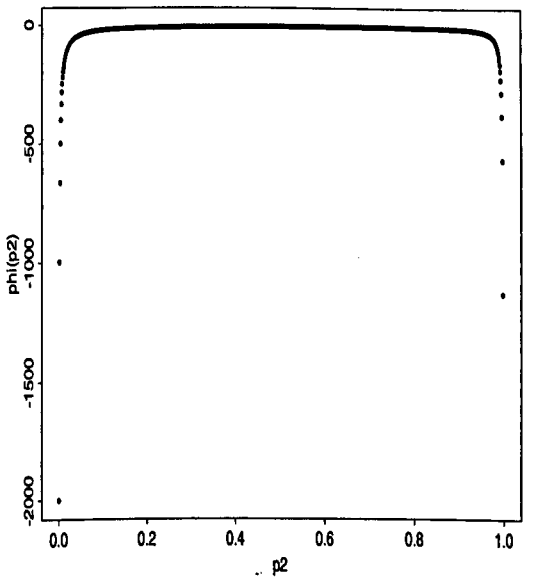


Figure 8.6. A-optimality: $\phi(p_2)$ $[= \phi(p)]$ versus p_2.

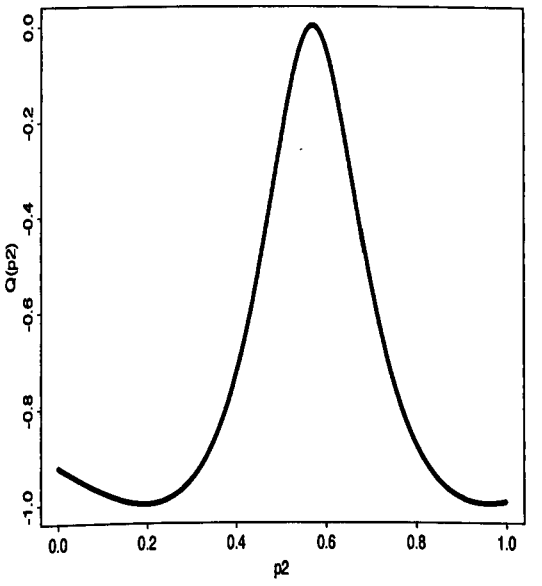


Figure 8.7. A-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = [R_3 - 1]$.

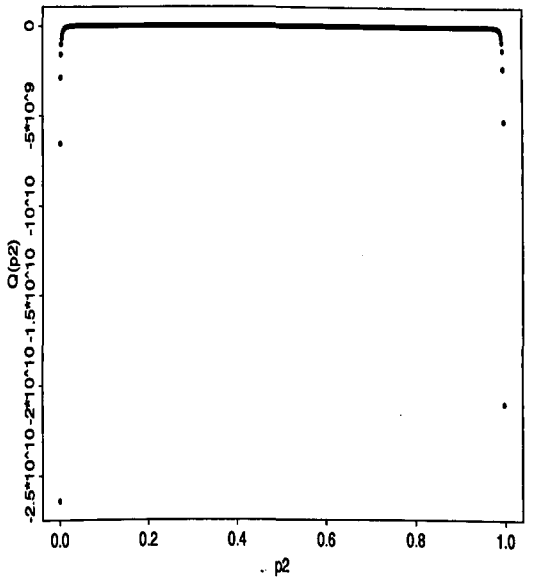


Figure 8.8. *D*-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = -[h^T h]$.

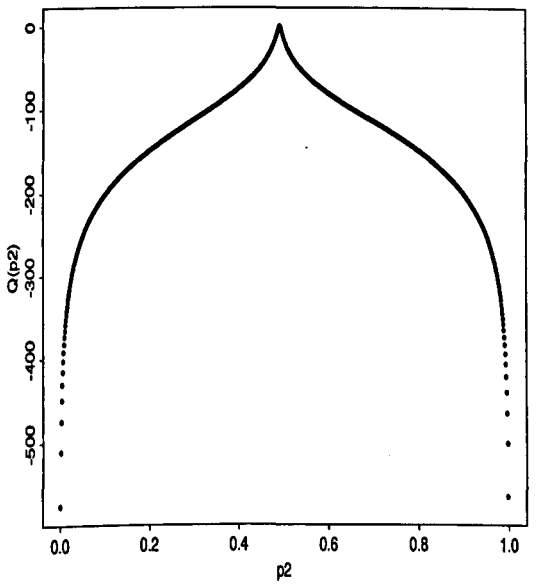


Figure 8.9. *D*-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = -[\log(1 + h^T h)]^2$.

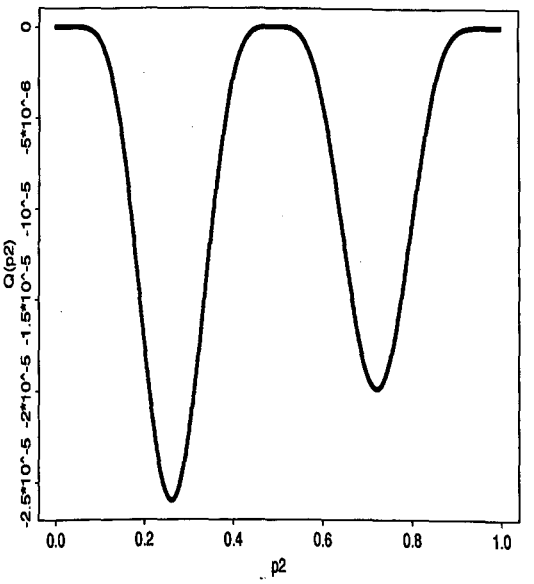


Figure 8.10. *D*-optimality: $Q(p_2)$ versus p_2, where $Q(p_2)$ = -[log{1 + $h(E)^T h(E)$}]^2.

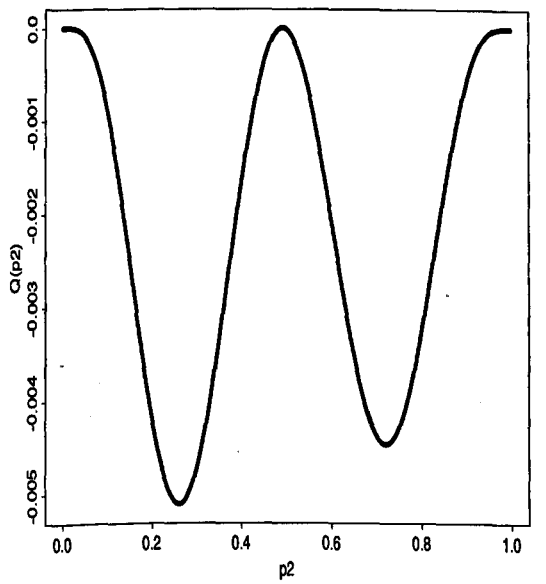


Figure 8.11. *D*-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = -[h(E)]^T [h(E)]$.

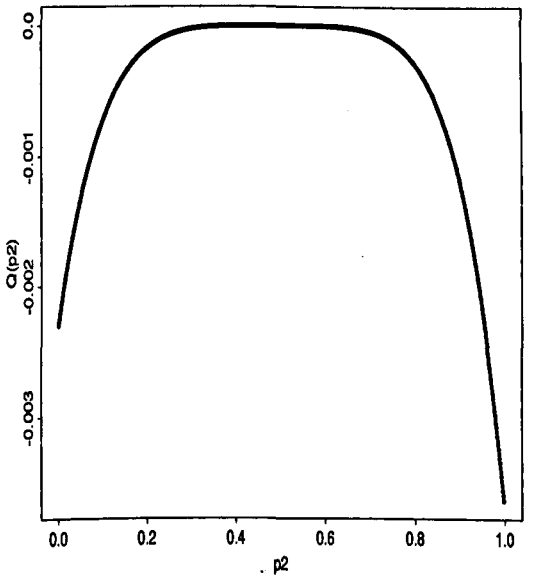


Figure 8.12. *D*-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = -[1 - R_1]^2$.

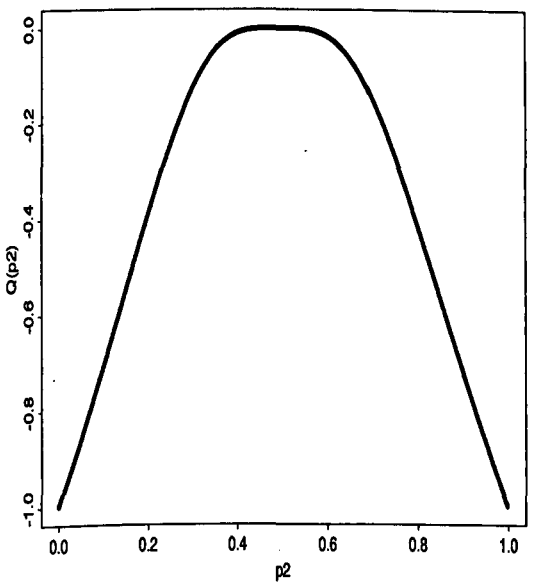


Figure 8.13. *D*-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = -[1 - R_2]^2$.

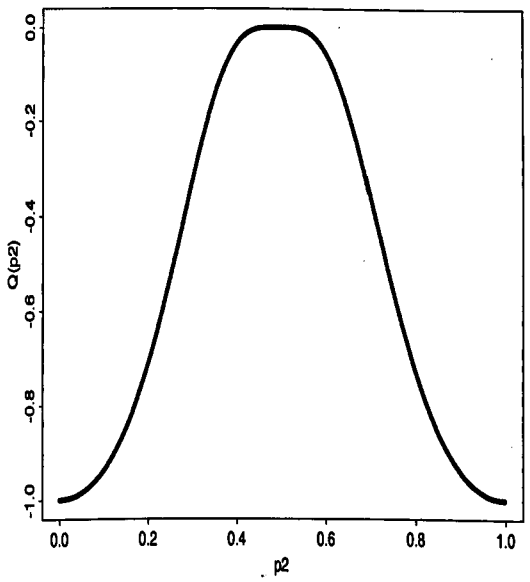


Figure 8.14. *D*-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = -[1 - R_3]^2$.

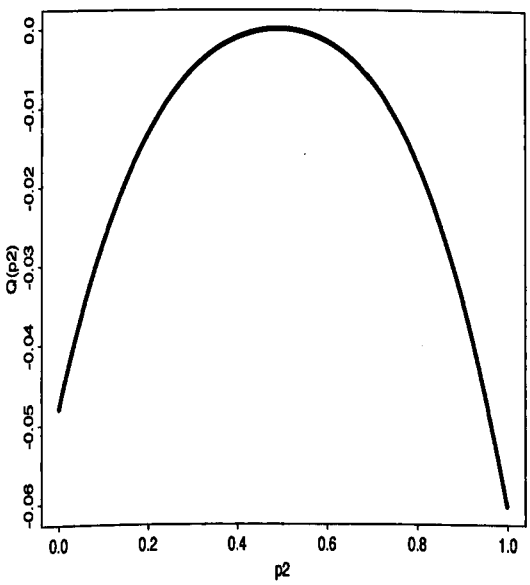


Figure 8.15. *D*-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = Q_2 = [R_1 - 1]$

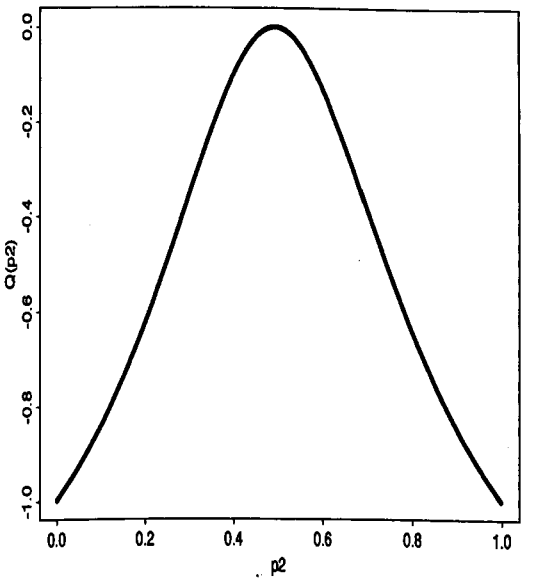


Figure 8.16. *D*-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = [R_2 - 1]$.

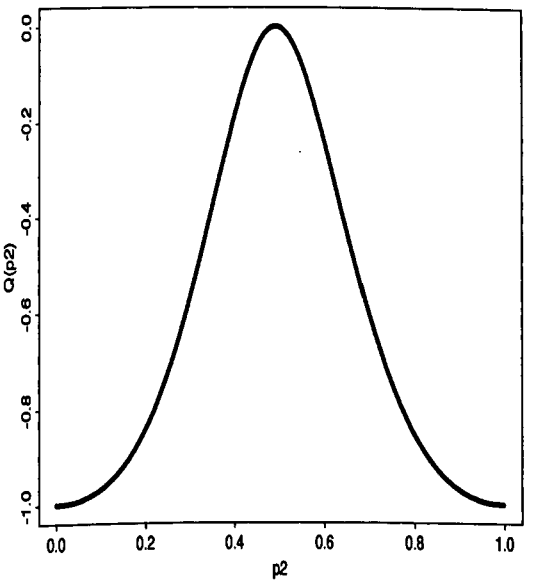


Figure 8.17. *D*-optimality: $Q(p_2)$ versus p_2, where $Q(p_2) = [R_3 - 1]$.

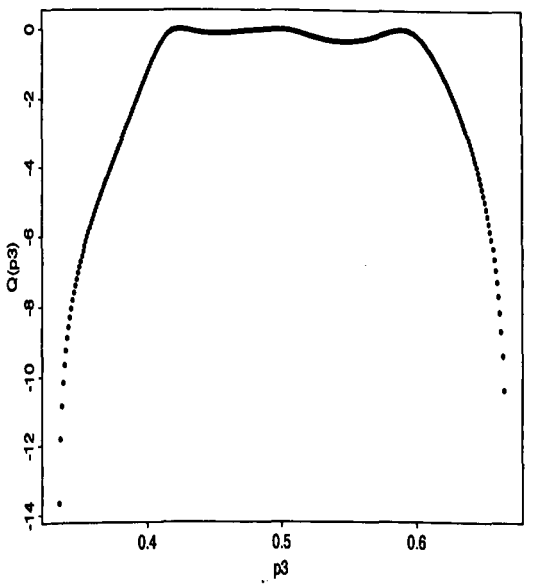


Figure 8.18. *D*-optimality: $Q(p_3)$ versus p_3, where $Q(p_3) = -[ln(1 + h^T P^2 h)]$.

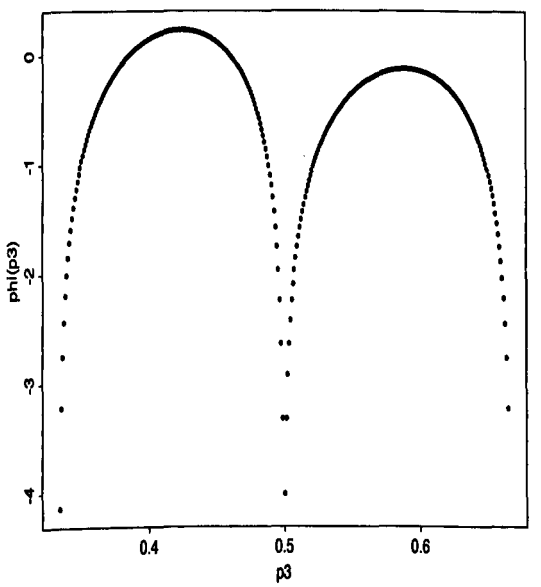


Figure 8.19. *D*-optimality: $\phi(p_3)$ versus p_3.

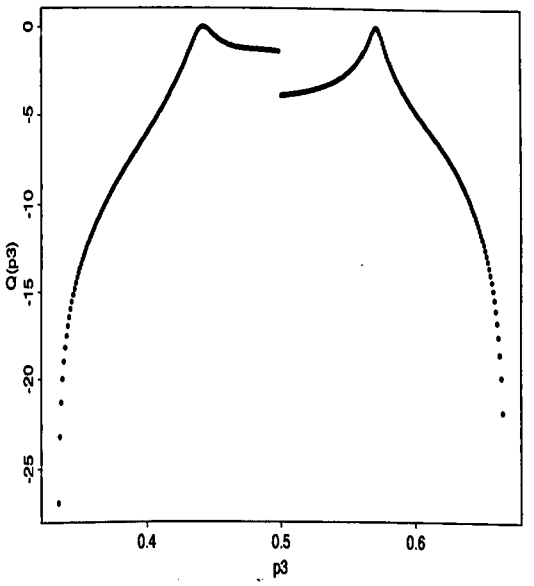


Figure 8.20. A-optimality: $Q(p_3)$ versus p_3, where $Q(p_3) = -[ln(1 + h^T P^2 h)]$.

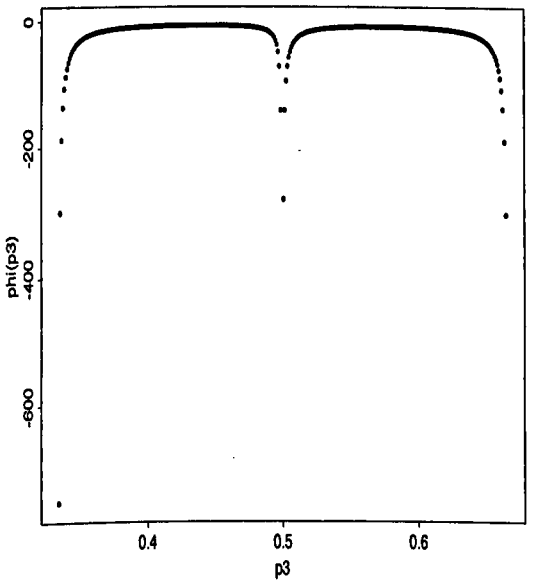


Figure 8.21. A-optimality: $\phi(p_3)$ versus p_3.

Chapter 9

Discussions and Future Work

9.1 Discussions - A Brief Review

This thesis has considered constructing optimising distributions with applications in estimation and optimal design by exploring a class of multiplicative algorithms (3.1), indexed by a function $f(.)$ which satisfies some conditions (positive and strictly increasing) and depends on one or more free parameters. Although each chapter has contained within it discussions of the issues raised and the conclusions drawn, it is useful to summarise the whole work together.

First we provided some basic introduction to the area of linear optimum design theory including properties of the information matrix and some design criteria. In chapter 2 we considered optimality conditions based on directional derivatives ($F\{p, q\}$) which play a basic role in our work. The properties of these derivatives were studied including those when the differentiability of the criterion $\phi(.)$ or $\psi(.)$ is defined.

In chapter 3 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 $F_j^{(r)}$ and $d_j^{(r)} = \partial\phi/\partial p_j$ while $F_j^{(r)} = d_j^{(r)} - \sum p_i^{(r)} d_i^{(r)}$ at $p = p^{(r)}$ and $f(.)$ satisfies some suitable properties and may depend on one or more free parameters. These iterations neatly submit to the constraints of problems (P1) and (P2). Properties of this algorithm were discussed.

Chapter 4 was devoted to the problem of determining maximum likelihood estimates under a hypothesis of marginal homogeneity for data in a square contingency table. This estimation problem could be transformed to an example of problem (P2). A particular example of data considered for this is a grading of the unaided distance vision of each eye of 3242 men (Placket (1974), p.77-79). We considered both a 3×3 and a 4×4 table derived from this data.

In chapter 5 we considered constructing optimal designs with some preliminary choices of $f(.)$ [in algorithm (3.1)]. These investigations also explored changing the argument of $f(.)$ from d_j to $[d_j - c]$ and F_j, where c is a suitably chosen constant. Convergence of the above algorithm seemed to be faster when choosing the argument to be F_j. Also convergence is improved by considering a class of functions $[H_\beta(F)]$ based on a distribution function $G(\delta F)$. The reason for good convergence rates using this function is that it changes at a reasonable rate at zero noting that $\sum_j p_j F_j = 0$, since $F_j = d_j - \sum_i p_i d_i$.

In chapter 6 we tried to improve convergence of algorithm (3.1) by introducing two approaches - approach I and approach II. Convergence is improved by considering both of these. The choices of $h(.)$ [in approach I] and $g(.)$ [in approach II] give almost the same number of iterations within each eaxmple. Convergence rates

also depend on the choice of δ. Overall for all choices of $h(.)$ and $g(.)$ [except for $g(x) = ln(e+\delta(x-1))$] the best values of δ (written in bold font inside the tables) seem to be somewhere between 1.5 and 2.0. Values higher than those presented in the tables were also investigated.

But our most surprising result occurred with the 'clustering approach' of chapter 7. This idea emerges if we run algorithm (3.1) in a design space which is a discretisation of a continuous space. If we produce plots of the weights and the variance function versus design points, we see that 'clusters' start forming in early iterations of the above algorithm. The idea was that, at an appropriate iterate $p^{(r)}$, the weights p_i are transformed to weights within clusters and total cluster weights. These are then found optimally by using algorithm (7.25). For each of the regression models considered, iteration results clearly demonstrate that, by using this approach convergence is much improved over using the raw form of algorithm (3.1).

In chapter 8 we considered the problem of finding a constrained 'approximate' design : one maximising a criterion (D-, D_A- or A-optimality) subject to the equality of variances of the estimates of two parametric functions. We approached the problem by initially formulating the Lagrangian, but removing the Lagrange parameter through a substitution and transforming the first order Lagrangian conditions to an optimisation problem. Ensuring the constraint was also transformed to an optimisation problem. Denoting the criteria of these two problems by Q and G respectively these are such that they have a common maximum of zero which is simultaneously attained at the constrained optimal design weights. Convergence was slow in the examples considered although G and Q can often

be both close to zero despite in some cases the directional derivatives (F_i^G and F_i^Q) being somewhat nonzero. However, we almost attain these derivatives to be zero with also zero values of the functions G and Q in the case of a quadratic model with $Q = [R_3 - 1]$ and $f(x) = \exp(\delta x)$ for D-optimality; in the case of example-1 with $Q = -[1 - exp(Q_5)]^2$ and $f(x) = \exp(\delta x)$ for A-optimality; and in the case of example-1 with $Q = [R_3 - 1]$, $f(x) = exp(\delta x)$ for D_A-optimality.

9.2 Future Work

This section comprises of the following three main topics:

9.2.1 Estimation Problems

Firstly, consider the estimation problem of chapter 4. Though in the beginning of this chapter we provided some possible examples of this type of problem, we considered only the problem of maximum likelihood estimation subject to the hypothesis of marginal homogeneity in a square contingency table.

We considered only two cases : 3×3 and 4×4 contingency tables. Further work on this would be to consider a multi-way square contingency tables.

Other possible work would be to compute the cell probabilities under an independence model in a two-way classification table of two attributes when the data can be viewed as incomplete, i.e., some of the cell frequencies are missing. When there is no missing data maximum likelihood estimates of the cell probabilities can be

obtained explicitly. But, certainly numerical techniques are needed to compute ML estimates when there are missing cell frequencies. In that case some of the terms in the likelihood may be linear functions of several probabilities. It is then not possible to have an analytic solution.

9.2.2 Optimisation with respect to Several Distributions

Secondly, consider the clustering approach of chapter 7. This approach transformed a problem of optimisation with respect to one distribution to one of optimisation with respect to several distributions. There are several naturally arising examples of this problem to which an appropriate version of algorithm (7.25) could be applied. One example of this arises in image processing or image labelling. This proves to be an optimisation problem with respect to several probability distributions, one for each of the pixels or nodes. We first describe the following problem outlined by Torsney (1988), and then discuss some suitable applications of algorithm (3.1) to this problem.

Let a picture or a graph comprise of I pixels or I nodes indexed by $i = 1, 2, \ldots, I$ where each pixel or node must belong to one of J states labelled $j = 1, 2, \ldots, J$.

A common problem in this is to have incomplete information about the true state of each pixel. An estimate or an initial approximation can be obtained from a measurement subject to error when the states form a continuum. However, these individual labels do not together produce a consistent picture or complexion. Thus it is necessary to improve on this first assessment.

An approach would be to consider this problem by permitting a solution which expresses pixel-state preferences via a probability distribution across the labels for each pixel. This introduces a variable p_{ij} which is the probability that pixel i belongs to state j, satisfying the following constraints

$$\begin{aligned} p_{ij} &\geq 0, \quad i = 1,2,\ldots,I, \quad j = 1,2,\ldots,J \\ \sum_{j=1}^{J} p_{ij} &= 1, \quad i = 1,2,\ldots,I. \end{aligned}$$

Boyce et al (1987) propose that such a probability distribution be chosen to maximise

$$\phi(p) = \sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij}\, ln(q_{ij})$$

where

$$q_{ij} = \left(\sum_{u \in N(i)} c_u \left[\sum_{t=1}^{J} p_{ij|ut}\, p_{ut} \right]^{\beta} \right)^{\frac{1}{\beta}}$$

$N(i)$ = a set of neighbours of pixel i,

$p_{ij|ut}$ = the conditional probability that pixel i has label j given that pixel u has label t, which is assumed to be known, or for which an approximation is available,

c_u are weights on neighbours satisfying $c_u \geq 0$, $\sum_u c_u = 1$,

β is a free positive variable.

Faugeras and Berthod (1981) propose that q_{ij} is an approximation to p_{ij} that could be obtained from a current estimate of the joint distribution of the labels of pixel i and of its neighbours in $N(i)$.

Here a compromise measure between consistency and ambiguity is the function $\phi(p)$. Consistency and ambiguity are concerned with the deviations of p_{ij} from q_{ij} and of the distribution for a pixel from the classification of the pixel to a single label respectively. Potential measures of these are respectively

$$-\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij}\, ln\left(\frac{q_{ij}}{p_{ij}}\right) \quad \text{and} \quad -\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij}\, ln(p_{ij}),$$

and [$-\phi(p)$] is the sum of these.

If we are choosing the p_{ij} optimally we wish to find several optimising distributions. One natural extension of algorithm (7.25) would be the following:

$$\underline{p}_i^{(n+1)} = \underline{p}_i^{(n)} f_{p_i}(d^{p_i(n)}) \Big/ \sum_{j=1}^{J} p_{ij}^{(n)} f_{p_i}(d_j^{p_i(n)}), \quad i = 1, 2, \ldots, I \qquad (9.1)$$

where

$$\underline{p}_i = (p_{i1}, p_{i2}, \ldots, p_{iJ})^T$$

$$d_j^{p_i} = \frac{\partial \phi}{\partial p_{ij}}, \quad j = 1, 2, \ldots, J$$

and $\underline{p}_i^{(n)}$, $d^{p_i(n)}$ stand for the values at n^{th} iteration.

The above algorithm will have similar properties like those listed in section 7.5.2 in relation to maximising $\phi_m(\underline{q}, \underline{r}_1, \underline{r}_2, \ldots, \underline{r}_m)$. In particular, an optimal solution

is a fixed point of it.

The solution must be such that $p_{ij} = 1$ for $j = j_i'$, say and $p_{ij} = 0$, $j \neq j_i'$ for each i.

9.2.3 Constrained Optimisation Problems

Lastly consider the constrained optimisation problem of chapter 8. There is clearly difficulty in attaining the first order conditions at least in respect of Q. This is not surprising in view of the plots of some choices of Q, especially in figures 8.2, 8.4 and 8.9. These functions seem virtually non-differentiable at the optimum while being convex with steep derivatives on either side. These are hostile conditions for iterations (8.35).

Further development is needed. Possibly further transformations of Q to a more concave shape would help e.g.

$$-Q \;=\; -\tilde{Q} \;=\; \begin{cases} \{-Q\}^t & \text{for } -Q < 1 \\ \{-Q\}^{1/t} & \text{for } -Q > 1 \end{cases}$$

for some t.

Modifications using the clustering approach of Mandal and Torsney (2000b) (see chapter 7) to deal with zero optimal weights is another option.

There is also the issue of multiple local maxima to be considered. This and the extension to the case of several equality constraints, which in principle is straight forward, will be the focus of future work.

Appendix A

Interation Results Summarised in Chapter 7

Trigonometric Regression				
	Number of iterations needed to achieve $\max_{1\leq j\leq J}\{F_j\} \leq 10^{-n}$ for			
δ	n=1	n=2	n=3	n=4
0.5	14	114	988	8793
1.0	7	58	494	4397
1.4	6	42	353	3141
1.5	6	39	330	2931
1.6	6	37	309	2748
1.7	6	35	291	2586
1.8	6	33	275	2443
1.9	**8**	**31**	**261**	**2314**

Table A.1. Iteration Results : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$

Quadratic Regression				
	Number of iterations needed to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$ for			
δ	n=1	n=2	n=3	n=4
0.5	18	96	993	9990
1.0	9	49	497	4996
1.4	7	35	355	3569
1.5	7	33	332	3331
1.6	7	31	311	3123
1.7	6	29	293	2939
1.8	6	27	277	2776
1.9	**7**	**27**	**262**	**2630**

Table A.2. Iteration Results : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$

Cubic Regression				
	Number of iterations needed to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$ for			
δ	n=1	n=2	n=3	n=4
0.5	16	96	970	8079
1.0	9	49	486	4040
1.4	7	35	347	2886
1.5	6	33	324	2694
1.6	6	31	304	2525
1.7	6	29	286	2377
1.8	7	27	270	2245
1.9	**9**	**27**	**256**	**2127**

Table A.3. Iteration Results : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$

Quartic Regression				
	Number of iterations needed to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$ for			
δ	n=1	n=2	n=3	n=4
0.5	14	97	988	9768
1.0	8	49	495	4885
1.4	6	35	354	3489
1.5	6	33	330	3257
1.6	6	31	310	3053
1.7	6	30	291	2874
1.8	8	27	275	2714
1.9	**11**	**29**	**261**	**2571**

Table A.4. Iteration Results : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$

Second-Order Model				
	Number of iterations needed to achieve $\max_{1\le j\le J}\{F_j\} \le 10^{-n}$ for			
δ	n=1	n=2	n=3	n=4
0.5	22	197	782	1310
1.0	12	99	392	656
1.5	8	66	261	437
1.8	7	56	218	365
1.9	7	53	206	345
2.0	7	50	196	328
2.1	6	48	187	313
2.2	6	46	178	298
2.3	**6**	**44**	**171**	**286**

Table A.5. Iteration Results : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$

Trigonometric Regression								
	Number of iterations needed to achieve $\max_{1\le j\le m+J}\{F_j^{q,r_1,\ldots,r_m}\} \le 10^{-n}$ for				Number of iterations needed to achieve $\max_{1\le j\le J}\{F_j^{p}\} \le 10^{-n}$ for			
δ	n=1	n=2	n=3	n=4	n=1	n=2	n=3	n=4
40.0	2	7	8	32	2	7	13	108
45.0	2	7	8	28	2	7	12	95
50.0	2	7	8	25	2	7	11	85
55.0	2	7	8	23	2	7	10	77
56.0	2	7	8	23	2	7	10	76
57.0	2	7	8	22	2	7	10	74
58.0	2	7	8	22	2	7	9	73
59.0	2	7	8	21	2	7	9	71
60.0	**2**	**6**	**7**	**22**	**2**	**7**	**9**	**71**

Table A.6. Clustering Approach : Iteration Results : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$.

Quadratic Regression								
	Number of iterations needed to achieve $\max_{1\le j\le m+J}\{F_j^{q,r_1,\ldots,r_m}\} \le 10^{-n}$ for				Number of iterations needed to achieve $\max_{1\le j\le J}\{F_j^{p}\} \le 10^{-n}$ for			
δ	n=1	n=2	n=3	n=4	n=1	n=2	n=3	n=4
40.0	1	3	5	43	2	3	13	126
45.0	1	3	5	38	2	3	12	112
50.0	1	3	4	34	2	3	11	101
55.0	1	3	4	31	1	3	10	92
60.0	1	3	4	29	2	3	9	84
65.0	1	3	4	27	2	3	8	78
70.0	1	3	4	25	2	3	8	72
75.0	1	3	4	23	2	3	7	68
76.0	1	3	4	23	2	3	7	67
77.0	1	3	4	23	2	3	7	66
78.0	**1**	**3**	**4**	**23**	**2**	**3**	**7**	**65**

Table A.7. Clustering Approach : Iteration Results : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$.

Cubic Regression								
	Number of iterations needed to achieve $\max_{1\le j\le m+J}\{F_j^{q,r_1,\ldots,r_m}\} \le 10^{-n}$ for				Number of iterations needed to achieve $\max_{1\le j\le J}\{F_j^{p}\} \le 10^{-n}$ for			
δ	n=1	n=2	n=3	n=4	n=1	n=2	n=3	n=4
40.0	2	3	4	31	2	3	13	102
45.0	2	3	4	27	2	3	12	91
50.0	2	3	4	25	2	3	11	82
55.0	2	3	4	23	2	3	10	74
56.0	2	3	4	22	2	3	10	73
57.0	2	3	4	22	2	3	9	72
58.0	2	3	4	22	2	3	9	71
59.0	2	3	4	21	2	3	9	69
60.0	2	3	4	21	2	3	9	68
61.0	**2**	**3**	**4**	**21**	**2**	**3**	**9**	**67**

Table A.8. Clustering Approach : Iteration Results : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$.

Quartic Regression								
	Number of iterations needed to achieve $\max_{1\le j\le m+J}\{F_j^{q,r_1,\ldots,r_m}\} \le 10^{-n}$ for				Number of iterations needed to achieve $\max_{1\le j\le J}\{F_j^{p}\} \le 10^{-n}$ for			
δ	n=1	n=2	n=3	n=4	n=1	n=2	n=3	n=4
30.0	1	3	4	34	2	3	17	163
35.0	1	3	4	29	2	3	15	139
40.0	1	3	4	26	2	3	13	122
45.0	1	3	4	23	2	3	12	108
50.0	1	3	4	21	2	3	11	98
51.0	1	3	4	21	2	3	10	96
52.0	1	3	4	20	2	3	10	94
53.0	**1**	**3**	**4**	**20**	**2**	**3**	**10**	**92**

Table A.9. Clustering Approach : Iteration Results : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$.

Second-Order Model								
	Number of iterations needed to achieve $\max_{1\le j\le m+J}\{F_j^{q,r_1,\ldots,r_m}\} \le 10^{-n}$ for				Number of iterations needed to achieve $\max_{1\le j\le J}\{F_j^{p}\} \le 10^{-n}$ for			
δ	n=1	n=2	n=3	n=4	n=1	n=2	n=3	n=4
10.0	1	2	9	38	1	8	38	65
20.0	1	2	5	20	1	3	20	33
25.0	1	2	5	16	1	2	16	27
30.0	1	2	5	14	1	2	14	22
31.0	1	2	5	13	1	2	13	22
32.0	1	2	5	13	1	2	13	21
33.0	**1**	**2**	**5**	**13**	**1**	**2**	**12**	**21**

Table A.10. Clustering Approach : Iteration Results : $f(d) = d^{\delta}$, $d = \frac{\partial \phi}{\partial p}$.

Appendix B

Derivation of $\frac{\partial \phi}{\partial p_j}$ and $\frac{\partial^2 \phi}{\partial p_i \partial p_j}$

B.1 A-optimality

Let

$$h(M) = \{AM^{-1}A^T\} \tag{B.1}$$

where M is a symmetric matrix of order $k \times k$ and A is a matrix of order $s \times k$, $s < k$, $k =$ number of parameters and rank$(A) = s$.

Then the criterion function is

$$\phi(p) = \psi_A(M(p)), \quad M(p) = \sum_j p_j \underline{v}_j \underline{v}_j^T$$

with

$$\psi_A(M) = -tr\{h(M)\}.$$

The first order partial derivatives of ψ_A can be derived by using the Gâteaux derivative of section 2.3.2 of chapter 2.

$$G_{\psi_A}\{M,N\} = -tr\{G_h\{M,N\}\}. \tag{B.2}$$

From the definition of Gâteaux derivative

$$G_h\{M,N\} = \lim_{\varepsilon \downarrow 0} \frac{h(M+\varepsilon N) - h(M)}{\varepsilon} \tag{B.3}$$

$$= \lim_{\varepsilon \downarrow 0} \frac{A(M+\varepsilon N)^{-1}A^T - AM^{-1}A^T}{\varepsilon}. \tag{B.4}$$

Now

$$\begin{aligned}(M+\varepsilon N)^{-1} &= (I+\varepsilon M^{-1}N)^{-1}M^{-1} \\ &= M^{-1/2}(I+\varepsilon M^{-1/2}NM^{-1/2})^{-1}M^{-1/2} \\ &= M^{-1/2}(I+\varepsilon B)^{-1}M^{-1/2} \end{aligned} \tag{B.5}$$

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

$$\begin{aligned}(I+\varepsilon B)^{-1} &= I - \varepsilon B + \varepsilon^2 B^2 - \varepsilon^3 B^3 + \varepsilon^4 B^4 - \cdots\cdots \\ &= I - \varepsilon B \quad \text{for } \varepsilon \text{ small.} \end{aligned} \tag{B.6}$$

Thus by using (B.5) and (B.6) above, $h(M+\varepsilon N)$ in (B.3) would be

$$\begin{aligned}h(M+\varepsilon N) &= A(M+\varepsilon N)^{-1}A^T \\ &= AM^{-1/2}(I-\varepsilon B)M^{-1/2}A^T \\ &= AM^{-1}A^T - \varepsilon AM^{-1/2}BM^{-1/2}A^T \\ &= AM^{-1}A^T - \varepsilon AM^{-1}NM^{-1}A^T. \end{aligned}$$

Hence from (B.3) and (B.2)

$$G_h\{M,N\} = -AM^{-1}NM^{-1}A^T \tag{B.7}$$

$$\text{and} \quad G_{\psi_A}\{M,N\} = -tr\{-AM^{-1}NM^{-1}A^T\}. \tag{B.8}$$

Then from (B.8) with $M = M(p)$,

$$\begin{aligned}\frac{\partial \phi}{\partial p_j} &= G_{\psi_A}\{M, \underline{v}_j \underline{v}_j^T\} \\ &= -tr\{-AM^{-1}\underline{v}_j \underline{v}_j^T M^{-1}A^T\} \\ &= \underline{v}_j^T M^{-1}A^T AM^{-1}\underline{v}_j.\end{aligned}$$

Let

$$\psi'_{jA}(M) = \underline{v}_j^T M^{-1}A^T AM^{-1}\underline{v}_j.$$

Now

$$\begin{aligned}\psi'_{jA}(M+\varepsilon N) &= \underline{v}_j^T (M+\varepsilon N)^{-1}A^T A(M+\varepsilon N)^{-1}\underline{v}_j \\ &= \underline{v}_j^T M^{-1/2}(I-\varepsilon B)M^{-1/2}A^T AM^{-1/2}(I-\varepsilon B)M^{-1/2}\underline{v}_j \\ &= \underline{v}_j^T [M^{-1}A^T - \varepsilon M^{-1}NM^{-1}A^T]\,[AM^{-1}\varepsilon AM^{-1}NM^{-1}]\underline{v}_j \\ &= \underline{v}_j^T M^{-1}A^T AM^{-1}\underline{v}_j \\ &\quad -\varepsilon[\underline{v}_j^T M^{-1}NM^{-1}A^T AM^{-1}\underline{v}_j + \underline{v}_j^T M^{-1}A^T AM^{-1}NM^{-1}\underline{v}_j] \\ &\quad +\varepsilon^2[\underline{v}_j^T M^{-1}NM^{-1}A^T AM^{-1}NM^{-1}\underline{v}_j].\end{aligned}$$

Hence

$$G_{\psi'_{jA}}\{M,N\} = -[\underline{v}_j^T M^{-1}NM^{-1}A^T AM^{-1}\underline{v}_j + \underline{v}_j^T M^{-1}A^T AM^{-1}NM^{-1}\underline{v}_j].$$

Hence with $M = M(p)$,

$$\begin{aligned}\frac{\partial^2 \phi}{\partial p_i \partial p_j} &= G_{\psi'_{jA}}\{M, \underline{v}_i \underline{v}_i^T\} \\ &= -[\underline{v}_j^T M^{-1} \underline{v}_i \underline{v}_i^T M^{-1} A^T A M^{-1} \underline{v}_j + \underline{v}_j^T M^{-1} A^T A M^{-1} \underline{v}_i \underline{v}_i^T M^{-1} \underline{v}_j] \\ &= -2 \left[\underline{v}_j^T M^{-1} A^T A M^{-1} \underline{v}_i\right] \left[\underline{v}_j^T M^{-1} \underline{v}_i\right].\end{aligned}$$

B.2 D_A-optimality

The criterion function is

$$\phi(p) = \psi_{D_A}(M(p))$$

with

$$\psi_{D_A}(M) = -log\, det\{A M^{-1} A^T\} = -log\, det\{h(M)\}.$$

As in the case of A-optimality, the first order partial derivatives can be derived from

$$G_{\psi_{D_A}}\{M, N\} = -tr\{h^{-1}(M)\, G_h\{M, N\}\}. \tag{B.9}$$

Hence from (B.7) and (B.9) with $N = \underline{v}_j \underline{v}_j^T$ and $M = M(p)$,

$$\begin{aligned}\frac{\partial \phi}{\partial p_j} &= -tr\{(A M^{-1} A^T)^{-1} (-A M^{-1} \underline{v}_j \underline{v}_j^T M^{-1} A^T)\} \\ &= \underline{v}_j^T M^{-1} A^T (A M^{-1} A^T)^{-1} A M^{-1} \underline{v}_j.\end{aligned} \tag{B.10}$$

Let

$$\psi'_{jD_A}(M) = \underline{v}_j^T M^{-1} A^T (AM^{-1}A^T)^{-1} AM^{-1} \underline{v}_j. \tag{B.11}$$

Now from (B.11)

$$\begin{aligned}
\psi'_{jD_A}(M+\varepsilon N) &= \underline{v}_j^T (M+\varepsilon N)^{-1} A^T [A(M+\varepsilon N)^{-1} A^T]^{-1} A(M+\varepsilon N)^{-1} \underline{v}_j \\
&= \underline{v}_j^T [M^{-1/2}(I-\varepsilon B)M^{-1/2} A^T][AM^{-1/2}(I-\varepsilon B)M^{-1/2} A^T]^{-1} \\
&\quad [AM^{-1/2}(I-\varepsilon B)M^{-1/2}]\, \underline{v}_j^T \\
&= \underline{v}_j^T [M^{-1}A^T - \varepsilon M^{-1} N M^{-1} A^T][E-\varepsilon F]^{-1} \\
&\quad [AM^{-1} - \varepsilon AM^{-1} N M^{-1}]\, \underline{v}_j^T
\end{aligned} \tag{B.12}$$

where

$$E = AM^{-1}A^T, \quad F = AM^{-1/2} B M^{-1/2} A^T.$$

Thus

$$\begin{aligned}
[E-\varepsilon F]^{-1} &= [E^{1/2}[I-\varepsilon E^{-1/2} F E^{-1/2}]E^{1/2}]^{-1} \\
&= [E^{-1/2}[I-\varepsilon E^{-1/2} F E^{-1/2}]^{-1} E^{-1/2}] \\
&= [E^{-1/2}[I+\varepsilon E^{-1/2} F E^{-1/2}] E^{-1/2}] \\
&\qquad [\text{by expanding similarly in (B.6)}] \\
&= [E^{-1} + \varepsilon E^{-1} F E^{-1}].
\end{aligned} \tag{B.13}$$

Thus from (B.12) and (B.13),

$$\begin{aligned}
\psi'_{D_A}(M+\varepsilon N) &= \underline{v}_j^T [M^{-1}A^T - \varepsilon M^{-1}NM^{-1}A^T][E^{-1} + \varepsilon E^{-1}FE^{-1}] \\
&\quad [AM^{-1} - \varepsilon AM^{-1}NM^{-1}]\underline{v}_j^T \\
&= \underline{v}_j^T M^{-1}A^T(AM^{-1}A^T)^{-1}AM^{-1}\underline{v}_j \\
&\quad + \varepsilon\Big[\underline{v}_j^T M^{-1}A^T(AM^{-1}A^T)^{-1}AM^{-1}NM^{-1}A^T(AM^{-1}A^T)^{-1}AM^{-1}\underline{v}_j \\
&\qquad -\underline{v}_j^T M^{-1}A^T(AM^{-1}A^T)^{-1}AM^{-1}NM^{-1}\underline{v}_j \\
&\qquad -\underline{v}_j^T M^{-1}NM^{-1}A^T(AM^{-1}A^T)^{-1}AM^{-1}\underline{v}_j\Big] \\
&\quad + \text{ higher order terms of } \varepsilon \\
&= \underline{v}_j^T H\underline{v}_j + \varepsilon\Big[\underline{v}_j^T HNH\underline{v}_j - \underline{v}_j^T HNM^{-1}\underline{v}_j - \underline{v}_j^T M^{-1}NH\underline{v}_j\Big] \\
&\quad + \text{ higher order terms of } \varepsilon \qquad \text{(B.14)}
\end{aligned}$$

where $H = M^{-1}A^T(AM^{-1}A^T)^{-1}AM^{-1}$. Hence from (B.14)

$$G_{\psi'_{jD_A}}\{M,N\} = \underline{v}_j^T HNH\underline{v}_j - \underline{v}_j^T HNM^{-1}\underline{v}_j - \underline{v}_j^T M^{-1}NH\underline{v}_j.$$

Hence with $M = M(p)$,

$$\begin{aligned}
\frac{\partial^2 \phi}{\partial p_i \partial p_j} &= G_{\psi'_{jD_A}}\{M, \underline{v}_i\underline{v}_i^T\} \\
&= \underline{v}_j^T H\underline{v}_i\underline{v}_i^T H\underline{v}_j - \underline{v}_j^T H\underline{v}_i\underline{v}_i^T M^{-1}\underline{v}_j - \underline{v}_j^T M^{-1}\underline{v}_i\underline{v}_i^T H\underline{v}_j \\
&= [\underline{v}_j^T H\underline{v}_i]^2 - 2[\underline{v}_j^T H\underline{v}_i][\underline{v}_j^T M^{-1}\underline{v}_i] \\
&= [\underline{v}_j^T M^{-1}A^T(AM^{-1}A^T)^{-1}AM^{-1}\underline{v}_i]^2 \\
&\quad -2[\underline{v}_j^T M^{-1}A^T(AM^{-1}A^T)^{-1}AM^{-1}\underline{v}_i][\underline{v}_j^T M^{-1}\underline{v}_i].
\end{aligned}$$

Bibliography

Alahmadi, A. M. (1993). *Algorithms for the construction of constrained and unconstrained optimal designs.* PhD thesis, University of Glasgow, Glasgow.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972). *Robust Estimates of Location.* Princeton University Press, Princeton.

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

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

Atwood, C. L. (1976). Convergent design sequences, for sufficiently regular optimality criteria. *Annals of Statistics*, 4:1124–1138.

Atwood, C. L. (1980). Convergent design sequences for sufficiently regular optimality criteria, II: singular case. *Annals of Statistics*, 8:894–912.

Beckenbach, E. F. and Bellman, R. (1961). *Inequalities.* Springer Verlag, Berlin.

Boyce, J. F., Feng, J., and Haddow, E. R. (1987). Relaxation labelling and the entropy of neighbourhood information. *Pattern Recognition Letters*, 6:225–234.

Bradley, R. A. (1965). Another interpretation of a model for paired comparisons. *Psychometrika*, 30:315–318.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs I. The method of paired comparisons. *Biometrika*, 39:324–345.

Chernoff, H. (1953). Locally optimal designs for estimating parameters. *Annals of Mathematical Statistics*, 24:586–602.

Cook, D. and Fedorov, V. (1995). Constrained optimization of experimental design (with discussion). *Statistics*, 26:129–178.

Davidson, R. R. (1969). On a relationship between two representations of a model for paired comparisons. *Biometrics*, 25:597–599.

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.

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

Eplett, W. J. R. (1980). An influence curve for two sample rank tests. *J. Roy. Statist. Soc. Series B*, 42:64–70.

Farrel, R. H., Kiefer, J., and Walbran, A. (1967). Optimum multivariate designs. *Proc. 5th Berkeley Symp.*, Vol. 1, pages 113-138, University of California Press, Berkeley.

Faugeras, O. D. and Berthod, M. (1981). Improving consistency and reducing ambiguity in stochastic labelling: An optimisation apprroach. *IEEE Trans. Pattern Anal. Machine Intell.*, PAMI-3:412–424.

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

Fellman, J. (1989). An empirical study of a class of iterative searches for optimal designs. *Journal of Statistical Planning and Inference*, 21:85–92.

Graybill, F. A. (1969). *Introduction to matrices with applications in statistics.* Wadsworth International Group, Belmont, California.

Hampel, F. R. (1968). *Contributions to the theory of robust estimation.* PhD thesis, University of California, Berkeley.

Hampel, F. R. (1971). A general qualitative definition of robustness. *Annals of Mathematical Statistics*, 42:1887–1896.

Hoel, P. G. (1958). Efficiency problems in polynomial regression. *Annals of Mathematical Statistics*, 29:1134–1145.

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

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

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

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

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

Mandal, S. and Torsney, B. (2000b). Construction of optimal designs using a clustering approach. (in preparation).

Molchanov, I. and Zuyev, S. (2000). Variational calculus in the space of measures and optimal design. *Optimum Design 2000*, 79-90, Kluwer Academic Publishers.

Murray, G. D. and Titterington, D. M. (1978). Estimation problems with data from a mixture. *J. Roy. Statist. Soc. Series C*, 27:325–334.

Pazman, A. (1986). *Foundations of optimum experimental design.* Reidel, Dordrecht.

Placket, R. L. (1974). *The analysis of categorical data.* Griffin, London.

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

Pukelsheim, F. and Rosenberger, J. L. (1993). Experimental designs for model discrimination. *J. Am. Statist. Assoc.*, 88:642–649.

Rhode, C. A. (1965). Generalised inverses of partitioned matrices. *J. Soc. Indust. Appl. Math.*, 13:1033–1053.

Rockafellar, R. T. (1970). *Convex Analysis.* Princeton University Press, Princeton.

Seber, G. A. F. (1977). *Linear regression analysis.* Wiley, New York.

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

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

Silverman, B. W. and Titterington, D. M. (1980). Minimum covering ellipses. *SIAM J. Sci. Stat. Comput.*, 1:401–409.

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

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

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

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

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.

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.

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

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical analysis of finite mixture distributions.* Wiley, Chichester.

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

Torsney, B. (1981). *Algorithms for a constrained optimisation problem with applications in statistics and optimum design.* PhD thesis, University of Glasgow, Glasgow.

Torsney, B. (1983). A moment inequality and monotonicity of an algorithm. *Proc. Internat. Symp. on Semi-Infinite Programming and Applications (Edited by Kortanek, K. O. and Fiacco, A. V.). Lecture Notes in Economics and Mathematical Systems*, vol. 215, pages 249-260, University of Texas, Austin.

Torsney, B. (1988). Computing optimising distributions with applications in design, estimation and image processing. *Optimal Design and Analysis of Experiments (Edited by Dodge, Y., Fedorov, V. V. and Wynn, H. P.)*, 361-370, Elsevier Science Publishers B. V., North Holland.

Torsney, B. and 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.)*, pages 121-129, Physica-Verlag.

Torsney, B. and Alahmadi, A. M. (1995). Designing for minimally dependent observations. *Statistica Sinica*, 5:499–514.

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

Von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Annals of Mathematical Statistics*, 18:309–348.

Whittle, P. (1971). *Optimisation Under Constraints.* Wiley-Interscience, London.

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

Wu, C. F. J. (1976). *Contributions to optimisation theory with applications to optimal design of experiments.* PhD thesis, University of California, Berkeley.

Wu, C. F. J. (1978). Some iterative procedures for generating nonsingular optimal designs. *Communications in Statistics A*, 7:1399–1412.

Wynn, H. P. (1969). *The Theory and Construction of Optimum Experimental Designs.* PhD thesis, Imperial College, London.

Wynn, H. P. (1970). The sequential generation of D-optimal experimental designs. *Annals of Mathematical Statistics*, 41:1655–1664.

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.