OPTIMAL STATIC AND SEQUENTIAL DESIGN:

A CRITICAL REVIEW

by

IAN FORD

A dissertation submitted to the

UNIVERSITY OF GLASGOW

for the degree of

Doctor of Philosophy

1976

ACKNOWLEDGEMENTS

I would like to express my thanks to my supervisor Professor S. D. Silvey for his advice and encouragement during the period of this research. I would also like to extend my gratitude to all of the members of the Statistics department of Glasgow University for their willingness to lend a sympathetic ear to my problems, during both my undergraduate and postgraduate careers. Finally, I would like to thank Mrs. L. Williamson for her patient typing of this thesis.

This research was carried out while I was in receipt of a Science Research Council grant.

SUMMARY

The aim of this thesis is to review and augment the theory and methods of optimal experimental design.

In Chapter 1 the scene is set by considering the possible aims of an experimenter prior to an experiment, the statistical methods one might use to achieve those aims and how experimental design might aid this procedure. It is indicated that, given a criterion for design, a priori optimal design will only be possible in certain instances and, otherwise, some form of sequential procedure would seem to be indicated.

In Chapter 2 an exact experimental design problem is formulated mathematically and is compared with its continuous analogue. Motivation is provided for the solution of this continuous problem, and the remainder of the chapter concerns this problem. A necessary and sufficient condition for optimality of a design measure is given. Problems which might arise in testing this condition are discussed, in particular with respect to possible non-differentiability of the criterion function at the design being tested. Several examples are given of optimal designs which may be found analytically and which illustrate the points discussed earlier in the chapter.

In Chapter 3 numerical methods of solution of the continuous optimal design problem are reviewed. A new algorithm is presented with illustrations of how it should be used in practice. It is shown that, for reasonably large sample size, continuously optimal designs may be approximated to well by an exact design. In situations where this is not satisfactory algorithms for improvement of this design are reviewed.

Chapter 4 consists of a discussion of sequentially designed experiments, with regard to both the philosophies underlying, and the application of the methods of, statistical inference.

In Chapter 5 we criticise constructively previous suggestions for fully sequential design procedures. Alternative suggestions are made along with conjectures as to how these might improve performance.

Chapter 6 presents a simulation study, the aim of which is to investigate the conjectures of Chapter 5. The results of this study provide empirical support for these conjectures.

In Chapter 7 examples are analysed. These suggest aids to sequential experimentation by means of reduction of the dimension of the design space and the possibility of experimenting semi-sequentially. Further examples are considered which stress the importance of the use of prior information in situations of this type. Finally we consider the design of experiments when semi-sequential experimentation is mandatory because of the necessity of taking batches of observations at the same time.

In Chapter 8 we look at some of the assumptions which have been made and indicate what may go wrong where these assumptions no longer hold.

TABLE OF CONTENTS

CHAPTER 1

MOTIVATION

1.1 Consider the following experimental setting. We have obtained a set of N independent observations $\underline{y} = (y_1,\ldots,y_N)^T$ from a probability distribution identified by the density function $p(y|\underline{x},\underline{\theta})$ where $\underline{x} \in \mathcal{X}$ is a variable subject to the experimenter's control in the design space $\mathcal{X}$, and $\underline{\theta}$ is a K-vector of unknown parameters.

Let us assume that $N = \sum_{i=1}^{r} n_i$, where n_i denotes the number of observations which were taken at the point $\underline{x}_i \in \mathcal{X}$ (that is $p_i = \frac{n_i}{N}$ is the proportion of observations taken at $\underline{x}_i$), and r is the number of distinct $\underline{x}_i$'s which were chosen. We shall call the set of points $\{ \underline{x}_i ; i=1,\ldots,r \}$ the spectrum of the experimental design and we shall call

$$\xi_N = \left\{ \begin{pmatrix} p_i \\ \underline{x}_i \end{pmatrix} ; \quad i=1,\ldots,r \right\} \text{ the design measure.}$$

We shall now consider several forms of $p(y|\underline{x},\underline{\theta})$ and how we might analyse the data subsequent to an experiment of the above design, with a view to estimating the unknown parameters $\underline{\theta}$ and making inferences on them.

1.2.1 Firstly take $p(y|\underline{x},\underline{\theta})$ to be $N(\sum_{i=1}^{K} f_i(\underline{x})\theta_i, \sigma^2)$. That is we have a regression situation which is linear in the parameters but not necessarily linear in $\underline{x}$, with a normal distribution of error; variance σ^2 assumed known, and independent of $\underline{x}$ and $\underline{\theta}$.

(i) The maximum likelihood estimates of $\underline{\theta}$ are well known to be given by $\hat{\underline{\theta}}_{ML} = (X^TX)^{-1} X^T\underline{y}$, where

$$X = \begin{bmatrix} f_1(\underline{x}_1) & \cdots & f_K(\underline{x}_1) \\ \vdots & & \vdots \\ f_1(\underline{x}_r) & \cdots & f_K(\underline{x}_r) \end{bmatrix}, \quad X \text{ is } (N \times K).$$

In the repeated sampling sense the above estimates are normally distributed in the form

$$\hat{\underline{\theta}}_{ML} \sim N(\underline{\theta}, (X^T X)^{-1} \sigma^2).$$

Therefore, if the design of our experiment is to have any effect on the distribution of $\hat{\theta}_{ML}$ then it must be through the matrix $X^T X$. In an analogy with the use of the term sufficient statistic in the field of statistical inference we might refer to $X^T X$ as being a sufficient design statistic. What function of the matrix $X^T X$ will be a necessary design statistic will, of course, depend on the reasons why the experiment is being carried out.

We note here an elementary matrix result which says that if X is of the above form and we write $\underline{f}(\underline{x}) = (f_1(\underline{x}), \ldots, f_K(\underline{x}))^T$, then

$$X^T X = N \sum_{i=1}^{r} p_i \, \underline{f}(\underline{x}_i) \, \underline{f}(\underline{x}_i)^T .$$

(ii) Suppose now that we wish to make inferences on the vector of parameters $\underline{\theta}$ and that we take an approach to statistical inference based solely on the likelihood function. We will be interested in the shape of the likelihood function as a function of $\underline{\theta}$. The likelihood $L_{\underline{y}_N}(\underline{\theta})$ in this example may be written as

$$\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp \left\{ - \frac{N}{2\sigma^2} \sum_{i=1}^{r} p_i \, (y_i - \underline{f}^T(\underline{x}_i) \, \underline{\theta})^2 \right\} .$$

A little algebraic manipulation reveals that, as a function of $\underline{\theta}$, $L_{\underline{y}_N}(\underline{\theta})$ may be written as proportional to

$$\exp\{-\frac{1}{2\sigma^2}(\underline{\theta}-\hat{\underline{\theta}}_{ML})^T X^T X (\underline{\theta}-\hat{\underline{\theta}}_{ML})\}.$$

We observe that the shape of the likelihood function is under the control of the matrix $X^T X$ and therefore this matrix will play a dominant role in the making of inferences on the unknown vector $\underline{\theta}$. We note here that the log likelihood $\ell(\underline{\theta}) = \log L_{\underline{y}_N}(\underline{\theta})$ is an elliptical function,

$$\text{constant} - K(\underline{\theta}-\hat{\underline{\theta}}_{ML})^T X^T X (\underline{\theta}-\hat{\underline{\theta}}_{ML}),$$

centred on $\hat{\underline{\theta}}_{ML}$ with length and orientation of its axes controlled by $X^T X$.

(iii) Let us assume that we are willing to adopt a Bayesian approach to statistical inference. Adding the extra assumption of prior information about $\underline{\theta}$ in the form of a multivariate normal prior distribution, that is $\pi(\underline{\theta}) \sim N(\underline{\theta}_o, \Omega^{-1})$, and applying Bayes formula it is relatively easy to obtain

$$\pi(\underline{\theta}|\underline{y}_N) \sim N(\hat{\underline{\theta}}_B, (\frac{1}{\sigma^2} X^T X + \Omega)^{-1}),$$

where $\hat{\underline{\theta}}_B$ denotes the mean of the Bayesian posterior distribution. Again we see that experimental design will affect this posterior distribution via the matrix $X^T X$. Obviously the effect of our design will depend on the matrix Ω. However, if we denote little or vague prior knowledge by $\Omega \to \underline{0}$, where $\underline{0}$ is the null matrix, and observe that for large N, $\frac{1}{\sigma^2} X^T X$ will dominate Ω, then it may be seen that in these instances the matrix $X^T X$ will have a dominating role.

To summarise (i), (ii) and (iii) above it may be said that if we are interested in the repeated sampling distribution of maximum likelihood estimates or in making inferences on the vector of parameters $\underline{\theta}$ according to the approaches of (ii) or (iii), then the three approaches would seem to be in agreement as far as a choice of direction in which to look for selection of a criterion for experimental design is concerned. That is, we would seem to wish to optimise some function, which we shall call ϕ, of

$$X^TX = N \Sigma p_i \underline{f}(x_i) \underline{f}(x_i)^T \quad .$$

1.2.2 We now turn our attention to a more general situation than the one considered in 1.2.1 above, namely that where the observations y come from any distribution $p(y|\underline{x},\underline{\theta})$ such that the asymptotic theory of maximum likelihood estimation of $\underline{\theta}$ carries through. This may be regarded as a theoretical restriction only.

For example we might have $p(y|\underline{x},\underline{\theta})$ as:-

(i) $N(\eta(\underline{\theta},\underline{x}),\sigma^2)$ where $\eta(\underline{\theta},\underline{x})$ may be a non-linear function of $\underline{x}$ and $\underline{\theta}$.

(ii) A binary response distribution of the form
$p(1|\underline{x},\underline{\theta}) = \eta(\underline{\theta},\underline{x})$, $p(0|\underline{x},\underline{\theta}) = 1 - \eta(\underline{\theta},\underline{x})$
where again $\eta(\underline{\theta},\underline{x})$ will be a non-linear function of the parameters $\underline{\theta}$ and $\underline{x}$.

(iii) The observations may come from a finite set of populations having distributions with different means which are related in a linear or non-linear fashion, for example we may have three populations

(a) $Po(\theta)$, (b) $Po(\theta + \lambda)$, (c) $Po(\theta / \lambda)$.

With these more general examples we shall have to appeal to large sample results to find directions for seeking criteria for experimental design. We will echo the treatment of the normal-linear

regression model and consider three different approaches.

(i) Consider firstly the use of maximum likelihood estimation and the repeated sampling distribution of these estimates. It is well known that for independent observations and subject to mild regularity conditions the distribution of $\hat{\underline{\theta}}_{ML}$ is asymptotically normal of the form $N(\underline{\theta}, \frac{1}{N}.M(\underline{\theta})^{-1})$ where

$$N.M(\underline{\theta}) = \mathbb{E}\{ - \frac{\partial^2}{\partial\theta_i \partial\theta_j} \ell(\underline{\theta}) \}$$

$$= N \sum_{i=1}^{r} p_i \mathbb{E}\{ - \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(y|\underline{x}_i,\underline{\theta}) \}$$

$$= N \sum_{i=1}^{r} p_i \; I(\underline{x}_i, \underline{\theta}) \;,$$

where $I(\underline{x},\underline{\theta}) = \mathbb{E}\{- \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(y|\underline{x},\underline{\theta}) \}$ is the Fisher information matrix at the point $\underline{x}$.

Again if we wish to see how experimental design will affect the nature of this asymptotic distribution of $\hat{\underline{\theta}}_{ML}$ we need only consider the matrix $M(\underline{\theta})$. If we continue the analogy of 1.2.1 then $M(\underline{\theta})$ might be thought of as an asymptotically sufficient statistic.

(ii) As we are dealing with large sample situations here we shall consider the Bayesian and pure likelihood approaches together. Although the methods of making inferences may be different in principle in the two situations the function being used to make these inferences will not, asymptotically. This is a result of the information in the likelihood function swamping the effect of the prior distribution as $N \to \infty$.

The asymptotic Bayes or pure likelihood approaches are based on the normalisation of the posterior distribution or the likelihood function. This is equivalent to saying that the log likelihood function can asymptotically be described by a second order Taylor expansion about $\hat{\underline{\theta}}_{ML}$. For conditions where this theory is applicable see for example Dawid (1970). The regularity conditions required are similar to those necessary for the asymptotic normality of the repeated sampling distribution

of $\hat{\underline{\theta}}_{ML}$. Expanding $\ell(\underline{\theta})$ about $\hat{\underline{\theta}}_{ML}$ we see that

$$\ell(\underline{\theta}) = \text{const} - \tfrac{1}{2}(\underline{\theta} - \hat{\underline{\theta}}_{ML})^T S(\underline{\theta} - \hat{\underline{\theta}}_{ML}) \quad ,$$

where $S = \{ - \frac{\partial^2}{\partial\theta_i \partial\theta_j} \ell(\underline{\theta}) \}_{\underline{\theta} = \hat{\underline{\theta}}_{ML}}$

We shall refer to S as the sample information matrix. In the situation of a designed experiment let us assume that $n_i \to \infty$ as $N \to \infty$ in such a way that $\frac{n_i}{N} \to p_i$.

Therefore, by a law of large numbers

$$S \to \mathbb{E}(S) = N \sum_{i=1}^{r} p_i I(\underline{x}_i, \hat{\underline{\theta}}_{ML}) \quad \text{as} \quad N \to \infty$$

$$= N.M(\hat{\underline{\theta}}_{ML}).$$

For large N we might presume $\hat{\underline{\theta}}_{ML}$ to be close to $\underline{\theta}$ suggesting $M(\underline{\theta})$ as an asymptotic criterion for experimental design.

1.2.3 To summarise the above two sections it may be said that the asymptotic repeated sampling theory and the theory of an approach to statistical inference based on the likelihood principle produce general criteria for design which are similar and which are related to the criterion which was obtained by exact results in normal-linear model theory. The criterion being to optimise some function ϕ of $M(\underline{\theta})$, where $N.M(\underline{\theta})$ denotes the Fisher information matrix of the experiment, $M(\underline{\theta})$ being a positive definite symmetric matrix, and ϕ denoting some property of $M(\underline{\theta})$ which will be related to inferences to be made on the parameters subsequent to the experiment.

At this point it is important to make the following observation, namely that although in the normal-linear example of 1.2.1 $M(\underline{\theta}) = \frac{1}{N} X^T X$ is independent of $\underline{\theta}$, this in general will not be the case and in the more general examples of 1.2.2 $M(\underline{\theta})$ will typically depend on $\underline{x}$ and $\underline{\theta}$ the

vector of unknown parameters. Therefore, although for the linear-model the problem is effectively reduced to an a priori optimisation, in general some form of sequential procedure will be necessary. However, it will be seen later that solution of the a priori design problem for given $\underline{\theta}$ will be of more than academic interest. As the sequential type of design introduces complications which would only be of a confusing nature at this juncture we postpone discussion of these factors until a later chapter, and, until otherwise mentioned, restrict ourselves to the a priori design which we shall hereafter refer to as a static design.

1.3 In previous sections we have formulated a very general experimental design problem, namely to optimise some function ϕ of $M(\underline{\theta})$, $M(\underline{\theta}) \in \mathcal{M}$ where $\mathcal{M}$ represents the set of all Fisher information matrices that experimental conditions permit. We now consider possible contenders for the function ϕ, these being functions which will fall naturally from the uses to which the estimates of $\underline{\theta}$ are to be put subsequent to experimentation. Of course, a major problem in practice will be to get the experimenter to express his wishes in a particular mathematical form. In what follows we shall assume that he has done so, and we shall reconsider this practical problem in the final chapter. The set of criteria which is discussed is in no way intended to be all-encompassing, but has been chosen to reflect the general properties of possible criteria and to highlight particular examples where solution of the optimal design problem might be complicated.

1.3.1 In this section we shall assume that the experimenter is interested in all of the parameters jointly. We shall consider four criteria.

(i) $\phi_1 = \log \det(M(\underline{\theta}))$

From 1.2(ii) it may be seen that joint confidence intervals for the vector of unknown parameters may be described by ellipsoids of the form $(\underline{\theta} - \hat{\underline{\theta}}_{ML})^T M(\underline{\theta}) (\underline{\theta} - \hat{\underline{\theta}}_{ML}) \leq q$. The surface of these ellipsoids will represent regions of equal 'confidence'. The volume

of the above ellipsoids is proportional to $\{ \det(M(\underline{\theta})) \}^{-\frac{1}{2}}$, so, maximising log det $(M(\underline{\theta}))$ would be equivalent to minimsing the volume of all confidence ellipsoids for $\underline{\theta}$ of the above form. That is, we are making our confidence regions, in some sense, as compact as is possible. We take ϕ_1 as log det for mathematical ease later on. The criterion ϕ_1 has become known, in the literature, as the D-optimality criterion, and has by far dominated the literature on optimal design. It should be said however that it would appear that its pre-eminence as a criterion owes as much to the fact that its mathematical form has enabled more nice analytic and geometric arguments to be put forward, as to the extent to which it might actually reflect an experimenters design wishes.

Wynn (1969) notes that maximising det $(M(\underline{\theta}))$ is equivalent to maximising the Gaussian curvature, at $\underline{\theta}_o$, of the power of the F-test with null hypothesis $\underline{\theta} = \underline{\theta}_o$. Lindley (1956) and Stone (1959), using an information theory argument, show that maximising det $(M(\underline{\theta}))$ is equivalent to maximising the expected gain in information about $\underline{\theta}$ assuming little prior knowledge. Bernardo (1976) extends this approach and shows that if one adopts Bayesian methods and if one is interested only in making very pure inferences about the unknown parameters, then one's approach to design should be based on something closely related to the Shannon information approach. However, there would appear to be a weakness in his argument relating to his definition of a pure inference problem and the consequences in defining a natural utility function to be used to construct a criterion for experimental design. Draper and Hunter (1967a) observe that maximising $\det(M(\underline{\theta}))$ maximises the posterior density function at the point of maximum posterior density. This, of course, might be deduced as a corollary to the above argument concerning the confidence ellipsoids.

Properties of ϕ_1

(a) ϕ_1 is an increasing function of the positive definite symmetric matrices. That is, for M_1 positive definite symmetric and M_2 positive semi-definite symmetric matrices

$$\phi_1 \{ M_1 + M_2 \} \geq \phi_1 \{ M_1 \} .$$

(b) ϕ_1 is a concave function of the positive definite symmetric matrices. That is, for M_1, M_2 positive definite symmetric matrices

$$\phi_1\{\alpha M_1 + (1-\alpha) M_2\} \geq \alpha\phi_1\{M_1\} + (1-\alpha)\phi_1\{M_2\} ,$$

$$0 \leq \alpha \leq 1, \quad \text{(see Appendix 3).}$$

(c) As a criterion for optimal design ϕ_1 is invariant under non-singular transformations of $\underline{\theta}$. To see this consider the non-singular transformation

$$\underline{\theta} = (\theta_1,\ldots,\theta_K)^T \rightarrow (\tau_1(\underline{\theta}),\ldots,\tau_K(\underline{\theta}))^T = \underline{\tau}(\underline{\theta}) .$$

Let J be the Jacobian of the transformation, then

$$J = \begin{bmatrix} \frac{\partial\tau_1(\underline{\theta})}{\partial\theta_1} & \cdots & \frac{\partial\tau_K(\underline{\theta})}{\partial\theta_1} \\ \vdots & & \vdots \\ \frac{\partial\tau_1(\underline{\theta})}{\partial\theta_K} & \cdots & \frac{\partial\tau_K(\underline{\theta})}{\partial\theta_K} \end{bmatrix}^{-1} \quad \text{and } M(\underline{\tau}(\underline{\theta})) = JM(\underline{\theta})\, J^T.$$

Therefore, $\phi_1\{M(\underline{\tau}(\underline{\theta}))\} = \log\det\{M(\underline{\tau}(\underline{\theta}))\}$

$$= 2\log\det\{J\} + \log\det\{M(\underline{\theta})\}$$

$$= \text{const} + \phi_1\{M(\underline{\theta})\}$$

Property (c) above has been cited by previous authors as being a reason for selecting ϕ_1 as a criterion in preference to others. However, as there would seem to be many situations where one would not wish a criterion to have the above property we merely note it as an interesting fact.

(ii) $$\frac{\phi_2}{\eta_{\underline{\theta}}(\underline{x},\underline{\theta})} = \max_{\underline{x}\in\mathcal{X}} \frac{\eta_{\underline{\theta}}^T(\underline{x},\underline{\theta})\, M(\theta)^{-1}\, \eta_{\underline{\theta}}(\underline{x},\underline{\theta})}{\left[\frac{\partial\eta(\underline{x},\underline{\theta})}{\partial\theta_1}, \ldots\ldots, \frac{\partial\eta(\underline{x},\underline{\theta})}{\partial\theta_K}\right]^T}$$

This criterion will typically be of interest where $\eta(\underline{x},\underline{\theta}) = E(y|\underline{x})$. We note that if $\eta(\underline{x},\underline{\theta}) = \sum_{i=1}^{K} f_i(\underline{x})\,\theta_i$, as in our example of 1.2.1, then

$$\eta_{\underline{\theta}}^T(\underline{x},\underline{\theta})\, M(\theta)^{-1}\, \eta_{\underline{\theta}}(\underline{x},\underline{\theta}) = \underline{f}^T(\underline{x})\, M(\underline{\theta})^{-1}\, \underline{f}(\underline{x})$$

is the variance of the estimated expected response at $\underline{x}$, namely var $(\eta(\underline{x}, \hat{\underline{\theta}}_{ML}))$. Similarly, if $\eta(\underline{x},\underline{\theta})$ is a non-linear function of $\underline{\theta}$ then $\eta_{\underline{\theta}}^T(\underline{x},\underline{\theta})\, M(\underline{\theta})^{-1}\, \eta_{\underline{\theta}}(\underline{x},\underline{\theta})$ may readily be seen to be a first order approximation to var $(\eta(\underline{x},\hat{\underline{\theta}}_{ML}))$. Therefore, a sensible design criterion would be to minimise over the set of possible designs the maximum over $\underline{x}\in\mathcal{X}$ of the variance of our expected response. That is, maximise $\phi_2\{M(\underline{\theta})\}$ over the set of possible designs.

Properties of ϕ_2

(a) ϕ_2 is an increasing function on the set of positive definite symmetric matrices.

(b) ϕ_2 is a concave function on the set of positive definite symmetric matrices. (The proof is analogous to that in Appendix 2).

(c) ϕ_2 is invariant under non-singular transformations of $\underline{\theta}$. Consider again the non-singular transformation $\underline{\theta} \to \tau(\underline{\theta})$.

$$\phi_2\{M(\underline{\tau}(\underline{\theta}))\} = -\max_{\underline{x}\in\mathcal{X}} \eta_{\underline{\theta}}^{\tau T}(\underline{x},\underline{\theta}) J^T \{JM(\underline{\theta})J^T\}^{-1} J\eta_{\underline{\theta}}(\underline{x},\underline{\theta})$$

$$= -\max_{\underline{x}\in\mathcal{X}} \eta_{\underline{\theta}}^T(\underline{x},\underline{\theta})\{M(\underline{\theta})\}^{-1}\eta_{\underline{\theta}}(\underline{x},\underline{\theta})$$

$$= \phi_2\{M(\underline{\theta})\}$$

(iii) $\phi_3 = - \int_{\underline{x}\in\mathcal{X}'} \eta_{\underline{\theta}}^T(\underline{x},\underline{\theta})\, M(\underline{\theta})^{-1}\, \eta_{\underline{\theta}}(\underline{x},\underline{\theta})\, p(\underline{x})\, d\underline{x}$

Here again $\eta_{\underline{\theta}}^T(\underline{x},\underline{\theta})M(\underline{\theta})^{-1}\,\eta_{\underline{\theta}}(x,\underline{\theta})$ will represent typically the variance or a first order approximation to the variance of the estimated expected response $\eta(\underline{x}, \hat{\underline{\theta}}_{ML})$. However, here our criterion is not based on the points with highest variance but on the expected value of the variance over some region $\mathcal{X}'$, and with respect to some measure $p(\underline{x})$. $p(\underline{x})$ might reflect the incidence with which the $\underline{x}$ values occur in nature and therefore the incidence with which the estimated response might be being used in a prediction situation. Alternatively $p(\underline{x})$ might be constructed by the experimenter using subjective weights to represent the relative importances of accuracy at points $\underline{x}\in\mathcal{X}'$. $p(\underline{x})$ will be such that $\int_{\underline{x}\in\mathcal{X}'} p(\underline{x})\, d\underline{x} = 1$. Note that we do not necessarily take $\mathcal{X}' \equiv \mathcal{X}$.

Properties of ϕ_3

(a) ϕ_3 is an increasing function on the set of positive definite symmetric matrices.

(b) ϕ_3 is a concave function on the set of positive definite symmetric matrices (The proof is analogous to that in Appendix 2).

(c) ϕ_3 is invariant under non-singular transformations of $\underline{\theta}$. The proof of this is analogous to that for ϕ_2.

(iv) $\phi_4 = - \max_{\underline{c}} \dfrac{\underline{c}^T M(\underline{\theta})^{-1} \underline{c}}{\underline{c}^T\underline{c}}$

Typically, with this criterion, one would be interested in estimating any linear combination of the parameters $\underline{\theta}$, and the criterion is such that we would wish to best estimate the worst estimated linear combination of the parameters, the linear combinations being normalised for obvious reasons.

A matrix result (see Graybill) tells us that if M^{-1} is a positive definite symmetric matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_K$, then $\lambda_K > 0$ and $\max_i \lambda_i = \lambda_1 = \max_{\underline{c}} \frac{\underline{c}^T M^{-1} \underline{c}}{\underline{c}^T \underline{c}}$. Therefore, our desire to minimise with respect to our design $\max_{\underline{c}} \frac{\underline{c}^T M(\underline{\theta})^{-1} \underline{c}}{\underline{c}^T \underline{c}}$

$\left(\equiv \text{maximise over design} - \max_{\underline{c}} \frac{\underline{c}^T M(\underline{\theta})^{-1} \underline{c}}{\underline{c}^T \underline{c}} = \phi_4(M(\underline{\theta})) \right)$ is equivalent, mathematically, to minimising the maximum eigenvalue of $M(\underline{\theta})^{-1}$, for which reason it has become known as the E-optimality criterion.

Properties of ϕ_4

(a) ϕ_4 is an increasing function over the positive definite symmetric matrices

(b) ϕ_4 is a concave function over the positive definite symmetric matrices. (The proof is analogous to that of Appendix 2).

1.3.2 In this section we shall assume that the experimenter is interested in only a subset of the parameters, but is interested in them jointly. We shall consider one criterion.

(i) $\phi_5 = \log \det \{M_s(\underline{\theta})\}$.

Let us assume, without loss of generality, that we are interested only in the first s of our parameters, $s < K$. Partition $M(\underline{\theta})$ as follows

$$M(\underline{\theta}) = \begin{matrix} \\ s \\ (K-s) \end{matrix} \begin{matrix} \begin{matrix} \quad s \qquad\qquad (K-s) \end{matrix} \\ \begin{bmatrix} M_{11} & M_{12} \\ M_{12}^T & M_{22} \end{bmatrix} \end{matrix}$$

The variance covariance matrix of the marginal distribution of $\underline{\theta}_s = (\theta_1, \ldots, \theta_s)^T$ is given by M_s^{-1} where $M_s = M_{11} - M_{12} M_{22}^{+} M_{12}^T$ and M_{22}^{+}

denotes the Moore-Penrose generalised inverse of M_{22}, to allow for the possibility of $\underline{\theta}_{K-s} = (\theta_{s+1}, \ldots, \theta_K)^T$ not being estimable. Motivation for the use of this criterion comes from similar reasons to those discussed under ϕ_1.

Properties of ϕ_5

(a) ϕ_5 is an increasing function over the positive definite symmetric matrices.

(b) ϕ_5 is a concave function over the positive definite symmetric matrices. (see Silvey(1974)).

Because the natural domain of definition of ϕ_5 may include a subset of the positive semi-definite symmetric matrices the above properties must be extended to cover this more general case (Silvey (1974)).

1.3.3 In this section we shall assume that the experimenter is interested in using his parameter estimates independently. Therefore,he will be interested only in the marginal distributions of each θ_i and not in the joint distribution of $\underline{\theta}$. Our design criteria will therefore be functions of the diagonal elements of $M(\underline{\theta})^{-1}$ only. We present two possible criteria.

(i) $\phi_6 = -$ trace $\{M(\underline{\theta})^{-1}\}$

Maximising ϕ_6 will be equivalent to minimising the sum of the marginal variances of the θ_i's. This criterion was considered by Elving (1952) and Chernoff (1953).

Properties of ϕ_6

(a) ϕ_6 is an increasing function over the set of positive definite symmetric matrices.

(b) ϕ_6 is a concave function over the set of positive definite symmetric matrices. (The proof is analogous to that of Appendix 2).

(ii) $\phi_7 = -\,\mathrm{mde}\{M(\underline{\theta})^{-1}\}$, (mde ≡ maximum diagonal element).

This criterion seems to have been almost universally ignored in the literature. Maximising ϕ_7 is seen to be equivalent to best estimating the worst estimated parameter in the sense of minimising its marginal variance. In the context where interest is in the parameters independently this criterion would seem to be of a very suitable type. Criteria similar to ϕ_6 might allow some of the parameters to be fairly poorly estimated although the average variance might be small.

Properties of ϕ_7

(a) ϕ_7 is an increasing function over the positive definite symmetric matrices.

(b) ϕ_7 is a concave function over the positive definite symmetric matrices. (see Appendix 2).

1.3.4

To summarise this section we note that all of the criteria which we have considered have the following two properties.

(a) ϕ is an increasing function over the positive definite symmetric matrices.

(b) ϕ is a concave function over the positive definite symmetric matrices.

Property (a) is a property which we would expect, intuitively, all criteria to have. It might be thought of as saying that an extra observation or set of observations will always make an experiment more informative, in any reasonable sense. As we shall see in the next chapter (b) is a very useful property for all our criteria to have and most of that chapter will depend on it.

We note here that although natural criteria for design will be of the form $\phi\{N.M(\underline{\theta})\}$ we take our criteria as $\phi\{M(\underline{\theta})\}$. This may readily be seen to be an equivalent form for solution of the design problem and will prove to be the most convenient form to use in what follows.

1.4

We now present a result which will be a simplifying factor in most of the examples considered in this thesis. It is presented as a Theorem.

1.4.1 Theorem

Let $p(y|\underline{x},\underline{\theta})$ be of the form $f(y,\underline{x},\eta(\underline{\theta},\underline{x}))$, that is, the probability density function of y depends on $\underline{\theta}$ only through the function $\eta(\underline{\theta},\underline{x})$ which is independent of y, then

$$\text{(i)}\quad I(\underline{x},\underline{\theta}) = \mathbb{E}_y \left\{ -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(y|\underline{x},\underline{\theta}) \right\}$$

$$= \frac{1}{a(\underline{x},\underline{\theta})} \cdot \eta_{\underline{\theta}}(\underline{x},\underline{\theta}).\eta_{\underline{\theta}}(\underline{x},\underline{\theta})^T$$

That is, $I(\underline{x},\underline{\theta})$ is of rank one.
If also $\eta(\underline{x},\underline{\theta}) = \mathbb{E}(y|\underline{x})$ and $p(y|\underline{x},\underline{\theta})$ is such that the Cramér-Rao lower bound for unbiased estimators of $\eta(\underline{x},\underline{\theta})$ is attained then

$$\text{(ii)}\quad I(\underline{x},\underline{\theta}) = \frac{1}{\mathrm{var}(y|\underline{x})} \eta_{\underline{\theta}}(\underline{x},\underline{\theta}).\eta_{\underline{\theta}}(\underline{x},\underline{\theta})^T$$

Proof

$$\text{(i)}\quad I(\underline{x},\underline{\theta}) = \mathbb{E}_y \left[-\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f \right] = \mathbb{E}_y \left[\frac{\partial \log f}{\partial\theta_i} \cdot \frac{\partial \log f}{\partial\theta_j} \right]$$

$$\frac{\partial \log f}{\partial\theta_i} = \frac{\partial \log f}{\partial\eta} \cdot \frac{\partial\eta(\underline{x},\underline{\theta})}{\partial\theta_i}$$

$$\therefore I(\underline{x},\underline{\theta}) = \mathbb{E}_y \left\{ \left(\frac{\partial \log f}{\partial\eta} \right)^2 \right\} \eta_{\underline{\theta}}(\underline{x},\underline{\theta}).\eta_{\underline{\theta}}(\underline{x},\underline{\theta})^T$$

$$= \frac{1}{a(\underline{x},\underline{\theta})} \eta_{\underline{\theta}}(\underline{x},\underline{\theta}).\eta_{\underline{\theta}}(\underline{x},\underline{\theta})^T$$

(ii) $y|\underline{x}$ itself is an unbiased estimate of $\eta(\underline{x},\underline{\theta})$. Therefore, if the Cramér-Rao lower bound is attained

$$\text{var}\,(y|\underline{x}) = \frac{1}{\sum_y \{(\frac{\partial \log f}{\partial \eta})^2\}}\,, \quad \text{Silvey (1970)}$$

$$\therefore\ a(\underline{x},\underline{\theta}) = \text{var}\,(y|\underline{x})$$

$$\therefore\ I(\underline{x},\underline{\theta}) = \frac{1}{\text{var}\,(y|\underline{x})}\ \eta_{\underline{\theta}}(\underline{x},\underline{\theta}).\ \eta_{\underline{\theta}}(\underline{x},\underline{\theta})^T$$

Two examples where the conditions of (ii) are satisfied are

(i) If $p(y|\underline{x},\underline{\theta}) \sim N(\eta\,(\underline{x},\underline{\theta}),\sigma^2\,)$, σ^2 assumed known, then,

$$I(\underline{x},\underline{\theta}) = \frac{1}{\sigma^2}\ \eta_{\underline{\theta}}(\underline{x},\underline{\theta}).\ \eta_{\underline{\theta}}(\underline{x},\underline{\theta})^T$$

(ii) If $p(y|\underline{x},\underline{\theta})$ is a general binary response model, that is

$$p(1|\underline{x},\underline{\theta}) = \eta(\underline{x},\ \underline{\theta})\ ,\ p(0|\underline{x},\underline{\theta}) = 1 - \eta(\underline{x},\underline{\theta})$$

then $$I(\underline{x},\underline{\theta}) = \frac{1}{\eta(\underline{x},\underline{\theta})\ \{1 - \eta(\underline{x},\underline{\theta})\}}\ .\ \eta_{\underline{\theta}}\ (\underline{x},\ \underline{\theta})\ .\ \eta_{\underline{\theta}}(\underline{x},\underline{\theta})^T$$

An obvious situation where the conditions of the theorem are not satisfied is if

$$p(y|\underline{x},\underline{\theta}) \sim N(\underline{f}^T\underline{\phi}\ ,\sigma^2) \text{ and } \underline{\theta} = (\underline{\phi}^T,\sigma^2\ )^T\ .$$

Here $$I(\underline{x},\underline{\theta}) = \frac{1}{\sigma^2}\cdot \begin{bmatrix} & & 0 \\ & \underline{f}\,\underline{f}^T & \\ & & 0 \\ 0 & 0 & \frac{1}{2}\sigma^2 \end{bmatrix}\ .$$

Obviously $I(\underline{x},\underline{\theta})$ has rank 2. Note however that in this example if we are actually interested in only the parameters $\underline{\phi}$, even if σ^2 is unknown, then, because $\underline{\phi}$ and σ^2 are estimated independently, one's design interests would be directed towards the leading $(K-1) \times (K-1)$ sub-matrix

of $I(\underline{x},\underline{\theta})$, namely $I(\underline{x},\underline{\phi}) = \underline{f}\,\underline{f}^T$ which has rank 1.

1.4.2

Suppose the conditions of Theorem 1.4.1 are upheld.

Let $\underline{v} = \dfrac{1}{\{a(\underline{x},\underline{\theta})\}^{\frac{1}{2}}} \cdot \eta_{\underline{\theta}}(\underline{x},\underline{\theta})$, $\underline{v} = (v_1,\ldots,v_K)^T$,

then our design problem may be written equivalently as, maximise

$$\phi\{M(\underline{\theta})\} \quad , \; M(\underline{\theta}) = \sum_{i=1}^{r} p_i \, \underline{v}_i \, \underline{v}_i^T ;$$

$\Sigma\, p_i = 1$, $\underline{v} \in V$, where V represents the domain of definition of $\underline{v}$: to distinguish it from the design space $\mathfrak{X}$ we shall call it the induced design space. Note that the only difference, in our static design problem, introduced by having different density functions or different design spaces is that the induced design space is altered. Therefore, it may be seen that it will be the geometry of V which will be the controlling factor in defining an optimal design for a given static design problem.

It is worth noting, at this point, that the induced design space V will, in our general model, typically be a function of $\underline{\theta}$, the vector of unknown parameters. Obviously it will be the manner in which the geometry of V depends on $\underline{\theta}$ which will characterise the dependence of the optimal design on $\underline{\theta}$.

As an example consider the following binary response model.

$$p(1|\underline{x},\underline{\theta}) = \frac{\exp(\theta_1 + \theta_2 x)}{\{1 + \exp(\theta_1 + \theta_2 x)\}} , \qquad \mathfrak{X} \equiv \left[-1, +1\right] .$$

It can easily be shown that

$$\underline{v}(\underline{\theta}) = \frac{\exp\{\frac{1}{2}(\theta_1 + \theta_2 x)\}}{\{1 + \exp(\theta_1 + \theta_2 x)\}} \cdot (1, x)^T .$$

Fig. 1.4.2. shows the induced design space for several values of $\underline{\theta}$, illustrating the effect on the geometry of V. Points in the induced design space corresponding to values of +1, 0 and -1 for x are highlighted.

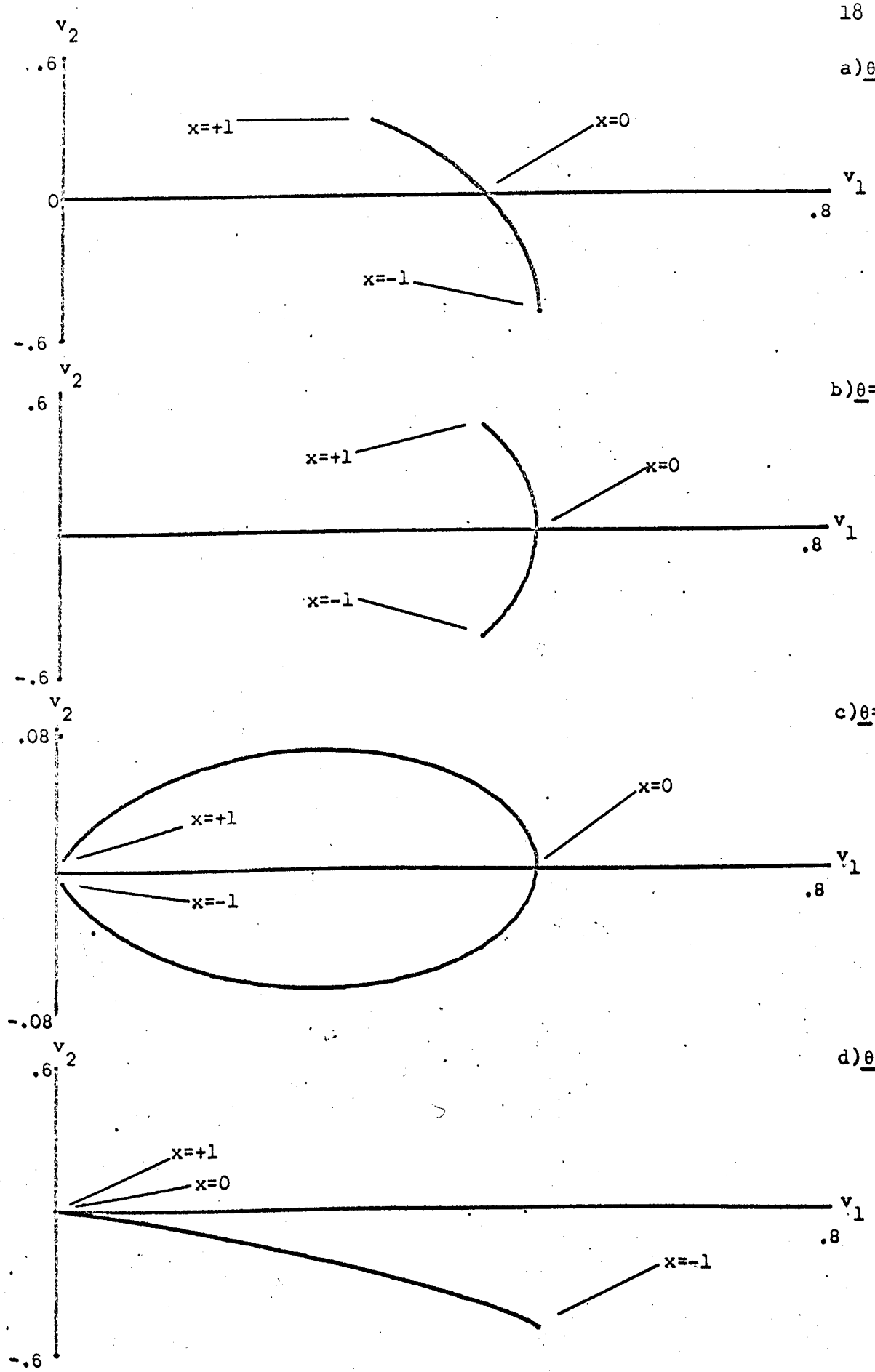

v2
.6
a)θ=(1,1)T
x=+1
x=0
0
v1
.8
x=-1
-.6
v2
b)θ=(0,1)T
.6
x=+1
x=0
v1
.8
x=-1
-.6
v2
c)θ=(0,10)T
.08
x=0
x=+1
v1
.8
x=-1
-.08
v2
d)θ=(10,10)T
.6
x=+1
x=0
v1
.8
x=-1
-.6

FIG. 1.4.2

CHAPTER 2

THE OPTIMAL STATIC DESIGN PROBLEM

2.1 In Chapter 1 we have provided motivation for the study of the following mathematical programming problem.

Take a function ϕ on the positive definite symmetric matrices with the following properties

P1: ϕ is an increasing function on the positive definite symmetric matrices.

P2: ϕ is a concave function on the positive definite symmetric matrices.

If a function, for example ϕ_5, is defined and finite on a wider class of matrices, then the above properties may be adjusted accordingly.

Problem 1: Given a set of points $\mathcal{X}$, we require to select a number, r say, of points from $\mathcal{X}$, and a set of positive weights associated with these points. We shall call our choice, namely

$$\begin{pmatrix} p_1, \ldots\ldots, p_r \\ \underline{x}_1, \ldots\ldots, \underline{x}_r \end{pmatrix},$$

a design measure and denote it by ξ_N. There will be restrictions on our choice of p_i's in that $p_i = \frac{n_i}{N}$, $\sum_{i=1}^{r} n_i = N$, where all the n_i's and N are positive integers. Our choice of ξ_N should be such that we maximise

$$\phi\{M(\xi_N)\}, \quad M(\xi_N) = \sum_{i=1}^{r} p_i \, I(\underline{x}_i),$$

where $I(\underline{x})$ is a positive semi-definite symmetric matrix defined on $\mathcal{X}$ and assumed to be continuous on $\mathcal{X}$. $\mathcal{X}$ will be assumed to be a compact set.

We compare the above practical exact problem with its continuous analogue.

Problem 2: Select a design measure ξ from the set Ξ of measures on the Borel sets of $\mathcal{X}$ to maximise

$$\phi\{ M(\xi) \} , \quad M(\xi) = \int_{\underline{x}\in\mathcal{X}} I(\underline{x})\ \xi(d\underline{x}), \quad \int_{\underline{x}\in\mathcal{X}} \xi(d\underline{x}) = 1.$$

Problem 2 will be the main item of study in this chapter.

Interest in this problem would seem to have its roots in a paper of Smith (1918). The problem would then seem to have been largely ignored until a revival of interest in the early 1950's, indicated by the papers of Elving (1952) and Chernoff (1953). Not much progress was made thereafter until the paper of Kiefer and Wolfowitz (1960), subsequent to which a great deal of attention has been paid to the solution of Problem 2 both in particular and general situations.

For the remainder of this section we consider justifications for this study. The following theorem has important repercussions in this direction.

Theorem 2.1.1: (Caratheodory's Theorem)

Each point s^* in the convex hull S^* of any subset S, of n-dimensional space, can be represented in the form,

$$s^* = \sum_{i=1}^{n+1} \alpha_i\ s_i, \quad \text{where} \quad \alpha_i \geq 0, \quad \sum_{i=1}^{n+1} \alpha_i = 1, \quad s_i \in S.$$

If s^* is a boundary point of the set S^* then α_{n+1} can be set equal to zero.

To see the importance of the above we note that any element of

$$\mathcal{M} = \{ M(\xi) = \int_{\underline{x}\in\mathcal{X}} I(\underline{x})\ \xi(d\underline{x}) , \ \xi \in \Xi \}$$

may be defined as a point in $\frac{1}{2}K(K+1)$ dimensional space, whose coordinates are defined to be the upper triangular elements of $M(\xi)$. It is also true that these points form a convex subset of $\frac{1}{2}K(K+1)$ dimensional space which is the convex hull of the set of points obtained from the set of matrices $\{ I(\underline{x}), \ \underline{x}\in \mathcal{X} \}$.

Note also that by P1, for $a > 1$, $\phi\{aM\} > \phi\{M\}$. Therefore, if M is an interior point of $\mathcal{M}$, then so is $M^1 = (1 + \varepsilon)M$, for sufficiently small $\varepsilon > 0$, and $\phi\{M^1\} > \phi\{M\}$. By hypothesising that an $M(\xi)$ which maximises ϕ is an interior point of $\mathcal{M}$ we see that there must be a matrix of greater ϕ value on the boundary, by increasing ε . Therefore, we have that ϕ attains its maximum value on the boundary of $\mathcal{M}$.

From this observation, and from Th.2.1.1, we have

(i) Any design matrix may be attained by a design measure which attaches positive weight to at most $\frac{1}{2}K(K+1)+1$ points.

(ii) An optimal design measure can be found which attaches positive weight to at most $\frac{1}{2}K(K+1)$ points.

NOTES: 1. Often the number of points required will be considerably less than $\frac{1}{2}K(K+1)$

2. Problem 1 will often be virtually impossible to find a solution to, whilst, as we shall see, Problem 2 will typically be solvable. The fact that Problem 2 has a solution which attaches positive weight to at most $\frac{1}{2}K(K+1)$ points suggests that for large N we may be able to approximate to our optimal design measure ξ^* by one of the form ξ_N.

3. Much of our motivation for study of this form of design problem came from large sample results, and, in such circumstances, we might presume that good approximations, as suggested in 2. above, will be obtained.

4. Even if we do not consider our approximation to be satisfactory, we might use the approximation as a starting point in an iterative algorithm to search for possible improvements. We shall consider this further in Chapter 3.

5. Even for small N, with our linear example of Chapter 1, solution of the continuous problem may be of interest, as some problems will have continuous solutions ξ^*, of the form ξ_N.

2.2: The most important result in this section is presented in the form of an equivalence theorem. The initial breakthrough in this area was made by Kiefer and Wolfowitz (1960), who proved an equivalence theorem with ϕ_1 as criterion. Later Karlin and Studden (1966) generalised this to criterion ϕ_5. Further improvements in the direction of more general concave criteria may be found in Fedorov (1972), Fedorov and Malyutov (1972), Whittle (1973), Silvey and Titterington (1974) and Kiefer (1974). In the statement of this theorem we will closely follow Whittle (1973). An appreciation of the generality of Whittle's theorem and the simplicity of its proof may be obtained from a comparison with Karlin and Studden (1966) and Kiefer (1974).

As an introduction to this theorem we make some definitions.

Definition 2.2.1

Define the directional derivative

$$\Phi\{M,N\} = \lim_{\varepsilon \to 0^+} \varepsilon^{-1}\{\phi\{(1-\varepsilon)M + \varepsilon N\} - \phi\{M\}\}$$

as the derivative of ϕ at M in the direction of another matrix N. As ϕ is concave, the quotient after the limit will be monotone increasing and will always exist if we allow a value of $+\infty$.
Note that $\Phi\{M,M\} = 0$.

Definition 2.2.2

We define the function ϕ to be differentiable at $M \in \mathcal{M}$ if

$$\Phi\{M, \Sigma\ p_j M_j\} = \Sigma\ p_j \Phi\{M,M_j\}, \Sigma\ p_j = 1, M_j \in \mathcal{M}, \forall_j.$$

A theorem ensures that this is in keeping with the normal definition of differentiability.

Definition 2.2.3

Define the maximal rate of ascent of ϕ from $M \in \mathcal{M}$ by

$$\Phi^*\{M\} = \sup_{N \in \mathcal{M}} \Phi\{M,N\}.$$

Note that if ϕ is differentiable at M then

$\Phi^*\{M\} = \sup_{x \in \mathcal{X}} \Phi\{M, I(\underline{x})\}$, as $N \in \mathcal{M}$ may be written in the form $N = \Sigma\ p_j I(\underline{x}_j)$.

Definition 2.2.4

A design matrix M^* is said to be ϕ-optimal if

$$\phi\{ M^*\} = \sup_{M \in \mathcal{M}} \phi\{ M \} .$$

Note that if ξ_1 and ξ_2 are design measures in the set of design measures Ξ then $(1-\alpha)\, \xi_1 + \alpha \xi_2 \in \Xi$ and

$$M\{(1-\alpha)\, \xi_1 + \alpha\xi_2\} = (1-\alpha)\, M\{\xi_1\} + \alpha\, M\{\xi_2\} .$$

Therefore, in the following theorem we might take the design measure ξ in place of the design matrix M, as in fact Whittle (1973) does. However, it will suit our purposes to consider ϕ as a function of symmetric matrices rather than as a function of design measures.

Theorem 2.2.1: (General Equivalence Theorem of Whittle (1973))

(a) If ϕ is concave then a ϕ-optimal design matrix M^* can equivalently be characterised by any of the three conditions:-

(i) M^* maximises ϕ

(ii) M^* minimises $\Phi^*\{ M \}$

(iii) $\Phi^*\{M^*\} = 0$.

(b) If ϕ is differentiable at M then (ii) and (iii) may be rewritten:-

(ii)' M^* minimises $\sup_{\underline{x} \in \mathcal{X}} \Phi\{M, I(\underline{x})\}$.

(iii)' $\Phi\{M^*, I(\underline{x})\} = 0$, for all $\underline{x}$ in the spectrum of a design which produces M^*.

The importance of the above theorem is that it provides us with a tool to test for ϕ-optimality of a given M. Note that section (a) of the above theorem provides us with a necessary and sufficient condition for ϕ-optimality of a matrix M even in the absence of differentiability at M. Note also that although we have this tool it may be difficult to use, as the test is equivalent to showing that $\Phi\{M,N\} \leq 0$, $\forall\ N \in \mathcal{M}$, which, in practice, may be extremely difficult to do. If we have differentiability of ϕ at M, then our test is much simpler, that being to show that $\Phi\{ M, I(\underline{x})\} \leq 0$, $\forall\ \underline{x} \in \mathcal{X}$. It may be seen that the

directional derivatives of the various criteria which we have proposed will play an important role in any problem we consider. In Fig.2.2.1 we list the directional derivatives of the seven criteria mentioned in Chapter 1.

Notes on Fig. 2.2.1.

1. For derivation of the directional derivative of ϕ_1 see for example Fedorov (1972). The directional derivatives of ϕ_2, ϕ_3, ϕ_4, ϕ_6 are produced in an analogous manner to that of ϕ_7, which is derived in Appendix 4. The directional derivative of ϕ_5 is due to Davies (1974).

2. The optimal design for criterion ϕ_5 may exist at matrices which are singular. The directional derivative is given under this assumption, the simplification, if we do not have singularity, being obvious. Note the following notation

M is partitioned as $\begin{matrix} & \begin{matrix} s & (K-s) \end{matrix} \\ \begin{matrix} s \\ (K-s) \end{matrix} & \begin{bmatrix} M_{11} & M_{12} \\ M_{12}^T & M_{22} \end{bmatrix} \end{matrix}$

$\begin{matrix} & \begin{matrix} s & (K-s) \end{matrix} \end{matrix}$

N is written XX^T, $X = \begin{matrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} & \begin{matrix} s \\ (K-s) \end{matrix} \end{matrix}$,and again

M^+ denotes the Moore-Penrose inverse of M.

3. Consider example (i) of 1.4.1, that is, non-linear regression with normal error, assumed independent of $\underline{x}$. We note that an optimal design for ϕ_1 will always exist at a non-singular M and that ϕ_1 will always be differentiable there. Note also that $I(\underline{x},\theta) = \frac{1}{\sigma^2} \eta_{\underline{\theta}}(\underline{x},\underline{\theta}).\eta_{\underline{\theta}}^T(\underline{x},\underline{\theta})$

Part (b) of Theorem 2.2.1. will apply and may be written in the following form, for criterion ϕ_1.

An optimal design matrix M* can equivalently be characterised by

(i) M* maximises log det {M}

(ii) M* minimises $\{ \frac{1}{\sigma^2} \max_{\underline{x}\in\mathcal{X}} \eta_{\underline{\theta}}^T(\underline{x},\underline{\theta})\, M^{-1}\, \eta_{\underline{\theta}}(\underline{x},\underline{\theta}) - K \}$.

(iii) $\frac{1}{\sigma^2} \max_{\underline{x}\in\mathcal{X}} \eta_{\underline{\theta}}^T(\underline{x},\underline{\theta})\, M^{*-1}\, \eta_{\underline{\theta}}(\underline{x},\underline{\theta}) = K$.

This is in fact the equivalence theorem of Kiefer and Wolfowitz (1960), and we note by comparing (i) and (ii) that finding a ϕ_1-optimal design, in this example, will be equivalent to finding a ϕ_2 optimal design. In other situations, where condition (ii) of Theorem 1.4.1 is satisfied, we may consider an alternative criterion to ϕ_2, namely

$$\phi_2^1 = \frac{1}{\mathrm{var}(y(\underline{x}))} \cdot \eta_{\underline{\theta}}(\underline{x},\underline{\theta})^T\, M^{-1}\, \eta_{\underline{\theta}}(\underline{x},\underline{\theta}))$$

$$\doteq \frac{\mathrm{var}(\eta(\underline{x},\hat{\underline{\theta}}))}{\mathrm{var}\,(y(\underline{x}))} \quad .$$

It may readily be seen that in this situation maximising ϕ_1 will be equivalent to maximising ϕ_2^1.

4. It has been mentioned above that problems may arise when functions are not differentiable at points in $\mathcal{M}$ at which the function may be maximised. By considering Fig.2.2.1 we may see that non-differentiability of ϕ_2, ϕ_4 and ϕ_7 may occur, due to non-uniqueness of the * values. This problem will be highlighted for ϕ_7 in Appendix 2. Kiefer (1974) has considered the problem for ϕ_4. For ϕ_2 the situation is potentially more difficult, but if there is equivalence with the ϕ_1 design problem it may be unnecessary to face the problem in practice. ϕ_5 will be seen, in the next section, to be non-differentiable at singular $M \in \mathcal{M}$. In the following section we consider problems raised by non-differentiability, with illustrations using ϕ_5 and ϕ_7.

i	$\phi_i\{M\}$	$\Phi\{M,N\}$
1	$\log\det\{M\}$	$\mathrm{tr}\{M^{-1}N\} - K$
2	$-\max_{\underline{x}\in\mathcal{X}} \eta_{\underline{\theta}}^T(\underline{x},\underline{\theta})M^{-1}\eta_{\underline{\theta}}(\underline{x},\underline{\theta})$	$\eta_{\underline{\theta}}^T(\underline{x}^*,\underline{\theta})[M^{-1}NM^{-1} - M^{-1}]\eta_{\underline{\theta}}(\underline{x}^*,\underline{\theta})$, $\underline{x}^*$ is the maximising value in col.2.
3	$-\int_{\underline{x}\in\mathcal{X}} \eta_{\underline{\theta}}^T(\underline{x},\underline{\theta})M^{-1}\eta_{\underline{\theta}}(\underline{x},\underline{\theta})p(\underline{x})d\underline{x}$	$\int_{\underline{x}\in\mathcal{X}} \eta_{\underline{\theta}}^T(\underline{x},\underline{\theta})[M^{-1}NM^{-1} - M^{-1}]\ \eta_{\underline{\theta}}^T(\underline{x},\underline{\theta})\ p(\underline{x})d\underline{x}$.
4	$-\max_{\underline{c}} \dfrac{\underline{c}^T M^{-1}\underline{c}}{\underline{c}^T\underline{c}}$	$\dfrac{\underline{c}^{*T}[M^{-1}NM^{-1} - M^{-1}]\underline{c}^*}{\underline{c}^{*T}\underline{c}}$, $\underline{c}^*$ is the maximising value in col.2
5	$\log\det\{M_s\}$	$\mathrm{tr}\{(X_1 - M_{12}M_{22}^+X_2)\{I - X_2^T\{(I-M_{22}M_{22}^+)X_2X_2^T(I - M_{22}M_{22}^+)\}^+X_2\}$ $(X_1-M_{12}M_{22}^+X_2)^T(M_{11}-M_{12}M_{22}^+M_{12}^T)\}-s$
6	$-\mathrm{tr}\{M^{-1}\}$	$\mathrm{tr}\{M^{-1}NM^{-1} - M^{-1}\}$
7	$-\mathrm{m.d.e}\{M^{-1}\}$	$\{M^{-1}NM^{-1} - M^{-1}\}_{ss}$, where $\{M^{-1}\}_{ss^*}$ is m.d.e $\{M^{-1}\}$

Fig. 2.2.1.

2.3: As mentioned above, if ϕ is differentiable at M, then all that is required, to check optimality of M, is to show that

$$\Phi\{M,I(\underline{x})\} \leq 0, \forall\ \underline{x} \in \mathcal{X} \quad .$$

When we have non-differentiability of ϕ at M the problem arises that

$$\Phi\{M,I(\underline{x})\} \leq 0, \forall\ \underline{x} \in \mathcal{X} \quad ,$$

does not necessarily imply that

$$\Phi\{M,N\} \leq 0, \forall\ N \in \mathcal{M}.$$

However, note that there is no problem if

$$\Phi\{M,I(\underline{x})\} > 0$$

for some $\underline{x} \in \mathcal{X}$ as this is sufficient to show non-optimality of M. The following examples will illustrate the above.

Example 1 (Silvey (1974)).

Suppose that the conditions of Theorem 1.4.1. hold and we have induced design space $V = \{ (0,0), (1,0), (4,1), (4,2) \}$.
Take ϕ_5 as criterion where we are interested in the first parameter only, that is, D_1 optimality.

Consider the design measure $\eta_1 = \begin{pmatrix} 1 \\ (1,0) \end{pmatrix}$

$$M(\eta_1) = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad M^{+}(\eta_1) = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

From Fig.2.2.1 we have

$$\Phi\{M(\eta_1), \underline{v}\,\underline{v}^T\} = \begin{cases} v_1^2 - 1 , & v_2 = 0 \\ -1 , & v_2 \neq 0 \end{cases} , \quad \underline{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

$$\therefore \quad \Phi\{M(\eta_1), \underline{v}\,\underline{v}^T\} \leq 0, \quad \forall\ \underline{v} \in V .$$

Consider now the design measure $\eta_2 = \begin{pmatrix} \frac{1}{2} & , & \frac{1}{2} \\ (4,1) & , & (4,2) \end{pmatrix}$,

then from Fig. 2.2.1 we have

$$\Phi\{ M(\eta_1), M(\eta_2) \} = \frac{3}{5} ,$$

illustrating the non-optimality of η_1 and the non differentiability at $M(\eta_1)$.

Example 2:

Again assume conditions of Theorem 1.4.1 to hold. Let the induced design space be

$$V = \{ (1,0), (0,1), (2,0), (0,2) \} .$$

Take ϕ_7 as criterion.
Consider the measure $\eta_1 = \begin{pmatrix} \frac{1}{2} & , & \frac{1}{2} \\ (0,1) & , & (1,0) \end{pmatrix}$.

$$M(\eta_1) = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} , \quad M(\eta_1)^{-1} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} ;$$ note coincident maximum diagonal elements.

From Fig.2.2.1. and Appendix 4.

$$\Phi\{M(\eta_1), \underline{v}\,\underline{v}^T \} = \{ M^{-1}\,\underline{v}\,\underline{v}^T M^{-1} - M^{-1} \}_{ss}$$

$$= -2, \forall\ \underline{v} \in V.$$

Take $\eta_2 = \begin{pmatrix} \frac{1}{2} & , & \frac{1}{2} \\ (0,2) & , & (2,0) \end{pmatrix}$, $M(\eta_2) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

We have $\Phi\{M(\eta_1), M(\eta_2)\} = 2$, implying non optimality of η_1 and non-differentiability of ϕ_7 at $M(\eta_1)$.

We again note that the convex set of matrices $\mathcal{M}$ may be represented by a convex sub-set of $\frac{1}{2}K(K+1)$ dimensional space, the coordinates of the members of which are given by the elements of the upper triangular parts of the matrices of which $\mathcal{M}$ is composed. Call this convex subset $\mathcal{M}'$. The following points are highlighted as Lemmas.

Lemma 2.3.1.

$$\phi\{ a\,\underline{m}'\} > \phi\{\underline{m}'\} \quad , \quad \underline{m}' \in \mathcal{M}', \quad a > 1.$$

This follows immediately from the fact that

$$\phi\{ aM \} > \phi\{M\} \quad , \quad M \in \mathcal{M}, \quad a > 1 .$$

This tells us that if we think of the convex set $\mathcal{M}'$ as being surrounded by the minimal convex cone which will contain it, then the only feasible points in $\mathcal{M}'$, corresponding to possible optimal M, lie on the upper surface of $\mathcal{M}'$ not hidden by the cone. In a three dimensional analogy with an ice-cream cone one might consider the ice-cream as representing $\mathcal{M}'$ and the surface of the ice-cream, which is visible, as representing the feasible region.

Lemma 2.3.2.

$$\Phi\{A, \lambda A +(1-\lambda)B\} = (1-\lambda)\,\Phi\{ A,B\},\ \lambda \in [0,1];$$

that is, $\Phi\{ A,B \} \leq 0 \Rightarrow \Phi\{ A, \lambda A +(1-\lambda)B\} \leq 0.$
This true whether ϕ is differentiable at A or not.

Proof:

$$\Phi\{A,\lambda A +(1-\lambda)B \} = \lim_{\varepsilon\to 0^+} \varepsilon^{-1}\{\phi\{(1-\varepsilon)A +\varepsilon\lambda A +\varepsilon(1-\lambda)B\} -\phi\{ A\} \}$$

$$= \lim_{\varepsilon\to 0^+} \varepsilon^{-1}\{\phi\{ (1-\varepsilon(1-\lambda))A + \varepsilon(1-\lambda)B\} - \phi\{A\} \}$$

$$= (1-\lambda) \lim_{\varepsilon\to 0^+} (1-\lambda)^{-1}\, \varepsilon^{-1}\{\phi\{(1-\varepsilon(1-\lambda))A +\varepsilon(1-\lambda)B\} -\phi\{ A\}\}$$

$$= (1-\lambda) \lim_{\varepsilon(1-\lambda)\to 0^+} (1-\lambda)^{-1}\varepsilon^{-1}\{\phi\{ (1-\varepsilon(1-\lambda))A+\varepsilon(1-\lambda)B\}-\phi\{ A\}\}$$

$$= (1-\lambda)\,\Phi\{ A,B \} .$$

Consider the effect the above has on Example 1 above. $\mathcal{M}'$ is the convex hull of the set $\{ A,B,C,D \} = \{(0,0,0), (1,0,0),(16,4,1),(16,8,4)\}$, see Fig. 2.3.1.

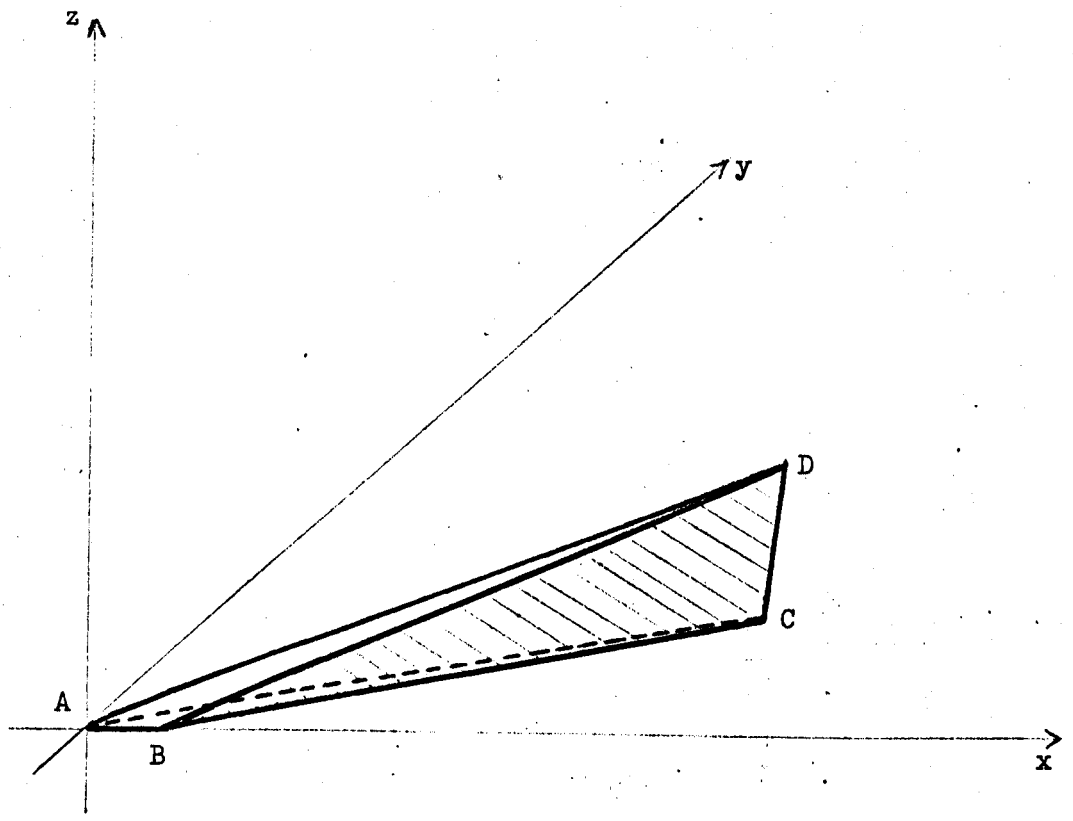


Fig. 2.3.1

By Lemma 2.3.1, the only feasible optimal points lie on triangle BCD. If we are testing $M(\eta_1)$ for optimality then by Lemma 2.3.2. we need only consider directional derivatives from point B to points on the line CD.

Any matrix represented by a point on the line CD may be written in the form

$$M(\alpha) = \alpha \begin{bmatrix} 16 & 4 \\ 4 & 1 \end{bmatrix} + (1-\alpha) \begin{bmatrix} 16 & 8 \\ 8 & 4 \end{bmatrix} = \begin{bmatrix} 16 & 8-4\alpha \\ 8-4\alpha & 4-3\alpha \end{bmatrix}$$

We note here that if M is non-singular then the directional derivative for ϕ_5, given in Fig.2.2.1, may be written as

$$\mathrm{tr}\ \{ M^{-1} N - M_{22}^{-1} N_{22} \} - s,$$

with the usual matrix partitioning notation.

If M is singular, but N is not, we might define the directional derivative $\Phi(M,N)$ alternatively as

$$\lim_{\varepsilon \to 0^+} \Phi\{(1-\varepsilon) M + \varepsilon N, N\} .$$

In this example $M(\alpha)$ is nonsingular, $\alpha \in (0,1)$.

$$(1-\varepsilon)M(\eta_1) + \varepsilon\, M(\alpha) = \begin{bmatrix} 1 + 15\varepsilon & \varepsilon(8-4\alpha) \\ \varepsilon(8-4\alpha) & \varepsilon(4-3\alpha) \end{bmatrix},$$

$$\{(1-\varepsilon)M(\eta_1) + \varepsilon\, M(\alpha)\}^{-1} = \frac{1}{\varepsilon\{(1+15\varepsilon)(4-3\alpha)-\varepsilon(8-4\alpha)^2\}} \begin{bmatrix} \varepsilon(4-3\alpha) & -\varepsilon(8-4\alpha) \\ -\varepsilon(8-4\alpha) & (1+15\varepsilon) \end{bmatrix},$$

$$\Phi\{(1-\varepsilon)M(\eta_1)+\varepsilon M(\alpha), M(\alpha)\} = \frac{1}{\varepsilon} \cdot \frac{(4-4\varepsilon-3\alpha-32\alpha^2\varepsilon+35\alpha\varepsilon)}{(4-4\varepsilon-3\alpha+19\alpha\varepsilon-16\alpha^2\varepsilon)} - \frac{1}{\varepsilon} \cdot \frac{(4-3\alpha)}{(4-3\alpha)} - 1$$

$$= \frac{-16\alpha^2 + 16\alpha}{4 - 4\varepsilon - 3\alpha + 19\alpha\varepsilon - 16\alpha^2\varepsilon} - 1$$

$$\therefore \Phi\{M(\eta_1), M(\alpha)\} = \lim_{\varepsilon \to 0^+} \left\{ \frac{16\alpha\,(1-\alpha)}{4-4\varepsilon - 3\alpha + 19\alpha\varepsilon - 16\alpha^2\varepsilon} - 1 \right\}$$

$$= \frac{16\alpha(1-\alpha)}{4-3\alpha} - 1 \begin{cases} > 0, & \alpha \in [\cdot 2735,\ \cdot 9139] \\ \leq 0, & \text{otherwise} \end{cases}$$

Therefore, in this simple example the problem has been reduced from one of scanning a triangle to one of scanning a line.

Lemmas 1 and 2 will of course apply to ϕ_7 as well. We now give two additional Lemmas which will be particular to ϕ_7. Their proofs are quite straightforward and may be found in Appendix 4.

It has already been noted that non-differentiability of ϕ_7 may occur when M^{-1} has coincident maximum diagonal elements. Let these be $(ss_1, \ldots\ldots, ss_r)$ and define a matrix N to have the ss_j property if $\Phi\{M,N\} = \{M^{-1} N M^{-1} - M^{-1}\}_{ss_j}$.

Lemma 2.3.3.

Let N_1, $N_2 \in \mathcal{M}$ have the ss_j property. Then $\lambda N_1 + (1-\lambda)N_2$ has the ss_j property.

Lemma 2.3.4.

Let $N \in \mathcal{M}$ have the ss_j property. Then σN has the ss_j property, $\sigma > 0$.

What Lemma 2.3.3. tells us is that the members of $\mathcal{M}$ which have the ss_j property, for a given M, form a convex subset of $\mathcal{M}$. We shall denote this subset by $\mathcal{M}ss_j$. Similarly the elements of $\mathcal{M}'$ with the ss_j property form a convex subset of $\mathcal{M}'$, which we shall denote by $\mathcal{M}'ss_j$.

Lemma 2.3.4 tells us that these subsets are formed by intersections of convex cones with $\mathcal{M}'$.
Note that M itself has all the ss_j properties, j=1,....,r. Therefore, M will lie on the intersection

$$\mathcal{M}ss_1 \cap \mathcal{M}ss_2 \cap \ldots\ldots\ldots \cap \mathcal{M}ss_r .$$

It will also be evident from Appendix 4 that

$$\Phi\{M,\lambda M_1 + (1-\lambda)M_2\} = \lambda\Phi\{M,M_1\} + (1-\lambda)\,\Phi\{M,M_2\} ,$$

$$\lambda \in [0,1] \quad ; \quad M_1, M_2 \in \mathcal{M}ss_j, \forall_j \quad .$$

Therefore, although we do not have complete differentiability as in Definition 2.2.2, we have a kind of partial differentiability and hence testing optimality of a given matrix might not always be quite as difficult as Theorem 2.2.1 might indicate. This is because, to check optimality of a matrix M, where ϕ is not differentiable at M, we do not need to calculate $\Phi\{M,N\}$, $\forall$ $N \in \mathcal{M}$, but only for the N which are the generators of the convex sets $\mathcal{M}ss_j$, j=1,....,r, and which are feasible optimal solutions.

In the same sense Lemma 2.3.2. shows that we have a form of partial differentiability in general, in that it shows that we need only look at extreme feasible points of $\mathcal{M}$ from M. Obviously the above will only be of real help when the upper surface of $\mathcal{M}$ is planar, or has planar regions.

Consider again the trivial Example 2 above. $\mathcal{M}'$ is the

convex set obtained using the generators { A,B,C,D } = { (0,0,1), (1,0,0), (4,0,0), (0,0,4)} . See Fig.2.3.2. below.

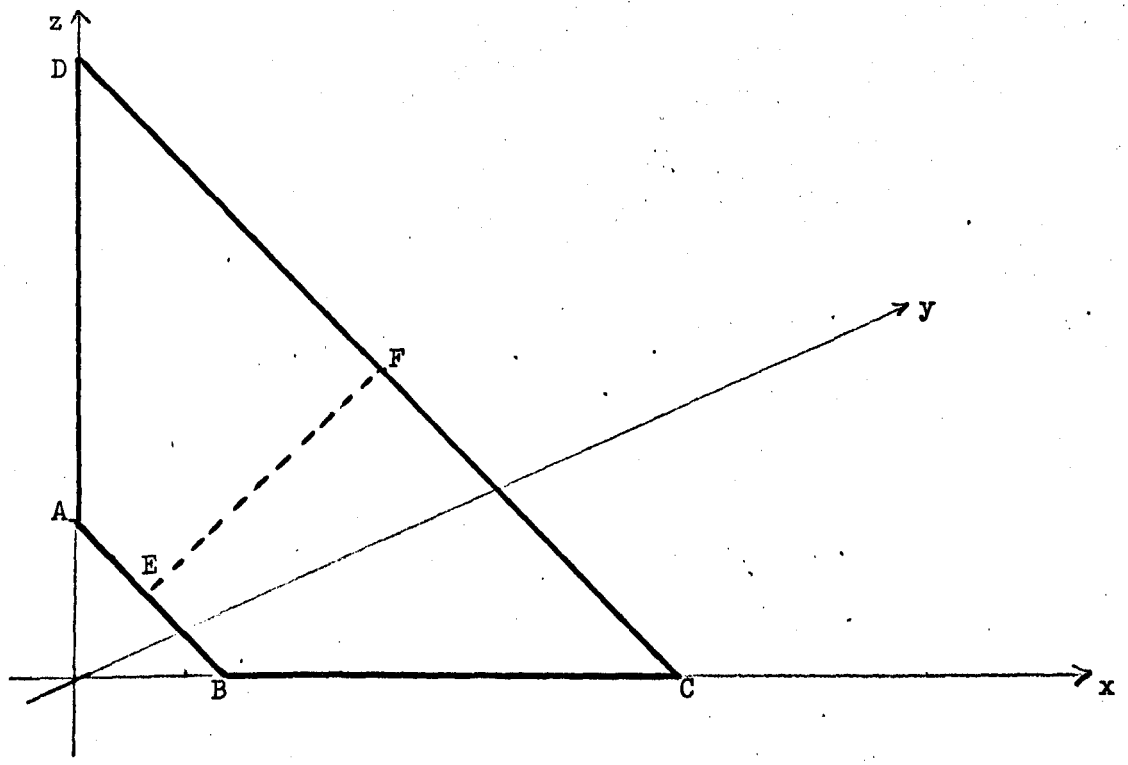


Fig. 2.3.2.

By symmetry, for matrix $M(\eta_1) = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$, we have

$$\mathcal{M}'_{ss_1} \equiv \text{BEFC}, \quad \mathcal{M}'_{ss_2} \equiv \text{AEFD}.$$

Therefore, we need only consider directional derivatives to points A,D,F,C,B. We have already seen that $\Phi\{M(\eta_1), M_F\} > 0$. That is, $M(\eta_1)$ is non-optimal.

2.4: Non-uniqueness of M*

Consider the following example.

Example:

Let the induced design space be $V = \{\underline{v} = (1,a)^T, a\varepsilon[-2. +2]\}$.

Take a design measure $\eta(p) = \begin{Bmatrix} p & , (1-2p) & , p \\ (1,-2) & , (1,0) & , (1,+2) \end{Bmatrix}$, $p \in [0, \frac{1}{2}]$.

$$M(\eta(p)) = \begin{bmatrix} 1 & 0 \\ 0 & 8p \end{bmatrix} , \quad M(\eta(p))^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{8p} \end{bmatrix} , \quad p \in [0, \tfrac{1}{2}]$$

$$M(\eta(o))^{+} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

In this example it may readily be shown that $\eta(p)$ is optimal in the ϕ_4, ϕ_5 (for estimation of the first parameter) and ϕ_7 senses, for $p \in [\frac{1}{8}, \frac{1}{2}]$. That is, our optimal design matrix is not unique. Although $\eta(p)$ is optimal in the above senses, for $p \in [\frac{1}{8}, \frac{1}{2}]$, a better design might be obtained by choosing a particular p value from $[\frac{1}{8}, \frac{1}{2}]$ to optimise some secondary criterion. For example taking $p = \frac{1}{2}$ will produce a design which is also ϕ_1 and ϕ_6 optimal.

This non-uniqueness arises because some of our functions are not strictly concave.

Define strict concavity of ϕ by

$$\phi\{ \lambda A + (1-\lambda)B \} \geq \lambda\phi\{ A\} + (1-\lambda)\phi \{B\} ,$$

with equality if and only if $A = B$.

Functions ϕ_1, ϕ_3, ϕ_6 will always have unique maxima because they are strictly concave functions (see Appendix 3). However ϕ_4, ϕ_5, ϕ_7 may not have unique maxima as illustrated above. ϕ_2, when taken in the form ϕ_2' of 2.2, will have a unique maximum because of its equivalence with ϕ_1. However, this may not always be true for general ϕ_2 (see Appendix 3).

To summarise this, we might say that when using one of ϕ_4, ϕ_5 or ϕ_7 as a primary criterion it would be prudent to be on the look out for possible non-uniqueness of M^*, and the possibility of choosing from the set of optimal designs in order to optimise some secondary criterion.

2.5: An alternative approach to the optimal design theory

Following remarks by Silvey (1972), in the discussion of papers by Wynn and Laycock, that the optimal design problem might be linked to a duality problem; the equivalence theorems were developed for D-optimality (Sibson (1972)), D_s - optimality (Silvey and Titterington (1973)) and for general concave ϕ-optimality (Silvey and Titterington (1974)) as corollaries of stronger duality theorems, using Strong Lagrangian methods. Although the general theory was developed under the assumption that $I(\underline{x})$ was of rank one, there are no complications if $I(\underline{x})$ is taken to have rank greater than one, as may be seen from Silvey and Titterington (1974). However, in order to exploit the above approach in practice by considering the design problem as a geometric covering problem, it is essential that $I(\underline{x})$ has rank one.

In what follows we shall assume that the assumptions of Theorem 1.4.1 hold and we shall denote the induced design space by V.

The geometric interpretations produce duality theorems for D- and D_s-optimality as stated below.

D-optimality

Define the minimal ellipsoid problem as that of finding the ellipsoid $\underline{v}^T M^{-1} \underline{v} \leq K$ of minimal content containing V.

Theorem 2.5.1 (Sibson)

If V is a compact set spanning R^K, the D-optimal design problem for V is the dual of the minimal ellipsoid problem for V and the two problems share a common extreme value.

D_s-optimality

Define the thinnest central cylinder problem as;

Let A be a positive definite sxs matrix and B be a s x(K-s) matrix. The thinnest central cylinder problem for V is that of finding a cylinder

$$(\underline{v}^{(1)} + B\,\underline{v}^{(2)})^T A(\underline{v}^{(1)} + B\,\underline{v}^{(2)}) \leq S, \quad \underline{v} = \begin{bmatrix} \underline{v}^{(1)} \\ \underline{v}^{(2)} \end{bmatrix} \begin{matrix} s \\ (K-s) \end{matrix},$$

containing V and such that the determinant of A is maximised.

Theorem 2.5.2 (Silvey and Titterington)

Let V be any compact subset of R^K which spans the leading s-dimensional co-ordinate subspace. Then, for V, the D_s-optimal design problem is the dual of the thinnest central cylinder problem and the two problems share a common extreme value.

The possibility of practical improvements, using the above geometric interpretations, arises because in some simple situations the geometry of V may enable one to spot optimal ellipsoids or cylinders and allow calculation of optimal designs without recourse to iterative algorithms.

With respect to the afore-mentioned problem of testing optimality of a singular matrix for criterion ϕ_5, it would appear that the above approach will only be of help in the simplest of situations, due to the difficulty in obtaining the optimal matrix B in the above (see Silvey and Titterington (1973)).

2.6: Examples

This section will be devoted to a set of examples of optimal designs which can be calculated explicitly. This, of course, will not always be possible in general. However, the examples given, although simple in form, may have practical applications, and will serve to illustrate sections of the theory described above.

2.6.1: D-optimal designs on K-points

One can imagine many practical situations where an experimenter will be interested in carrying out a designable experiment at the minimum number of design points in the design space, in order to economise on time, money and resources. Therefore, it is of interest to investigate the best such design, and when it will be optimal. We restrict the problem to D-optimal designs in situations where $I(\underline{x})$ is of rank one. Again we represent the design space by V.

Let our design measure be $\xi = \begin{pmatrix} p_1, \ldots\ldots, p_K \\ \underline{v}_1, \ldots\ldots, \underline{v}_K \end{pmatrix}$

$$M(\xi) = \sum_{i=1}^{K} p_i \underline{v}_i \underline{v}_i^T = U \Lambda U^T, \quad U = (\underline{v}_1 \ldots\ldots \underline{v}_K), \quad \Lambda = \text{diag}\{p_i\} .$$

$$|M(\xi)| = |U^2| \prod_{i=1}^{K} p_i , \quad \sum_{i=1}^{K} p_i = 1.$$

$|M(\xi)|$ is maximised with respect to the p_i's when $p_i = \frac{1}{K}, \forall_i$. That is, the best K-point design, in the D-optimum sense, puts weights $\frac{1}{K}$ at each design point, independently of the points chosen.

Suppose now that we have selected a set of K-points $\{\underline{v}_1, \ldots\ldots, \underline{v}_K\}$ with optimal weights.
That is, our design measure is given by $\xi = \begin{pmatrix} \frac{1}{K}, \ldots\ldots, \frac{1}{K} \\ \underline{v}_1, \ldots\ldots, \underline{v}_K \end{pmatrix}$

$$M(\xi) = U \Lambda U^T = \frac{1}{K} U U^T .$$

By Theorem 2.2.1., a necessary and sufficient condition for D-optimality of ξ is

$$\Phi\{M(\xi), N\} \leqslant 0, \forall N \in \mathcal{M} .$$

That is, our K-point design is D-optimal

iff $\quad \Phi\{\frac{1}{K} U U^T, \underline{v} \underline{v}^T\} \leqslant 0, \forall \underline{v} \in V .$

iff $\quad K \underline{v}^T (UU^T)^{-1} \underline{v} - K \leqslant 0 , \forall \underline{v} \in V .$

iff $\quad \underline{v}^T U^{T-1} U^{-1} \underline{v} \leqslant 1 , \forall \underline{v} \in V .$

iff $\quad (U^{-1}\underline{v})^T (U^{-1}\underline{v}) \leqslant 1 , \forall \underline{v} \in V .$

iff $\quad U^{-1}$ transforms every point $\underline{v}$ in V inside the unit hypersphere.

If the design space V is discrete, consisting of m points, then one could compute $|U_{\alpha_i}|$, i=1,....., ${}^{m}C_K$, U_{α_i} representing the matrix obtained from the αi'th selection of K points from the m available. Observing the maximum, the existence of a K-point D-optimal design might be checked using the above necessary and sufficient condition. In certain simple situations the best K-point design may be calculated analytically.

Example 1:

Let V be $\{ \underline{v};\ \underline{v}^T A \underline{v} = 1 \}$, A symmetric, positive definite. That is, V is the surface of an ellipsoid in K dimensions. As we have said the problem is reduced by knowing that the optimal weights are $\frac{1}{K}$. We therefore wish to maximise

$$\left| \sum_{i=1}^{K} \underline{v}_i \underline{v}_i^T \right| = \left| U \right|^2,$$

subject to the constraints

$$\underline{v}_i^T A \underline{v}_i = 1 \ , \forall \ \underline{v}_i \ .$$

Let $A = BB^T$, and $\underline{y} = B^T \underline{v}$.

The problem is equivalent to finding K vectors $\underline{y}_1, \ldots, \underline{y}_K$ to maximise

$$\left| \sum_{i=1}^{K} \underline{y}_i \underline{y}_i^T \right| = \left| Y \right|^2 = \Delta^2 \ ,$$

or equivalently maximise $|Y| = \Delta$, subject to $\underline{y}_i^T \underline{y}_i = 1$.

Introduce Lagrange multipliers $\lambda_1, \ldots, \lambda_K$.

We require to maximise the Lagrangian form

$$|Y| - \sum_{i=1}^{K} \lambda_i (\underline{y}_i^T \underline{y}_i - 1)$$

$$= \begin{vmatrix} y_{11} \cdots y_{1K} \\ \vdots \quad \vdots \\ y_{K1} \cdots y_{KK} \end{vmatrix} - \Sigma \lambda_i (\underline{y}_i^T \underline{y}_i - 1)$$

$$\frac{\partial}{\partial \underline{y}_i} = 0 \Rightarrow \begin{bmatrix} c_{i1} \\ \vdots \\ c_{iK} \end{bmatrix} = 2\lambda_i \underline{y}_i \ , \quad \text{where } c_{ij} \text{ denotes the co-factor of the } (i,j)\text{'th element of Y.}$$

$$\Rightarrow \lambda_i = \frac{\Delta}{2} \ , \forall_i \ , \text{ as } \underline{y}_i^T \begin{bmatrix} c_{i1} \\ \vdots \\ c_{iK} \end{bmatrix} = \Delta = 2\lambda_i$$

Also $$\underline{y}_j^T \begin{bmatrix} c_{i1} \\ \vdots \\ c_{iK} \end{bmatrix} = 0 \quad , \quad i \neq j$$

$$\Rightarrow \quad y_j^T y_i = 0$$

That is, the problem is solved by taking any K orthogonal points on the unit sphere and transforming them back into V space via $\underline{v} = (B^T)^{-1} \underline{y}$.

Example 2.

Consider the problem of allocating weights with which to take observations from 3 populations in order to estimate two unknown parameters best in the D-optimal sense. Let the observations come from Poisson distributions

$$(1) \ \text{Po}(\theta_1) \ , \ (2) \ \text{Po}(\theta_2) \ , \ (3) \ \text{Po}(\theta_1 + \theta_2) \ .$$

The conditions of Theorem 1.4.1 hold, giving an induced design space

$$V = \left\{ \left(\frac{1}{\sqrt{\theta_1}}, 0 \right), \left(0, \frac{1}{\sqrt{\theta_2}} \right), \left(\frac{1}{\sqrt{\theta_1 + \theta_2}}, \frac{1}{\sqrt{\theta_1 + \theta_2}} \right) \right\} .$$

Intuition suggests that taking observations only from populations (1) and (2) might be D-optimal.

$$U^{-1} = \begin{pmatrix} \sqrt{\theta_1} & 0 \\ 0 & \sqrt{\theta_2} \end{pmatrix} .$$

U^{-1} transforms V to $V' = \{(0,1), (1,0), \left(\frac{\sqrt{\theta_1}}{\sqrt{\theta_1 + \theta_2}}, \frac{\sqrt{\theta_2}}{\sqrt{\theta_1 + \theta_2}} \right)\}$

The three points in V' lie on the unit circle for all (θ_1, θ_2) implying D-optimality of the design which allocates observations equally to populations (1) and (2).

Example 3:

It is interesting to compare the above necessary and sufficient condition for the existence of a K-point D-optimal design with the sufficient condition of White (1975). A very simple example illustrates the weakness of White's result.

Let V be the three points on the unit circle

$$\{(0,1),\ (1,0),\ \left(\frac{1}{\sqrt{2}},\ \frac{1}{\sqrt{2}}\right)\}\ .$$

From Example 2 above the optimal two point design, namely

$$\begin{pmatrix} \frac{1}{2} & , & \frac{1}{2} \\ (0,1), & & (1,0) \end{pmatrix} ,$$

is also D-optimal.

White's condition for D-optimality of the K-point design is that

$$\left| U_{\alpha_\ell} \right|^2 \geq \sum_{\substack{i=1 \\ i \neq \ell}}^{t} \left| U_{\alpha_i} \right|^2 , \quad t = {}^{m}C_{K} ,$$

where again α_i denotes the i'th selection of K from m points, α_ℓ being the selection under scrutiny.

Number the three points $(0,1)$, $(1,0)$, $\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$ as 1,2,3 respectively.

$$\alpha_1 = (1,2) \quad , \quad \alpha_2 = (1,3) \quad , \quad \alpha_3 = (2,3).$$

$$\left| U_{\alpha_1} \right|^2 = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix}^2 = 1$$

$$\left| U_{\alpha_2} \right|^2 = \begin{vmatrix} 1 & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{vmatrix}^2 = \frac{1}{2}$$

$$\left| U_{\alpha_3} \right|^2 = \begin{vmatrix} 0 & 1 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{vmatrix}^2 = \frac{1}{2}$$

Observing that $\left| U_{\alpha_1} \right|^2 = \left| U_{\alpha_2} \right|^2 + \left| U_{\alpha_3} \right|^2 = 1$, we note that White's condition confirms D-optimality of α_1. However, the weakness of White's condition is at once evident. If V were to consist of four or more points lying inside the unit circle, with points 1, 2 and 3 being included, then White's condition will cease to hold although α_1 will still trivially be D-optimal by what has been given above.

2.6.2: D_s optimal design to estimate the difference between two population means.

Suppose we have two populations A, B supplying observations with means $\lambda, \lambda + \mu$ respectively. The parameter of interest is μ the difference between the population means. Let the variances of the observations be v_a and v_b. In what follows we shall assume that one of the two following conditions holds.

(i) λ and μ are the only unknown parameters and the full conditions of Theorem 1.4.1 hold.

Examples: The observations may come from the following distributions. (a) normal, (b) exponential, (c) Poisson.

(ii) v_a and v_b may be unknown as well as λ and μ but are estimated independently of them. The full conditions of Theorem 1.4.1 will be assumed to hold if v_a and v_b are known.

Example: The observations may come from normal distributions with unknown variances.

Allocate observations according to the design $\begin{pmatrix} p & , & 1-p \\ A & , & B \end{pmatrix}$

$$M(p) = \begin{bmatrix} \frac{p}{v_a} + \frac{1-p}{v_b} & \frac{1-p}{v_b} \\ \frac{1-p}{v_b} & \frac{1-p}{v_b} \end{bmatrix}$$

The criterion for best estimating μ is $\dfrac{|M(p)|}{|M_{11}(p)|} = \dfrac{\frac{p(1-p)}{v_a \cdot v_b}}{\frac{p}{v_a} + \frac{1-p}{v_b}}$

The above can readily be shown to have its maximum when $p = \dfrac{\sqrt{v_a}}{\sqrt{v_a} + \sqrt{v_b}}$.

That is, the optimal allocation of observations is in the ratio $\dfrac{\sqrt{v_a}}{\sqrt{v_b}}$

2.6.3: Simple linear prediction.

Suppose that observations are distributed as $N(\alpha + \beta x, \sigma^2)$, σ^2 assumed known. Suppose that in the laboratory we can obtain observations only at points $x \in [a,b]$, but that in the future predictions may be required at points possibly outside this interval. Let us assume that we know the distribution $p(x)$ on x with which predictions will be required in the future. The criterion for design will be the average variance of the estimated expected response.

That is we will want to

$$\min_{\xi} \int (1 \; x) \, M^{-1}(\xi) \begin{pmatrix} 1 \\ x \end{pmatrix} p(x) \, dx$$

$$= \min_{\xi} \; \mathrm{tr} \{ M^{-1}(\xi) \int \begin{pmatrix} 1 & x \\ x & x^2 \end{pmatrix} p(x) \, dx \}$$

$$= \min_{\xi} \; \mathrm{tr}\{ M^{-1}(\xi) \, A\} \;, \quad A = \begin{bmatrix} 1 & \mu \\ \mu & \mu^2 + \sigma^2 \end{bmatrix}$$, where μ and σ^2

denote the mean and variance of $p(x)$.

It will be sufficient to consider the interval $[-1, +1]$ for x, as the criterion $\mathrm{tr}\{ M^{-1}(\xi) \, A \}$ will obviously be invariant under linear transformations on x if the density $p(x)$ is adjusted accordingly. Consider $\mathcal{M} = \{(1,x,x^2) \; ; \; x \in [-1, +1]\}$, see shaded region in Fig.2.6.1.

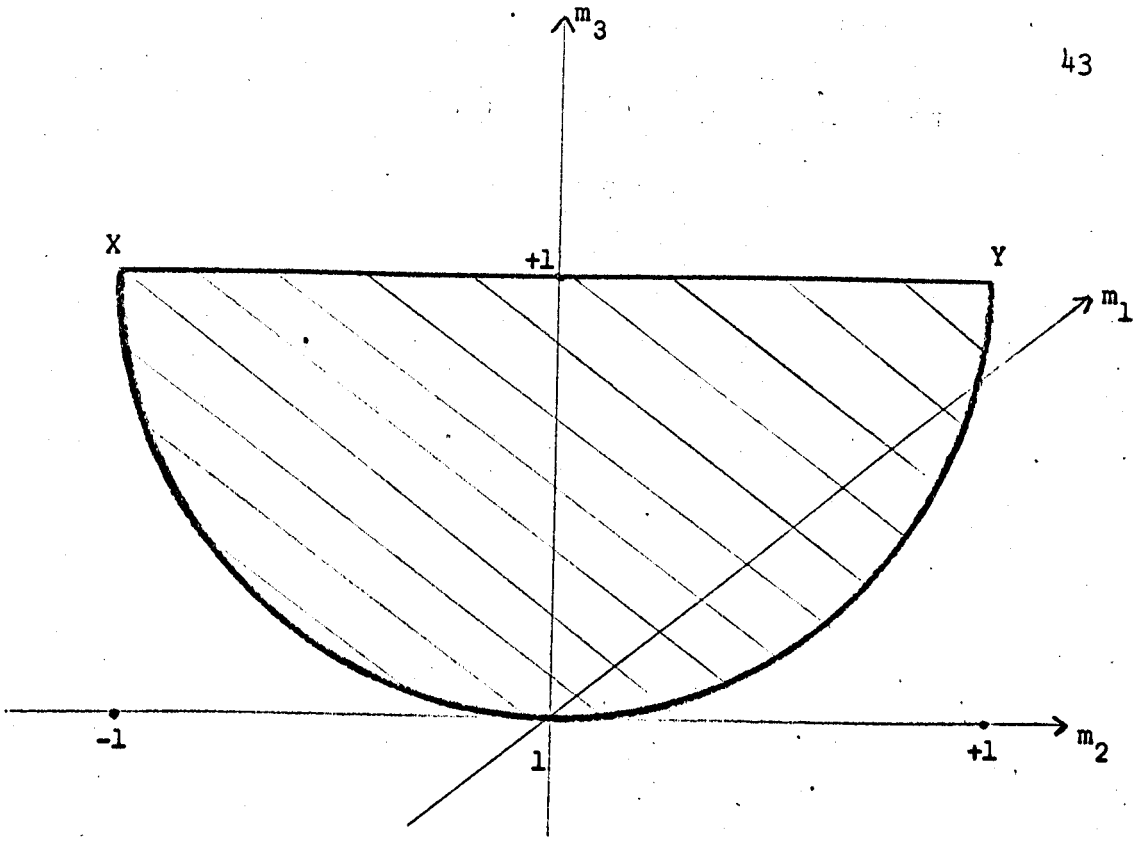


Fig. 2.6.1.

Because our criterion is an increasing function, the optimal point in $\mathcal{M}'$ must lie on the line XY, that is, the optimal design measure must put positive weight only at points -1, +1.

Take as design measure $\begin{pmatrix} 1-p & , & p \\ -1 & , & +1 \end{pmatrix}$.

$$M(p) = \begin{bmatrix} 1 & 2p-1 \\ 2p-1 & 1 \end{bmatrix}, \quad M(p)^{-1} = \frac{1}{4p(1-p)} \begin{bmatrix} 1 & 1-2p \\ 1-2p & 1 \end{bmatrix}.$$

$$\operatorname{tr}\{ M(p)^{-1} A\} = \frac{1}{4p(1-p)} \left\{(1+ \mu)^2+\sigma^2 - 4p\mu\right\}.$$

This is maximised where

$$p = \frac{((1+\mu)^2+\sigma^2) - ((1+\mu)^2 +\sigma^2)^{\frac{1}{2}} . ((1-\mu)^2 + \sigma^2)^{\frac{1}{2}}}{4\mu}, \quad \mu \neq 0$$

$$= \tfrac{1}{2}, \quad \mu = 0.$$

2.6.4: Linear Calibration Design

Suppose we have two factors u and y. In the laboratory we may observe y subject to stimuli $u \in [a,b]$. These responses are known to be subject to an error which is known to be normally distributed, the expected value of y given u being a linear function of u.

That is, $p(y|u) \sim N(\alpha + \beta u, \tau^2)$; α, β, τ^2 being unknown parameters. The distribution of u values arising in nature is known to be of the form $p(u) \sim N(\lambda, \sigma^2)$, λ, σ^2 known.

A simple application of Bayes formula to the above density functions will reveal that $p(u|y)$ is of the form

$$N\left(\frac{\frac{\lambda}{\sigma^2} - \frac{\alpha\beta}{\tau^2}}{\frac{\beta^2}{\tau^2} + \frac{1}{\sigma^2}} + \frac{\frac{\beta}{\tau^2}}{\frac{\beta^2}{\tau^2} + \frac{1}{\sigma^2}} \cdot y \;,\; \frac{1}{\frac{\beta^2}{\tau^2} + \frac{1}{\sigma^2}} \right)$$

The object of the experiment is to investigate, in some sense, the distribution of u for given y, that is $p(u|y)$. Therefore we have a simple calibration type experiment. Let us suppose initially that we are interested in best estimating the parameters α, β, τ^2 in the D-optimal sense.

$$I(u) = \frac{1}{\tau^2} \begin{bmatrix} 1 & u & 0 \\ u & u^2 & 0 \\ 0 & 0 & \frac{1}{2\tau^2} \end{bmatrix} .$$

It is well known that the D-optimal design for estimating α, β, τ^2 is $\begin{pmatrix} \frac{1}{2} & , & \frac{1}{2} \\ a & , & b \end{pmatrix}$.

Now suppose we are interested in $p(u|y)$ and that parameter estimation is our criterion for design. In this case the natural parameters of interest will be, not (α, β, τ^2), but

$$m = \frac{\frac{\lambda}{\sigma^2} - \frac{\alpha\beta}{\tau^2}}{\frac{\beta^2}{\tau^2} + \frac{1}{\sigma^2}} \quad , \quad n = \frac{\frac{\beta}{\tau^2}}{\frac{\beta^2}{\tau^2} + \frac{1}{\sigma^2}} \quad , \quad o = \frac{1}{\frac{\beta^2}{\tau^2} + \frac{1}{\sigma^2}} \quad .$$

If D-optimality is our criterion, then, because (m,n,o) may be obtained by a non-singular transformation of (α,β,τ^2), the D-optimal design for estimating (m,n,o) will be the same as that for estimating (α,β,τ^2) , namely $\begin{pmatrix} \frac{1}{2} & , & \frac{1}{2} \\ a & , & b \end{pmatrix}$.

If however we are only interested in a subset of the new parameters, (m,n) say, then the problem is not so simple.

Let the Jacobian of the transformation be J.

Let $m = \phi_1(\alpha,\beta,\tau^2)$, $n = \phi_2(\alpha,\beta,\tau^2)$, $o = \phi_3(\alpha,\beta,\tau^2)$, then

$$J = \begin{bmatrix} \frac{\partial\alpha}{\partial\phi_1} & \frac{\partial\beta}{\partial\phi_1} & \frac{\partial\tau^2}{\partial\phi_1} \\ \frac{\partial\alpha}{\partial\phi_2} & \frac{\partial\beta}{\partial\phi_2} & \frac{\partial\tau^2}{\partial\phi_2} \\ \frac{\partial\alpha}{\partial\phi_3} & \frac{\partial\beta}{\partial\phi_3} & \frac{\partial\tau^2}{\partial\phi_3} \end{bmatrix}$$

The criterion for (α,β,τ^2) was $|N(\xi)|$, $N(\xi) = \begin{bmatrix} M(\xi) & \begin{matrix} 0 \\ 0 \end{matrix} \\ \begin{matrix} 0 & 0 \end{matrix} & \frac{1}{2\tau^2} \end{bmatrix}$

For (m,n,o) it will be $|J\,N(\xi)\,J^T|$

For (m,n) the D_s-optimal criterion will be

$$\frac{|J\,N(\xi)\,J^T|}{|J_2\,N(\xi)\,J_2^T|} \quad , \quad \text{where } J_2 = \begin{pmatrix} \frac{\partial\alpha}{\partial\phi_3} & \frac{\partial\beta}{\partial\phi_3} & \frac{\partial\tau^2}{\partial\phi_3} \end{pmatrix} .$$

Maximising the above will be equivalent to maximising

$$\frac{|M(\xi)|}{\begin{pmatrix} \frac{\partial\alpha}{\partial\phi_3} & \frac{\partial\beta}{\partial\phi_3} \end{pmatrix} M(\xi) \begin{pmatrix} \frac{\partial\alpha}{\partial\phi_3} \\ \frac{\partial\beta}{\partial\phi_3} \end{pmatrix} + \frac{1}{2\tau^2}\left(\frac{\partial\tau^2}{\partial\phi_3}\right)^2} .$$

By an argument analogous to that of 2.6.3 the optimal design will be concentrated at the end points of our interval $[a,b]$. The problem is thus reduced to calculating the optimal weights. This problem will be unaltered by transforming the u-axis to z such that the interval $[a,b]$ is transformed to $[-1, +1]$.

$$p(z) \sim N(\lambda^*,\sigma^{*2}), \quad \lambda^* = \frac{\lambda - \frac{b+a}{2}}{\frac{b-a}{2}}, \quad \sigma^{*2} = \frac{\sigma^2}{\frac{(b-a)^2}{4}}$$

$$p(y|z) \sim N(\alpha^* + \beta^* z, \tau^2), \quad \alpha^* = \alpha + \frac{b+a}{2}\cdot\beta, \quad \beta^* = \frac{b-a}{2}\beta$$

Let the design measure be $\begin{pmatrix} p & , & 1-p \\ (1,1) & , & (1,-1) \end{pmatrix}$.

$$M(p) = \begin{bmatrix} 1 & 2p-1 \\ 2p-1 & 1 \end{bmatrix}$$

The design problem is equivalent to solving the following problem.

$$\max_{p\varepsilon(0,1)} \frac{|M(p)|}{\begin{bmatrix} \frac{\partial\alpha^*}{\partial\phi_3} & \frac{\partial\beta^*}{\partial\phi_3} \end{bmatrix} M(p) \begin{bmatrix} \frac{\partial\alpha^*}{\partial\phi_3} \\ \frac{\partial\beta^*}{\partial\phi_3} \end{bmatrix} + \frac{1}{2\tau^2}\left(\frac{\partial\tau^2}{\partial\phi_3}\right)^2}$$

That is,

$$\max_{p\varepsilon(0,1)} \frac{4p(1-p)}{\left[\left(\frac{\partial\alpha^*}{\partial\phi_3}\right)^2 + \left(\frac{\partial\beta^*}{\partial\phi_3}\right)^2 + \frac{1}{2\tau^2}\left(\frac{\partial\tau^2}{\partial\phi_3}\right)^2\right] + 2(2p-1)\frac{\partial\alpha^*}{\partial\phi_3}\cdot\frac{\partial\beta^*}{\partial\phi_3}}$$

That is,

$$\max_{p\varepsilon(0,1)} \frac{4p(1-p)}{r + 2(2p-1)s}$$

This is maximised when $p = \dfrac{-(r-2s) + (r-2s)^{\frac{1}{2}}(r+2s)^{\frac{1}{2}}}{2s}$,

where; $r = \left(\dfrac{\partial\alpha^*}{\partial\phi_3}\right)^2 + \left(\dfrac{\partial\beta^*}{\partial\phi_3}\right)^2 + \dfrac{1}{2\tau^2}\left(\dfrac{\partial\tau^2}{\partial\phi_3}\right)^2$, $s = \dfrac{\partial\alpha^*}{\partial\phi_3} \cdot \dfrac{\partial\beta^*}{\partial\phi_3}$.

2.6.5: Linear Quantal Response Model

Consider the following binary response model.

$$p(1|x,\theta_1,\theta_2) = F(\theta_1 + \theta_2 x),\ p(0|x,\theta_1,\theta_2) = 1 - F(\theta_1 + \theta_2 x).$$

$F(\cdot)$ is a distribution function of a random variable whose density function is symmetric about the origin. $\mathcal{X}$ is a continuous subset of the real line. This example is also considered by White (1975).

The conditions of Theorem 1.4.1 apply, giving

$$I(x) = \lambda(z)\begin{bmatrix} 1 & x \\ x & x^2 \end{bmatrix}, \quad \lambda(z) = \frac{\{f(z)\}^2}{F(z)(1-F(z))}, \quad z = \theta_1 + \theta_2 x.$$

Note that the above comments imply that $\lambda(z)$ is symmetric about the origin. Taking D-optimality as criterion, symmetry suggests consideration of a two-point design of the form

$$\begin{pmatrix} \frac{1}{2} & , & \frac{1}{2} \\ \dfrac{a-\theta_1}{\theta_2} & , & \dfrac{-a-\theta_1}{\theta_2} \end{pmatrix} .$$

$$|M(a)| = \{a\,\lambda(a)\}^2/\theta_2 .$$

For the linear logistic and probit models the function $|M(a)|$ has unique maxima at $a = a^*$, a^* being approximately 1.54 and 1.14 respectively. If $\pm\dfrac{a^* - \theta_1}{\theta_2}$ lie in $\mathcal{X}$ then this two point design can be shown to be D-optimal. Also if $\mathcal{X}$ is symmetric about $-\dfrac{\theta_1}{\theta_2}$ and $\pm\dfrac{a-\theta_1}{\theta_2}$ do not

lie in $\mathcal{X}$ then it can be shown that the D-optimal design puts weights $(\frac{1}{2},\frac{1}{2})$ at the end points of $\mathcal{X}$. This example will be reconsidered in a later chapter.

2.6.6. It is of interest to consider some classical experimental designs and to utilise the equivalence theorem of 2.2. to investigate optimality of these designs with respect to some criteria. Three examples are investigated.

Example 1.

The first example is the Latin square. This type of design is generally applied in an agricultural setting, p treatments being allocated to a pxp matrix of plots in such a way that each treatment appears once and only once in each row and column.

A possible model, for observations resulting from such an experiment, is as follows.

$$y_{ijK} = \alpha_i + \beta_j + \gamma_K + \varepsilon_{ijK} \quad , \quad \varepsilon_{ijK} \sim N(0,\sigma^2), \; i,j,K=1,\ldots,p.$$

In the above y_{ijK} denotes an observation in the (i,j)th position of the matrix to which a treatment K has been applied. The parameters α_i, β_j and γ_K denote underlying constants which reflect, respectively, the effects due to position of the plot in the matrix and to the treatment applied. The ε_{ijK}'s are error terms distributed as above, which are assumed independent of position in the plot and treatment. The model has dimension 3p. Due to identifiability considerations the model is usually reduced to the following form.

$$y_{ijK} = \mu + \alpha_i + \beta_j + \gamma_K + \varepsilon_{ijK}, \quad \Sigma\,\alpha_i = \Sigma\,\beta_j = \Sigma\,\gamma_K = 0 .$$

The dimension of the model is now 3p-2.

The parameters are well known to be estimated orthogonally between effects as follows.

$$\hat{\mu} = \frac{G}{p^2} \; , \; \hat{\alpha}_i = \frac{R_i}{p} - \frac{G}{p^2} \; , \; \hat{\beta}_j = \frac{C_j}{p} - \frac{G}{p^2} \; , \; \hat{\gamma}_K = \frac{A_K}{p} - \frac{G}{p^2} \quad .$$

G denotes the grant total of the observations and R_i, C_j and A_K denote, respectively, the row, column and treatment totals.

The variances of the parameter estimates are,

$$\text{var}(\hat{\mu}) = \frac{\sigma^2}{p^2}, \quad \text{var}(\hat{\alpha}_i) = \text{var}(\hat{\beta}_j) = \text{var}(\hat{\gamma}_K) = \sigma^2\left(\frac{1}{p} - \frac{1}{p^2}\right), \forall\ i,j,K.$$

The design space in this example can be denoted by the set of triplets (i,j,K); $i,j,K = 1,\ldots,p$.

Take D-optimality as criterion. The directional derivative from a Latin square information matrix to any of the one point design information matrices, that is $\Phi\{M_{LS}, I(i,j,K)\}$, can easily be shown, from Fig.2.2.1, to be,

$$\begin{aligned}\Phi\{M_{LS}, I(i,j,K)\} &= \frac{p^2}{\sigma^2}\ \text{var}(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_K) - 3p + 2 \\ &= \frac{p^2}{\sigma^2}\left(\frac{\sigma^2}{p^2} + 3\sigma^2\left(\frac{1}{p} - \frac{1}{p^2}\right)\right) - 3p + 2 \\ &= 0, \quad \forall\ i,j,K.\end{aligned}$$

Therefore, the Latin square is a D-optimal design. Now consider D_s-optimality as criterion, where interest lies in the γ_K's. Because the parameters are estimated orthogonally it may again be easily shown from Fig.2.2.1 that,

$$\begin{aligned}\Phi\{M_{LS}, I(i,j,K)\} &= \frac{p^2}{\sigma^2}\ \text{var}(\hat{\gamma}_K) - p + 1 \\ &= \frac{p^2}{\sigma^2} \cdot \sigma^2\left(\frac{1}{p} - \frac{1}{p^2}\right) - p + 1 \\ &= 0.\end{aligned}$$

Therefore, the Latin square is D_s-optimal for estimating the treatment contrasts. Obviously the Latin square will also be D_s-optimal for estimating the row and column contrasts.

Example 2.

Consider now a classical block design. Suppose we have p treatments to be allocated to q blocks. We shall define a balanced design as a design in which every treatment appears once and only once in each block.

The model considered is as follows

$$y_{ij} = \alpha_i + \beta_j + \varepsilon_{ij} \ , \quad \varepsilon_{ij} \sim N(0,\sigma^2) \ , \quad \begin{matrix} i=1,\ldots,p, \\ j=1,\ldots,q. \end{matrix}$$

The notation is similar to that of Example 1. Again because of identifiability considerations the above (p+q) dimensional model is reduced to the following.

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \ , \ \Sigma \alpha_i = \Sigma\beta_j = 0, \quad \begin{matrix} i=1,\ldots,p, \\ j=1,\ldots,q. \end{matrix}$$

This model has dimension (p+q-1).

The parameters are estimated orthogonally between effects as follows,

$$\hat{\mu} = \frac{G}{pq} \ , \ \hat{\alpha}_i = \frac{T_i}{q} - \frac{G}{pq} \ , \ \hat{\beta}_j = \frac{B_j}{p} - \frac{G}{pq} \ , \quad \begin{matrix} i = 1,\ldots,p, \\ j = 1,\ldots,q. \end{matrix}$$

In the above T_i and B_j denote the treatment and block totals respectively. The variances of the estimates are,

$$\text{var}\,(\hat{\mu}) = \frac{\sigma^2}{pq} \ , \quad \text{var}\,(\hat{\alpha}_i) = \sigma^2\left(\frac{1}{q} - \frac{1}{pq}\right) \ , \quad \text{var}\,(\hat{\beta}_j) = \sigma^2\left(\frac{1}{p} - \frac{1}{pq}\right).$$

The design space can be denoted by the set of ordered pairs (i,j), i=1,....,p; j=1,....,q.

Take D-optimality as criterion. It can easily be shown that the directional derivative from the information matrix of a balanced design to any of the one point design information matrices is,

$$\begin{aligned} \Phi\,\{M_B, I(i,j)\} &= \frac{pq}{\sigma^2}\ \text{var}(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j) - (p+q-1) \\ &= \frac{pq}{\sigma^2}\left(\frac{\sigma^2}{pq} + \sigma^2\left(\frac{1}{q} - \frac{1}{pq}\right)\right) - (p+q-1) \\ &= 0, \ \forall \ i,j. \end{aligned}$$

Therefore, the balanced design is D-optimal. Now consider D_s-optimality as criterion, where interest is in the treatment contrasts. Because the parameters are estimated orthogonally it is easy to show that

$$\begin{aligned}\Phi\{ M_B, I(i,j)\} &= \frac{pq}{\sigma^2} \text{ var } (\hat{\alpha}_i) - (p-1) \\ &= \frac{pq}{\sigma^2} \cdot \sigma^2\left(\frac{1}{q} - \frac{1}{pq}\right) - (p-1) \\ &= 0, \ \forall \ i,j.\end{aligned}$$

Therefore, the design is D_s-optimal for estimation of the α_i's. It may easily be seen that the design will also be D_s-optimal for estimation of the block contrasts.

Example 3.

We now consider the optimality of balanced incomplete block designs (BIBD). The simplest of these can be thought of as reduced designs of the type considered in Example 2 above, where each pair of treatments appears once and only once in any block. To investigate the optimality of these designs we must first consider the design space. The design space will in fact be exactly that of Example 2, that is, the set of ordered pairs (i,j), i=1,....,p; j=1,....,q. With this in mind it will be obvious that the BIBD's will not be optimal in the design measure sense.

However, the optimality of BIBD's in the design measure sense is not really a relevant factor, as the need for these types of designs arises when there are restrictions on the number of treatments which can be applied in each block, and this type of restriction can not be readily included in the optimal design measure theory considered above. The problem is really one of proving optimality in the exact design sense, and problems of this type have been considered in several papers by Kiefer, the most comprehensive of which is Kiefer (1975).

2.6.7. In the literature of design theory there are many other examples of analytic calculation of optimal designs. Most examples are of D-optimal designs, particularly in the area of polynomial regression. These designs have been obtained by, for example, Smith (1918), Guest (1958) and Hoel (1958). The interested party is referred to Fedorov (1972) who devotes a chapter to polynomial regression design with D-optimality as criterion. In other areas and for other criteria Fedorov (1972) and White (1975) give some interesting examples. Titterington (1975) illustrates how geometric arguments allow analytic solutions to some D_s-optimal design problems in simple situations.

Despite the above however it is necessary, in general, to resort to numerical algorithms in order to obtain designs, and it is to this topic that Chapter 3 is devoted.

CHAPTER 3

ALGORITHMS

As we have seen, it may be possible to obtain, for some design problems, optimal designs by analytic means. However, in general this will not be possible and iterative numerical methods of solving design problems are necessary. Several types of algorithm have been put forward, all closely linked, yet each having its own distinctive features. Of course, as previously mentioned, the solution of the continuous design problem which we have been considering up to now may only be the precursor to actual production of an exact design which might be applied in practice. Therefore, at the end of this chapter, we shall also consider algorithms which have been put forward to produce exact designs ab initio, or to improve upon designs which have been produced by approximating to continuously optimal ones.

3.1. Wynn's Algorithm

We consider this algorithm firstly, despite the fact that it would appear likely that similar algorithms were developed by Fedorov and his co-workers in Russia in advance of Wynn. As publications of the earlier work of Fedorov appear only in Russian this is difficult to verify. In any case, this is irrelevant to the development of the theory in this thesis. Wynn's algorithm is considered first because it represents the simplest to apply in practice.

As Wynn's algorithm was produced initially to solve the D-optimal design problem in a case where $I(\underline{x})$ is of rank 1, we will restrict ourselves to this case, again with induced design space V. Extension of the algorithm to other situations will be obvious from the discussion which follows.

The algorithm may be summarised in the following manner

(1) Let ξ_K be a non-singular K point design measure. That is $M(\xi_K) = \sum_{i=1} \frac{1}{K} Y_i Y_i^T$, $|M(\xi_K)| > 0$. Set n=K.

(2) Find $\underline{v}^*$ such that $\max_{\underline{v}\in V} \underline{v}^T M(\xi_n)^{-1} \underline{v} = \underline{v}^{*T} M(\xi_n)^{-1} \underline{v}^*$

(3) Add $\underline{v}^*$ to the design spectrum and allocate weights $\frac{1}{n+1}$ to the (n+1), not necessarily distinct, points in the spectrum.

(4) Set n = n+1. Go to (2) and repeat.

Wynn (1970) proves convergence of the above algorithm to an optimal design measure. His proof is simplified by Pazman (1974).

We make the following observations on Wynn's algorithm.

(i) The computationally most exacting part of the algorithm is (2), where an optimisation problem must be solved. This may often be satisfactorily carried out by approximating to V by a finite grid of points and obtaining the maximum by direct search. Fedorov (1969) gives very useful formulae for updating $M(\xi_n)^{-1}$ as the algorithm progresses.

(ii) The algorithm is not necessarily monotonic, that is, $|M(\xi_{n+1})|$ is not necessarily greater than $|M(\xi_n)|$. This makes Wynn's proof of convergence very difficult, and, in application, may affect convergence rates of the algorithm.

(iii) The number of distinct points in the spectrum at stage n may become very large, the design becoming cluttered with points having very small weight (possibly relics of poor initial choices), thus slowing convergence to a design measure on a small number of points.

In order to compare Wynn's algorithm with others it is useful to create a 'picture' of what the algorithm is actually doing in a general setting. Let us return to the convex set of matrices $\mathcal{M}=\{ M(\xi), \xi\in \Xi \}$, on which is defined a concave function ϕ which we want to maximise. Corresponding to step (1) in the above procedure we take as starting point, a point $M(\xi_n)$ in $\mathcal{M}$ at which ϕ is differentiable. The next step is to look for the direction from $M(\xi_n)$ in which there is maximal rate of increase in ϕ, that is, for the direction in which the directional derivative is greatest. In the above this may be seen to be in the direction of the matrix $\underline{v}^* \underline{v}^{*T}$, which corresponds to the design matrix of a design measure

which puts weight unity at the point $\underline{v}^* \in \underline{V}$. The next action is to move a predetermined distance along the line of steepest ascent. In the above a move is made to the point

$$M(\xi_{n+1}) = (1 - \alpha_n)\, M(\xi_n) + \alpha_n \underline{v}^* \underline{v}^{*T}, \quad \alpha_n = \frac{1}{n+1} .$$

The process is repeated until suitable convergence occurs. Fedorov (1972) notes that any sequence of α_i's satisfying

$$\sum_{n=1}^{\infty} \alpha_n = \infty , \quad \lim_{n \to \infty} \alpha_n = 0 , \quad \alpha_n > 0, \forall n,$$

will guarantee suitable convergence of the algorithm.

Viewed in this manner as a steepest ascent type algorithm there is an obvious alteration to Wynn's algorithm which might improve convergence.

3.2. Fedorov's Algorithm.

The possible improvement in the above algorithm, which becomes evident, is to drop the restriction that the α_n's be predetermined, and to choose them in order to attain the maximum increase of ϕ at each stage. This leads to the algorithm of Fedorov (1972), which might be summarised as follows.

Repeat Wynn's algorithm with (3) changed to (3)' .

(3)' Let $M(\xi_{n+1}) = (1-\alpha_n)\, M(\xi_n) + \alpha_n \underline{v}^* \underline{v}^{*T}$,

$$\alpha_n = \frac{\underline{v}^{*T} M(\xi_n)^{-1} \underline{v}^* - K}{K(\underline{v}^{*T} M(\xi_n)^{-1} \underline{v}^* - 1)}$$

We make the following comments on this alteration.

(i) Although, for the case of D-optimality, the α_n may be found analytically, as given above, this will not typically be true for other criteria and may lead to substantial extra numerical work.

(ii) The procedure is now monotonic, simplifying convergence proofs, and possibly, but not necessarily, accelerating rate of convergence.

(iii) The problem still remains of poor points remaining in the design spectrum for long periods, with non trivial weights.

Atwood (1973) has suggested improvements to Fedorov's algorithm. They were presented as improvements to the algorithm for D and D_s optimality, but the generalisation of the application of the most important of these, to more general criteria, is at once apparent. Atwood's point in a more general situation is best appreciated by returning to the pictorial representation. Again imagine starting from a point $M(\xi_n) \in \mathcal{M}$. Fedorov's algorithm takes us in the direction of steepest ascent from this point. We note that there must be points $\underline{x} \in \mathcal{X}$ such that we are looking 'downhill' from $M(\xi_n)$ to $I(\underline{x})$, and therefore 'uphill' if we about turn and face the opposite direction, because of the differentiability of ϕ at $M(\xi_n)$. Atwood's improvement is a result of the fact that a greater increase in ϕ may result from moving away from one of these points rather than in the direction of steepest ascent. This is equivalent to allowing α_n to be negative in (3)' above. Note that we need only consider points $\underline{x}$ which are in the design spectrum at stage n for possible alternative directions, as bringing in new points with negative weight would take ξ_{n+1} outside the set of feasible design measures. The most natural point to use would seem to be that for which the directional derivative is smallest, and, in fact, Atwood (1973) shows that for D and D_s optimality this will give the maximal increase in ϕ.

The importance of this improvement is that it serves to 'weed out' the bad design points discussed above, thereby aiding convergence.

For Atwood's other improvements, which are much more trivial, see Atwood (1973).

We note that improved convergence has only been obtained at the expense of increased computation and the need for a more complex computer program.

3.3: Algorithms of Silvey and Titterington (1973) and Atwood (1976)

We first summarise the details of the algorithm of Silvey and Titterington, and then compare it with that of Fedorov.

(1) Let $M(\xi_K)_K$ be a differentiable K point design matrix in $\mathcal{M}$, $M(\xi_K) = \sum_{i=1} \frac{1}{K} I(\underline{x}_i)$, set n=K.

(2) Find $\underline{x}^*$ such that $\Phi\{M(\xi_n), I(\underline{x})\}$ is a maximum.

(3) Add the point $\underline{x}^*$ to the design spectrum. Find the optimum design measure on the finite set of m(n) distinct points in the design spectrum. Let this be ξ_{n+1}

(4) Goto (2) and repeat.

The only difference between the above algorithm and that of Fedorov lies in section (3).

Consider the convex hull of the one point design matrices obtained from the m(n) distinct design points at stage n, one of which will be $I(\underline{x}^*)$. This convex hull will, of course, contain $M(\xi_n)$. Fedorov's algorithm moves the procedure from $M(\xi_n)$ to a point $M(\xi_{n+1})$ on the line between $M(\xi_n)$ and $I(\underline{x}^*)$, such that $M(\xi_{n+1})$ is maximised. The algorithm described above is free to move the procedure to any point in the convex hull in order to maximise ϕ. Obviously the algorithm of Silvey and Titterington will give at least as great an increase in ϕ, at each stage, as that of Fedorov. However, it does so at the expense of a great deal of added computation. Fedorov's algorithm requires an optimisation at stage (3) with respect to one variable, while that of Silvey and Titterington requires an optimisation with respect to m(n)-1 variables. However, if the spectrum of the optimal design is contained in the m(n) distinct points at stage n the solution will be obtained in one more step. This suggests the use of Fedorov's or Wynn's algorithms initially to produce a set of points which might contain the support points of the optimal design and then a switch to the algorithm of Silvey and Titterington to complete the coup de grâce. However, although this plan seems worthy of consideration, it would appear to be difficult to decide, in general, when the change-over should best be made.

The Silvey and Titterington algorithm throws up another mathematical programming problem. That being, given a finite number of design points, how do we find the optimal weights which should be allocated to these points. Silvey and Titterington (1973) suggest the use of the Newton-Raphson technique for this problem. Sibson and Kenny (1975) suggest the use of the dual-simplex methods of the Kelley cutting plane (Kelley (1960), Wolfe (1961)). A problem which may arise in practice is that these algorithms may lead to non-feasible solutions, that is, solutions which attach negative weights to some points. The problem then becomes a more difficult one of constrained optimisation. An alternative algorithm suggested by Silvey, Titterington and Torsney (1976) would seem to solve this problem, at least for D-optimality, and possibly in other situations, although this remains to be proved.

Consider the situation at stage n in the algorithms of Silvey and Titterington. The procedure is situated at a point $M(\xi_n)$. A point $\underline{x}^*$ is added to the design spectrum.

Part (3) of the algorithm of Silvey and Titterington might be thought of as finding a measure η_n over the set of m(n) distinct points in the design spectrum in order to maximise

$$\phi\{M_{\cdot}(\xi_{n+1})\} = \phi\{M(\xi_n + \eta_n)\} ,$$

with the constraints

$$\text{(a)} \quad \int_{\underline{x}\in\mathcal{X}} \eta_n(d\underline{x}) = 0 ,$$

(b) ξ_{n+1} gives negative weight to no point.

Atwood (1976) suggests an algorithm which instead of trying to find an exact solution to the above problem, a procedure which may involve lengthy calculations, obtains an approximate solution to the above problem and then continues directly to the next step. What Atwood does is to find the maximising η_n, η_n^* say, for a second order Taylor expansion of $\phi\{M(\xi_{n+1})\}$ about $M(\xi_n)$. The next step is to move in the direction of this solution to a point which maximises the original function. That is, find an α to maximise $\phi\{M(\xi_n + \alpha\eta_n^*)\}$, again

with the obvious constraint that negative weight should be attached to no point. Atwood shows that with mild regularity conditions on ϕ, the above sequence of design measures will converge to an optimal design, convergence being monotonic. He also shows that asymptotically his quadratic approximation design will perform at least as well as Fedorov's at any given step. Although Atwood's algorithm will involve less computation than that of Silvey and Titterington at each stage, it will involve considerably more than that of Fedorov, particularly for criteria other than D-optimality.

3.4. Yet another algorithm.

We now present an algorithm which would seem not to have been presented before.

Let us return again to the convex hull of the m(n) distinct design points at stage n. We have noted that the algorithm of Fedorov moves the procedure from a point $M(\xi_n)$ along the line of steepest ascent towards another point $I(\underline{x}_1^*)$ say. The improvement of Atwood (1973) suggests that a greater increase may sometimes be obtained by moving along the line of steepest descent, that is, away from $I(\underline{x}_2^*)$ say. The essence of the algorithm to be presented here is that we carry out both processes at once by moving to a point in the plane containing the lines of steepest ascent and descent, in order to maximise ϕ. The algorithm is presented for the situation where $I(\underline{x})$ is of rank one and the criterion is D-optimality. The optimisation problem may be solved explicitly in this case.

Note the following notation.

$$I(\underline{x}_1^*) = \underline{v}_1 \underline{v}_1^T \quad , \quad I(\underline{x}_2^*) = \underline{v}_2 \underline{v}_2^T \; ;$$

$$d_1 = \underline{v}_1^T M(\xi_n)^{-1} \underline{v}_1 \; , \quad d_2 = \underline{v}_2^T M(\xi_n)^{-1} \underline{v}_2 \; ,$$

$$d_{12} = \underline{v}_1^T M(\xi_n)^{-1} \underline{v}_2 \; .$$

The algorithm differs from that of Fedorov only in section (3). In section (3) the procedure moves from $M(\xi_n)$ to $M(\xi_{n+1})$, $M(\xi_{n+1}) = (1 - \alpha_1^* - \alpha_2^*) M(\xi_n) + \alpha_1^* \underline{v}_1 \underline{v}_1^T + \alpha_2^* \underline{v}_2 \underline{v}_2^T$, where α_1^* and α_2^* are

chosen to maximise

$$|(1 - \alpha_1 - \alpha_2)\, M(\xi_n) + \alpha_1 \underline{v}_1 \underline{v}_1^T + \alpha_2 \underline{v}_2 \underline{v}_2^T| \qquad *$$

Typically it would be expected that $\alpha_1^* > 0$ and $\alpha_2^* < 0$.

We now obtain formulae for α_1^* and α_2^*.

Lemma 3.4.1.

$$|M + \underline{v}_1 \underline{v}_1^T + \underline{v}_2 \underline{v}_2^T| = |M| \{1 + d_1 + d_2 + d_1 d_2 - d_{12}^2\}$$

Proof

$$\begin{aligned} |M + \underline{v}_1 \underline{v}_1^T + \underline{v}_2 \underline{v}_2^T| &= |M + \underline{v}_1 \underline{v}_1^T| \{1 + \underline{v}_2^T (M + \underline{v}_1 \underline{v}_1^T)^{-1} \underline{v}_2\} \\ &= |M| \{1 + \underline{v}_1^T M^{-1} \underline{v}_1\} . \{1 + \underline{v}_2^T M^{-1} \underline{v}_2 - \frac{(\underline{v}_2^T M^{-1} \underline{v}_1)^2}{1 + \underline{v}_1^T M^{-1} \underline{v}_1}\} \\ &= |M| \{1 + d_1 + d_2 + d_1 d_2 - d_{12}^2\} \end{aligned}$$

By Lemma 3.4.1 we have

$$* = |M| (1 - \alpha_1 - \alpha_2)^K \{1 + \frac{\alpha_1 d_1}{1-\alpha_1-\alpha_2} + \frac{\alpha_2 d_2}{1-\alpha_1-\alpha_2} + \frac{\alpha_1 \alpha_2 d_1 d_2}{(1-\alpha_1-\alpha_2)^2} - \frac{\alpha_1 \alpha_2 d_{12}^2}{(1-\alpha_1-\alpha_2)^2}\}$$

We require to maximise * with respect to α_1 and α_2, $\alpha_1 + \alpha_2 < 1$.
Fix $\alpha_1 + \alpha_2 = x$, and introduce a Lagrange multiplier λ.
We require to maximise the Lagrangian form

$$\{1 + \frac{\alpha_1 d_1}{1-x} + \frac{\alpha_2 d_2}{1-x} + \frac{\alpha_1 \alpha_2}{(1-x)^2}(d_1 d_2 - d_{12}^2)\} - \lambda (\alpha_1 + \alpha_2 - x)$$

Equating the usual partial derivatives to zero we have

$$\frac{d_1}{1-x} + \frac{\alpha_2^*}{(1-x)^2}(d_1 d_2 - d_{12}^2) = \lambda$$

$$\frac{d_2}{1-x} + \frac{\alpha_1^*}{(1-x)^2}(d_1 d_2 - d_{12}^2) = \lambda$$

$$\alpha_1^* + \alpha_2^* = x$$

Solving the above equations for α_1^* and α_2^* we have

$$\alpha_1^* = \tfrac{1}{2}\{ - \frac{(d_2 - d_1)}{d_1 d_2 - d_{12}^2} \cdot (1-x^*) + x^* \}$$

$$\alpha_2^* = \tfrac{1}{2}\{ + \frac{d_2 - d_1}{d_1 d_2 - d_{12}^2} \cdot (1-x^*) + x^* \}$$

x^* in the above is the maximising value of x in the following

$$(1-x)^K \{1 + \frac{1}{2(1-x)} \left[(d_1+d_2)x + \frac{(d_2-d_1)^2}{d_1 d_2 - d_{12}^2} \cdot (1-x) \right] + \frac{d_1 d_2 - d_{12}^2}{4(1-x)^2} \cdot \left[x^2 - (1-x)^2 \cdot \frac{(d_2-d_1)^2}{(d_1 d_2 - d_{12}^2)^2} \right] \} .$$

Equating the partial derivative of the above, with respect to x, to zero, we have

$$- (1-x)^{K-3} f(x) = 0, \quad \text{where } f(x) = A x^2 + Bx + C$$

$$A = a + b + c, \quad B = -(2a + b), \quad C = a,$$

$$a = K + \frac{K}{4} \cdot \frac{(d_2-d_1)^2}{d_1 d_2 - d_{12}^2} - \frac{d_1 + d_2}{2},$$

$$b = - (K-1) \cdot \frac{d_1 + d_2}{2} + \frac{d_1 d_2 - d_{12}^2}{2},$$

$$c = (K-2) \cdot \frac{d_1 d_2 - d_{12}^2}{4} .$$

Solution of the above quadratic equation gives

$$x^* = \frac{2a + b \pm \sqrt{b^2 - 4ac}}{2a + 2b + 2c},$$

whichever solution is suitable. A piece of tedious but straightforward algebra reveals that real solutions to the quadratic will always exist

as $b^2 - 4ac = \frac{1}{4}(d_1 + d_2 - d_1 d_2 + d_{12}^2)^2 + K(K-2)\, d_{12}^2 > 0$.
The suitable solution may be found by substitution in the breakdown of * above.

We note that the extra computation involved in the calculation of α_1^* and α_2^* will be trivial. d_1 and d_2 will already have been computed, leaving only computation of d_{12} and substitution in the formulae above.

We now consider two examples, which will illustrate points which may be aids in the use of this algorithm, and indicate its potential.

Example 1

We take as an example a problem used by Wynn (1969).

Take as induced design space $V = \{(1,1,-1),(1,-1,1),(1,-1,-1),(1,2,2)\}$.
Let ξ_o allocate weights as $(\frac{1}{3},\frac{1}{3},\frac{1}{3},0)^T$.

$$d_1(\underline{v}_1) = d_1(\underline{v}_2) = d_1(\underline{v}_3) = 3 \quad , \quad d_1(\underline{v}_4) = 25.5.$$

Therefore, we select $\underline{v}_4$ for introduction to the design spectrum. That is, $d_1 = 25.5$.

There is a choice of points to select as a direction of steepest descent. However, remembering that intuitively we would expect $\alpha_1^* > 0$ and $\alpha_2^* < 0$, it may be seen from the reduction of * that the best point to introduce is the one which maximises d_{12} in modulus.

$$d_{12}(\underline{v}_1) = d_{12}(\underline{v}_2) = 4.5 \quad , \quad d_{12}(\underline{v}_3) = -6.$$

Therefore, choose $\underline{v}_3$, giving $d_{12} = -6$, $d_2 = 3$.

Substitution in the above formulae gives

$$a = 3\left(1 + \frac{(22.5)^2}{4 \times 40.5}\right) - 14.25 = -1.875$$

$$b = -2 \times 14.25 + \frac{40.5}{2} = -8.25$$

$$c = \frac{40.5}{4} = 10.125$$

$$a + b + c = 0 \implies x^* = \frac{a}{2a + b} = \frac{5}{32}$$

$$\implies \alpha_1^* = \frac{10}{32} \quad , \quad \alpha_2^* = -\frac{5}{32}$$

If we denote the design measures which put weight one at points $\underline{v}_3$, $\underline{v}_4$ by $\xi_{\underline{v}_3}$, $\xi_{\underline{v}_4}$, then we have

$$\xi_1 = (1 - \alpha_1^* - \alpha_2^*)\, \xi_o + \alpha_1^* \xi_{\underline{v}_4} + \alpha_2^* \xi_{\underline{v}_3} .$$

$$= \left(\frac{9}{32} , \frac{9}{32} , \frac{4}{32} , \frac{10}{32} \right)^T .$$

This is in fact the D-optimal design measure. That is, we have convergence in one iteration. This is obviously due to the symmetric nature of the design space and the starting design ξ_o. However, although the above example is a little flattering to the new algorithm, it does illustrate very well the possibilities offered by an algorithm which has a wider area of search. Wynn's algorithm converges to the above optimal design in thirty-two iterations.

Example 2

Take as induced design space $V = \{(1,0,0),(0,1,0),(0,0,1),(\frac{1}{2},\frac{1}{2},\frac{1}{2})\}$.
Let ξ_o allocate weights as $(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4})^T$.

$$d_1(\underline{v}_1) = d_1(\underline{v}_2) = d_1(\underline{v}_3) = \frac{24}{7} , \quad d_1(\underline{v}_4) = \frac{12}{7} .$$

There is a choice of direction of steepest ascent here. Due to the symmetry of the situation there would appear to be no reason for preferring any particular point, therefore, we take $\underline{v}_1$ for the direction of steepest ascent, giving $d_1 = \frac{24}{7}$.

Take $\underline{v}_4$ for the direction of steepest descent, giving

$$d_{12} = \frac{8}{7} , \quad d_2 = \frac{12}{7} .$$

Substitution in the formulae gives

$$a = 3\left\{1 + \frac{12^2}{4(12 \times 24 - 64)}\right\} - \frac{18}{7} = .9107$$

$$b = -\frac{2 \times 18}{7} + \frac{12 \times 24 - 64}{2 \times 49} = -2.8571$$

$$c = \frac{12 \times 24 - 64}{4 \times 49} = 1.428$$

$$\Rightarrow x^* = -.6$$

$$\Rightarrow \alpha_1^* = 0 , \quad \alpha_2^* = -.6.$$

With similar notation to Example 1 we have

$$\xi_1 = (1 - \alpha_1^* - \alpha_2^*)\ \xi_o + \alpha_1^*\ \xi_{\underline{v}_1} + \alpha_2^*\ \xi_{\underline{v}_4}$$

$$= (\cdot 4,\ \cdot 4,\ \cdot 4,\ -\cdot 2)^T.$$

That is, the solution is not feasible. Because of concavity of the criterion the obvious thing to do in this situation is to set the negative weight to zero and normalise the remaining vector. In this example we have

$$\xi_1 = \left(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}, 0\right) .$$

This is the D-optimal design.

In general the best policy would seem to be to move from ξ_1 back towards ξ_o until the design just becomes feasible.

3.5.

Before moving on to discuss the relative merits of the algorithms described above, a little should be said about the possibility of breakdown of the algorithms.

It may be seen, from the outlines of the algorithms, that a crucial factor in their use is that ϕ should be differentiable at $M(\xi_n)$, for all n. This is also a crucial factor in proofs of covergence of the algorithms. To illustrate where problems may arise we consider again the seven examples of criteria used in chapters one and two.

For functions ϕ_1, ϕ_2, ϕ_3, ϕ_4, ϕ_6 and ϕ_7 it will be immediately apparent that the functions will not attain their maxima at singular matrices. The singular matrices may be thought of as forming the set of extreme points of $\mathcal{M}$. We denote the set of non-singular matrices by $\mathcal{M}^+$, and define the above functions to take the value $-\infty$ at singular points in $\mathcal{M}$.

ϕ_1, ϕ_3 and ϕ_6 are differentiable everywhere in $\mathcal{M}^+$. Therefore, if we take as starting point $M(\xi_o) \in \mathcal{M}^+$ and use a monotonic algorithm, then $M(\xi_n) \in \mathcal{M}^+, \forall n$. Also, for algorithms of the Wynn type, it may be seen

for $\alpha_n \in (0,1), \forall n$, and $M(\xi_o) \in \mathcal{M}^+$ then $M(\xi_n) \in \mathcal{M}^+, \forall n$.

For criterion ϕ_5 it is possible that the optimal design matrix may be singular. It is interesting to note here that again, with a Wynn type algorithm, and non-singular starting point, $M(\xi_n) \in \mathcal{M}^+, \forall n$. That is, the algorithm will converge towards a singular optimum design without actually ever settling on a singular matrix. For this criterion it would seem to be difficult to set up a general rule for what to do, in one of the monotonic algorithms, if the process were to lead to a singular matrix. In simple cases it might be feasible to use some of the ideas of Chapter 2 to test for optimality.

At this point we introduce two theorems (Silvey (1974)).

Theorem 3.5.1.

If ϕ is differentiable at $M(\xi)$, $\delta \geq 0$, and $\phi\{ M(\xi), I(\underline{x})\} \leq \delta$, $\underline{x} \in \mathcal{X}$ then $\phi\{ M(\xi)\} \geq \phi\{ M(\xi^*) \} - \delta$, where ξ^* is ϕ-optimal.

Theorem 3.5.2.

If $0 < \alpha_n < 1$, $\alpha_n \to 0$ and $\Sigma \alpha_n \to \infty$ as $n \to \infty$ and ϕ is differentiable at $M(\xi_n)$, then for a sequence $M(\xi_n)$ produced by a Wynn type algorithm $\inf_n \max_{\underline{x} \in \mathcal{X}} \phi\{ M(\xi_n), I(\underline{x})\} = 0$.

Theorem 3.5.1. suggests a useful stopping rule in practice. Although it is not a strong enough result to guarantee optimality of a design at stage n, it may be sufficient to indicate that we are as near to an optimal design as is important in practice. Theorem 3.5.2. tells us that, with the given conditions on the design sequence, we can get as near as we want to an optimal design matrix at some point in the process. For criteria ϕ_2, ϕ_4 and ϕ_7, as we have noted, it may happen that there are non-differentiable points which are interior to $\mathcal{M}^+$. Now, even Wynn type procedures may break down. However, sensible ad hoc procedures should prove effective in practice, though they must remain difficult to justify rigorously as far as guaranteed convergence is concerned.

For example, consider the following procedure. Suppose that an algorithm leads to a point $M(\xi_n)$ at which ϕ is not differentiable and $\Phi\{ M(\xi_n), I(\underline{x})\} \leq 0, \forall \underline{x} \in \mathcal{X}$. Move a very small distance from $M(\xi_n)$ back along the line of approach to it, to a point, $M(\xi_n^\epsilon)$ say, at which ϕ is differentiable. If $\Phi\{M(\xi_n^\epsilon), I(\underline{x})\} \leq \delta, \forall \underline{x} \in \mathcal{X}$, for suitably small δ, then stop the algorithm with justification of Theorem 3.5.1 that we are suitably close to an optimum design. If $\Phi\{ M(\xi_n^\epsilon), I(\underline{x})\} > \delta$ for some $\underline{x} \in \mathcal{X}$, restart the algorithm with $M(\xi_n^\epsilon)$ as starting point. If the procedure continues to converge on $M(\xi_n^\epsilon)$ then restart the algorithm with a different starting point and compare outcomes.

3.6.

Comparison of the algorithms described, as to how they perform in practice must remain a virtually impossible task, the reasons for this being that performance will depend very much on the problem at hand, computing facilities and the programming abilities of the problem solver, not to mention the time and money available to solve the problem. The following, however, might indicate how the algorithms could be used in practice.

Consider firstly properties of the algorithms which might affect convergence rates.

Algorithms of the Wynn type, with pre-set α_n's, do not take into account local knowledge of the function at each stage, and by their very nature may take very long routes up the 'hill'.

Steepest ascent algorithms of the Fedorov type, although they take the direction in which the function is increasing most steeply and go to the maximising point in that direction, do not in general guarantee maximal increase in ϕ at each step. The obvious next step is to widen the area of search at each stage. The improvement of Atwood (1973) does this to a certain extent and the algorithm of 3.4. will improve matters even more. The algorithm of Silvey and

Titterington (1973) searches over a wider area than any of the above. These points are immediately obvious from the pictorial representation. The amount of computation and the complexity of the computer program necessary for the use of the above algorithms, at each stage, will increase in the same order.

Where to place the algorithm of Atwood (1976) is a little more difficult. Atwood shows that asymptotically, his algorithm will do at least as well as that of Fedorov at each stage, although what will happen for small n will obviously depend on how good the quadratic approximation is. The algorithm of Silvey and Titterington will still give the maximal increase in ϕ for a given stage and computationwise will still use most resources with Atwood (1976) coming second last.

In a situation with modest computing facilities, it would seem likely that an algorithm at the Wynn end of the scale would perform satisfactorily in practice, leaving open the option of a switch to a more complex algorithm, if problems were to arise in relation to convergence.

3.7. Exact Designs

3.7.1. As was mentioned in Chapter 2 the motivation for studying the continuous design problem comes from the fact that, for large samples, we may be able to approximate closely to a continuous optimal design measure with a design putting rational weights at the design points. We now justify this using an argument of Fedorov (1972).

To add uniformity to this section we shall redefine ϕ_1 as $|M|$ and ϕ_5 as $|M_s|$.

Note the following definitions.

(i) Let ξ^* be the optimal continuous design measure.

(ii) Let ξ_N^* be the optimal exact design measure for N observations.

(iii) Let $\tilde{\xi}_N$ be the exact design measure having the same spectrum as ξ^* (containing n distinct points say).

At each point x^* in the spectrum take $r_i^+ = \lceil (N-n)\, p_i^* \rceil^+$ observations, where $\lceil C \rceil^+$ denotes the least integer satisfying $\lceil C \rceil^+ \geq C$, the remaining $N - \Sigma \lceil (N-n)\, p_i^* \rceil^+$ observations are arbitrarily distributed.

The following theorem may easily be proved.

Theorem 3.7.1

$$\gamma\left(\frac{N-n}{N}\right) \phi(\xi^*) \leq \phi(\tilde{\xi}_N) \leq \phi(\xi_N^*) \leq \phi(\xi^*) \quad , \quad \text{where}$$

$$\gamma\left(\frac{N-n}{N}\right) \to 1 \quad \text{as } N \to \infty \ .$$

For, ϕ_1 , $\gamma\left(\frac{N-n}{N}\right) = \left(\frac{N-n}{N}\right)^K$,

ϕ_5 , $\gamma\left(\frac{N-n}{N}\right) = \left(\frac{N-n}{N}\right)^s$,

$\phi_2, \phi_3, \phi_4, \phi_6, \phi_7$, $\gamma\left(\frac{N-n}{N}\right) = \frac{N}{N-n}$.

The sandwiching inequality of Theorem 3.7.1 will ensure that for large N a design of the form $\tilde{\xi}_N$ will be almost certainly adequate in practice. However, if improvements are desired, or if, in the case of the linear model, we require an exact design for small N, then the following algorithms may be useful.

3.7.2.

Very little progress has been made in the area of finding analytic solutions to exact optimal design problems. A brief review of some problems which have been tackled will be given in 3.7.3. In the present subsection we will assume that no such solution is available and that some iterative numerical method is required.

The problem is restated.

Let $\mathcal{X}$ be the design space and N the sample size. It is required to select N points from $\mathcal{X}$ in order to maximise

$$\phi\{ M(\xi_N)\} \quad , \quad M(\xi_N) = \frac{1}{N} \sum_{i=1}^{N} I(\underline{x}_i), \quad \underline{x}_i \in \mathcal{X}, \forall i \ .$$

Note that if $\mathcal{X}$ is discrete (of size m say), then theoretically all selections of N from m points could be computed with corresponding ϕ values, the maximising selection being obtained by observation. Continuous $\mathcal{X}$ could also be approximated by a finite grid of points and the above procedure carried out. However, for large m, the above method of attack on the problem will be computationally infeasible, and some other method of solution will be required. We consider two algorithms, the first due to Fedorov (1972), the second to Wynn (1972) with improvements by Mitchell (1974).

In the two algorithms to be considered an initial design is required. An initial design of the form $\tilde{\xi}_N$ described in 3.7.1 would seem to be sensible for relatively large N. However, for small N, and where the geometry of the problem does not provide any inspired initial guesses, an arbitrarily chosen initial design will suffice.

The algorithm of Fedorov is an exchange algorithm, iteratively exchanging design points, one at a time, in order to obtain the maximum increase in ϕ. It may be described as follows.

(1) Select an initial design ξ_N.

(2) Define $\Delta(\underline{x}_j,\underline{x}) = \phi\{\frac{1}{N}\sum_{i=1}^{N} I(\underline{x}_i) + \frac{1}{N}(I(\underline{x}_j) - I(\underline{x}))\} - \phi\{\frac{1}{N}\sum_{i=1}^{N} I(\underline{x}_i)\}$

$\underline{x}_j \in \{\underline{x}_1,\ldots,\underline{x}_N\} \equiv$ spectrum of ξ_N.

Select $\underline{x}$ and $\underline{x}_j$ which satisfy

$$\max_{\underline{x}_j} \max_{\underline{x}\in\mathcal{X}} \Delta(\underline{x}_j,\underline{x}).$$

(3) Replace $\underline{x}_j$ by $\underline{x}$ in ξ_N.

(4) Go to (2) and repeat until no increase in ϕ is being observed.

The Wynn algorithm does not attempt to obtain the maximal increase at each step in the procedure, but has a close link with the continuous design algorithm of Wynn. Although presented originally for D-optimality, the extension to more general criteria is fairly obvious.

(1) Select an initial design ξ_N.

(2) Find an $\underline{x}_{N+1}$ such that

$$\Phi\{ M(\xi_N), \; I(\underline{x}_{N+1}) \} = \max_{\underline{x}\in\mathcal{X}} \Phi\{M(\xi_N), \; I(\underline{x})\}$$

(3) Add $\underline{x}_{N+1}$ to the design spectrum and compute directly the best N point design from the spectrum. Call it ξ_N.

(4) Go to (2) and repeat.

Improvements to the above have been suggested by Mitchell (1974a). If insufficient increase in ϕ is being obtained Mitchell allows the number of points in the design spectrum to either increase beyond N+1 or decrease below N, always returning eventually to an N point design.

The main drawback in the exact optimal design problem is that there are no strong results, such as the equivalence theorem of Chapter 2, which enable one to test the optimality of a given design. Also, although the above algorithms are monotonic and will converge, since they are bounded above by the continuous optimal design, they may converge to different designs, given different starting points.

3.7.3. Some exact N-point designs have been found, mainly with D-optimality as criterion and $I(\underline{x})$ of rank 1. M.J. Box (1968a) points out that when N=K, there is a geometrical interpretation of the design problem. That is, to find the set of points

$$\{\underline{v}_1,\ldots,\underline{v}_K\}, \quad \underline{v}_i \in V, \; i=1,\ldots,K,$$

such that the simplex, with the K-points and the origin as vertices,

has maximum volume. Note that the matrix V^{-1}, $V = (\underline{v}_1 \ldots\ldots \underline{v}_K)$, will transform the K points onto the K unit vectors along the axes of K dimensional space. Note also that if the set V is not transformed inside the hypercube of side 2, centred on the origin and with faces which cut the axes at right angles, then a simple exchange of design points will lead to an increase in the determinant of interest. This geometrically inspired algorithm is, in fact, exactly the same as Fedorov's when N=K and D-optimality is the criterion. By imagining the worst possible situation when the above algorithm stops, the possible weakness of one point exchange algorithms is at once obvious. Note that this algorithm only guarantees to produce a V^{-1} which will transform V inside the hypercube described above, whilst if there is actually a K-point continuous D-optimal design, then there is a V^{-1} which will transform V inside the unit hypersphere.

M.J. Box (1968a) also shows that, if the V^{-1} for the best K-point design actually transforms V inside the right (p+1) hedron contained in the unit sphere, then the D-optimal N point design is the design which gives near equal replications at the points of the best K-point design.

For the normal-linear model with

$$\eta(x,\underline{\theta}) = \theta_o + \theta_1 x + \theta_2 x^2 + \ldots\ldots + \theta_{K-1} x^{K-1}, \mathcal{X} = [-1, +1],$$

Wynn (1972) shows that a near equal replication of the best K-point design is in fact the best N point design for K=2 and K=3.

As mentioned previously, Kiefer (1975) has investigated the exact optimality of generalised Youden square designs.

In general, however, it is necessary to resort to the numerical algorithms described above.

CHAPTER 4.

SEQUENTIALLY DESIGNED EXPERIMENTS

The possible necessity for sequentially designed experiments was indicated in Chapter 1. Also mentioned was the fact that the inferences to be made after such an experiment would require special consideration, possibly leading to differences in design procedures suggested by different schools of inferential thought. In this short chapter we consider these problems, starting by describing a general form of sequential experiment.

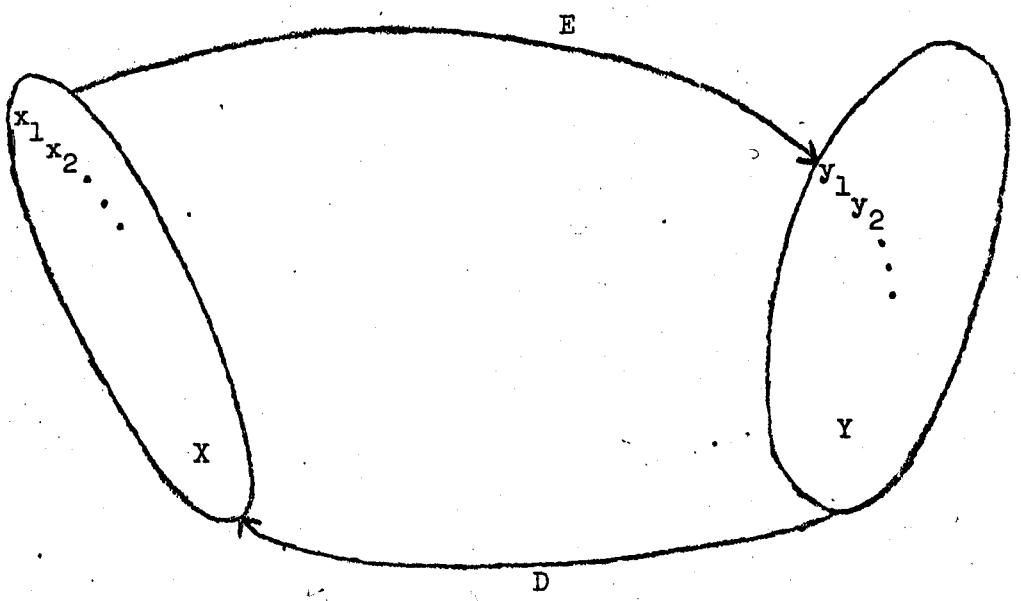


Fig. 4.1.1.

Consider the following sequentially designed experiment. Let the design space be $\mathcal{X}$ (for simplicity we shall regard $x \in \mathcal{X}$ as being scalar), at points in which an experiment E may be performed and an observation $y|x$, in the observation space Y, obtained. At stage n in the experiment a choice of the (n+1)st design point x_{n+1} is made according to a design procedure D. D will typically make use of the

vector of observations $\underline{y}_n$ and the vector of design points $\underline{x}_n$. In the context of this thesis it will suffice to consider only procedures D which are deterministic in nature. It will also often be true that the design space $\mathcal{X}$ will be rich in points relative to the observation space Y, and that there will be a one to one relationship, via the procedure D, between a design point x_{n+1} and the observation y_n, given the vectors $\underline{x}_n$ and $\underline{y}_{n-1}$. The situation prior to the(n+1)st stage of the experiment will be that observations $\underline{y}_n$ will have been obtained and a vector of design points $\underline{x}_{n+1}$ will have been defined. If there exists a one to one relationship as described above then, since D is deterministic, either $\underline{y}_n$ or $\underline{x}_{n+1}$ will be a sufficient statistic of the observations obtained. We shall assume that observations $y|x$ arise according to a probability distribution identified by the density function $p(y|x,\underline{\theta})$, which is known up to a vector of unknown parameters $\underline{\theta}$.

Let the likelihood of a set of observations obtained according to such a procedure, with a sample size of N, be $L_{\underline{y}_N}(\underline{\theta},D)$ then,

$$
\begin{aligned}
L_{\underline{y}_N}(\underline{\theta},D) &= p(y_1,\ldots,y_N|\underline{\theta},D) \\
&= p(\underline{y}_{N-1}|\underline{\theta},D).p(y_N|\underline{y}_{N-1},\underline{\theta},D) \\
&= p(\underline{y}_{N-1}|\underline{\theta},D).p(y_N|x_N,\underline{\theta},D), \text{ as D is deterministic.} \\
&= \prod_{i=1}^{N} p(y_i|x_i,\underline{\theta},D) \\
&= \prod_{i=1}^{N} p(y_i|x_i,\underline{\theta}).
\end{aligned}
$$

We write the likelihood as a function of D as D will typically form an integral part of the probability distribution of $\underline{y}_N$. Note that the likelihood of a particular realisation of the experiment, $(\underline{y}_N,\underline{x}_N)$ say, will be identical to that of a set of independent

observations $\underline{y}_N$ obtained at a set of predetermined design points $\underline{x}_N$. In the above experiment, however, each design point will be a function of the preceding observations and therefore, in repetitions of the experiment, different design sequences $\underline{x}_N$ will occur. This will be a major factor in separating the schools of inference to be considered.

4.2. In this section we will discuss maximum likelihood estimation of the parameters and the repeated sampling distribution of such estimates. Because the observations in the experiment are not independent, there are now no well known asymptotic results to appeal to, as there were in Chapter 1. The problem of proving consistency and asymptotic normality of maximum likelihood estimates in non-standard cases has been considered by Silvey (1961), Bar-Shalom (1971) and Bhat (1974). However, as pointed out by White (1975), the conditions imposed by the above would seem to be impossible to verify in practical problems of the type considered here.

Fedorov and Malyutov (1972) using a result of Jennrich (1969) observe that, in regression situations with normal error, if the design measure ξ_N tends to a non-singular limit as $N \to \infty$ then the least squares estimates of the parameters and hence, in this case, the maximum likelihood estimates of the parameters, will be consistent and asymptotically normal. White (1975) shows that if the sequence of estimates $\hat{\underline{\theta}}_N$ is consistent then, using a design procedure for D-optimality, the design sequence ξ_N will converge to a non-singular limit, namely the D-optimal design measure for the true value of the parameters $\underline{\theta}$. These facts, although interesting within themselves, do not really further the solution of the asymptotic problem, because the assumptions of the theorems are at least as strong as the results proved.

No real progress has been made towards the solution of this asymptotic problem. Instead we turn to the problem of what a statistician might do in practice. In fact, the asymptotic problem

mentioned above will be regarded as something of a red herring, the more important problem being, how well the approximations we shall make hold good for reasonably sized N and how we might design our experiment to make them better.

Let us consider the approach of Silvey, Bar-Shalom and Bhat to the proof of the asymptotic normality of maximum likelihood estimators. The standard procedure is to expand the log-liklihood function using Taylor's theorem, giving,

$$\frac{\partial \log L_{\underline{y}_N}(\hat{\underline{\theta}})}{\partial \underline{\theta}} = \frac{\partial \log L_{\underline{y}_N}(\underline{\theta})}{\partial \underline{\theta}} + \left\{\frac{\partial^2 \log L_{\underline{y}_N}(\underline{\theta})}{\partial \theta_i \, \partial \theta_j}\right\}\left(\hat{\underline{\theta}}_N - \underline{\theta}\right) + \underline{\varepsilon}_N.$$

$$\frac{\partial \log L_{\underline{y}_N}(\hat{\underline{\theta}})}{\partial \underline{\theta}} = 0 \text{ , by definition of } \hat{\underline{\theta}}_N.$$

The first assumption made is that the elements of $\underline{\varepsilon}_N$ are small relative to the other terms given.

Therefore, an approximation for $(\hat{\underline{\theta}}_N - \underline{\theta})$ might be,

$$(\hat{\underline{\theta}}_N - \underline{\theta}) \doteq \left\{\frac{-\partial^2 \log L_{\underline{y}_N}(\underline{\theta})}{\partial \theta_i \, \partial \theta_j}\right\}^{-1} \frac{\partial \log L_{\underline{y}_N}(\underline{\theta})}{\partial \underline{\theta}} ,$$

and, assuming that $\hat{\underline{\theta}}_N$ is approximately unbiased for $\underline{\theta}$, an approximation for var $(\hat{\underline{\theta}}_N)$ might be,

$$\text{var}\,(\hat{\underline{\theta}}_N) \doteq \mathop{E}_{\underline{y}_N}\left[(\hat{\underline{\theta}}_N - \underline{\theta})(\hat{\underline{\theta}}_N - \underline{\theta})^T\right]$$

$$\doteq \mathop{E}_{\underline{y}_N}\left[\left\{\frac{\partial^2 \log L_{\underline{y}_N}(\underline{\theta})}{\partial \theta_i \, \partial \theta_j}\right\}^{-1} \cdot \frac{\partial \log L_{\underline{y}_N}(\underline{\theta})}{\partial \underline{\theta}} \cdot \frac{\partial \log L_{\underline{y}_N}(\underline{\theta})^T}{\partial \underline{\theta}} \cdot \left\{\frac{\partial^2 \log L_{\underline{y}_N}(\underline{\theta})}{\partial \theta_i \, \partial \theta_j}\right\}^{-1}\right]$$

For suitably large N this approximation would be expected to hold well if the estimator $\hat{\underline{\theta}}_N$ had the nice properties of asymptotic normality

hoped for. A major problem lies in calculating it. The expectation is over all sequences $\underline{y}_N$ which might occur. This will depend in a complex fashion on D and $\underline{\theta}$. However, for given $\underline{\theta}$, computer simulations of a large number of similarly designed experiments of size N will enable one, at least theoretically, to obtain a reasonable estimate of the above. Note that if such simulations were actually carried out, then the actual values of $\hat{\underline{\theta}}_N$ could be computed and the sampling distribution of $\hat{\underline{\theta}}_N$ estimated. Comparison of the estimates of $\text{var}(\hat{\underline{\theta}}_N)$ obtained by these two methods suggests an empirical means of testing the stability of the design procedure being used.

White (1975) suggests that $\frac{1}{N} M(\xi_N)^{-1}$ might be a suitable approximation to $\text{var}(\hat{\underline{\theta}}_N)$, where $M(\xi_N)$ is the Fisher information matrix of a design measure ξ_N actually attained in a given sequential experiment. White obtains motivation for this approximation from the fact that if the same design measure ξ_N is replicated m times then as $m \to \infty$, $\hat{\underline{\theta}}_{N,m} \sim N(\underline{\theta}, \frac{1}{Nm} M(\xi_N)^{-1})$. This, of course, ignores the fact that, in replications of the actual sequentially designed experiment, radically different ξ_N can be expected, even for relatively large N, therefore causing great variation in $M(\xi_N)^{-1}$. However, despite the fact that the degree of approximation may not be particularly good, it would appear to be the only alternative to a large amount of computing, if we desire to estimate $\text{var}(\hat{\underline{\theta}}_N)$.

To summarise this section, it would seem that one does not have definite backing from theory for the assumption of normality of maximum likelihood estimates with the type of design procedure described above. Also, even if in practice approximate normality of the form $\hat{\underline{\theta}}_N \sim N(\underline{\theta}, \text{var}(\hat{\underline{\theta}}_N))$ were to be assumed, $\text{var}(\hat{\underline{\theta}}_N)$ would appear to be very difficult to calculate or even estimate.

4.3. Consider now an inferential approach based on the likelihood principle. One of the essential differences between the approaches of followers of the likelihood principle and the repeated sampling schools of thought is that the former base their inferences only on the probability of the events which actually take place in an experiment, as opposed to the

'what might have happened' attitude of the repeated sampling school. Therefore, in a sense, the inference problem for a believer in the likelihood principle is simpler. He need not concern himself with the other possible sequences of design points which a complex design procedure D might produce, as to him these are irrelevant.

Let us assume that inferences are to be made on $\underline{\theta}$ directly from the likelihood $L_{\underline{y}_N}(\underline{\theta}) = \prod_{i=1}^{N} p(y_i|x_i,\underline{\theta})$. A Bayesian argument will take the same course as what follows, as the posterior distribution $\pi(\underline{\theta}|\underline{y}_n)$ will, as a function of $\underline{\theta}$, differ from the likelihood only by a multiplying factor $\pi(\underline{\theta})$, the prior distribution on $\underline{\theta}$.

Note that, although we can write down explicitly a function which may be used to express relative degrees of belief in values of $\underline{\theta}$, we may not be able to use it directly for making inferences on $\underline{\theta}$, because of its complex nature as a function of $\underline{\theta}$. What is usually done in practice is to normalise the function. As noted in Chapter 1 this is equivalent to assuming that the log-likelihood function can adequately be approximated in a neighbourhood of $\hat{\underline{\theta}}_N$ by a second order Taylor expansion.

$$\log L_{\underline{y}_N}(\underline{\theta}) = \ell_{\underline{y}_N}(\underline{\theta}) \doteq \text{const} - \frac{1}{2}(\hat{\underline{\theta}}_N - \underline{\theta})^T \left\{ - \frac{\partial^2 \ell_{\underline{y}_N}(\underline{\theta})}{\partial\theta_i \partial\theta_j} \right\} (\hat{\underline{\theta}}_N - \underline{\theta}) .$$

As in the previous section we are lacking in asymptotic theoretical backing for this approximation. The conditions for asymptotic normality of posterior distributions are similar to those required for asymptotic normality of maximum likelihood estimates, however, the suitability of the above approximation would be relatively easy to check in any particular case.

4.4. The following example is intended to highlight the possible differences in the making of inferences based on repeated sampling distributions of estimators and on likelihood principle methods.

Example

A parameter θ may take only two values, θ_1 or θ_2. An informative experiment is carried out according to the following design.

(1) Take 5 observations. If the observations satisfy a condition, A say, let $\hat{\theta}$ be θ_1. If they satisfy another condition, B say, let $\hat{\theta}$ be θ_2. Otherwise continue.

(2) Take 5000 more observations and estimate θ.

Suppose that if 5005 observations are taken then the true value of the parameter is estimated with probability near 1. Suppose also that for the first 5 observations $p(A|\theta_2) = p(B|\theta_1) = \cdot 2$ and $p(A|\theta_1) \neq p(B|\theta_2)$.

That is, $p(\hat{\theta} = \theta_2|\theta_2) = \cdot 8$, $p(\hat{\theta} = \theta_1|\theta_2) = \cdot 2$

$p(\hat{\theta} = \theta_2|\theta_1) = \cdot 2$, $p(\theta = \theta_1|\theta_1) = \cdot 8$.

Therefore, based on this repeated sampling distribution of $\hat{\theta}$ alone, if $\hat{\theta} = \theta_2$ one might be tempted to lay odds of 4:1 on this being the true value of θ .

The above approach would seem to ignore the fact that subsequent to the experiment one would know the sample size and, if it were 5005, one would intuitively be thinking of putting odds of higher than 4:1 on $\theta = \theta_2$ being correct. The repeated sampling distribution of $\hat{\theta}$ would seem not to be a suitable vehicle for making inferences on the true value of $\hat{\theta}$, a method of inference which allows for conditioning on the events which actually take place being necessary. We note that, in the above example, the sample size is not necessarily an ancillary statistic. Despite this, if one were to condition on the design pathway which the experiment actually took, namely 5 or 5005 observations, then the conditional distributions of $\hat{\theta}_5$ or $\hat{\theta}_{5005}$ might give information which would lead to more sensible odds.

However, in the more complex design situation envisaged in 4.1., if one were to condition on the design path followed there would be no variability left in the experiment for repeated sampling methods. This is because the design path in the form of x_{n+1} is a sufficient statistic.

The lesson to be learned from the above would seem to be that the distributions of estimators in sequential experiments will depend not only on the natural variation in the populations being sampled from, but on the design procedure D being used. If the sequential actions embodied in a procedure D are independent of the data observed, then it may be possible to invoke the idea of ancillarity to justify the use of conditional repeated sampling inferences. In some situations, such as the example above, it might seem reasonable to make repeated sampling inferences after conditioning on a statistic which is not ancillary. However, this idea would seem to be extremely artificial and difficult to justify in general. As has been indicated above, in the fully sequential experiment envisaged in 4.1, any form of conditional repeated sampling method of inference would seem to be unsuitable. If one accepts the need for conditioning, in this instance, then essentially all one is left with is the likelihood function and a likelihood principle approach to inference.

It would seem that motivation for experimental design comes from two sources, which must lead to, at least in interpretation, different design criteria. In the first instance consider an experiment which is being carried out for the purpose of obtaining a point estimate. In this situation the repeated sampling distribution of the estimator being used, possibly in conjunction with a suitable loss function, would seem to be of prime importance for study with regards to experimental design. Practical examples of this type of 'decision with utility' problem are difficult to imagine, however, we do not exclude the possibility of their existence. Secondly, suppose that subsequent to an experiment we wish to make some kind of statement on the parameter space via an interval or subjective probability statement. In this situation the above example and ensuing discussion would appear to indicate the use

of an inferential approach based on the likelihood principle and, with regards to experimental design, an approach similarly based. That is, at a given point in a sequential experiment, one will be interested in conducting the remainder of the experiment on the basis of what has already taken place as opposed to what might have taken place.

The two approaches given above will not be independent. A design method of the second type, by which one attempts to make as precise statements as possible about the true parameter values in every realisation of the experiment will intuitively lead to a design method D which, for a sensible estimator $\hat{\underline{\theta}}$, will be 'good' in the repeated sampling sense. Where a priori design is possible the two motivations for design will essentially lead to the same type of criteria for design, as one has the opportunity to take the optimum design pathway in each realisation of the experiment (c.f. Chapter 1).

CHAPTER 5.

SOME SEQUENTIAL DESIGN PROCEDURES

As was seen in Chapter 1 the optimal static design will typically be dependent on $\underline{\theta}$. As the object of the experiment is to estimate the vector $\underline{\theta}$, in some sense, it may be seen that a designed experiment must necessarily be carried out sequentially, gradually building up knowledge about $\underline{\theta}$. We shall define a fully sequential experiment to be an experiment in which each stage consists of using only one experimental unit. In Chapter 7 it will be suggested that, in certain circumstances, a fully sequential experiment may be unnecessary, or even impossible to achieve, due to practical restrictions. However, in this instance, it will be assumed that a fully sequential design is possible. In the present chapter we shall consider contenders for a sequential procedure and motivation for their use. In Chapter 6, via a simulation study, an attempt will be made to compare the performance of some of these procedures in practice.

Note the following remarks on notation:

(i) Design points $x \in \mathcal{X}$ will again be written as scalars for simplicity.

(ii) We shall revert to earlier notation by writing the Fisher information matrix $I(x)$ as a function of $\underline{\theta}$, that is, as $I(x,\underline{\theta})$.

(iii) The Fisher information matrix for a set of independent observations taken at points in the vector $\underline{x}_n$ is written as $M(\underline{x}_n,\underline{\theta}) = \sum_{i=1}^{n} I(x_i,\underline{\theta})$.

(iv) The sample information matrix at stage n in a sequential experiment is written as $S(\underline{x}_n,\underline{y}_n,\underline{\theta})$.

5.1. Consider the situation at stage $n < N$ in an experiment where a total of N observations are to be taken, and the criterion for design is ϕ.

Let us first consider an approach to experimental design based on the repeated sampling distribution of maximum likelihood estimates. At stage n observations have been taken at design points $(x_1,\ldots,x_n) = \underline{x}_n^T$. At the end of the experiment it might be assumed that $\hat{\underline{\theta}}_{N,ML}$ is is approximately distributed as $N(\underline{\theta}, \text{var}(\hat{\underline{\theta}}_{N,ML}))$. The design problem is to decide at stage n how the remaining (N-n) observations should be taken, in order to maximise $\phi\{\text{var}(\hat{\underline{\theta}}_{N,ML})\}$. $\text{var}\,(\hat{\underline{\theta}}_{N,ML})$ will be a function of the unknown vector $\underline{\theta}$ and because of this the best plan would seem to be to look, not (N-n) steps ahead, but only one step ahead, and to select the (n+1) st point in some optimum fashion based on the knowledge of $\underline{\theta}$ obtained up to stage n. That is, to adopt a fully sequential experiment as defined above.

As noted in Chapter 4 Fedorov and White have used the following approximation for $\text{var}\,(\hat{\underline{\theta}}_n)$,

$$\text{var}(\hat{\underline{\theta}}_n) \doteq M(\underline{x}_n,\underline{\theta})^{-1} .$$

White (1975) considers the case where the criterion is D-optimality and obtains motivation for a design procedure from Wynn's iterative algorithm as described in Chapter 3. That is,she selects as x_{n+1} the point which corresponds to the direction of maximal rate of ascent of ϕ at $M(\xi_n,\underline{\theta})$, ξ_n being the design measure putting weights $\frac{1}{n}$ at each of $x_1,\ldots,x_n$. White substitutes $\hat{\underline{\theta}}_{n,ML}$ in $M(\xi_n,\underline{\theta})$ at each stage. Fedorov and Malyutov (1972) put forward a generalisation of this procedure which might be described as follows.

(1) Take K initial observations. Estimate $\underline{\theta}$ by $\hat{\underline{\theta}}_{K,ML}$. Set n=K.

(2) Find x_{n+1} such that $\phi\{M(\xi_n,\hat{\underline{\theta}}_{n,ML}), I(x_{n+1},\hat{\underline{\theta}}_{n,ML})\}$ is maximised, $x_{n+1} \in \mathcal{X}$.

(3) Take an observation at x_{n+1}, re-estimate $\underline{\theta}$ by $\hat{\underline{\theta}}_{n+1,ML}$. n= n+1. Go to (2) if $n \leq N$.

It should be noted that with the above procedure the maximum increase in ϕ will not necessarily be obtained. That is,

$$\phi\{M(\underline{x}_{n+1}, \hat{\underline{\theta}}_{n,ML})\} - \phi\{M(\underline{x}_n, \hat{\underline{\theta}}_{n,ML})\}$$

is not necessarily maximised. However, with D-optimality as criterion, it may readily be shown that the above procedure will give the maximal increase in ϕ.

At this juncture it would seem to be worthwhile to make a comparison between the sequential experimentation being considered in the present chapter and the iterative algorithms of Chapter 3. In the general discussion of the iterative algorithms of Chapter 3 it was observed that it might be sensible to use an algorithm which might take a large number of steps to get close to an optimal design if the computational work involved in each step was small. In an actual sequential experiment, however, the use of an experimental unit will typically represent a far greater outlay of resources than the taking of an extra step in a computer program. It would therefore seem to be more economical to attempt to get the maximum increase in ϕ at each stage in the design. In this sense, therefore, the iterative algorithms of Chapter 3 are possibly the wrong places to look for motivation for sequential design. The following is suggested as an alternative to Fedorov's procedure described above.

(1) Take K initial observations and estimate $\underline{\theta}$ by $\hat{\underline{\theta}}_{K,ML}$. $n = K$.

(2) Find x_{n+1} such that $\phi\{M(\underline{x}_n, \hat{\underline{\theta}}_{n,ML}) + I(x_{n+1}, \hat{\underline{\theta}}_{n,ML})\}$ is a maximum, $x_{n+1} \in \mathcal{X}$.

(3) Take an observation at x_{n+1} and re-estimate $\underline{\theta}$ by $\hat{\underline{\theta}}_{n+1,ML}$. $n = n+1$. Go to (2) if $n \leq N$.

5.2. A likelihood principle approach to experimental design will now be considered. At the end of an experiment it might be assumed that the likelihood function can be approximated to, in a neighbourhood of $\hat{\underline{\theta}}_{n,ML}$, by a second order Taylor expansion. A natural design criterion would seem to be to attempt to maximise some function ϕ of $S(\underline{x}_n, \underline{y}_n, \hat{\underline{\theta}}_{n,ML})$, a matrix which will describe the local shape of the likelihood function

at $\underline{\theta} = \hat{\underline{\theta}}_{n,ML}$. Again a fully sequential experiment would seem to be appropriate. Given $(\underline{x}_n, \underline{y}_n)$ at stage n one would wish to select $x_{n+1} \in \mathcal{X}$ to maximise $\phi\{ S(\underline{x}_{n+1}, \underline{y}_{n+1}, \hat{\underline{\theta}}_{n,ML})\}$. Note that $S(\underline{x}_{n+1}, \underline{y}_{n+1}, \hat{\underline{\theta}}_{n,ML})$ will typically be a function of y_{n+1}, the observation to be obtained from x_{n+1}. For this reason the most suitable design criterion would seem to be to maximise

$$\mathop{\mathbb{E}}_{y_{n+1}} \{ \phi \{ S(\underline{x}_{n+1}, \underline{y}_{n+1}, \hat{\underline{\theta}}_{n,ML}) \} \}.$$

Although this would seem to be the best design criterion in theory, the practical optimisation problem might be fairly difficult. G.E.P. Box and W.G. Hunter (1965a) have treated the case of non-linear regression with normally distributed error using Bayesian methods. They note that if the regression function $\eta(x,\underline{\theta})$ can be approximated, in a neighbourhood of $\hat{\underline{\theta}}_{n,ML}$, by a first order Taylor expansion, that is

$$\eta(x,\underline{\theta}) = \eta(x, \hat{\underline{\theta}}_{n,ML}) + (\underline{\theta} - \hat{\underline{\theta}}_{n,ML})^T . \eta_{\underline{\theta}}(x, \hat{\underline{\theta}}_{n,ML}),$$

then $S(\underline{x}_{n+1}, \underline{y}_{n+1}, \hat{\underline{\theta}}_{n,ML})$ may be written as

$$\sum_{i=1}^{n+1} \eta_{\underline{\theta}}(x_i, \hat{\underline{\theta}}_{n,ML}) . \eta^T_{\underline{\theta}}(x_i, \hat{\underline{\theta}}_{n,ML}),$$ which is independent of the vector of observations $\underline{y}_{n+1}$. Therefore, in the non-linear regression situation, this approximation reduces the criterion for design to that of White and Fedorov described in 5.1. It should be added however that this quasi-linearisation of $\eta(x,\underline{\theta})$ cannot be extended to other models in general as may be seen by considering the case of binary observations with $p(1|x,\underline{\theta}) = \eta(x,\underline{\theta})$.

Another approach is applicable to all models and also removes the awkward expectation from

$$\mathop{\mathbb{E}}_{y_{n+1}} \{\phi\{S(\underline{x}_{n+1}, \underline{y}_{n+1}, \hat{\underline{\theta}}_{n,ML})\}\}.$$

ϕ is a concave function, therefore, making use of Jensen's inequality,

we have,

$$\mathop{E}_{y_{n+1}} \{\phi\{ S(\underline{x}_{n+1}, \underline{y}_{n+1}, \hat{\underline{\theta}}_{n,ML}) \} \}$$

$$= \mathop{E}_{y_{n+1}} \{\phi\{ S(\underline{x}_n, \underline{y}_n, \hat{\underline{\theta}}_{n,ML}) + S(x_{n+1}, y_{n+1}, \hat{\underline{\theta}}_{n,ML})\} \}$$

$$\leq \phi\{ S(\underline{x}_n, \underline{y}_n, \hat{\underline{\theta}}_{n,ML}) + I(x_{n+1}, \hat{\underline{\theta}}_{n,ML})\} .$$

The design criterion suggested by the above is to try to maximise the given upper bound for

$$\mathop{E}_{y_{n+1}} \{\phi\{ S(\underline{x}_{n+1}, \underline{y}_{n+1}, \hat{\underline{\theta}}_{n,ML}) \} \}$$

with respect to x_{n+1}.

A sequential procedure may be obtained by substitution of

$$\mathop{E}_{y_{n+1}} \{\phi\{ S(\underline{x}_{n+1}, \underline{y}_{n+1}, \hat{\underline{\theta}}_{n,ML})\} \} ,$$

or its upper bound, in the second procedure of 5.1. in place of

$$\phi\{ M(\underline{x}_{n+1}, \hat{\underline{\theta}}_{n,ML}) \} .$$

5.3. The sequential procedures of 5.1 and 5.2 may be summarised as follows.

(1) Take K observations to start the process. Estimate $\underline{\theta}$ by $\hat{\underline{\theta}}_{K,ML}$. $n = K$.

(2) Choose $x_{n+1} \in \mathcal{X}$ to maximise some function of x_{n+1} and $\hat{\underline{\theta}}_{n,ML}$, say $f(x_{n+1}, \hat{\underline{\theta}}_{n,ML})$.

(3) Take an observation at x_{n+1}. Re-estimate $\underline{\theta}$ by $\hat{\underline{\theta}}_{n+1,ML}$. $n=n+1$. Go to (2) and repeat if $n \leq N$.

The following comments and conjectures are made concerning this form of procedure.

(i) The parameter $\underline{\theta}$ is estimated by $\hat{\underline{\theta}}_{n,ML}$ at each stage.

(ii) The sequence of estimates $\{\hat{\underline{\theta}}_n, i=K,\ldots,n\}$ will be subject to a fair amount of fluctuation for n small relative to K.

(iii) This excitability of $\{\hat{\underline{\theta}}_n\}$ in the early stages of the experiment will manifest itself, in step (2), in an erratic sequence of x_{n+1}'s.

(iv) The erratic nature of the x_{n+1}'s may introduce a feedback of excitability into the $\{\hat{\underline{\theta}}_n\}$ sequence affecting the rate at which the sequence is settling down to give a consistent estimate for $\underline{\theta}$.

These comments are made, for the moment, with only intuition as justification. Assuming the conjectures to be valid we shall consider possible means of improving the situation. As analytic comparison of design procedures of this type would seem to be impossible, it will be necessary to resort to a computer simulation study in Chapter 6. In this study an attempt will be made to compare the above procedures with alternatives which will be suggested and thereby justify the assertions made above, at least in the example to be considered.

Comment (ii) above concerns the stability of the sequence of estimates $\hat{\underline{\theta}}_{n,ML}$. For relatively small n the nature of the likelihood function may change greatly after only one additional observation. By its very nature the point $\hat{\underline{\theta}}_{n,ML}$, giving the maximising value of the likelihood function, will be very sensitive to such changes. As n becomes large and as long as observations have been taken at reasonably informative points in the design space then each observation will have diminishing effect on the shape of the likelihood function, thereby introducing a natural stability to $\hat{\underline{\theta}}_{n,ML}$. The type of estimator for $\underline{\theta}$ which the above comments would seem to indicate would be one which was more stable than $\hat{\underline{\theta}}_{n,ML}$ for small n and for which any bias introduced

by the stabilising process would be naturally overcome by the weight of information as n increased. The term 'estimator for $\underline{\theta}$' in the above sentence is used rather loosely, a more suitable expression might be 'value to be substituted for $\underline{\theta}$ in $f(x_{n+1},\underline{\theta})$', there being no compelling reason why the value used should be one which might actually be used as a point estimate of $\underline{\theta}$ if the experiment were to stop at stage n, for n small. A procedure which uses the set of rules given above with a sequence of estimates given by $\{\hat{\underline{\theta}}_{n,ML}\}$ shall be referred to as a Type 1 process.

Because of the comments made above it seems reasonable to suppose that improvement in design might be achieved by using a sequence $\{\hat{\underline{\theta}}_n\}$ which is less erratic than that produced by the maximising value of the likelihood function. Another contender which comes to mind is some form of weighted average of $\underline{\theta}$ over the likelihood function. For example one could take

$$\hat{\underline{\theta}}_n = \frac{\int \underline{\theta}\, L_{\underline{y}_n}(\underline{\theta})\, d\,\underline{\theta}}{\int L_{\underline{y}_n}(\underline{\theta})\, d\,\underline{\theta}},$$

which is immediately recognisable, in Bayesian terminology, as the mean of the posterior distribution on $\underline{\theta}$ with an improper uniform prior. This suggests considering a prior distribution $\pi(\underline{\theta})$ on $\underline{\theta}$ and using the posterior mean, that is

$$\hat{\underline{\theta}}_n = \frac{\int \underline{\theta}\, \pi(\underline{\theta})\, L_{\underline{y}_n}(\underline{\theta})\, d\,\underline{\theta}}{\int \pi(\underline{\theta})\, L_{\underline{y}_n}(\underline{\theta})\, d\,\underline{\theta}}.$$

One of the main criticisms of Bayesian methods is that the prior distribution $\pi(\underline{\theta})$ must be constructed by the experimenter, and therefore might introduce bias into inferences being made subsequent to an experiment, particularly for small samples. Let us suppose that a sequence of the above type is used and that the prior distribution does have a biasing effect on $\hat{\underline{\theta}}_n$. Even if this is true, it might be suggested that a more stable design procedure would result, not only because an averaging process is being used instead of maximisation, but

because the prior distribution will have the effect of an extra set of observations in slowing the rate of change of $\hat{\underline{\theta}}_n$ as the procedure progresses. As n increases the effect of the prior distribution will, of course, diminish, giving the sequence $\{\hat{\underline{\theta}}_n\}$ the type of properties hoped for. As a bonus, if the prior distribution has the the effect of concentrating the posterior distribution in a neighbourhood of the true value of $\underline{\theta}$ then the stabilising effect and therefore the 'optimality' of the design method would be expected to be improved even more.

An alternative process would be to use the posterior mode, that is, the maximising value of $\pi(\underline{\theta}).L_{\underline{y}_n}(\underline{\theta})$. This would also be expected to be more satisfactory than the pure maximum likelihood estimate. However, in the simulation study of Chapter 6 only a method using the posterior mean will be considered and a process which uses this form of estimator will be referred to as a Type 2 process.

By consideration of the general procedure described at the start of this section it may be seen that the reason for estimating $\underline{\theta}$ is to enable one to obtain an approximation of the value of x_{n+1} which maximises $f(x_{n+1},\underline{\theta})$ by maximising $f(x_{n+1},\hat{\underline{\theta}}_n)$. The value of x_{n+1} maximising $f(x_{n+1},\underline{\theta})$ will typically be a function of $\underline{\theta}$, $g(\underline{\theta})$ say. In practice we may not be able to write down this function explicitly, but, given a $\underline{\theta}$, $g(\underline{\theta})$ could be computed. Therefore,the reason for estimating $\underline{\theta}$ is to give a selection of the best next design point as $g(\hat{\underline{\theta}}_n)$. An alternative approach might be to consider x_{n+1} as a function of $\underline{\theta}$, namely $g(\underline{\theta})$, and to use as the (n+1)st design point, not $g(\hat{\underline{\theta}}_n)$, but the expectation of $g(\underline{\theta})$ over the posterior distribution for $\underline{\theta}$. That is,

$$x_{n+1} = \frac{\int g(\underline{\theta})\ \pi(\underline{\theta})\ L_{\underline{y}_n}(\underline{\theta})\ d\ \underline{\theta}}{\int \pi(\underline{\theta})\ L_{\underline{y}_n}(\underline{\theta})\ d\ \underline{\theta}}$$

For the reasons given above this procedure might also be expected to be more stable than the one using $x_{n+1} = g(\hat{\underline{\theta}}_{n,ML})$ as design point

at each stage. This third procedure will be referred to as a Type 3 process.

It should be noted that numerical calculations involved in Type 2 and Type 3 processes will typically be more demanding than those for the Type 1 process.

We now move on to investigate how these three processes work in practice using a simulation study on a simple example.

CHAPTER 6

A SIMULATION STUDY

The aim of this study will be to compare the effects of using design methods of Types 1, 2 and 3 in practice. In order to make this study feasible computationally the simplest form of example has been taken, that being a situation where there is only one unknown parameter in the model.

6.1. The probability model made use of in this study is of the following form.

$$p(1|x,\theta) = \exp(-\theta x), \quad p(0|x,\theta) = 1 - \exp(-\theta x), \quad x \in \mathcal{X} \equiv [a,b], \quad \theta > 0.$$

It may easily be shown that the following is true.

$$S(x,y,\theta) = \begin{cases} 0 & , \quad y=1 \\ \dfrac{x^2.\exp(-\theta x)}{(1-\exp(-\theta x))^2} & , \quad y=0 \end{cases}.$$

$$I(x,\theta) = \frac{x^2}{\exp(\theta x)-1}$$

Because K=1 the optimal static design will exist at one point, the optimal point being the value of x, x^* say, maximising $\dfrac{x^2}{\exp(\theta x)-1}$.

It can be shown that $x^* \in \{\frac{c}{\theta}, a, b\}$, $c \doteqdot 1.59$, where x^* takes the value $\frac{c}{\theta}$ if $\frac{c}{\theta} \in [a,b]$ and otherwise a or b according as which maximises $\dfrac{x^2}{\exp(\theta x)-1}$.

Therefore, the optimal static design is θ dependent and a sequential type procedure is indicated.

Because K=1 all of the criteria considered in Chapter 1 will be equivalent, reducing to $\dfrac{1}{\sum_{i=1}^{N} I(x_i,\theta)}$ for a repeated sampling type

criterion or to $\frac{-1}{\sum_{i=1}^{N} S(x_i, y_i, \theta)}$ for a likelihood type approach.

Consider first a repeated sampling type procedure and how the processes of Types 1, 2 and 3 will progress.

(1) At stage n in a Type 1 process x_{n+1} will be chosen to maximise

$$\frac{-1}{\sum_{i=1}^{n} I(x_i, \hat{\theta}_n) + I(x_{n+1}, \hat{\theta}_n)} ,$$

which is obviously maximised by $x^*_{n+1} \varepsilon \{\frac{c}{\hat{\theta}_{n,ML}}, a, b\}$,with rules as given above.

(2) A Type 2 process will be similar to a Type 1 process, the difference being that x_{n+1} will be given by $x^*_{n+1} \varepsilon \{\frac{c}{\hat{\theta}_{n,B}}, a, b\}$, where $\hat{\theta}_{n,B}$ denotes a Bayes type estimate.

(3) A Type 3 process will be relatively easy to apply in this example as $g(\theta)$ can be written explicitly as $\frac{c}{\theta}$, and x_{n+1} will be given by,

$$x_{n+1} = \frac{\int \frac{c}{\theta} \cdot \pi(\theta) \cdot L_{\underline{y}_n}(\theta)\, d\theta}{\int \pi(\theta)\, L_{\underline{y}_n}(\theta)\, d\theta} .$$

If a full likelihood principle-type design were to be attempted then the process would be a little more complex, as $g(\theta)$ can not now be written down explicitly and will be dependent on $S(\underline{x}_n, \underline{y}_n, \theta)$.

$g(\theta)$ will be the value of x_{n+1} maximising ,

$$\mathbb{E}_{y_{n+1}} \left\{ \frac{-1}{(S(\underline{x}_n, \underline{y}_n, \theta) + S(x_{n+1}, y_{n+1}, \theta))} \right\} .$$

$$= \frac{-p(1|x_{n+1},\theta)}{S(\underline{x}_n,\underline{y}_n,\theta)} - \frac{p(0|x_{n+1},\theta)}{S(\underline{x}_n,\underline{y}_n,\theta) + S(x_{n+1},0,\theta)}$$

$$= - \frac{S(\underline{x}_n,\underline{y}_n,\theta) + p(1|x_{n+1},\theta).S(x_{n+1},0,\theta)}{S(\underline{x}_n,\underline{y}_n,\theta).\{ S(\underline{x}_n,\underline{y}_n,\theta) + S(x_{n+1},0,\theta)\}} .$$

The Type 1,2 and 3 processes will proceed as described in Chapter 5, only with the function $g(\theta)$ as defined above. To minimise computing only a procedure such as that arrived at via the repeated sampling type method was used in the simulation. It should be noted, however, that this procedure might have been arrived at using the alternative likelihood principle process suggested in 5.2. That is, let $g(\theta)$ be the value of x_{n+1} maximising

$$- \frac{1}{\{S(\underline{x}_n,\underline{y}_n,\theta)+I(x_{n+1},\theta)\}} .$$

This gives again $g(\theta) = \frac{c}{\theta}$.

6.2. The simulation study was carried out according to the rules given in 6.1. Relevant details of the computer simulation study are discussed in Appendix 5.

The design space $\mathcal{X}$ was set to be the interval $[\cdot 5,30]$, and the simulation study was repeated with data being generated from distributions having parameter values $\theta = 1.0$ and $\theta = \sqrt{1.59} \doteq 1.2639$.

With processes Type 2 and Type 3 the prior distribution was taken to be uniform over the positive half of the real line. It was hoped that this would avoid biasing the study in favour of these processes.

All statistics generated in the study are based on 500 independent sequentially generated experiments. Statistics generated in the study, by which it was hoped to compare the processes, are listed below.

(i) For each process type the maximum likelihood estimates $\hat{\theta}_{n,ML}$ were obtained for n=10, 15, 20, 25, 30, 35, 40, 45, in each replication of the experiment. Relative frequency histograms were drawn up to investigate the effect of increasing n on the repeated sampling distribution of $\hat{\theta}_{n,ML}$. A typical set of histograms, for $\theta = 1.0$ and a Type 1 process, is given in Fig.6.2.1.. The sample means and variances of $\hat{\theta}_{n,ML}$ were computed for each process and n = 10, 15, 20, 25, 30, 35, 40, 45.

These are listed in Fig.6.2.2.

(ii) In order to investigate any possible bias in taking $\pi(\theta)$ to be the improper uniform prior, the repeated sampling distributions of $\hat{\theta}_{n,B}$ were investigated in the same way as the maximum likelihood estimates were. Sample means and variances of $\hat{\theta}_{n,B}$ are given in Fig.6.2.3.

(iii) In 4.2. it was suggested that a suitable method of investigating the stability of a process might be to compare estimates of var $(\hat{\theta}_{n,ML})$ obtained via the parameter estimates and via an estimate of

$$\mathop{E}_{\underline{y}_n}\left\{\left[\frac{\partial\, \ell_{\underline{y}_n}(\theta)}{\partial\theta}\right]^2 \Big/ \left[\frac{\partial^2 \ell_{\underline{y}_n}(\theta)}{\partial\theta^2}\right]^2\right\} .$$

The first are contained in Fig.6.2.2 and the second are tabulated in Fig.6.2.4.

(iv) In order to compare the processes in a Bayesian sense it was thought that comparisons of

$$\mathop{E}_{\underline{y}_n}\left\{\frac{1}{S(\underline{x}_n, \underline{y}_n, \theta)}\right\} ,$$

between process types, might be informative. Estimates of the above expectations are given in Fig.6.2.5.

(v) Another Bayes type comparison was obtained by recording, for each simulated experiment, the number of observations that were required for $S(\underline{x}_n, \underline{y}_n, \theta)$ to attain a fixed value. In this case the value used was 35.0. A typical relative frequency histogram for this type of data, for $\theta = 1.0$ and a Type 1 process, may be seen in Fig.6.2.6. The sample means and variances of the number of observations until absorption by this upper barrier are given in Fig.6.2.7, for each process type.

Note: In Fig.6.2.2. $\bar{\theta}_{ML}$ and S^2_{ML} denote the sample mean and variance of the maximum likelihood estimates. In Fig.6.2.3. $\bar{\theta}_B$ and S^2_B denote the sample mean and variance of the Bayes type estimates. In Fig. 6.2.7. $\bar{n}$ and S^2_n denote the sample mean and variance of the number of steps until absorption.

Distribution of $\hat{\theta}_{n,ML}$ for a Type 1 process, $\theta = 1.0$

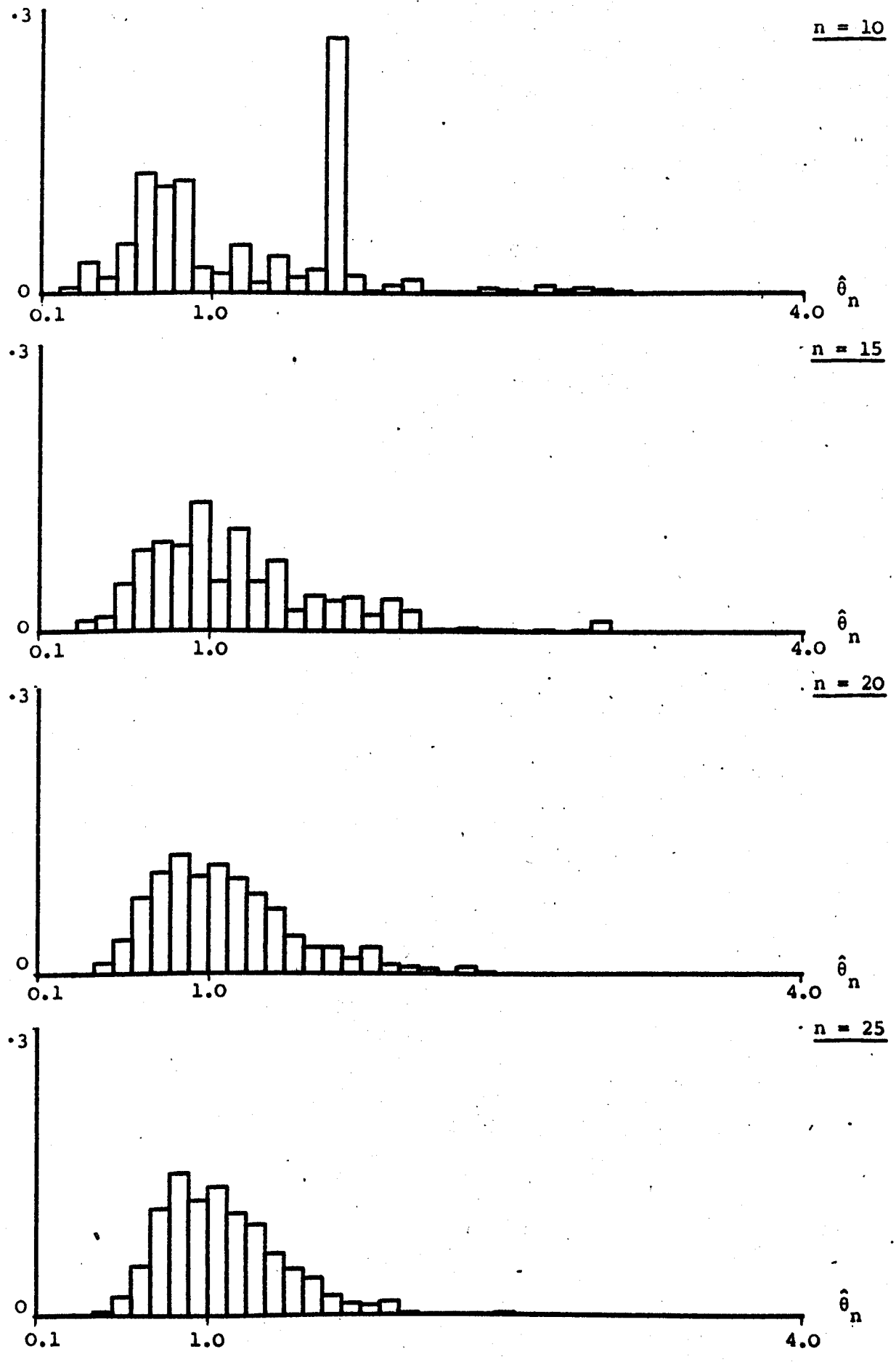


Fig. 6.2.1.a.

Distribution of $\hat{\theta}_{n,ML}$ for a Type 1 process, $\theta = 1.0$

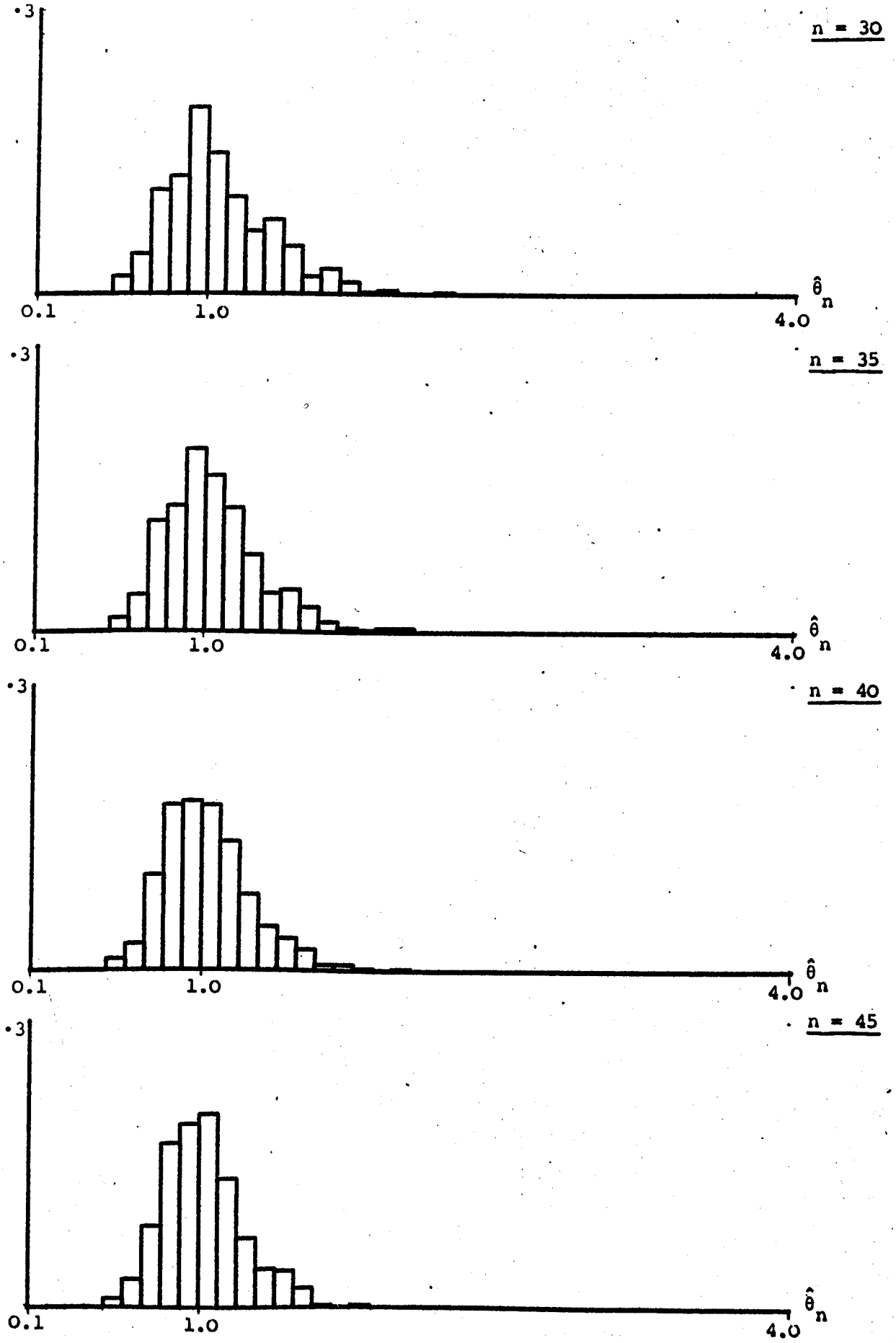


Fig. 6.2.1.b.

		$\theta = 1.0$		$\theta = 1.2639$	
	n	$\bar{\theta}_{ML}$	S^2_{ML}	$\bar{\theta}_{ML}$	S^2_{ML}
Type 1.	10	1.1901	.3229	1.3892	.3814
	15	1.1429	.2377	1.3576	.2895
	20	1.0906	.1402	1.3362	.2022
	25	1.0739	.0970	1.3010	.1325
	30	1.0524	.0745	1.2972	.1177
	35	1.0377	.0623	1.3026	.0917
	40	1.0303	.0523	1.3032	.0763
	45	1.0222	.0433	1.2964	.0661
Type 2.	10	1.1241	.2785	1.3704	.3379
	15	1.1054	.1807	1.3445	.2347
	20	1.0781	.1199	1.3239	.1679
	25	1.0630	.0843	1.3073	.1189
	30	1.0427	.0660	1.2915	.0980
	35	1.0276	.0549	1.2881	.0819
	40	1.0230	.0508	1.2826	.0653
	45	1.0146	.0423	1.2811	.0591
Type 3.	10	1.1390	.2977	1.4007	.3794
	15	1.1074	.1781	1.3504	.2408
	20	1.0719	.1216	1.3258	.1650
	25	1.0627	.0945	1.3032	.1204
	30	1.0469	.0753	1.2943	.1028
	35	1.0366	.0557	1.2900	.0859
	40	1.0251	.0456	1.2851	.0693
	45	1.0207	.0407	1.2789	.0613

Estimated means and variances of $\hat{\theta}_{n,ML}$

Fig. 6.2.2.

$\theta = 1.0$

	n	$\bar{\theta}_B$	S_B^2
Type 2	5	1.9010	.8107
	10	1.4333	.4252
	15	1.2940	.2537
	20	1.2085	.1571
	25	1.1611	.1030
	30	1.1275	.0773
	35	1.0953	.0634
	40	1.0769	.0561
	45	1.0606	.0472

$\theta = 1.2639$

	n	$\bar{\theta}_B$	S_B^2
Type 2	5	2.0844	.8153
	10	1.7241	.4763
	15	1.5583	.3162
	20	1.4694	.2063
	25	1.4292	.1494
	30	1.3935	.1162
	35	1.3715	.0966
	40	1.3518	.0732
	45	1.3416	.0667

Estimated means and variances of $\hat{\theta}_{n,B}$

Fig. 6.2.3.

$\theta = 1.0$

n	Type 1	Type 2	Type 3
10	34.9297	.4981	.5714
15	.7473	.2330	.2980
20	.1856	.1348	.2282
25	.1069	.0974	.1281
30	.0827	.0745	.0827
35	.0687	.0608	.0646
40	.0590	.0558	.0526
45	.0507	.0475	.0453

$\theta = 1.2639$

n	Type 1	Type 2	Type 3
10	7.9649	.7535	1.1719
15	.7781	.3210	.4669
20	.3424	.2296	.2409
25	.2018	.1719	.1780
30	.1371	.1334	.1418
35	.1047	.1022	.1159
40	.0850	.0797	.0930
45	.0747	.0691	.0777

Estimates of $\underset{y_n}{E}\left\{\left[\frac{\partial \ell_{y_n}(\theta)}{\partial\theta}\right]^2 \Big/ \left[\frac{\partial^2 \ell_{y_n}(\theta)}{\partial\theta^2}\right]^2\right\}$

Fig. 6.2.4.

$\theta = 1.0$

n	Type 1	Type 2	Type 3
10	.3584	.2254	.2326
15	.1585	.1337	.1364
20	.1046	.0948	.0961
25	.0782	.0732	.0736
30	.0629	.0598	.0597
35	.0525	.0505	.0502
40	.0451	.0437	.0434
45	.0395	.0384	.0382

$\theta = 1.2639$

n	Type 1	Type 2	Type 3
10	.5287	.3456	.3453
15	.2537	.2086	.2109
20	.1685	.1491	.1488
25	.1266	.1157	.1152
30	.1009	.0945	.0941
35	.0837	.0798	.0793
40	.0716	.0689	.0685
45	.0628	.0606	.0604

Estimates of $\underset{\underline{y}_n}{\mathbb{E}} \left\{ \frac{1}{S(\underline{x}_n, \underline{y}_n, \theta)} \right\}$

Fig. 6.2.5.

Number of steps until absorption of $S(\underline{x}_n, \underline{y}_n, \theta)$ by 35.0

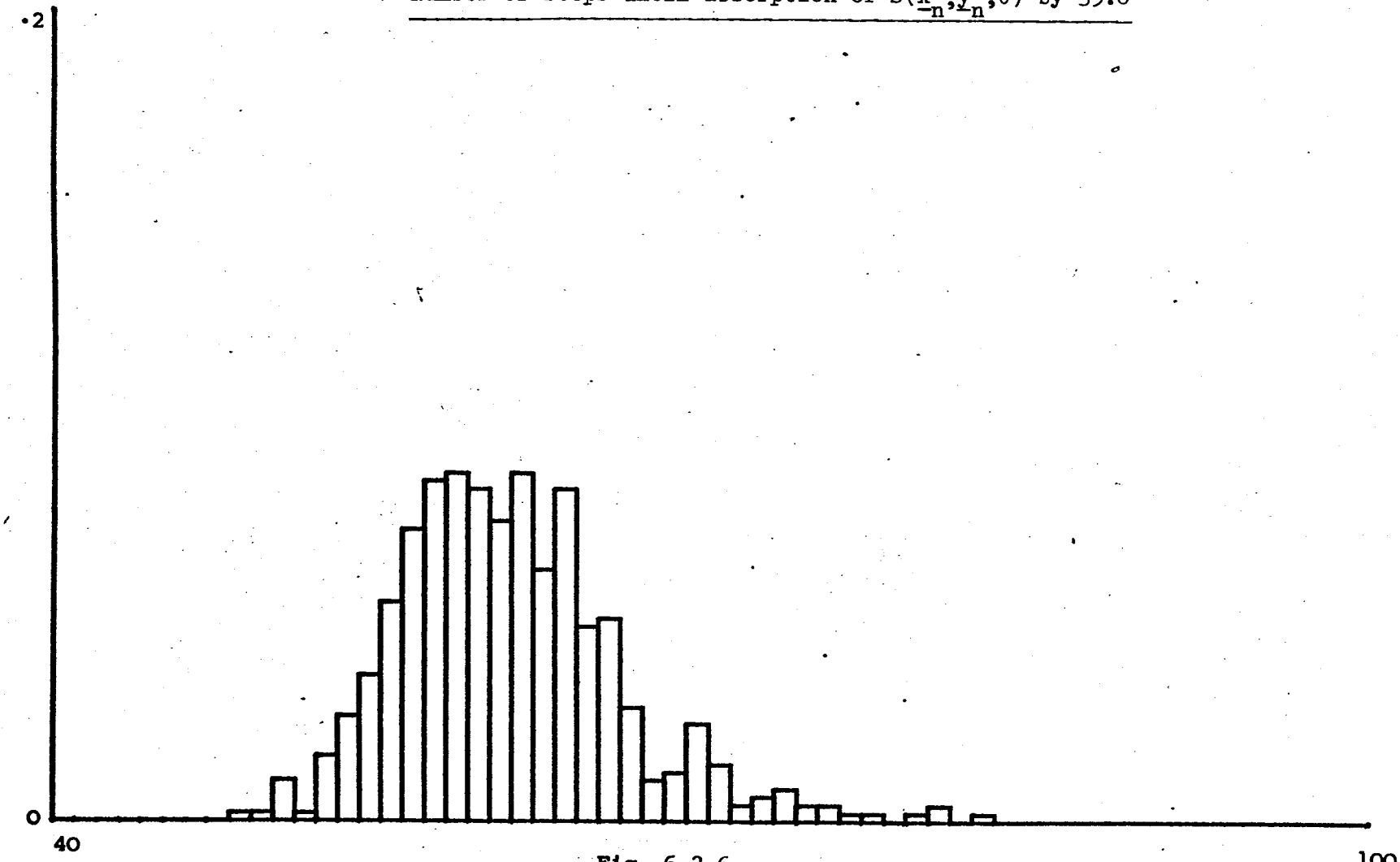


Fig. 6.2.6.

$\theta = 1.0$

Type 1		Type 2		Type 3	
$\bar{n}$	S_n^2	$\bar{n}$	S_n^2	$\bar{n}$	S_n^2
60.40	26.33	59.48	18.76	58.95	24.29

$\theta = 1.2639$

Type 1		Type 2		Type 3	
$\bar{n}$	S_n^2	$\bar{n}$	S_n^2	$\bar{n}$	S_n^2
92.95	38.57	91.76	27.02	91.39	35.98

Estimated means and variances of number of steps to absorption of $S(\underline{x}_n, \underline{y}_n, \theta)$

Fig. 6.2.7.

6.3 Comments on the results of the simulation study are listed below, followed by a general discussion of the results with regard to comparison of performances of the different process types.

Fig. 6.2.1.

The histograms illustrate the rather erratic estimation for small n, with gradual convergence towards a more normal-like distribution of estimates, centred on the true parameter value, as n becomes larger.

Fig. 6.2.2.

The Types 2 and 3 processes appear to be performing almost uniformly better than the Type 1 process with regards to bias and variance of the maximum likelihood estimates. Types 2 and 3 would seem to be a little more difficult to separate, for example at n = 45. Type 2 seems to be less biased but have larger variance for $\theta = 1.0$, the roles being reversed for $\theta = 1.2639$.

Fig. 6.2.3.

The data in this table would seem to indicate that there is substantial bias of $\hat{\theta}_B$ away from the true parameter values, particularly in the early stages of the process.

Fig. 6.2.4.

Comparison of the estimates of

$$\mathop{E}_{\underline{y}_n}\left\{\left[\frac{\partial \ell_{\underline{y}_n}(\theta)}{\partial\theta}\right]^2 \Big/ \left[\frac{\partial^2 \ell_{\underline{y}_n}(\theta)}{\partial\theta^2}\right]^2\right\},$$

in this table, with the S^2_{ML} values in Fig.6.2.2. indicates fairly wild differences, for small n, with the Type 1 process. However, even after 45 observations, the values are not particularly comparable, for any process type, notably for $\theta = 1.2639$.

Fig. 6.2.5.

Again comparison of estimates of

$$\mathop{E}_{\underline{y}_n}\left\{\frac{1}{S(\underline{x}_n, \underline{y}_n, \theta)}\right\}$$

would seem to indicate that Types 2 and 3 are performing uniformly better than the Type 1 process, and again Types 2 and 3 would seem to be difficult to separate.

Fig. 6.2.7.

Again Types 2 and 3 would seem to be dominant over the Type 1 process. Comparing Types 2 and 3, the Type 3 process would appear to have the smaller average number of steps to absorption but the larger variance.

Although the differences between the statistics generated by the simulation study for the three process types are not dramatic, there is one fact which is evident, that being that the Type 2 and 3 processes appear to dominate the Type 1 process almost uniformly over the different comparison methods. This, allied with the fact that Fig. 6.2.3. would appear to indicate that the improper uniform prior assumed is producing a bias away from the true parameter value, would seem to provide empirical backing for the conjectures made in Chapter 5, despite the fact that evidence has been presented only for two parameter values in one probability model. In conclusion, rather than claim that the conjectures of Chapter 5 have been completely vindicated, it would seem reasonable, from both the intuitive and empirical evidence above, to advise strongly that alternative procedures to those of the Type 1 class should be investigated in any practical situation, if possible. In situations where this type of sequential experiment is applicable, that is, where the probability model is assumed to be known, it would seem highly probable that some form of real prior knowledge about θ would be present, and this being the case it would seem to be essential that this knowledge should be utilised in any sequential experiment.

It should be noted that, although computing time was comparable for the three process types in the above one parameter model, the amount of computing necessary for the Type 2 and 3 processes will quickly become restrictive as the dimension of the parameter vector increases. In higher dimensions it may be possible to maintain some of the flavour of the Type 2 and Type 3 processes, but at the same time ease computing difficulties, if a posterior mode type process is used as suggested in 5.2..

CHAPTER 7

SEMI-SEQUENTIAL EXPERIMENTATION

Up to this point we have considered only fully sequential design procedures. Often experimental conditions and economic considerations will enforce the need for a less time consuming process. For example, a real problem was encountered in which observations were to be allocated to two treatments. The optimal allocation rule was a function of a set of unknown parameters. Unfortunately, early hopes that it would be possible to apply some of the preceding ideas to a real problem were dashed when it was revealed that each observation took three weeks to process, thus making a fully sequential design impractical. In fact only a two stage experiment was possible. Half of the observations were allocated evenly in stage one and, based on the results obtained, the rest were allocated in the second stage.

The aim of this chapter will be to investigate circumstances in which less than fully sequential designs will be expected to be quite satisfactory in practice, and to consider how the fully sequential processes of the preceding chapters might be adapted to allow for the taking of batches of observations at each stage. Initially a mélange of problems will be considered with a view to possible deviation from fully sequential procedures without definite restrictions on the nature in which these must take place. In 7.5. the more definite problem of design with fixed batch size will be investigated.

7.1. It will be useful, at this point, to consider a partitioning of the type of design problem which might be encountered. The partitioning process will be carried out according to the nature in which the optimal static design measure, for a given problem, depends on the vector of unknown parameters $\underline{\theta}$. Only models where a sequential design procedure would be expected to be necessary will be considered (that is where the Fisher information matrix is a function of $\underline{\theta}$).

Four categories of problems (P_1, P_2, P_3, P_4) will be used. For each category an example will be given and suggestions made as to how the sequential design procedure may be curtailed due to the nature of the problem.

P_1 : Both the spectrum of the optimal design and the optimal design weights are independent of $\underline{\theta}$

Example. In Example 2. of 2.6.1. a problem was analysed where observational units had to be allocated to three populations which would produce observations with probability distributions (1) $Po(\theta_1)$, (2) $Po(\theta_2)$, (3) $Po(\theta_1 + \theta_2)$. It was shown that the observational units should be allocated equally to populations (1) and (2) independently of (θ_1, θ_2), if D-optimality were the criterion. Therefore, even though the Fisher information matrix for each population is (θ_1, θ_2) dependent, the D-optimal design measure is not, thereby removing the need for a sequential experiment. This situation would not be expected to occur often in practice, but it is interesting to note that it may occur.

P_2 : The optimal design spectrum is independent of $\underline{\theta}$ but the optimal weights are not.

Examples: Problems which fall into this category are the comparison of means example of 2.6.2. and the D_2 optimal design problem in the calibration example of 2.6.4.

In the above examples the spectrum of the optimal design consists of only two points, which are independent of $\underline{\theta}$. This suggests that in a sequential experiment the design space should be reduced to these two points. After stage n an observation should be taken to make $(p_{1,n+1}, p_{2,n+1})$ as close as possible to $(p_1^*(\hat{\underline{\theta}}_n), p_2^*(\hat{\underline{\theta}}_n))$, where $(p_1^*(\hat{\underline{\theta}}_n), p_2^*(\hat{\underline{\theta}}_n))$ are the optimal allocations given $\underline{\theta} = \hat{\underline{\theta}}_n$ and $(p_{1,n+1}, p_{2,n+1})$ are the allocations which would actually be in use at stage (n+1).

It should be noted that, in the above type of process, the instability of the sequential process for small n will largely be removed. Because the set of design points which may be used has been reduced to a small set of points which are known to be informative for all $\underline{\theta}$, the possibility of the design procedure moving erratically across the design space has been removed. Also, the possibility of achieving a design measure close to the optimum design measure, for reasonable sized N, will be greatly increased because all of the observations are being taken at points in the optimal design spectrum. Because of this, inferences subsequent to such an experiment will be made with more confidence that approximations being made are good, and also possibly with less computation. One would expect the approximation of Fedorov and White to var $(\hat{\underline{\theta}}_N)$ to be more satisfactory and that the sample information matrix would be well approximated by a suitable Fisher information matrix as the number of observations at each design point will typically be large and a law of large numbers may be invoked.

P_3 : <u>The optimal design is dependent on $\underline{\theta}$ but its dimension and the optimal design weights are not</u>

<u>Example</u>: Examples which fall into this category are the quantal response types of problem of <u>2.1.5.</u>. It was shown that the optimal design measure was of the form

$$\begin{pmatrix} \frac{1}{2} & , & \frac{1}{2} \\ \frac{a-\theta_1}{\theta_2} & , & \frac{-a-\theta_1}{\theta_2} \end{pmatrix} ,$$

if these design points lie in $\mathcal{X}$.

In many practical situations it would be expected that a design procedure should, by necessity, be simple to carry out with the minimum of computation. This example suggests a possibly useful method of carrying out some sequential procedures, particularly if a reasonable

number of observations are to be taken at each stage of the procedure and design is really only feasible between stages. The procedure suggested is to take observations at stage n according to the design measure

$$\left(\begin{array}{cc} \frac{1}{2} & \frac{1}{2} \\ \frac{a-\hat{\theta}_{1,n}}{\hat{\theta}_{2,n}} & \frac{-a-\hat{\theta}_{1,n}}{\hat{\theta}_{2,n}} \end{array} \right) .$$

Again, the fact that a reasonable number of observations are being taken at each design point might make approximations more satisfactory .

P_4: Both the optimal design spectrum and the optimal design weights are dependent on $\underline{\theta}$.

In situations such as this a fully sequential procedure would seem to be indicated, if possible. However, it may be possible that the position can be improved upon in situations where there exists some form of prior knowledge about the possible values that the unknown parameters might take. In Chapter 5 it was suggested how this knowledge might be utilised in fully sequential procedures. Now we consider how it might be used to suggest suitable semi-sequential procedures. It will be useful, at this point, to consider an example.

7.2. Let us reconsider the quantal response example of 2.1.5. in the particular case where the model is logistic. That is,

$$p(1|x,\underline{\theta}) = \frac{\exp(\theta_1 + \theta_2 x)}{1 + \exp(\theta_1 + \theta_2 x)} \quad , \quad p(0|x,\underline{\theta}) = \frac{1}{1 + \exp(\theta_1 + \theta_2 x)}$$

The design space $\mathcal{X}$ will be taken to be $[-1, +1]$.

$$I(x,\underline{\theta}) = \frac{\exp(\theta_1+\theta_2 x)}{\{1+\exp(\theta_1+\theta_2 x)\}^2}\begin{bmatrix} 1 & x \\ x & x^2 \end{bmatrix} = V(x,\underline{\theta}).\begin{bmatrix} 1 & x \\ x & x^2 \end{bmatrix}$$

As was noted in 2.1.5., the D-optimal design measure for estimating $\underline{\theta} = (\theta_1,\theta_2)^T$ can be shown to be

$$\begin{pmatrix} \frac{1}{2} & , & \frac{1}{2} \\ \frac{a-\theta_1}{\theta_2} & , & \frac{-a-\theta_1}{\theta_2} \end{pmatrix},$$

where $a \doteq 1{\cdot}5434$, if $\frac{a-\theta_1}{\theta_2}$ and $\frac{-a-\theta_1}{\theta_2}$ both lie in $[-1,+1]$.

It was thought to be of interest to investigate what the optimum design measure would be if the above optimum design points were not contained in $[-1, +1]$. To this end the optimal design was computed for a large number of points in a scan of $\underline{\theta}$ space, using an iterative algorithm. What was found was that, in each case, the D-optimal design measure existed at two points. It is therefore of interest to investigate the D-optimal two point design for this model as a function of $\underline{\theta}$. That is, to find (x_1, x_2) to maximise

$$|M(\xi)| = \tfrac{1}{4}\, V(x_1, \underline{\theta}).V(x_2,\underline{\theta}).(x_1-x_2)^2 , \quad x_1,x_2 \in [-1, +1].$$

Let the optimum design for (θ_1,θ_2) be (x_1^*,x_2^*). Use of the symmetry of $V(x,\underline{\theta})$ with respect to $\underline{\theta}$ $(V(x,\underline{\theta}) = V(x,-\underline{\theta}))$ and the symmetry of the design space about the origin reveals the following, (x_1^*,x_2^*) is the D-optimal 2-point design for $(-\theta_1, -\theta_2)$, $(-x_1^*,-x_2^*)$ is the D-optimal 2-point design for $(\theta_1,-\theta_2)$.

Therefore, essentially it is only necessary to consider the problem for $\theta_1,\theta_2 \geq 0$.

By differentiation of $|M(\xi)|$ above it can be shown that a global maximum is attained when $2\theta_1 + \theta_2(x_1^* + x_2^*) = 0$, the x_i^*'s being given by the two solutions of the equation

$$\exp(\theta_1 + \theta_2 x^*) = - \frac{1 + (\theta_1 + \theta_2 x^*)}{1 - (\theta_1 + \theta_2 x^*)} .$$

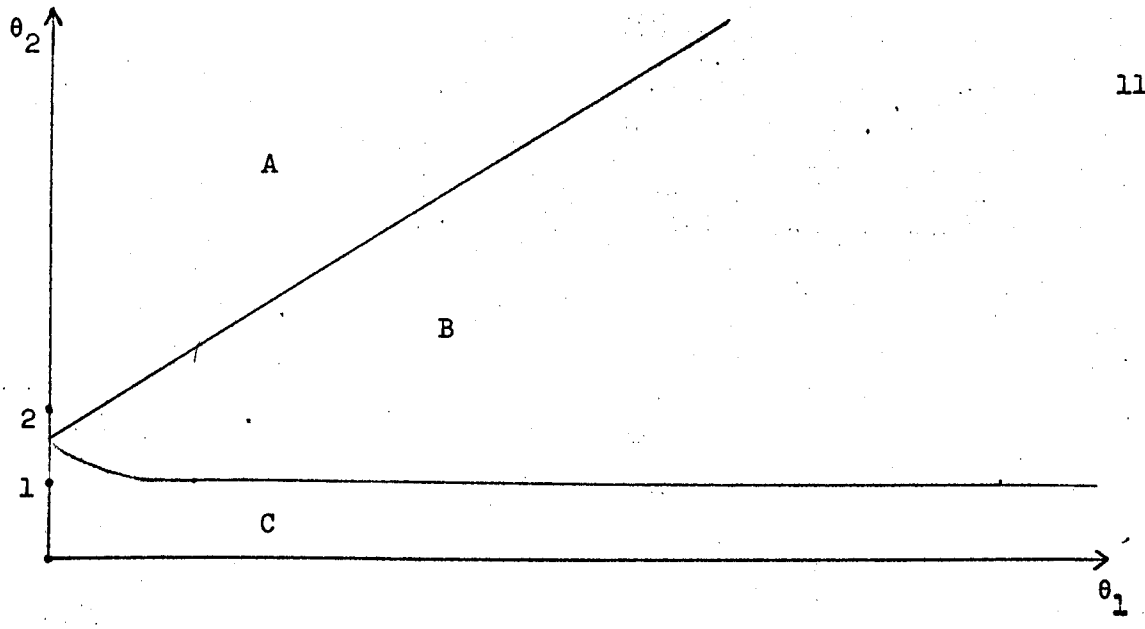
That is, where $\theta_1 + \theta_2 x^* = \pm a$, a being the solution to $\exp(a) = \frac{a+1}{a-1}$ giving $a \doteqdot 1.5434$.

If $x_1^*, x_2^* \in [-1, +1]$ then this is the optimal two point design. If not, then the solution must occur on the boundary of the permissible region. For $\theta_1, \theta_2 \geq 0$ it can be shown that one point must then be -1 and the second is given by the solution to

$$\exp(\theta_1 + \theta_2 x) = \frac{2 + (x+1)\theta_2}{-2 + (x+1)\theta_2} .$$

If the solution to the above does not lie in $[-1, +1]$ then the optimal two point design is given by $x_1^* = -1$, $x_2^* = +1$.

These optimum 2-point designs are identical to the continuous D-optimum designs evaluated using numerical methods. The above analysis of the problem suggests considering how the optimum static design varies as a function of $\underline{\theta}$. Fig.7.2.1. shows how the parameter space may be broken down into three regions according to whether the optimum design falls into one of the three categories described above. Again it is only necessary to consider $\theta_1, \theta_2 \geq 0$ as the other quadrants may be obtained by the symmetries previously mentioned.



Region A: $\theta_2 - \theta_1 \geq 1.54$, Region B: upper boundary, $\theta_2 - \theta_1 = 1.54$

lower boundary, $\exp(\theta_1 + \theta_2) = \frac{1+\theta_2}{-1+\theta_2}$

Fig. 7.2.1.

A, B and C correspond to the three regions described above. If the true parameter value lies in A then the optimal static design is given by

$$\begin{pmatrix} \frac{1}{2} & , & \frac{1}{2} \\ \frac{a-\theta_1}{\theta_2} & , & \frac{-a-\theta_1}{\theta_2} \end{pmatrix} .$$

If the true parameter value lies in B then the optimal static design is given by

$$\begin{pmatrix} \frac{1}{2} & , & \frac{1}{2} \\ -1 & , & x \end{pmatrix} ,$$

where x is the solution of,

$$\exp(\theta_1 + \theta_2 x) = \frac{2 + (x+1)\theta_2}{-2 + (x+1)\theta_2} .$$

If the true parameter value lies in C then the optimal static design is given by

$$\begin{pmatrix} \frac{1}{2} & , & \frac{1}{2} \\ -1 & , & +1 \end{pmatrix} \quad .$$

This example suggests possible alternatives to fully sequential experimentation in situations where prior information about parameter values is available and time spent on experimentation is an important factor.

In Chapter 5 it was suggested that prior information might be utilised by constructing a prior distribution on $\underline{\theta}$ space. Possibly an easier thing for an experimenter to do would be to express his prior information via a region in $\underline{\theta}$ space in which he was 'certain' that the true parameter values lay. If the experimenter was certain that the true parameter value lay in C then a sequential experiment would be unnecessary, the optimal static design being independent of $\underline{\theta}$, given $\underline{\theta}$ in C. If he were certain that the true value lay in C or B then a semi-sequential experiment is suggested, taking half of the available observations at $x = -1$, and using the information obtained to design sequentially with the other half of the observational units. If the experimenter can not say, a priori, in which region the true parameter values lie then a fully sequential procedure might be embarked upon, with the possibility of changing over to a semi-sequential procedure if the a posteriori probabilities of regions B and/or C became suitably high.

7.3. Reconsider the example used in the simulation study of Chapter 6.

The design space was taken to be $[\cdot 5, 30]$, and it was noted that the optimal static design would take all the observations at one point, that being $x^* = \frac{1.59}{\theta}$. Let us assume that prior information can be produced in the interval form discussed in 7.2., and that, in this case, an interval in which the experimenter is certain that the true parameter value lies is $[1.59, 3.18]$ say. This is equivalent to saying that he is confident that $x^* \in [\frac{1}{2}, 1]$. We shall call this region the informative design space and denote it by $\mathcal{X}'$. A sensible

plan might be to use not $\mathcal{X}$ as design space but $(\mathcal{X} \cap \mathcal{X}')$, that is, in this example $[\cdot 5, 1]$. This action might prove useful in two ways. Firstly it may avoid the possibly more awkward construction of a prior distribution and yet might utilise the prior information in increasing the stability of a sequential procedure by restricting the design space to more informative points. It should be noted that the improvement in design gained by this strategy will be analogous to that gained in the examples of P_2 in 7.1., although in 7.1 no prior knowledge of $\underline{\theta}$ was required. It should also be noted that this process of reducing the dimension of $\mathcal{X}$ may serve to make the design more robust to outlying observations.

7.4. In this chapter we have considered, rather informally, possible situations where designs which are less than fully sequential may be satisfactory. In considering a practical problem this informal approach would seem to be essential as each particular problem is likely to have its own peculiarities. However, the above sections would seem to indicate a general structure for investigation of a new design problem which might be built upon in the particular situation on hand. This might be summarised as follows.

(1) Investigate the optimal static design measure as a function of $\underline{\theta}$, if possible.

(2) Attempt, with the aid of the experimenter, to obtain some form of prior assessment as to what the true parameter values might be.

(3) Investigate the possibility of less than fully sequential procedures suggested by (1). Is the prior information of any help?

(4) If a fully sequential design seems to be necessary then prior information may be useful. Regular re-assessment of the possibility of semi-sequential procedures may also be useful. It may be possible, with the aid of prior knowledge, to reduce the volume of the design space in which the experiment is allowed to take place.

7.5. In the above we have considered, generally, methods of 'cutting corners' in sequential procedures, in order to speed up the progress of an experiment. Often it is to be expected that circumstances will dictate the nature in which a design procedure must be less then fully sequential. Particularly in industrial settings it will be necessary to experiment by taking large batches at each stage, and it is to this problem which we turn our attention now.

Let us assume that we have a total of $N = rn$ observational units at our disposal, where r is the number of stages available and n is the number of observational units in each batch. We take a likelihood principle approach to design. Consider the situation after stage $i < r$. We have $\hat{\underline{\theta}}_i$ as estimate of the parameters. n design points are to be selected, at which to take observations in the $(i+1)$st stage. Echoing the methods of Chapter 5 the natural criterion for design would seem to be to select $\underline{x}_{i+1} = (x_{i+1,1}, \ldots, x_{i+1,n})^T$ to maximise

$$\int_{\underline{y}_{i+1}} \{\phi\{\sum_{t=1}^{i} S(\underline{x}_t, \underline{y}_t, \hat{\underline{\theta}}_i) + S(\underline{x}_{i+1}, \underline{y}_{i+1}, \hat{\underline{\theta}}_i)\}\} ,$$

with the obvious notation. As was mentioned in Chapter 5 this is likely to be an awkward problem to solve exactly, even when $\underline{x}_{i+1}$ is scalar. However, for large n, the alternative procedure suggested in Chapter 5 may be useful; the alternative procedure being to select $\underline{x}_{i+1}$ to maximise an upper bound for the above criterion, that is

$$\phi\{\sum_{t=1}^{i} S(\underline{x}_t, \underline{y}_t, \hat{\underline{\theta}}_i) + \sum_{j=1}^{n} I(x_{i+1,j}, \hat{\underline{\theta}}_i)\}$$

$$= \phi\{S + n\, M(\xi_n)\}, \quad S = \sum_{t=1}^{i} S(\underline{x}_t, \underline{y}_t, \hat{\underline{\theta}}_i) ,$$

$$M(\xi_n) = \frac{1}{n} \cdot \sum_{j=1}^{n} I(x_{i+1,j}, \hat{\underline{\theta}}_i) .$$

This suggests solving the following continuous design problem.

Find ξ to maximise $\phi\{ S + n.M(\xi) \}$.

This problem is analogous to the type of problem considered in Chapter 2,as it may readily be seen that the positive definite matrix S will have no complicating effect on the general structure of the problem. If this continuous problem has a solution with a design spectrum of dimension small relative to n then the optimal design measure may be approximated to using the methods of 3.7.1. If a reasonable number of observations is being taken at each design point then a law of large numbers will ensure that we can get close to the upper bound given above.

CHAPTER 8

EPILOGUE

In Chapters 1 to 3 we have presented the theory and assumptions underlying the optimal experimental design problem and have reviewed methods of solution of this problem. In Chapters 4 to 7 we have investigated possible sequential design procedures which might prove useful when formulation of an a priori optimal design is impossible because the optimal static design is a function of a vector of unknown parameters.

All statistical methods are founded on certain assumptions. As Statistics is essentially a practical subject the strength of any of its methods, in application, must lie not only in their theoretical background, but also in the practising statistician's appreciation of their dependence on given assumptions and how they will perform when these assumptions no longer hold. With this in mind we first consider some of the assumptions of the preceding chapters and briefly indicate some of the work which has been done on experimental designs for situations when these assumptions no longer hold.

To round off this thesis suggestions for directions of further research will be made.

8.1.

Undoubtedly the most crucial assumption which has been made, in the theory and methods which have been discussed so far, is that there is a known model. Possible justifications for this are twofold. Firstly, sometimes theoretical considerations in the situation on hand will enable derivation of a mathematical model which is known up to a vector of unknown fundamental constants. In this situation interest will often lie in gaining knowledge about these constants. Alternatively, past experience in similar situations may suggest that a particular model will well describe the results of the experiment to be carried out.

In this situation experimentation will typically be for the purpose of prediction or some related reason.

If, as will often be the case, the exact form of the model is unknown, optimal experimentation of the nature considered in Chapters 1 to 3 will be, not only often impossible, but also extremely unwise. The following example will illustrate this point.

Suppose the following assumptions hold.

$$p(y|x,\alpha,\beta) \sim N(\alpha + \beta\phi(x),\sigma^2),\ \sigma^2 \text{ assumed known},\ \mathcal{X} \in [a,b] .$$

Also suppose that ϕ is a function known only in the sense that it has one of the two following properties.

$$(1)\quad \sup_{x\in\mathcal{X}} \phi(x) = \phi(a) \quad , \quad \inf_{x\in\mathcal{X}} \phi(x) = \phi(b) .$$

$$(2)\quad \sup_{x\in\mathcal{X}} \phi(x) = \phi(b) \quad , \quad \inf_{x\in\mathcal{X}} \phi(x) = \phi(a) .$$

From the geometric approach of 2.5. it will be obvious that the D-optimal design for estimating (α,β) will be to allocate observations evenly at a and b, for all ϕ with one of the above two properties.

If the above design were to be used, and if, independently of the experiment being carried out, one were to discover the true function ϕ then one would have carried out a D-optimal experiment. However, as must be the case in practice, the experiment itself must attempt to discriminate between the models available, for which purpose the above design will have zero power.

The above example illustrates a very important point, that being that optimal designs of the type investigated in Chapters 1 to 3

may often be useless for detecting departures from the assumed model, even within a set of models having individually the same optimal designs with respect to a given criterion.

Ironically, it is in situations where one feels that the strongest design is being applied (that is, where an a priori optimal static design can be found) that one is in the weakest possible position for detecting deviations from the assumed model. In this sense perhaps the application of the tag 'optimal' to the designs of Chapters 1 to 3 is a little presumptuous. In the sequential designs suggested in Chapters 4 to 7 it would be possible to sequentially reassess model assumptions in the light of data obtained and adjust design accordingly.

With regards to the problem of discrimination between a finite set of possible models notable work has been done by Hunter and Reiner (1965), G.E.P. Box and Hill (1967), Atkinson and Cox (1974) and Atkinson and Fedorov (1975a, 1975b).

8.2.

In 8.1. we considered briefly possible problems which might arise when the probability model underlying the observations is unknown. Sometimes theoretical considerations may take us only part of the way to knowing this model fully. For example, in regression type situations giving rise to observations with the following distribution $p(y|\underline{x},\underline{\theta}) \sim N(\eta(\underline{\theta},\underline{x}), v(y|\underline{x}))$, it may be that the form of $\eta(\underline{\theta},\underline{x})$ is given by theory but the nature of the error in the observations $v(y|\underline{x})$, as a function of $\underline{x}$, is unknown. Remembering that points in the induced design space V may be written in the form $\frac{1}{\sqrt{v(y|\underline{x})}} \cdot \eta_{\underline{\theta}}(\underline{\theta},\underline{x})$, $\underline{x}\in\mathcal{X}$, it may be appreciated that the shape of the induced design space could be moulded to almost any form by suitable adjustment of $v(y|\underline{x})$. It would appear that little work has been done on the design of experiments where the nature of $v(y|\underline{x})$ is unknown with the possible exception of Box, M.J. and Draper (1968). From the above it would seem that the optimal design will be very sensitive to changes in $v(y|\underline{x})$. In situations where bounds could be put on $v(y|\underline{x})$

minimax type designs might be applicable, although these are likely to be conservative in nature. Generally a sequentially designed experiment would seem to be the only alternative.

Of course, even the assumption of the normal distribution of error in the above regression situation may be unjustified. For regression situations other authors (for example Fedorov) have restricted themselves to non-probabilistic methods of estimation, such as least-squares. In the context of estimation it must be said that all that the assumption of a probability model is doing is to define a loss function which might not be suitable for the true model, and in this sense any fixed non-probabilistic method such as least squares is doing no better.

8.3.

Given an experimental design criterion, it has been shown that there exist fairly powerful methods for computing optimal design measures where this is possible prior to an experiment, and a number of sensible sequential design procedures when this is not the case. In practice, as was suggested in 1.3., decision on a suitable criterion for design may be difficult. This problem will depend on the experimenter's ability to express his wishes clearly, and the statistician's dexterity in translating these wishes into mathematical form.

In general there can be no solution to the above problem. However, if the experimenter's wishes appear to be rather vague then a sensible approach would be to select a set of criteria which seem to be vaguely suitable. Having selected this set, the optimal design could be computed for each criterion and performance of that optimal design investigated with regard to the other criteria. In this way a design might be obtained which is robust against the uncertain experimenter.

8.4.

With regards to the solution of a continuous optimal design problem, with concave criterion, it would seem that the theory and

methods are in a fairly healthy condition. There still remain many problems of analytic derivation of optimal designs and possible problems which might arise in pathological situations, for example where the criterion may possibly be non differentiable. However, possibly these are problems more for the purist than for the practising statistician merely interested in producing a design and not in the nicest way of achieving that end. With reasonable sample size it has been shown in Chapter 3 that it will be possible to approximate well to a continuous optimal design using an exact design. The optimal exact design problem per se will surely always remain a difficult problem to solve, with explicit solutions only being possible in simple or symmetric situations.

Where an a priori design is impossible to compute, we have taken the attitude that every design problem is likely to have its own peculiarities and therefore its own peculiar method of solution. In this thesis we have considered a number of methods of experimentation in a variety of situations, and it would seem likely that, however difficult a design problem might appear, there will always be something sensible which might be done as an alternative to a purely random design.

An area of possible application which has not been touched upon in this thesis is in the field of Control Theory. In this field all of the problems which have been encountered in the above are present, and in most situations seem to be magnified, even the derivation of information matrices being a formidable task. This would seem to be an area where further work might be done. With regards to the general background to the Control type problem the review paper of Astrom and Eykhoff (1971) would appear to be useful, with fairly recent reviews of work done on the Control design problem being contained in Mehra (1974a, 1974b) and Keviczky (1975). There would appear to be a certain degree of inconsistency in the above papers. However, a substantial amount of background work will be required by the writer before any authoritative argument can be put forward on relevant points.

APPENDIX 1

The following lemma will be useful in the appendices which follow. The proof is given for completeness.

Lemma A.1.

Inverse is a convex operation on the positive definite symmetric matrices. That is, given A,B positive definite symmetric then,

$$\{\alpha A + (1-\alpha)B\}^{-1} \leq \alpha A^{-1} + (1-\alpha)B^{-1}, \alpha \in [0,1] .$$

The inequality above indicates that the difference of the two sides (RHS - LHS) is positive semi-definite. The inequality may be replaced by an equality only if A = B.

Proof:

A and B are positive definite symmetric matrices and therefore may be written in the following form (Graybill),

$$A = PP^T \quad , \quad P \text{ is of full rank.}$$
$$B = P \Lambda P^T \quad , \quad \Lambda \text{ is diagonal } \{\lambda_i\} \quad , \quad \lambda_i > 0, \ \forall i .$$

$$\Rightarrow \quad A^{-1} = RR^T \quad , \quad B^{-1} = R \Lambda^{-1} R^T \quad , \quad R^T = P^{-1} \quad ,$$

$$\{\alpha A + (1-\alpha)B\} = P(\alpha I + (1-\alpha)\Lambda)P^T \quad ,$$

$$\{\alpha A + (1-\alpha)B\}^{-1} = R(\text{diag}\{\frac{1}{\alpha+(1-\alpha)\lambda_i}\})R^T \quad ,$$

$$\alpha A^{-1} + (1-\alpha)B^{-1} = R(\text{diag}\{\alpha + \frac{(1-\alpha)}{\lambda_i}\})R^T \quad ,$$

$$\Rightarrow \quad \alpha A^{-1} + (1-\alpha)B^{-1} - \{\alpha A + (1-\alpha)B\}^{-1} = R(\text{diag}\{\alpha + \frac{(1-\alpha)}{\lambda_i} - \frac{1}{\alpha+(1-\alpha)\lambda_i}\})R^T$$

$$= R(\text{diag}\{\mu_i\})R^T ,$$

$$\mu_i = \frac{(\alpha + (1-\alpha)\lambda_i)(\alpha\lambda_i + (1-\alpha)) - \lambda_i}{\lambda_i(\alpha + (1-\alpha)\lambda_i)}$$

$$= \frac{\alpha(1-\alpha)(1-\lambda_i)^2}{\lambda_i(\alpha + (1-\alpha)\lambda_i)}$$

This gives the required result as μ_i will be non-negative for all i and non-zero for some i unless A = B.

APPENDIX 2

Let $\phi_7 \equiv -$ m.d.e(M^{-1}) , M a positive definite symmetric matrix.

(i) $\phi_7 (M_1 + M_2) \geq \phi_7 (M_1)$, M_1 a positive definite symmetric matrix and M_2 a positive semi-definite symmetric matrix.

(ii) ϕ_7 is a concave function on the positive definite symmetric matrices. That is,

$$\phi_7 (\alpha M_1 + (1-\alpha)M_3) \geq \alpha \phi_7(M_1) + (1-\alpha)\phi_7(M_3), \alpha \in [0,1] ,$$

M_3 positive definite symmetric.

Proof.

(i) It is well known (see Graybill) that, for M and N positive definite symmetric, if $M \geq N$ then $M^{-1} \leq N^{-1}$.

Let $N = M_1$, $M = M_1 + M_2$

$$\Rightarrow M_1^{-1} \geq (M_1 + M_2)^{-1}$$

$$\Rightarrow \underline{x}^T M_1^{-1} \underline{x} \geq \underline{x}^T (M_1 + M_2)^{-1} \underline{x} \quad , \forall \underline{x} .$$

Take $\underline{x} = \begin{pmatrix}1\\0\\0\\\vdots\\0\end{pmatrix}$, $\begin{pmatrix}0\\1\\0\\\vdots\\0\end{pmatrix}$, etc..

$$\Rightarrow \text{mde}\,(M_1^{-1}) \geq \text{mde}\,((M_1 + M_2)^{-1})$$

$$\Rightarrow \phi_7 (M_1) \leq \phi_7 (M_1 + M_2) .$$

(ii) By Lemma A.1.

$$\alpha M_1^{-1} + (1-\alpha) M_3^{-1} \geq \{\alpha M_1 + (1-\alpha) M_3\}^{-1} , \alpha \in [0, 1] .$$

$$\Rightarrow \alpha\, \underline{x}^T M_1^{-1} \underline{x} + (1-\alpha)\, \underline{x}^T M_3^{-1} \underline{x} \geq \underline{x}^T \{ \alpha M_1 + (1-\alpha) M_3 \}^{-1} \underline{x} \,, \forall\, \underline{x} \,.$$

Take $\underline{x} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$, $\begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$, etc..

Let $\alpha\, M_1^{-1} + (1-\alpha) M_3^{-1} = \{ \sigma_{ij} \}$, $(\alpha M_1 + (1-\alpha) M_3)^{-1} = \{ \tau_{ij} \}$,

$M_1^{-1} = \{ a_{ij} \}$, $M_3^{-1} = \{ b_{ij} \}$.

Then $\sigma_{ii} \geq \tau_{ii}$, $\forall\, i$

$$\Rightarrow \max_i \tau_{ii} = \tau_{ss} \leq \alpha\, a_{ss} + (1-\alpha)\, b_{ss}$$

$$\leq \alpha \max_i a_{ii} + (1-\alpha) \max_i b_{ii}$$

$\Rightarrow$ mde is a convex function

$\Rightarrow$ ϕ_7 is a concave function.

APPENDIX 3

(i) $\phi_1(M) = \log |M|$ is strictly concave over the set of positive definite symmetric matrices.

(ii) $\phi_6(M) = -\text{trace}(M^{-1})$ is strictly concave over the set of positive definite symmetric matrices.

(iii) $\phi_3(M) = -\int_{\underline{x}\in\mathcal{X}'} \eta_{\underline{\theta}}(\underline{x},\underline{\theta})^T M^{-1} \eta_{\underline{\theta}}(\underline{x},\underline{\theta})\, p(\underline{x})d\underline{x}$ is strictly concave over the set of positive definite symmetric matrices if $\int_{\underline{x}\in\mathcal{X}'} \eta_{\underline{\theta}}(\underline{x},\underline{\theta}) \cdot \eta_{\underline{\theta}}(\underline{x},\underline{\theta})^T p(\underline{x})\, d\underline{x}$ is positive definite.

(iv) $\phi_2 = -\max_{\underline{x}\in\mathcal{X}} \eta_{\underline{\theta}}(\underline{x},\underline{\theta})^T M^{-1} \eta_{\underline{\theta}}(\underline{x},\underline{\theta})$ is not necessarily strictly concave.

Proof.

(i) Because log is strictly increasing it will be sufficient to show that

$$|\alpha M_1 + (1-\alpha) M_2| > |M_1|^{\alpha} |M_2|^{1-\alpha}, \quad \alpha\in[0,1], \quad M_1, M_2$$

positive definite symmetric.

As M_1, M_2 are positive definite symmetric matrices they may be written as follows

$$M_1 = R \Lambda R^T, \quad M_2 = RR^T, \quad \Lambda = \text{diag}\{\lambda_i\}.$$

Now, $|\alpha M_1 + (1-\alpha)M_2| > |M_1|^{\alpha} |M_2|^{1-\alpha}$

$\iff |R|^2|\alpha\Lambda + (1-\alpha)I| > |R|^2 |\Lambda|^{\alpha}|I|^{1-\alpha}$

$\iff \prod(\alpha\lambda_i + (1-\alpha)) > \prod \lambda_i^{\alpha} \prod 1^{1-\alpha}$

But, $\alpha\lambda_i + (1-\alpha) \geq \lambda_i^{\alpha}$, $\forall i$, $\lambda_i \neq 1$, by the Arithmetic, Geometric mean inequality.

$$\Rightarrow \quad |\alpha M_1 + (1-\alpha)M_2| > |M_1|^{\alpha} |M_2|^{1-\alpha}$$

$$\Rightarrow \quad \text{strict concavity of } \phi_1 .$$

(ii) By Lemma A.1., for M_1, M_2 positive definite symmetric matrices, we have,

$$\alpha M_1^{-1} + (1-\alpha) M_2^{-1} \geq (\alpha M_1 + (1-\alpha)M_2)^{-1} \text{ , with equality only if } M_1 = M_2.$$

The eigenvalues of (LHS - RHS) must be non-negative with at least one non zero unless $M_1 = M_2$, the strict concavity of ϕ_6 follows from the fact that the trace of a matrix is equal to the sum of its eigenvalues.

(iii) $\int_{\underline{x}\in\mathcal{X}} \eta_{\underline{\theta}}(\underline{x},\underline{\theta})^T M^{-1} \eta_{\underline{\theta}}(\underline{x},\underline{\theta}) \, p(\underline{x}) \, d\underline{x}$ can be written in the form

$$\text{tr}\,(M^{-1}A), \; A = \int_{\underline{x}\in\mathcal{X}} \eta_{\underline{\theta}}(\underline{x},\underline{\theta}) \, \eta_{\underline{\theta}}(\underline{x},\underline{\theta})^T p(\underline{x}) \, d\underline{x}.$$

If A is positive definite the result follows by an argument similar to (ii) above using the fact that we may write A as BB^T, where B is of full rank, and $\text{tr}\,(M^{-1}A) = \text{tr}(B^T M^{-1} B)$.

(iv) We merely note here that if the set of vectors $\eta_{\underline{\theta}}(\underline{x},\underline{\theta})$, $\underline{x}\in\mathcal{X}$ coincides with the set of vectors on the unit sphere then ϕ_2 will coincide with ϕ_4 which has already been shown to be not strictly concave in 2.4.

APPENDIX 4.

Let $\phi(M) = -\,\mathrm{mde}(M^{-1})$. The directional derivative at M in the direction of N is given by,

$$\Phi\{M,N\} = \{M^{-1} N M^{-1} - M^{-1}\}_{ss} \quad , \text{ where } \{M^{-1}\}_{ss} \text{ is the}$$

biggest diagonal element of M^{-1}.

If there are r coincident biggest diagonal elements $ss_1,\ldots,ss_r$ say, then we choose s such that

$$\{M^{-1} N M^{-1}\}_{ss} = \min_{ss_i} \{M^{-1} N M^{-1}\}_{ss_i} \quad .$$

Proof.

$\Phi\{M,N\}$ is defined as $\lim_{\varepsilon\to 0^+} \varepsilon^{-1}\{\phi((1-\varepsilon) M + \varepsilon N) - \phi(N)\}$.

Define the Gateaux derivative $\Phi_*\{M,N\}$ by

$$\lim_{\varepsilon\to 0^+} \varepsilon^{-1}\{\phi(M + \varepsilon N) - \phi(M)\} .$$

For simplicity we derive the Gateaux derivative first and then the directional derivative by noting that,

$$\Phi\{M,N\} = \Phi_*\{M,N-M\} .$$

$$(M + \varepsilon N)^{-1} = M^{-1}(I + \varepsilon NM^{-1})^{-1}$$

$$= M^{-1} - \varepsilon M^{-1} NM^{-1} + o(\varepsilon) ,$$

$$\therefore \phi_7(M + \varepsilon N) - \phi_7(M) = -\,\mathrm{mde}\{(M + \varepsilon N)^{-1}\} + \mathrm{mde}\{M^{-1}\}$$

$$= \varepsilon\{M^{-1} N M^{-1}\}_{ss} + o(\varepsilon) ,$$

for small enough ε, where the ss th element of M^{-1} is its biggest diagonal element, or in the case of r coincident maximum diagonal elements $ss_1,\ldots,ss_r$ say, the ss th element is the one such that

$$\{M^{-1} N M^{-1}\}_{ss} = \min_{ss_i} \{M^{-1} N M^{-1}\}_{ss_i} .$$

$$\Rightarrow \quad \Phi_* \{M,N\} = \{M^{-1} N M^{-1}\}_{ss}$$

$$\Rightarrow \quad \Phi \{M,N\} = \{M^{-1} N M^{-1} - M^{-1}\}_{ss}$$

Notes (1) If there are coincident maximum diagonal elements of M^{-1}, then to have differentiability of ϕ_7 at M it would be necessary to have (by Defn. 2.2.2)

$$\Phi\{M, \lambda N_1 + (1-\lambda)N_2\} = \lambda \Phi\{M,N_1\} + (1-\lambda) \Phi\{M,N_2\} .$$

From the above it may be seen that this is equivalent to having,

$$\min_{ss_i} \{M^{-1} (\lambda N_1 + (1-\lambda)N_2)M^{-1}\}_{ss_i} = \lambda \min_{ss_i} \{M^{-1} N_1 M^{-1}\}_{ss_i} + (1-\lambda) \min_{ss_i} \{M^{-1} N_2 M^{-1}\}_{ss_i}, \quad N_1, N_2 \in \mathcal{M}.$$

This will not typically be true.

(2) If ss_j is the minimum diagonal element of $\{M^{-1} N M^{-1}\}$ over $\{ss_j, j = 1,\ldots,r\}$, then it will also be the minimum of $\{\sigma M^{-1} N M^{-1}\}_{ss_j}$, $\sigma > 0$.

(3) If $\min_{ss_i} \{M^{-1} N_1 M^{-1}\}_{ss_i}$ and $\min_{ss_i} \{M^{-1} N_2 M^{-1}\}_{ss_i}$ are both the ss_jth elements then the ss_jth element will be the minimum of $\{M^{-1} (\lambda N_1 + (1-\lambda)N_2) M^{-1}\}_{ss_i}$, $\lambda \in [0,1]$.

$\Rightarrow \Phi\{M, \lambda N_1 + (1-\lambda)N_2\} = \lambda\Phi\{M,N_1\} + (1-\lambda)\Phi\{M,N_2\}$, for such matrices.

Notes (2) and (3) prove Lemmas 3 and 4 of 2.3.

APPENDIX 5

COMPUTING METHODS

A.5.1. Simulation study

The important details of the computer programs which were used in the simulation study of Chapter 6 are given below.

A different program was used for each process type. However, the structure of the three programs was essentially the same. Each program consisted of a MAIN program which interacted with a set of subroutines as illustrated in Fig. A.5.1.

The subroutines used in the programs are listed below, with brief details of their function and any numerical techniques used in them.

MAIN

This part of the program acts as a controller, calling up subroutines to perform the simulations of the sequential experiment, and accumulating, at each stage, relevant information which it eventually outputs in the form of the statistics listed in Chapter 6.

As illustrated in Fig. A.5.1. MAIN interacts directly with RANDU, EXPT and DESIGN for a Type 1 process, and directly with RANDU, EXPT, DESIGN and F.MAX for Type 2 and Type 3 processes.

RANDU

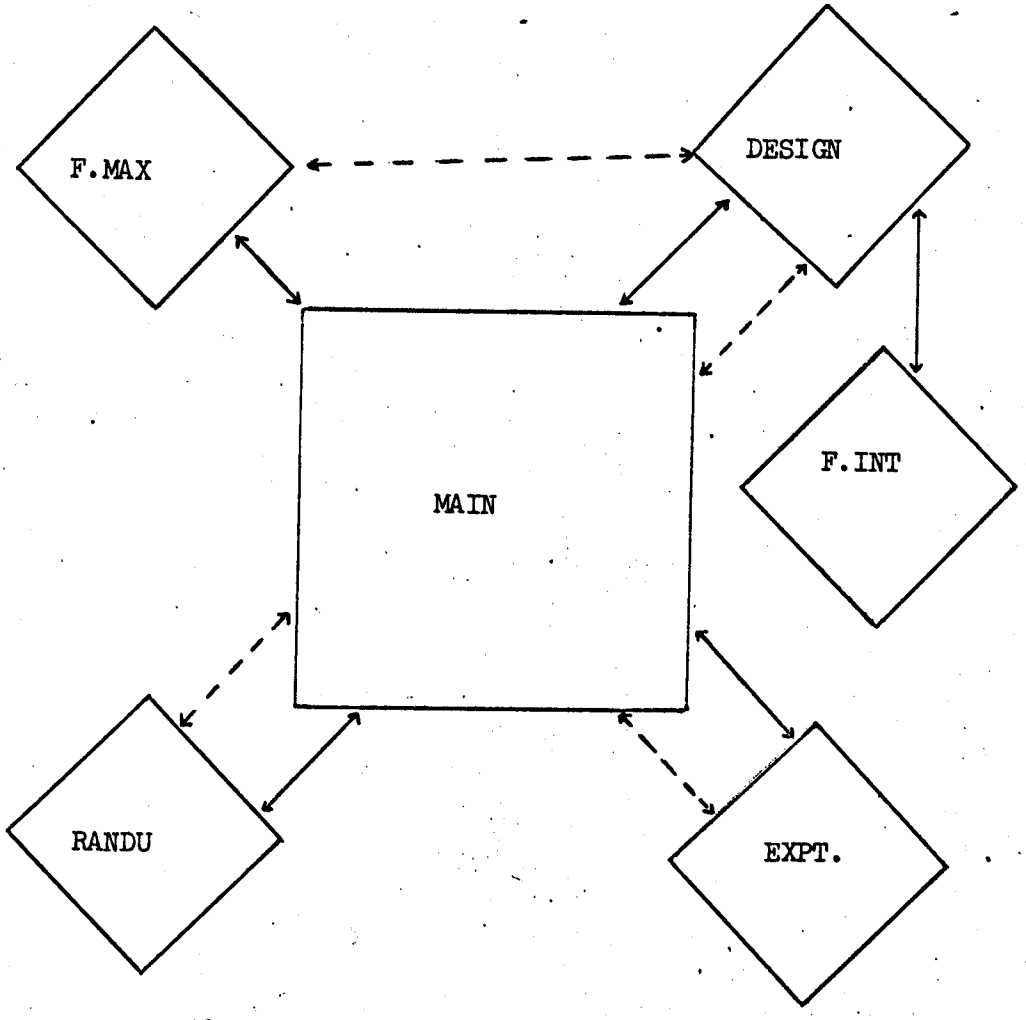
This is the standard I.B.M. pseudo-random number generator which generates random variates from a $U[0,1]$ distribution.

EXPT.

Given a $U[0,1]$ random variate u and an x and a θ from MAIN this subroutine generates binary variables according to the distribution on hand. The rule for generation of the binary variable y is,

$$y = \begin{cases} 1 & , \quad u < \exp(-x\theta) \\ 0 & , \quad \text{otherwise.} \end{cases}$$

Interaction between MAIN program and subroutines



Type 2 and Type 3 processes

Type 1 process.

Fig. A.5.1.

DESIGN

Given the vectors of past design points and observations this subroutine computes the next design point according to the rules of Chapter 6.

For a Type 1 process DESIGN interacts with a maximisation subroutine F.MAX.

For Type 2 and Type 3 processes DESIGN interacts with a numerical integration subroutine F.INT.

F.MAX

Given the vectors of design points and observations this subroutine maximises the corresponding likelihood function. The Newton-Raphson method of function maximisation was found to be perfectly adequate in this situation, particularly as during a sequential experiment good initial estimates of $\hat{\theta}_{n+1}$ are available in the form of $\hat{\theta}_n$.

F.INT.

This subroutine was used to numerically integrate the functions required for the Type 2 and Type 3 processes, these being of the following form.

$$f(\theta) = g(\theta).\exp(-\theta \sum_{i=1}^{r} x_i) \prod_{j=r+1}^{n} (1 - \exp(-\theta y_j)) .$$

$x_1,\ldots,x_r$ denote the design points at which 1's were observed, and $y_{r+1},\ldots,y_n$ denote the design points at which zeros were observed.

$$g(\theta) \in \{\theta, \frac{1}{\theta}, 1\} .$$

The range of integration is $\theta \in [0, \infty)$.

A substantial simplification of the problem is obtained by the following transformation.

$$\phi = \frac{1}{1+\theta} , \text{ giving}$$

$$f^*(\phi) = g^*(\phi)\exp\left(-\sum_{i=1}^{r} x_i\left(\frac{1}{\phi}-1\right)\right)\prod_{j=r+1}^{n}\left(1-\exp\left(-y_j\left(\frac{1}{\phi}-1\right)\right)\right)$$

$$g^*(\phi) \in \left\{\frac{1-\phi}{\phi^3}\ ,\ \frac{1}{\phi(1-\phi)}\ ,\ \frac{1}{\phi^2}\right\}.$$

The range of integration is now $\phi \in [0,1]$.
It was found that Simpson's Rule was a suitable method of numerical integration, the interval $[0,1]$ being divided into 50 segments.

A.5.2. Graphplotting

Diagrams and histograms in this thesis were drawn on a Hewlett-Packard graphplotter linked to a Hewlett-Packard 9810A desk-top computer.

REFERENCES

Only references highlighted with asterisks actually appear in the text. However, the remainder do represent an important set of papers both with regards to the development of the theory and to the application of optimal experimental designs to practical problems.

REFERENCES

*Astrom, K.J. and Eykhoff, P. (1971). System identification - A survey. Automatica, Vol.7, 123-162.

Atkinson, A.C. (1969). Constrained maximisation and the design of experiments. Technometrics, Vol.11, 616-618.

*Atkinson, A.C. and Cox, D.R. (1974). Planning experiments for discriminating between models. J.R.S.S.B., Vol.36, 321-348.

*Atkinson, A.C. and Fedorov, V.V. (1975a). The design of Experiments for discriminating between two rival models. Biometrika, Vol.62, 57-70.

*Atkinson, A.C. and Fedorov, V.V. (1975b). Optimal design. Experiments for discriminating between several models. Biometrika, Vol.62, 289-303.

Atwood, C.L. (1969). Optimal and efficient designs of experiments. Ann. Math. Stats., Vol.40, 1570-1602.

*Atwood, C.L. (1973). Sequences converging to D-optimal designs of experiments. Ann. Stats., Vol.1, 342-352.

*Atwood, C.L. (1976). Convergent design sequencies for sufficiently regular optimality criteria. Ann. Stats., to appear.

*Bar-Shalom, Y. (1971). On the asymptotic properties of the maximum likelihood estimate obtained from dependent observations. J.R.S.S.B., Vol.33, 72-77.

*Bernardo, J.M. (1976). Expected information as expected utility. Technical report, Universidad de Valencia.

*Bhat, B.R. (1974). On the method of maximum - likelihood for dependent observations. J.R.S.S.B., Vol.36, 48-53.

*Box, M.J. (1968a). The occurrence of replications in optimal designs to estimate parameters in non-linear models. J.R.S.S.B., Vol.30, 290-302.

Box, M.J. (1968b). The use of designed experiments in non-linear model-building. In 'The Future of Statistics', D.G. Watts (Ed.), Academic Press, Inc., New York, 241-257.

*Box, M.J. and Draper, N.R. (1968). Non-linear model-building under non-homogeneous variance assumptions. I.C.I. Ltd., Central Instrument Research Lab., Research note 68/9.

Box, M.J. (1970). Some experiences with a non-linear experimental design criterion. Technometrics, Vol.12, 569-589.

Box, M.J. (1971). An experimental design criterion for precise estimation of a subset of the parameters in a non-linear model. Biometrika, Vol.58, 149-153.

Box, G.E.P. and Lucas, H.L. (1959). Design of experiments in non-linear situations, Biometrika, Vol.46, 77-90.

Box, G.E.P. and Draper, N.R. (1959). A basis for the selection of a response surface design. JASA, Vol.54, 622-653.

*Box, G.E.P. and Hunter, W.G. (1965a). Sequential design of experiments for non-linear models. Proc. IBM. Sc. Comp. Symp. 1963, IBM, New York.

Box, G.E.P. and Hunter,W.G. (1965b). The experimental study of physical mechanisms. Technometrics, Vol.7, 23-42.

*Box, G.E.P. and Hill, W.J. (1967). Discrimination among mechanistic models. Technometrics, Vol.9, 57-71.

*Chernoff, H. (1953). Locally optimal designs for estimating parameters. Ann. Math. Stats., Vol.24, 586-602.

Cochran, W.G. (1973). Experiments for non-linear models. JASA, Vol.68, 771-781.

*Davies, M. (1974). Derivation of the directional derivative for D_s optimality. Private communication.

*Dawid, A.P. (1970). On the limiting normality of posterior distributions. Proc. Camb. Phil. Soc., Vol.67, 625-633.

*Draper, N.R. and Hunter, W.G. (1966). Design of experiments for parameter estimation in multi-response situations. Biometrika, Vol.53, 525-533.

Draper, N.R. and Hunter, W.G. (1967a). The use of prior distributions in the design of experiments for parameter estimation in non-linear situations. Biometrika, Vol.54, 147-153.

Draper, N.R. and Hunter, W.G. (1967b). The use of prior distributions in the design of experiments for parameter estimation in non-linear situations: multiresponse case. Biometrika, Vol.54, 662-665.

*Elfving, G. (1952). Optimum allocation in linear regression theory. Ann. Math. Stats., Vol.23, 255-262.

Fedorov, V.V. and Pazman, A. (1968). Design of physical experiments. Fortschritte der Physik. Vol.16, 325-356.

*Fedorov, V.V. (1972). Theory of Optimal Experiments. Academic Press, New York.

*Fedorov, V.V. and Malyutov, M.B. (1972). Optimal designs in regression experiments. M.O.S., Vol.14, 237-324.

*Graybill, F.A. Introduction to matrices with applications in statistics. Wadsworth Pub.Co. Inc., Belmont, Calif..

*Guest, P.G. (1958). The spacing of observations in polynomial regression. Ann. Math. Stats. Vol.29, 294-299.

Hill, W.J. and Hunter, W.G. (1974). Design of experiments for subsets of parameters. Technometrics. Vol.16, 425-434.

*Hoel, P.G. (1958). Efficiency problems in polynomial estimation. Ann. Math. Stats., Vol.29, 1134 - 1145.

*Hunter, W.G. and Reiner, A.M. (1965). Designs for discriminating between two rival models. Technometrics, Vol.7, 307-323.

Hunter, W.G., Hill, W.J. and Henson, T.L. (1969). Designing experiments for precise estimation of all or some of the constants in a mechanistic model. Can. Journ. Chem. Eng., Vol.47, 76-80.

*Jennrich, R.I. (1969). Asymptotic properties of non-linear least squares estimators. Ann. Math. Stats., Vol.40, 633-643.

*Karlin, S. and Studden, W.J. (1966). Optimal experimental designs. Ann. Math. Stats. Vol.37, 783-815.

*Kelley, J.E. Jr. (1960). The cutting-plane method for solving convex programs. J.Soc. Indust. Appl. Math., Vol.8, 703-712.

*Keviczky, L. (1975). Design of experiments for the identification of linear dynamic systems. Technometrics, Vol.17, 303-308.

Kiefer, J. (1958). On the non-randomised optimality and randomised non-optimality of symmetrical designs. Ann. Math. Stats., Vol.20, 675-699.

Kiefer, J. (1959). Optimum experimental designs. JRSSB, Vol.21, 272-319.

Kiefer, J. (1961a). Optimum designs in regression problems II. Ann. Math. Stats., Vol.32, 298-325.

Kiefer, J. (1961b). Optimum experimental designs V, with applications to rotatable designs. Proc. Fourth Berk. Symp., Vol.1, 381-405.

*Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). Ann. Math. Stats., Vol.2, 849-879.

*Kiefer, J. (1975). Generalised Youden designs. In 'A survey of statistical design and linear models'. North-Holland Pub.Co., Amsterdam.

Kiefer, J. and Wolfowitz, J. (1959). Optimum designs in regression problems. Ann. Math. Stats., Vol.30, 271-294.

*Kiefer, J. and Wolfowitz, J. (1960). The equivalence of two extremum problems. Can. J. Math., Vol.12, 363 - 366.

Lauter, E. (1974). A method of designing experiments for non-linear models. M.O.S., Vol.5, 697-708.

Laycock, P.J. and Silvey, S.D. (1968). Optimal designs in regression problems with a general convex loss function. Biometrika, Vol.55, 53-56.

*Lindley, D.V. (1956). On a measure of information provided by an experiment. Ann. Math. Stats., Vol.27, 986-1005.

*Mehra, R.K. (1974a). Synthesis of optimal inputs for multiinput - multioutput (MIMO) systems with process noise. Technical Report No.649, Harvard University.

*Mehra, R.K. (1974b). Optimal input signals for parameter estimation in dynamic systems - Survey and new results. IEEE. Trans. on. Auto. Control, Vol.19, 753-768.

*Mitchell, T.J. (1974a). An algorithm for construction of D-optimal experimental designs. Technometrics, Vol.16, 203-211.

Mitchell, T.J. (1974b). Computer construction of D-optimal first-order designs. Technometrics, Vol.16, 211-220.

*Pazman, A. (1974). A convergence theorem in the theory of D-optimum experimental designs. Ann. Stats., Vol.2, 216-218.

*Sibson, R. (1972). Contribution to the discussion of the papers of Wynn and Laycock. JRSSB, Vol.34, 181-183.

*Sibson, R. and Kenny, A. (1975). Coefficients in D-optimal experimental design. JRSSB, Vol.37, 288-291.

*Silvey, S.D. (1961). A note on maximum-likelihood in the case of dependent random variables. JRSSB, Vol.23, 444-452.

*Silvey, S.D. (1970). Statistical Inference. Penguin Books Ltd.

*Silvey, S.D. (1972). Discussion of papers by Wynn and Laycock. JRSSB, Vol.34, 174-175.

*Silvey, S.D. (1974). Some aspects of the theory of optimal linear regression design with a general concave criterion function. Technical report No.75, Series 2, Princeton University.

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

*Silvey, S.D. and Titterington, D.M. (1974). A Lagrangian approach to optimal design. Biometrika, Vol.61, 299-302.

*Silvey, S.D., Titterington, D.M. and Torsney, B. (1976). An algorithm for D-optimum design on a finite design space. Preprint.

*Smith, K. (1918). On the standard deviations and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. Biometrika, Vol.12, 1-85.

*Stone, M. (1959). Application of the measure of information to the design and comparison of regression experiments. Ann. Math. Stats., Vol. 30, 55-70.

*Titterington, D.M. (1975). Optimal design: Some geometrical aspects of D-optimality. Biometrika, Vol.62, 313-320.

*White, L. (1973). An extension of the general equivalence theorem to non-linear models. Biometrika, Vol.60, 345-348.

*White, L. (1975). Ph.D. Thesis. Imperial College, London.

*Whittle, P. (1973). Some general points in the theory of optimum experimental design. JRSSB, Vol.35, 123-130.

*Wolfe, P. (1961). Accelerating the cutting-plane method for non-linear programming. J. Soc. Indust. Appl. Math., Vol.9, 481-488.

*Wynn, H.P. (1969). Ph.D. Thesis. Imperial College, London.

*Wynn, H.P. (1970). The sequential generation of D-optimum experimental designs. Ann. Math. Stats., Vol.41, 1655-1664.

*Wynn, H.P. (1972). Theory and construction of D-optimum experimental designs. JRSSB, Vol.34, 133-147.