LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Department of Statistical and Mathematical Sciences

OPTIMUM EXPERIMENTAL DESIGN FOR MODEL DISCRIMINATION AND GENERALIZED LINEAR MODELS

by

Antonio Carlos Monteiro Ponce de Leon

Thesis submitted for the degree of
Doctor of Philosophy in the University of London

June 1993

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

The main subject of this thesis concerns the optimum design of experiments for discriminating between two rival mathematical models. In addition, optimality of designs for parameter estimation is investigated although restricted to binary response models. Optimal design theory and generalized linear models form the background for this work. The former provides the tools for construction of the optimum designs whereas the latter provides the framework in which the methods are developed.

For model discrimination the procedures which are proposed may not only be applied to compare two regression models but also to compare two generalized linear models as long as they belong to the same subclass. The principle of the so called T–optimality criterion, originally introduced for discriminating between two regression models, is extended to other classes such as generalized linear models. Within each context a theorem based on the General Equivalence Theorem from the optimal design theory is shown to hold thus allowing both constructing and checking optimum designs.

Optimum experimental designs to estimate the parameters of a binary response model is the other subject of this thesis. Initially, well known link functions such as logit, probit and complementary log–log are considered. Later, this range is widenned by introducing a family of link functions which includes the logit and the complementary log–log links as particular members.

One common feature of these two problems is that classical optimal designs depend on the unknown values of the model parameters. Therefore, only locally optimal designs can be obtained unless observations may be taken sequentially, in which case several methods to search for the optimum are available in the literature. As an alternative to locally and sequentially optimal experiments, Bayesian designs are introduced for both model discrimination and parameter estimation.

ACKNOWLEGMENTS

From the start, when the idea of pursuing a PhD degree was still so distant, to the very end, when the idea of finishing it seemed so close, I could count with the incentive and support of many people to whom I am very grateful. Unfortunately, due to space limitations only a few of these people can be mentioned here.

Firstly, I would like to thank my supervisor, Professor Anthony C. Atkinson, who not only proposed the subject of this thesis but also guided me throughout the research work with valuable suggestions and constructive criticisms. His knowledge and experience on problems of optimum experimental designs, especially for model discrimination, were essential for my understanding of the theoretical subtleties related to this field. I am also grateful for his consistent encouragement which helped me overcome the difficult stages of writing a PhD thesis.

Secondly, I am indebted to CAPES (Ministry of Education, Brazil) for its financial support and to UERJ (Rio de Janeiro, Brazil) for agreeing to a leave of absence, without either of which I would not have been able to complete this work. I would also like to thank all colleagues, both from the Institute of Mathematics and Statistics (UERJ, Brazil) and from the Institute of Pure and Applied Mathematics (CNPq, Brazil), who encouraged me to pursue a PhD degree in the United Kingdom.

Finally, I would like to dedicate this thesis to my wife, Malin, whose support and understanding, especially in the last few months when I spent most of the weekends and evenings in front of the computer, played an important part in the process which led to this thesis, and to my son, Andreas, whose presence provided me with plenty of motivation.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 6

CHAPTER 7

LIST OF FIGURES

Page

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 7

CHAPTER 1. INTRODUCTION

1.1. BASIC OBJECTIVES OF THE THESIS

The main subject of this thesis is the optimum design of experiments for discriminating between two rival mathematical models. The basic problem with which we are concerned occurs when two models are supposed to be potentially capable of providing good fits for the responses arising from a controlled experiment. Generally speaking, we consider pairs of models whose deterministic mechanisms generate patterns which agree closely with that of the underlying experiment. Consequently, they also agree closely with each other. However, such a restrictive assumption is not crucial for the formulation and solution of the problem. The only reason for this assumption is that otherwise the problem does not seem to be relevant. Having said that, it is important to point out that other interests originating in a similar framework could give rise to more specific assumptions.

Under these cirscumstances, the experimenter faces the problem of choice as to which one of the two models should be adopted for general purposes such as inference about the model parameters, parameter interpretability, prediction, etc. Our approach for this problem is based upon the principle of providing a more adequate setting for making the right decision rather than that of developing the means for such decisions. In other words, we concentrate on matters pertaining to the planning of the experiment, with emphasis on maximizing the efficiency with which the choice of model is made. An essential requirement for this approach is that the experimental conditions must be under the full control of the experimenter, thus, allowing a choice of the "optimum" levels for the explanatory variables at which the responses will be taken. This leads to the definition of a criterion of optimality which is the key point for the theory and illustrations presented in the first part of this thesis, which comprises Chapters 3, 4, and 5.

Optimum design theory and generalized linear models form the background for this work. The former provides the tools for construction of the optimum designs whereas the latter provides the framework in which the methods are developed. The procedures which are proposed may not only be applied to compare two regression models but also to compare two generalized linear models as long as they belong to the same subclass. The common factor leading to such a wide scope for the applications of optimum design for model discrimination is the principle of T–optimality (which was originally introduced for discriminating between two regression models) and its natural extensions to other classes such as generalized linear models.

In addition to the problem of model discrimination, optimality of designs for parameter estimation is investigated although restricted to binary response models. This subclass of generalized linear models has been the subject of intensive research recently. Certainly, the advent of the theory of generalized linear models, the increase in the scope for its applications, its fast accessibility in a number of statistical computer packages, and its easiness of implementation and interpretation are among the reasons for such an increasing interest. Following this trend, some advances have also been made in the planning of experiments giving rise to binary responses. Nevertheless, several branchs of this theory are still in need of further developments.

So, optimum experimental designs to estimate the parameters of a binary response model form the second subject of this thesis, comprising Chapters 6 and 7. Initially, binary response models with well known link functions such as logit, probit, and complementary log–log are considered. Later, the range of link functions is widened by introducing a family of link functions which includes the logit and the complementary log–log links as particular members. Generally speaking, our interest lies in searching for experimental designs that maximize the precision with which the model parameters are estimated. As in the first part of the thesis, optimum design theory and generalized linear models underlie the methodology developed in the second part of the thesis.

Another common feature of the two general problems tackled in this thesis is that classical optimum designs depend on the unknown values of the model parameters. Therefore, only locally optimum designs can be obtained unless observations may be taken sequentially. As an alternative to local and sequentially optimum experiments, Bayesian designs are introduced for both model discrimination and parameter estimation. Within each context a theorem based on the General Equivalence Theorem from optimum design theory is shown to hold thus allowing both construction and checking of optimum designs.

1.2. PLAN OF THE THESIS

Chapter 2 contains two summary sections; a review of the literature on discrimination between models as well as on parameter estimation for binary data models and a review of optimum design theory. The aim of this chapter is not only to list the relevant references for the thesis, some of which are quoted in the main text, but also to present an overview of the basic concepts concerning optimum design theory, which underlie the results of the thesis. In the literature review, the references have been grouped by either of the two main subjects of the thesis and by methods presented such as local optimum designs, sequential designs, Bayesian optimum designs, etc. In the latter section, some of the most important concepts of optimum design theory are formalized or quoted. These will be used throughout the thesis in the theoretical developments contained in Chapters 3 to 7. It is important to emphasize that the contents of this chapter are not intended to cover completely either the literature on or the main results of optimum design theory.

Chapters 3 to 7 describe the main results of the thesis. Our presentation in each of these chapters is similar. First, a brief introduction describes the specific problem to be tackled. This is followed by a section of background where not only works related to the problem are quoted but also other important concepts are introduced. Then, in the next two or three sections, the criteria of optimality are introduced as well as the related derivative functions. A longer section containing examples, plots, and results, together with brief analyses and interpretations, illustrates the methods described in the previous sections. Finally, a section of discussion where the main conclusions are summarized and further topics for future research are suggested.

The first part of the research is covered, as mentioned before, in Chapters 3, 4, and 5. Chapter 3 deals with the problem of optimum designs to discriminate between two regression models. Initially, the criterion of T–optimality, giving rise to local optimum designs, is introduced and then extended to cases when prior distributions are available for the sets of parameters. Accordingly, the extended criterion is called Bayesian T–optimality. Two widely used regression models are used as illustration, the quadratic and the double exponential models.

In Chapter 4, a similar methodology is applied in determining optimum designs for discriminating between two binary data models. Again, not only is the concept of T–optimality used as the criterion of optimality but also local and Bayesian optimum designs are defined for this class of generalized linear models. As illustration, two binary data models with different link functions are used. Link functions studied are the logit, probit, and the complementary log–log.

Concluding the subject of model discrimination, Chapter 5 extends the results of Chapters 3 and 4 to the class of generalized linear models. The deviance plays the key role for a further generalization of the T–optimality criterion. Again, because of the dependence of the optimum designs on the true value of the parameters, Bayesian optimum designs are proposed for more general situations. The criteria utilized in previous chapters are shown to be particular cases of the generalized T–optimality criterion. The subclass of Poisson models with reciprocal and logarithmic link functions is used as illustration.

The second part of the research work developed for this thesis is contained in Chapters 6 and 7. In the former, simple binary data models such as logit and probit are considered. We show how to obtain both local and Bayesian optimum designs with the purpose of parameter estimation in this framework. Several examples illustrate the methods. In the latter, a family of link functions is introduced. The aim, then, can vary between optimum designs to estimate the model parameters, the link function parameter, the complete set of parameters, or a combination of the first two aims.

We conclude in Chapter 8 with both a summary of the main results obtained in this thesis and suggestions for future research. As each chapter of research contains its own section of discussions and conclusions, some of these have been omitted in the final chapter. The reader can refer to the last section of each chapter for more specific conclusions.

CHAPTER 2. REVIEW AND BACKGROUND

2.1. INTRODUCTION

This chapter contains both a review of the literature related to the subjects of the thesis and a concise summary of the results from optimum design theory that are relevant to the development of the methodology described in Chapters 3 to 7. The former, presented in Section 2.2, is intended to cover the main contributions to design of experiments for model discrimination as well as parameter estimation for binary data models. The latter, presented in Section 2.3, consists of basic results from optimum design theory and convexity theory that are crucial for the definitions and proofs presented in the subsequent chapters.

Section 2.2 is subdivided into two subsections consisting of reviews for model discrimination and parameter estimation for binary data models. Even though some papers contain more than one approach, the discussions presented in each subsection are motivated by which approach was adopted such as sequential designs, Bayesian optimum designs, etc. Most references considered are based on optimum design theory although there are also important contributions which are based on other approaches.

Section 2.3 is based on the books of Silvey (1980) and Atkinson and Donev (1992). The first book together with that of Fedorov (1972) are classical references for the study of optimum design, and until recently were the only ones available in the English language. Atkinson and Donev's book, published last year, covers not only the general theory but also the advances as well as some of the applications of optimum design in the last twenty years.

2.2. REVIEW OF THE LITERATURE

2.2.1. MODEL DISCRIMINATION

Because optimum designs for model discrimination depend on which one of the models is true and on its unknown parameters, only locally optimum experiments can be found if classical techniques are to be used. An important alternative is to adopt sequentially designed experiments which consist of optimizing a certain criterion (over the design region) based on the information provided by previous observations.

Procedures for obtaining sequentially designed experiments can vary according to the number of points that are chosen at each stage. Although most widely used methods are based on the choice of a single point, there are indications that methods in which more than one point is chosen might be more efficient in terms of convergence. In either case, the criterion to locate the next design point(s), at which the new response(s) will be measured, uses the information from the responses provided by previous experiments similarly designed.

A second variant regards the manner in which the point(s) chosen at a given stage will be incorporated into the current design. The first alternative is simply to add the selected point(s) to the current design. The second is to replace the least informative point(s) belonging to the current design by the point(s) selected according to the adopted criterion. Either way, the process is repeated until a certain criterion of convergence is satisfied.

Most methods proposed to obtain optimum designs for model discrimination are in fact sequential methods as can be seen in the following summary of these methods and their subsequent developments.

In this thesis, the major reference for the subject of optimum designs for model discrimination is the paper by Atkinson and Fedorov (1975a) in which methods to obtain locally optimum, sequential, and Bayesian designs, based on the criterion of T–optimality for discriminating between two regression models, linear or nonlinear in the parameters, are developed and illustrated. For the case of linear models, T–optimum and the equal interest D_s–optimum designs, proposed by Atkinson and Cox (1974), are compared by means of simulation techniques and measures of efficiency. Further, because T–optimality requires the prior knowledge of the true model and the true values of its parameters, other criteria such as Bayesian T–optimality, the maximin criterion, and a combination of the these two, are suggested for the general problem.

Possibly the first formalization of the problem of designing experiments for model discrimination is due to Chernoff (1959) who, initially, considered the problem of testing two simple hypotheses (H_0: $\theta = \theta_0$ vs H_1: $\theta = \theta_1$), under the assumptions that only one experiment is available and that, under either hypothesis, the probability distribution for the response y is known to be $f_i(y)$, i=0,1. Given an initial set of responses $\{y_1,y_2,\cdots,y_n\}$, the Wald sequential likelihood–ratio test provides a decision rule for choosing between further sampling, or accepting either hypothesis. To be more specific, the decision is based on the loglikelihood–ratio $S_n=\sum_i \log\{f_1(y_i)/f_0(y_i)\}$ and the following rule: If $S_n \geq A$, H_0 is rejected in favour of H_1; if $S_n \leq B$, H_0 is accepted; and if $B < S_n < A$, the sequential sampling continues. The problem, then, becomes one of finding suitable numbers A and B. For this, other components involved are prior probabilities for H_0 and H_1 (p_0 and $1-p_0$), the two probabilities of error α and β, a fixed cost c per observation sampled, the loss r_0 due to rejecting H_0 when it is true, the loss r_1 due to accepting H_1 when it is false, and the expected numbers of experiments, under hypotheses H_0 and H_1, prior to stopping the process, say $E(N|H_0)$ and $E(N|H_1)$. Thus, a sensible criterion to determine the numbers A and B is to minimize the expected risk $[p_0R_0 + (1-p_0)R_1]$, where $R_0 = r_0\alpha + cE(N|H_0)$ and $R_1 = r_1\beta + cE(N|H_1)$. An approximate solution for this minimization problem, for the case of c approaching zero (hence, large samples are allowed), involves the calculation of the Kullback–Leibler information numbers. Further, the method is extended to situations in which there is a finite set of experiments and both hypotheses are composite.

Also for the problem of designing experiments to discriminate between two models, Hunter and Reiner (1965) suggested a sequential method for selecting, at each stage, a single point in the design region. First, they suppose that a set of responses $\{y_1,y_2,\cdots,y_n\}$ based on n points $\{x_1,x_2,\cdots,x_n\}$ in the design region, is available. The criterion of optimality for selecting the optimum point in the subsequent stage, say $\overset{*}{x}_{n+1}$, is to maximize the residual sum of squares for the fit of the false model to the pseudodataset $\{y_1,y_2,\cdots,y_n,\hat{y}_{n+1}\}$, where $\hat{y}_{n+1}$ denotes the predicted value, based on the least squares estimates obtained from the n available observations, for the correct model at any feasible point x_{n+1}. Then, a further observation is taken at the optimum point $\overset{*}{x}_{n+1}$, and the process is repeated. This sequential procedure gives rise to two possible optimizations, corresponding to which model is assumed correct. Also suggested was a simpler version of this method which consists of maximizing the squared difference between the predicted values (for the correct and false models) using the least squares estimates for each respective set of parameters, that are obtained from the n available observations. The latter method leads to designs which are asymptotically T–optimum, as mentioned in Atkinson and Fedorov (1975a).

Box and Hill (1967) criticized this approach with the argument that a criterion of optimality should take into consideration not only the absolute difference between the estimated expected responses for the competing models, obtained from a given sample, but also the variances associated with these estimates. As an alternative, they proposed a Bayesian approach in which, initially, each of the competing models is assigned a prior probabilitity for being correct (for instance, equal probabilities). Then, based on an initial set of experiments a sample of responses is obtained, and the posterior probability for each model to be correct is calculated. For every subsequent stage, a single experiment is selected from the design region, following the criterion of maximizing an upper bound for the expected change in entropy, a concept used in communication systems, which is a function of the prior and posterior probabilities for the competing models. This process is repeated until the posterior probabilities show the superiority of one model with respect to the others. The method presupposes normality (although this assumption is not crucial), second order assumptions, and known equal variances for the random component of the competing models.

Following Box and Hill's approach, Hill and Hunter (1969) extended this method for the case of unknown equal variances. Basically, a noninformative prior distribution is assumed for the logarithm of the standard deviation σ, and after a number of observations, say $\{y_1,y_2,\cdots,y_{n-1}\}$, both the posterior distribution of σ and the distribution of y_n are obtained. The latter is substituted into the criterion function proposed by Box and Hill (1967), for which an approximation is obtained, and the same steps of the original method apply thereafter.

Andrews (1971) approached the problem of optimum designs for model discrimination from another point of view. First, k competing models are considered and from these a general model is constructed by including all terms corresponding to distinct independent variables so that each original model can be generated by suitably constraining the full parameter set. Given that a sample $\{y_1,y_2,\cdots,y_n\}$ is available, the purpose of this sequential method is to detect inadequate models based on the evidence showed by the resulting p–values, $\{\alpha_i(n);i=1,\cdots,k\}$, for testing each individual model against the full model. Under second order assumptions, and normal errors the above are simple F–tests so that the computation of the p–values $\{\alpha_i(n)\}$ is straightforward. To initiate the sequential design strategy, a predicted value $\hat{y}_{n+1}$ is required so that the new set of p–values, say $\{\hat{\alpha}_i(n+1)\}$, can be obtained from the pseudodataset $\{y_1,y_2,\cdots,y_n,\hat{y}_{n+1}\}$. It is proposed that the initial n observations are taken at fewer design points, say $\{x_1,x_2,\cdots,x_{n'}\}$, where $n' < n$ and each point is replicated at least twice so that the mean response at each of these points could then be used as the predicted responses $\{\hat{y}_{n+1}(x_i);\ i=1,\cdots,n'\}$. Then, the criterion is defined as minimizing a suitable function of the $\{\hat{\alpha}_i(n+1)\}$, over the set of initial design points $\{x_1,x_2,\cdots,x_{n'}\}$. A simulated example shows that the method is quite efficient for discarding unsuitable models. However, both the strong dependence on normality and the restrictions on the design region are clear disadvantages of this method.

Atkinson and Cox (1974) proposed the equal efficiency designs for discriminating between several models. As in Andrews (1971), an extended model containing the linearly independent terms from all rival models is constructed, so that each individual model is a particular case of the extended one. Then, following the approach proposed by Atkinson (1972), exact and/or continuous optimum designs can be determined to detect departures from individual models to the general one, by applying the criterion of D_s–optimality for estimating the complement of the vector of parameters for the ith model with respect to that for the combined model. Therefore, when it is feasible to assign relative importances for each model, maximization of a linear combination (whose coefficients are equal to the relative importances) of all the D_s–optimality criteria could be used as a criterion of optimality. Otherwise, they propose maximizing a measure of efficiency associated with each "continuous" D–optimum design, restricted to the case of equal efficiencies for all m competing models.

Atkinson and Fedorov (1975b) extended the concept of T–optimality to the case of experimental designs for discriminating between more than two models. Similarly to the case for two models, the T–optimum design to discriminate between v models ($v>2$) depends upon which is the true model and the values of its parameters. Under the assumption that these are known, the problem reduces to finding a design maximizing a linear combination of the noncentrality parameters corresponding to the closest models. Such a design is called T_p–optimum, where p is used to stress the dependence of the optimum design on the set of unknown weights $\{p_j\}$ (coefficients of the linear combination). The case in which there is only one closest model can be solved by applying the criterion of T–optimality. In practice, however, neither the true model nor its parameters are known a priori so that other approaches need be adopted. Atkinson and Fedorov (1975b) suggest sequential procedures leading to T–optimality. Another possibility is a Bayesian approach generalizing that presented in Chapter 3 of this thesis.

Motivated by the maximin approach proposed by Atkinson and Fedorov (1975a), Jones and Mitchell (1978) introduced two further criteria of optimality for the problem of designing experiments to discriminate between two nested linear models in which the restricted model is specified by a vector of parameters β_1 whereas the full model is specified by the extended vector $(\beta_1^t\ \beta_2^t)^t$. Both criteria are based on functions of the noncentrality parameter when the full model is assumed correct. Regarding the restricted model as a class of models, a measure of the distance between the full model and the members of this class is used to restrict the search corresponding to the minimization step of the maximin approach. This is equivalent to putting constraints in the β_2–space so that only inadequate models are regarded in the minimization. Wijesinha and Khuri (1987) extended these criteria to the case of discriminating between nested linear multiresponse models.

Ramsey and Chesher (1976) proposed alternative measures of the "difference" between two regression models with the purpose of providing the means for the experimenter to make a decision whether or not the circumstances require methods of model discrimination.

Using the same framework as that of Atkinson and Fedorov (1975a) and further assumptions, Fedorov (1978) proved that there always exists a T–optimum design that contains no more than (m_2+1) supporting points, where m_2 is the number of parameters for the assumed false model. Further, for the maximin approach also proposed in the former paper, he proved that there always exists a maximin design containing no more than (m_1+m_2+1) supporting points. These results hold for the case where the true model and the true values of its parameters are known.

Ponce de Leon and Atkinson (1991) implemented the criterion of Bayesian T–optimality to discriminate between two regression models, either linear or nonlinear, as suggested by Atkinson and Fedorov (1975a). Basically, the assumptions of prior probabilities for each model to be correct and conditioned on these events, prior probabilities for each set of parameters are made, and then incorporated in the formulation of the problem to obtain optimum designs for model discrimination. It is shown that an extension of the General Equivalence Theorem, due to Kiefer and Wolfowitz (1960), holds under these circumstances, and thus, "continuous" Bayesian T–optimum designs can be constructed and checked for optimality. Discrete prior distributions are used to illustrate the procedures.

Given that a handful of sensible sequential procedures for determining optimum designs for model discrimination has been available for about eight years, it is surprising that so little has been done so far with respect to comparisons between their performances, resulting optimum designs, convergence properties, etc. However, some attempts have been made in this respect. These are given below.

A modified version of Chernoff's method and the method of Box and Hill (1967) are compared by Meeter, Pirie, and Blot (1970) in three examples. They conclude that the former method is asymptotically optimal for experiments with sufficiently small costs of sampling. The latter also performs well in the examples studied, but as its optimal properties are unknown further research on the subject is suggested.

Atkinson (1981) investigated the alleged weakness of the sequential method proposed by Hunter and Reiner (1965), which does not take the variances of the predicted values into consideration, as opposed to the method of maximum expected change in entropy proposed by Box and Hill (1967). For the case of two linear models it is shown that, as the number of steps of the sequential procedure tends to infinity and provided that both methods converge to designs nonsingular for each model, the methods are asymptotically equivalent in the sense that Box and Hill's criterion function reduces to Hunter and Reiner's.

Huang (1991) presented several algorithms to implement some of the available sequential procedures for model discrimination such as those proposed by Box and Hill (1967), Atkinson and Fedorov (1975a,b), and Atkinson and Cox (1974). There is a considerable improvement in the performance of these algorithms as compared to classical procedures when applied to variations of Examples 5.2 and 5.3 of Box and Hill (1967), both of which have a choice of four feasible models, and Example 2 of Atkinson and Fedorov (1975b), which has a choice of three similar linear models. As far as the results of these comparisons are concerned the number of iterations taken by each algorithm to converge to the optimum design is used to illustrate the algorithm performances, whereas for T–optimum and D_s–optimum designs, optimality for the resulting designs is illustrated with plots of the derivative functions. Further illustrations are provided by comparisons between optimum designs arising from the utilization of different optimality criteria. Similar comparisons between the so called classical, the Vertex–exchange, and the Global substitution algorithms were also reported by Pronzato, Huang and Walter (1991).

Methods of designing experiments for model discrimination, sequential or nonsequential, do not take into consideration the accuracy associated with the estimates for the underlying parameters of the correct model. On the other hand, specific optimization procedures for parameter estimation assume that the correct model is known a priori. Therefore, it seems natural to think of the dual problem of model discrimination and parameter estimation. For instance, in a sequential experiment, at each stage, either purpose could be emphasized according to whether or not the current set of responses points to the correctness of a specific model.

This problem was first considered by Hill, Hunter, and Wichern (1968) who introduced a sequential procedure combining the criterion of Box and Hill (1967) for model discrimination with that of Box and Lucas (1959) for parameter estimation in nonlinear models (i.e. D–optimality applied to the linearized nonlinear model about an estimate for the parameters). The criterion function underlying the sequential procedure is defined as $C = w_1D + w_2E$; where D is a standardized value of the Box and Hill's criterion function; E is the expected standardized value of the D–optimality criterion function (expectation is over the distribution of posterior probabilities provided by Box and Hill's method); and the pair of complementary weights w_1 and w_2 is given by a monotonic function of π_b, the posterior probability associated with the best current model. This function is such that the weight w_1 decreases as π_b increases, thus placing more emphasis on parameter estimation as the best model is more clearly identified. Further, this function allows the rate of decrease of w_1 with respect to π_b to be controlled by the experimenter.

Borth (1975) proposed a generalization of Box and Hill's criterion for model discrimination, in which is incorporated the expected change in entropy measuring the uncertainty about the parameters for each model. As does the original version, this criterion (called total entropy) aims to choose, sequentially, single points in the design region so that the expected change in entropy is maximized. Its advantage lies in the fact that both problems of parameter estimation and model discrimination are tackled simultaneously. This generalization is implemented by regarding prior distributions for the vectors of parameters of the m competing models, say $\{p(\theta_i);\ i=1,\cdots,m\}$, and subsequently partitioning the parameter spaces $\{\Theta_i\}$ so as to define sub–models of the m original models. Then, the concept of entropy is applied to the probabilities $\pi_{ij}(n-1) = \pi_i(n-1)\ \pi_{j|i}(n-1)$;

where $\pi_i(n-1)$ is the probability of the ith model prior to performing the nth experiment, $\pi_{ij}(n-1)$ is the probability that the jth sub–model of the ith model is correct, and $\pi_{j|i}(n-1)$ is the probability that the jth sub–model is correct given that the ith model is correct. This gives rise to two separate terms; one represents the entropy measuring the uncertainty about which model is correct (i.e. the original entropy as in Box and Hill (1967)) whilst the other represents the weighted average (with weights equal to the probabilities $\{\pi_i(n-1)\}$) of the entropies measuring the uncertainty about the vector of parameters corresponding to each model. The same steps of the original method follow thereafter.

Hill (1978b) pointed out that in the example used by Atkinson and Fedorov (1975b) to illustrate a procedure for obtaining sequential designs to discriminate between more than two models, the three–point design putting equal weights at the endpoints and the midpoint of the design region $\mathcal{X} = [-1,1]$ is D–optimum for all three models regarded. Thus, he observes that the multipurpose design criterion of Hill, Hunter, and Wichern (1968) would be simplified since whichever model were chosen by a discriminatory experiment, the D–optimum design to estimate its parameters would be the same. However, if a pure discrimination procedure were utilized, it was suggested that the resulting optimum design should also incorporate trials at the three points of the D–optimum design. In this case, the problem of how much weight should be assigned to these trials has to be addressed. Finally, for the case of linear models, he suggested that the D–optimum designs should be determined beforehand so that any similarity could be considered whichever sequential procedure is subsequently adopted.

For linear regression models, Fedorov and Khabarov (1986) showed the equivalence between optimum designs for parameter estimation and discrimination between two models, one of which is taken to be the null model. More specifically, they showed that the solution to the problem of finding an optimum design for estimating a linear combination of the parameters corresponding to the nondegenerate model, say $c^t\theta$, and that for testing the significance of $c^t\theta$ are the same. Intuitively, this equivalence is not surprising since the accuracy of the estimator $c^t\hat{\theta}$ is crucial for both problems.

Chambers and Cox (1967) suggested a criterion of optimality for determining nonsequential designs with the purpose of discriminating between the logit and probit models for binary data. Yanagisawa (1988) extended Chambers and Cox's criterion. Ponce de Leon and Atkinson (1992a,b) proposed an extension of the T–optimality criterion to discriminate between two generalized linear models, with emphasis on the subclass of binary data models.

Finally, further references on the subject of experimental designs for model discrimination can be found in reviews by Pereira (1977), Hill (1978a), Atkinson (1982), Atkinson (1986), and Atkinson and Donev (1992).

2.2.2. PARAMETER ESTIMATION FOR BINARY DATA MODELS

Not long ago, Crump (1979) pointed out that the subject of experimental designs for dose–response problems in carcinogenesis needed further development. For instance, the following are extracted from his paper: "two dose–response models may fit experimental data about equally well and yet predict responses that differ by many orders of magnitude at low doses"; "Unfortunately, the mode of actions of carcinogenesis is not sufficiently well understood. There is considerable disagreement among scientists as to the shape of the dose response curve in the low dose region"; and "It is unlikely that an experiment would ever be designed solely for the purpose of model discrimination or solely for the purpose of risk extrapolation. Nevertheless, it is still important to study design problems of these kinds". Although the subject of binary data models, as a whole, has attracted more interest since 1979, the above extracts stress the importance of further research on specific methods of determining optimum experimental designs for binary data models in fields such as model discrimination and/or parameter estimation.

In the last subsection, references related to model discrimination were discussed whereas in this subsection, we concentrate on references related to problems of experimental designs for parameter estimation and to families of link functions, both for binary response models. However, as the most important contributions and recent advances in the former subject are well covered by Morgan (1992, Chapter 8) in a comprehensive survey of the literature, only references on the latter subject are discussed here.

Prentice (1976) suggested a four–parameter link function family as an alternative to logit and probit models for fitting dose–response data. Several well known models are members of this family such as logit, probit, extreme minimum value, extreme maximum value, double exponential, etc. The inclusion of at least one of the two extra parameters in the model showed significant improvements to the fit of Bliss' classical data (mortality of adult beetle after five hours exposure to gaseous carbon disulphide, Bliss (1935)) relative to the logit model as shown by the maximized loglikelihood function and the loglikelihood ratio test. On the other hand, it is inevitable that the variances for the estimates of percentage points, for instance, will be larger as a result of more parameters having to be estimated. A further restriction on the use of this family of link functions is related to the difficulties in the computation of the maximum likelihood estimates for the two extra parameters.

Pregibon (1980) proposed another four–parameter family of link functions which, like Prentice's, generates symmetric and skewed dose–response curves. Moreover, in Section 4 of his paper a three–parameter family of link functions which contains both the logit and the complementary log–log links is presented. Experimemtal design problems involving this family of link functions is the subject of Chapter 7 of this thesis.

Aranda–Ordaz (1981) proposed two three–parameter families of link functions giving rise, among others, to the logistic and complementary log–log links. The former is a family of symmetric transformations, in the sense that the probabilites provided by the members of this family are not affected by interchanging successes and failures. Two particular members of this family are the logistic and the identity links. The latter is a family of asymmetric transformations (in fact, this is the second family suggested in Pregibon (1980)), with the exception of the logistic case which corresponds to the additional parameter equal to one. Two datasets are analysed to illustrate the benefits obtained by including an extra parameter in the model. In the first case, where the symmetric family of links is fitted to the data, a slight improvement is observed as compared to the logit and probit fits. In the second example, where the asymmetric family of models is fitted to the data from Bliss (1935), the results are similar to those of Prentice (1976), showing a significant improvement with respect to the logit fit. The loglikelihood is maximized for the complementary log–log link.

Rocke (1993) suggests a three–parameter beta transformation family which includes link functions (or transformations) such as the arcsin square–root, the logit, and the identity. Further, a comprehensive survey of other families of link functions is presented together with an example in which the profile loglikelihood functions for the beta, folded–power, and Aranda–Ordaz transformations are compared, revealing that the respective maxima are similar although they are achieved at different values of the additional parameter estimates.

2.3. REVIEW OF OPTIMUM DESIGN THEORY

In this section, some important theoretical results from optimum design are mentioned briefly. Most of them are extracted from Silvey (1980) and Atkinson and Donev (1992).

Suppose that the responses related to a well defined experiment may be described by the following model

$$Y_i = \eta(\mathbf{x}_i, \boldsymbol{\theta}) + \epsilon_i \qquad (i=1,\cdots,N) \tag{2.3.1}$$

where the errors obey the second order assumptions; $\boldsymbol{\theta}$ is a column vector of p unknown parameters, the values of $\{\mathbf{x}_i\}$ are such that $\mathbf{x}_i \in \mathcal{X}$, a prespecified design region; and $\eta(.,.)$ is a known function which may be linear or nonlinear in the parameters. For linear models, (2.3.1) simplifies as follows

$$\eta(\mathbf{x}_i, \boldsymbol{\theta}) = \mathrm{f}(\mathbf{x}_i)^t \boldsymbol{\theta} \qquad (i=1,\cdots,N) \tag{2.3.2}$$

where $\mathrm{f}(\mathbf{x}_i)^t = (\mathrm{f}_1(\mathbf{x}_i),\cdots,\mathrm{f}_p(\mathbf{x}_i))$; and $\{\mathrm{f}_j(.); j=1,\cdots,p\}$ are known functions. Hence, we can summarize the N equations in (2.3.2) by writing $\eta(\mathbf{x},\boldsymbol{\theta}) = F\boldsymbol{\theta}$, where F denotes the design matrix.

Assume that the interest lies in estimating a function of the vector of parameters $\boldsymbol{\theta}$ and that the experimenter can select the values $\{\mathbf{x}_i; i=1,\cdots,N\}$ in the design region $\mathcal{X}$ so as to optimize the estimation of this function of $\boldsymbol{\theta}$, based on a suitable criterion. Depending on the underlying interest of the experiment, different criteria can be adopted such as D–optimality, E–optimality, A–optimality, etc. Here, we are concerned with any criterion whose associated criterion function is defined as a concave function of the Fisher information matrix for the vector of unknown parameters $\boldsymbol{\theta}$. This is further explained below.

Denote the sample of N observations, $\{\mathbf{x}_i; i=1,\cdots,N\}$, by a design ξ_N which is represented as

$$\xi_N = \begin{Bmatrix} \mathbf{x}_1 \cdots \mathbf{x}_n \\ p_1 \cdots p_n \end{Bmatrix} \tag{2.3.3}$$

where $p_i = r_i/N$, and r_i is the number of replications of the design point $\mathbf{x}_i$ in the sample of N observations. Then, let us denote the Fisher information matrix for the vector of parameters $\boldsymbol{\theta}$, based on N observations, by $M(\xi_N,\boldsymbol{\theta})$ for a nonlinear model, and by $M(\xi_N)$ for a linear model. To simplify the notation, only the linear case is regarded in what follows. The case of nonlinear models is analogous.

A design as described in (2.3.3) is called a discrete design measure since it can be viewed as a discrete probability distribution on $\mathfrak{X}$. Thus, let $\mathfrak{D}$ be the class of all discrete probability distributions in $\mathfrak{X}$, let $\mathscr{M}_N$ be the set of matrices $\{M(\xi_N)\colon \xi_N \in \mathfrak{D}\}$ so that $M(.)\colon \mathfrak{D} \to \mathscr{M}_N$, and let $\phi(.)\colon \mathscr{M}_N \to \mathbb{R}$ denote the function associated with a specific criterion of optimization. Then, the optimization problem, regarded as a maximization one, can be described as

$$\phi(M(\xi_N^*)) = \sup_{\xi_N \in \mathfrak{D}} \phi(M(\xi_N)) \tag{2.3.4}$$

Here, it is worth noticing that the corresponding problem of maximization for the case of nonlinear models would only be possible to solve for fixed values of the unknown vector of parameters θ. In this case the optimum design is called locally optimum.

Optimization problems such as (2.3.4) are called exact problems of optimum design. Because of the complex features of the set of matrices $\mathscr{M}_N$, Kiefer (1974) suggested the approximate theory of optimum design which, technically, consists of extending the class $\mathfrak{D}$ of all discrete probability distributions on $\mathfrak{X}$ to the class $\mathscr{H}$ of all probability distributions on $\mathfrak{X}$. Accordingly, any element of the class $\mathscr{H}$ is called a design measure and is denoted by ξ. Further, the new set of matrices $\{M(\xi)\colon \xi \in \mathscr{H}\}$ is denoted by $\mathscr{M}$, and the previously described optimization problem becomes

$$\phi(M(\xi^*)) = \sup_{\xi \in \mathscr{H}} \phi(M(\xi)) \tag{2.3.5}$$

All problems of optimum design proposed in this thesis utilize the approximate theory and the corresponding criteria of optimization are described as in (2.3.5).

Under this more general framework, some important results, establishing conditions which an optimum design must satisfy, can be proven. For this we need to introduce the notion of the Fréchet derivative.

Definition 2.3.1. The Fréchet derivative of ϕ at $M(\xi_1)$ in the direction of $M(\xi_2)$ is

$$F_\phi(M_1,M_2) = \lim_{\alpha \to 0} \frac{1}{\alpha}\left[\phi\{(1-\alpha)M_1 + \alpha M_2\} - \phi(M_1)\right]$$

where $M_1 = M(\xi_1)$ and $M_2 = M(\xi_2)$.

Silvey (1980) summarized in a series of three Theorems and a Corollary the most essential results for the theory of optimum designs. These are reproduced below.

Theorem 2.1. Provided that ϕ is concave on $\mathscr{M}$, ξ^* is ϕ–optimum if and only if

$$F_\phi\{M(\xi^*),M(\xi)\} \le 0 \text{ for all } \xi \in \mathscr{H}.$$

Theorem 2.2. Let $M(\xi_x)$ be a design measure putting all mass at the point $x \in \mathfrak{X}$. If ϕ is concave on $\mathscr{M}$ and differentiable at $M(\xi^*)$, then ξ^* is ϕ–optimum if and only if

$$F_\phi\{M(\xi^*),f(x)f(x)^t\} \le 0 \text{ for all } x \in \mathfrak{X}.$$

Note: For linear models $M(\xi_x) = f(x)f(x)^t$ whilst for nonlinear models $M(\xi_x)$ depends upon the true value of θ.

Theorem 2.3. Provided that ϕ is differentiable at all points of $\mathscr{M}^+$, the subset of $\mathscr{M}$ where $\phi(M) > -\infty$, and a ϕ–optimum measure exists, then ξ^* is ϕ–optimum if and only if

$$\sup_{x\in\mathfrak{X}} F_\phi\{M(\xi^*),f(x)f(x)^t)\} = \inf_{\xi\in\mathscr{H}} \sup_{x\in\mathfrak{X}} F_\phi\{M(\xi),f(x)f(x)^t)\}$$

Corollary 2.1. If ξ^* is ϕ–optimum and ϕ is differentiable at $M(\xi^*)$ then we have both

$$\sup_{x\in\mathfrak{X}} F_\phi\{M(\xi^*),f(x)f(x)^t)\} = 0$$

and

$$E[F_\phi\{M(\xi^*),f(\underset{\sim}{x})f(\underset{\sim}{x})^t)\}] = 0$$

where $\underset{\sim}{x}$ is a random vector with distribution ξ^*. Of course, this can happen only if $F_\phi\{M(\xi^*),f(\underset{\sim}{x})f(\underset{\sim}{x})^t)\} = 0$ with probability one. Thus, if ξ^* is discrete with finite support points $x(1),\cdots,x(I)$, then

$$F_\phi\{M(\xi^*),f(x(i))f(x(i))^t\} = 0 \qquad (i=1,\cdots,I)$$

The above Theorems form the celebrated General Equivalence Theorem of the approximate theory for optimum designs.

CHAPTER 3. OPTIMUM EXPERIMENTAL DESIGN FOR DISCRIMINATING BETWEEN TWO RIVAL REGRESSION MODELS

3.1. INTRODUCTION

This chapter is concerned with the design of experiments for discriminating between two regression models, one or both of which may be nonlinear in the parameters. Atkinson & Fedorov (1975a) describe T–optimum designs for this purpose which are optimum when it is known which one of the models is true. The designs, which satisfy an equivalence theorem of optimum design theory, are locally optimum, in the sense that they depend upon the values of the unknown parameters of the true model. Here the theory is extended to situations in which there is a specified prior probability that each model is true and, conditionally on this probability, prior distributions for the parameters in the two models. Our central result is that such designs again satisfy an equivalence theorem which can be used both for the construction of designs and for checking the optimality of a proposed design.

In the next section the background for the problem of discriminating between regression models is presented. In Section 3.3 we give a description of the problem and introduce the criterion of T–optimality. The equivalence theorem for T–optimum designs with prior distributions is presented in Section 3.4. Examples are in Section 3.5. Finally, some further topics about the problem are discussed in Section 3.6.

3.2. BACKGROUND

Designing experiments to discriminate between two models, and more than two, is a fairly recent subject of concern to both applied and theoretical statisticians. A substantial amount of research has so far been developed within the framework of optimal design theory although other approaches such as fully sequential methods, which originated with the articles by Hunter and Reiner (1965) and Box and Hill (1967), have also been the sources of important contributions to the field.

In this problem the aim is to find a design for which the expected values under each model, are as far apart as possible according to an appropriate criterion. It is an intuitive argument to say that some points in the design region are surely more informative than others regarding the purpose of model discrimination.

For instance, let us take the differences between the expected responses from two models at any point from the design region. Points with large such differences clearly assist the purpose of model discrimination. In contrast, points with small expected differences in response will only tend to confound the models. Generally speaking, the larger the expected difference in responses under the two models at a given design point the more informative such point should be considered.

Therefore, experiments to discriminate between models should ideally take observations at points of the former type while avoiding the latter. The difficulties begin with the fact that the differences in expected responses are unknown. Indeed, they depend on the unknown values of the parameters of the true model which is also unknown.

There are many questions which arise naturally in this context. For example, how should one combine points with large differences in response so that the entire design is effective w.r.t. discriminating between the models ? How many points are to be included in the optimum design ? Should some points have more weight than others ? All these questions are related to the criterion which will be used. We adopt the criterion of T–optimality, proposed by Atkinson and Fedorov (1975a) to discriminate between two models. T–optimality can also be applied to the problem of discriminating between more than two models; see Atkinson and Fedorov (1975b) for details.

Whichever criterion is chosen we must bear in mind that, in practice, any data set yielded by a designed experiment will not always show conclusive evidence that either model is correct. This is due, of course, to the impossibility of taking the random effect into consideration, i.e. the procedures which will be introduced are only for discriminating between the systematic parts of the models while the random effect is disregarded.

3.3. THE CRITERION OF T–OPTIMALITY

Assume that both models and sets of parameters have prior probabilities associated with them. These probability distributions are incorporated into the framework in which the problem is considered. One can think of these prior probability distributions as reflecting the knowledege acquired in previous experiments, or alternatively as the experimenter's prior belief.

To be more precise we introduce our notation which is based on that of Silvey (1980, Chapter 3). Let $\mathcal{X}$, a compact set, be the design region; let $\mathcal{H}$ be the class of all probability distributions on $\mathcal{X}$; let $\xi \in \mathcal{H}$ be a design measure, or just a design, and suppose the model is written as

$$E(Y) = \eta_t(x), \quad x \in \mathcal{X}$$

where the true model $\eta_t(x)$ is one of two known functions $\eta_1(x,\theta_1)$ and $\eta_2(x,\theta_2)$ with respective prior probabilities π_{01} and $\pi_{02} = 1 - \pi_{01}$. The set of parameters θ_j, of dimension m_j, has prior probability distribution $p_{0j}(\theta_j)$ and $\Theta_j \subset \mathbb{R}^{m_j}$ is the parameter space of the set θ_j {j = 1, 2}. In what follows, we define the necessary elements in order to introduce the criterion of T–optimality.

Definition 3.1. The quantities

$$\Delta_1(\xi,\theta_2) = \inf_{\theta_1 \in \Theta_1} \int_{\mathcal{X}} \left\{\eta_2(x,\theta_2) - \eta_1(x,\theta_1)\right\}^2 \xi(dx) \tag{3.3.1.a}$$

and

$$\Delta_2(\xi,\theta_1) = \inf_{\theta_2 \in \Theta_2} \int_{\mathcal{X}} \left\{\eta_1(x,\theta_1) - \eta_2(x,\theta_2)\right\}^2 \xi(dx) \tag{3.3.1.b}$$

are the non–centrality parameters for the first model when the second is true, and vice versa.

The reason for this terminology is that for a given set of responses, (3.3.1.a) is proportional to the non–centrality parameter of the χ^2 distribution of the residual sum of squares corresponding to the fit of the first model to the data. This holds, of course, under linear regression theory, normal assumptions and the assumption that the second model is true. Analogously, so is (3.3.1.b) w.r.t. the fit of the second model to the data.

They are thus the residual sums of squares for the false models in the absence of experimental error. The values of these non–centrality parameters depend both on the design and on the unknown values, respectively, of θ_2 and θ_1. However, by assuming prior probabilities for θ_2 and θ_1, the dependence is transferred to $p_{02}(\theta_2)$ and $p_{01}(\theta_1)$.

We now can define the criterion of T–optimality, firstly as it was originally proposed and later to include situations for which prior information is incorporated.

Definition 3.2. Assume that the jth model is true and its set of parameters θ_j is known. A *local T–optimum design*, or just *T–optimum*, ξ^* is such that

$$\Delta_{\bar{j}}(\xi^*) = \sup_{\xi \in \mathcal{H}} \Delta_{\bar{j}}(\xi) \tag{3.3.2}$$

where $\Delta_{\bar{j}}(\xi) = \Delta_{\bar{j}}(\xi,\theta_j)$, j = 1,2 and $\bar{j}$ = not j.

In our notation this corresponds to the case in which $\pi_{0j} = 1$ and $p_{0j}(\theta_j)$ has mass function at the point θ_j, $\{j = 1, 2\}$. T–optimality was proposed by Atkinson & Fedorov (1975a). They used properties of T–optimum designs to propose a sequential procedure leading to designs which are asymptotically T–optimum. As a consequence of T–optimum designs depending on the model truth and the true value of its parameters they are called locally optimal. Chernoff (1953) first used the term "locally optimal" in the context of designing experiments to estimate parameters of a given probability distribution.

Obviously, locally T–optimum designs are implausible in practice, as if the true model and its parameters are known, there is no need for model discrimination. Nevertheless, the concept is essential to the development of both sequential and Bayesian designs.

<u>Definition 3.3.</u> Assume that the jth model is true and that the only information about θ_j is the prior distribution $p_{0j}(\theta_j)$. Then, a *partially Bayesian T–optimum design* ξ^* is such that

$$\gamma_{\bar{j}}(\xi^*) = \sup_{\xi \in \mathscr{H}} \gamma_{\bar{j}}(\xi) \qquad (3.3.3)$$

where $\gamma_{\bar{j}}(\xi) = E_{\theta_j}\left\{\Delta_{\bar{j}}(\xi, \theta_j)\right\}$, $j = 1, 2$ and $\bar{j}$ = not j.

In words, ξ^* maximizes the expected non–centrality parameter for model $\bar{j}$ when model j is true. In practice, partially Bayesian T–optimum designs, as opposed to locally T–optimum ones, are more realistic in the sense that situations in which it is known which is the true model, although the only information about its parameters is represented by a prior distribution, are more likely to occur.

However, in this case the interpretation given to the purpose of the experiment ought to be slightly modified. Here, the aim becomes that of finding a design which purely discriminates against the false model, known a priori, rather than, as stated in the general problem, to find a design which discriminates between the models having no information about the model truth except that either model is true. Potential applications occur, for instance, when expected responses provided by a model other than the true, hence taken to be the false, need to be avoided due to any specific reason.

Definition 3.4. Assume that there is a prior probability that each model is true and, conditionally on this probability, prior distributions for the parameters in the two models. Then, a *fully Bayesian T–optimum design* ξ^* is such that

$$\Gamma(\xi^*) = \sup_{\xi \in \mathscr{H}} \Gamma(\xi) \tag{3.3.4}$$

where $\Gamma(\xi) = \sum_j \pi_{0j} \gamma_{\bar{j}}(\xi) = \pi_{01} \mathrm{E}_{\theta_1}\left\{\Delta_2(\xi,\theta_1)\right\} + \pi_{02} \mathrm{E}_{\theta_2}\left\{\Delta_1(\xi,\theta_2)\right\}$ and $\bar{j} =$ not j.

Hence, in the full problem, the aim of the experiment is to maximize the expected non–centrality parameter of the false model, the expectation being taken over models and over the prior distributions of the parameters. Therefore, one can interpret such a design as the result of a compromise process over the regions in which the prior distributions of θ_1 and θ_2 are defined.

For all the above criteria to be appropriate we must assume that the models are not nested. For if one model is nested within the other, one or both of the non–centrality parameters will be zero. A discussion of designs for nested models is given by Atkinson (1972) and Atkinson & Fedorov (1975a). If one model is nested within the other, a possibility for recovering the present framework is to impose constraints on the values of the parameters, which ensure that the models are separate.

Alternatively, a natural approach to this problem is to adopt the criterion of D_s–Optimality, where the parameters of interest are those belonging to the full model but absent in the rescrited one. Designs obtained in such a way would provide estimates of the extra parameteres with minimum variance, and therefore, would assist the purpose of discriminating between the two nested models.

In the next section we investigate properties of optimum designs ξ^* which maximize $\Delta_{\bar{j}}(\xi)$, $\gamma_{\bar{j}}(\xi)$ or $\Gamma(\xi)$.

3.4. AN EXTENSION OF THE GENERAL EQUIVALENCE THEOREM TO BAYESIAN T-OPTIMUM DESIGNS

The most important benefit which arises from the Bayesian approach is that optimal designs no longer depend on specific values of the parameters of the true model but only on their prior distributions. Once expectations are taken over these priors, standard optimization techniques may be used to determine the optimum design. Although the complexity involved in the search increases considerably, the problems are still tractable in terms of the computational burden.

An advantage of optimum design theory is that we are able to check whether or not the design produced by the standard optimization methods is T–optimum. This tool makes the task of searching for optimum solutions much simpler than it usually is. Often, in a typical optimization problem a local optimum emerges after the search as the actual optimum solution of the problem, although there is no guarantee that it indeed is. Fortunately, this is not the case in the present context where if such local optima are found they will be certainly discovered and then discarded. Ultimately, a new search for the global optimum starts and the checking procedure is repeated until the global optimum is determined.

In order to establish procedures for checking the optimality of a proposed Bayesian T–optimum design, it is necessary to extend some of Atkinson and Fedorov's results on non–Bayesian designs, which are based on the same theoretical results that lead to the proof of the celebrated General Equivalence Theorem of Kiefer and Wolfowitz (1960).

First let us introduce the following notation for the minimization problems whose solutions are given by the noncentrality parameters (3.3.1.a) and (3.3.1.b). For j = 1,2 ; $\bar{j}$ = not j ; $\xi \in \mathscr{H}$; and $\theta_j \in \Theta_j$ denote

$$\int_{\mathcal{X}} \left\{\eta_j(x,\theta_j) - \eta_{\bar{j}}(x,\theta^{\dagger}_{\bar{j}})\right\}^2 \xi(dx)$$

$$= \inf_{\theta_{\bar{j}}\in\Theta_{\bar{j}}} \int_{\mathcal{X}} \left\{\eta_j(x,\theta_j) - \eta_{\bar{j}}(x,\theta_{\bar{j}})\right\}^2 \xi(dx) \qquad (3.4.1)$$

Now, when $\xi = \xi^*$ assume that (3.4.1) has unique solutions, denoted by θ^*_1, over all $\theta_2 \in \Theta_2$, relevant to the prior $p_{02}(\theta_2)$, and θ^*_2 over all $\theta_1 \in \Theta_1$, relevant to the prior $p_{01}(\theta_1)$. Then, we may state the following theorem :

THEOREM 3.1

(i) a necessary and sufficient condition for a design ξ^* to be Bayesian T–optimum is fulfilment of the inequality

$$\psi(x,\xi^*) \leq \Gamma(\xi^*)\text{ , for all } x \in \mathcal{X}$$

where $\psi(x,\xi^*) = \sum_j \pi_{0j} E_{\theta_j}\left\{\eta_j(x,\theta_j) - \eta_{\bar{j}}(x,\theta^*_{\bar{j}})\right\}^2$;

(ii) at the points of the Bayesian T–optimum design $\psi(x,\xi^*)$ achieves its upper bound ;

(iii) for any non–optimum design ξ, that is a design for which $\Gamma(\xi) < \Gamma(\xi^*)$,

$$\sup_{x\in\mathcal{X}} \psi(x,\xi) > \Gamma(\xi^*) \text{ ;}$$

(iv) the set of Bayesian T–optimum designs is convex.

Conditions (i) and (ii) can be used to check whether or not a supposed Bayesian T–optimum design is indeed optimum. If, for one of the two models, the parameter estimates are not unique, but only belong to some set, the theorem can be suitably modified as in Theorem 1 of Atkinson & Fedorov (1975a). Otherwise the results remain the same. Such a situation might occur when a p parameter model is false and can be disproved by a T–optimum design on p+1 points. If the alternative model contains p+2 or more parameters, these parameters will not be uniquely estimable. In practice the problem does not arise as, in the numerical construction of the design, any singular information matrix will be regularized by the addition of a small multiple of the identity matrix. Theorem 3.1 in the form given here applies to the regularized problem. The main steps required to prove Theorem 3.1 are given in Appendix A.

In the next section numerical examples of locally, partially Bayesian and fully Bayesian T–optimum designs are presented to illustrate the theory. Concommitantly, a sensitivity analysis is carried out so as to assess the effect of prior distributions upon the structure of the resulting T–optimum designs.

3.5. EXAMPLES

The example we analyse is the same as Atkinson & Fedorov (1975a) present using a non–Bayesian approach. It consists of designing an experiment over the interval $\mathcal{X} = [-1,1]$ to discriminate between the two regression models :

$$\text{1st model : } \quad \eta_1(x,\theta_1) = \theta_{10} + \theta_{11}e^x + \theta_{12}e^{-x}$$

$$\text{2nd model : } \quad \eta_2(x,\theta_2) = \theta_{20} + \theta_{21}x + \theta_{22}x^2$$

Such a choice of models is justified by their similarity in the design region considered, so that problems of model specification could arise.

The purpose of the following series of examples is to illustrate the theory: parameter values and prior probabilities are chosen to this end. To learn more about partially and fully Bayesian T–optimum designs a rather informal sensitivity analysis is carried out in the examples. We start with a very simple case.

EXAMPLE 3.1. Suppose that the first model is true and its parameters are known. Their values are $\theta_{10} = 2.0$, $\theta_{11} = -2.0$ and $\theta_{12} = -1.0$. Thus, we should search for a design satisfying (3.3.1.b), i.e. a design which is optimum to this specific set of parameters, therefore, locally T–optimum.

The resulting design is shown below. The notation is that the design support points are in the first line whereas their weights are in the second. The maximized value of the criterion function $\Delta_2(\xi)$ is $\Delta_2(\xi^*) = 2.478 \times 10^{-3}$.

$$\xi^* = \begin{Bmatrix} -1.0 & -0.2961 & 0.6231 & 1.0 \\ 0.0987 & 0.2690 & 0.4013 & 0.2310 \end{Bmatrix} \qquad (3.5.1)$$

There are some interesting features in design (3.5.1). Firstly, the first and third points, and therefore the second and fourth, share half of the weight. Secondly, even though the classical upper limit for the number of support points in the optimum design, given by Carathéodory's Theorem to be p(p + 1)/2, where p is the number of unknown parameters, has not been proven to hold in this context, in the present example such a limit is obeyed as there are three unknown parameters and four support points in the optimum design.

Figure 3.1.1 shows the plot of the derivative function over the design region. Because the peaks occur at the support points of the design, condition (ii) of Theorem 3.1 is satisfied. Condition (i) is also verified, i.e. the maximized value of the criterion function is the upper bound of the derivative function over the design region. Hence, design (3.5.1) is the actual locally T–optimum design.

Further analysis of Figure 3.1.1 reveals that the troughs occur at the points of the design region providing the least information about discriminating between the models. In the present example, all three of them have null derivative functions, which means that the fitted values of the expected responses of model 2 coincide with the expected responses for model 1 at these points, where the fitted values are provided by fitting model 2 to the expected responses of model 1 at the points of the optimum design.

Continuing this example we now ilustrate the effect of replacing the hypothesis of knowing the true values of θ_{10}, θ_{11} and θ_{12} by that of knowledge of a prior distribution for θ_1. Initially a rather concentrated discrete prior distribution is considered as shown in Table 3.1. Our objective is to find a design that attains condition (3.3.3).

Each set of parameters in the prior distribution, beginning with the second, represents a vertex of a cube centered at the point (2.0,–2.0,–1.0), the set of true values of the parameters in the first part of the example. All vertices are placed so that the volume of the cube is small, reflecting concentrated prior information. Each vertex has the same probability whereas the centre point is twice as likely. Such a distribution of points in the parameter space causes only small variations on the expected responses for model 1 from one set of parameters to another, so that one should expect the partially Bayesian T–optimum design for the prior of Table 3.1 to be very similar to design (3.5.1).

TABLE 3.1 – CONCENTRATED 9–POINT PRIOR PROBABILITY DISTRIBUTION FOR θ_1

θ_{10}	θ_{11}	θ_{12}	$p_{01}(\theta_1)$
2.0	–2.0	–1.0	0.2
1.9	–2.1	–0.9	0.1
1.9	–2.1	–1.1	0.1
1.9	–1.9	–0.9	0.1
1.9	–1.9	–1.1	0.1
2.1	–2.1	–0.9	0.1
2.1	–2.1	–1.1	0.1
2.1	–1.9	–0.9	0.1
2.1	–1.9	–1.1	0.1

The resulting partially Bayesian T–optimum design is shown below. The maximized value of the criterion function $\gamma_2(\xi)$ is $\gamma_2(\xi^*) = 2.504 \times 10^{-3}$. Figure 3.1.2 shows the plot of the derivative function.

$$\xi^* = \left\{\begin{matrix} -1.0 & -0.2984 & 0.6224 & 1.0 \\ 0.0993 & 0.2695 & 0.4007 & 0.2305 \end{matrix}\right\} \qquad (3.5.2)$$

Again not only is the number of support points in the optimum design equal to four but also the first and third support points share 50% of the weight. A

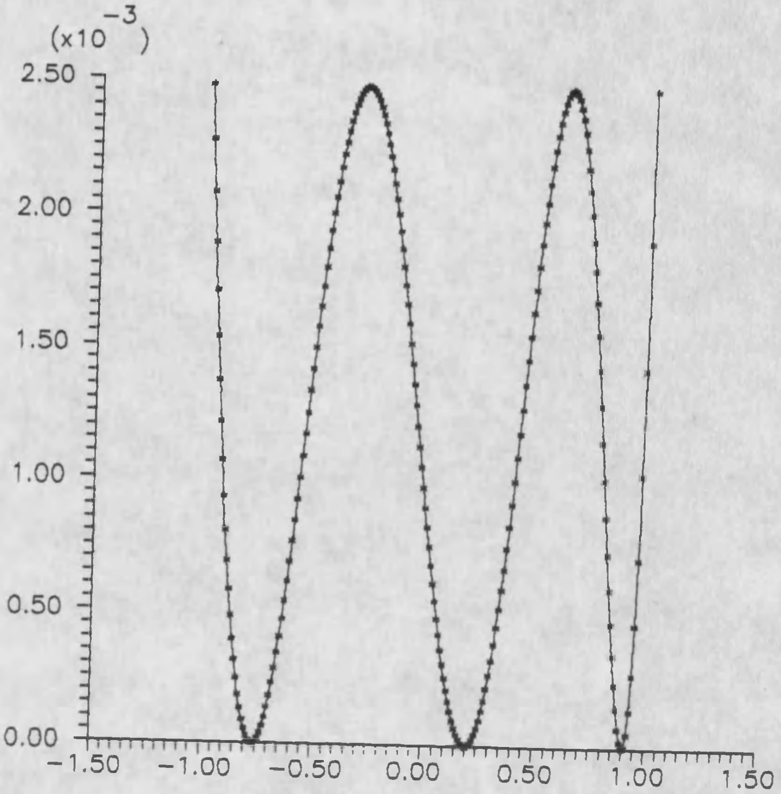


Figure 3.1.1. Derivative Function for design (3.5.1)

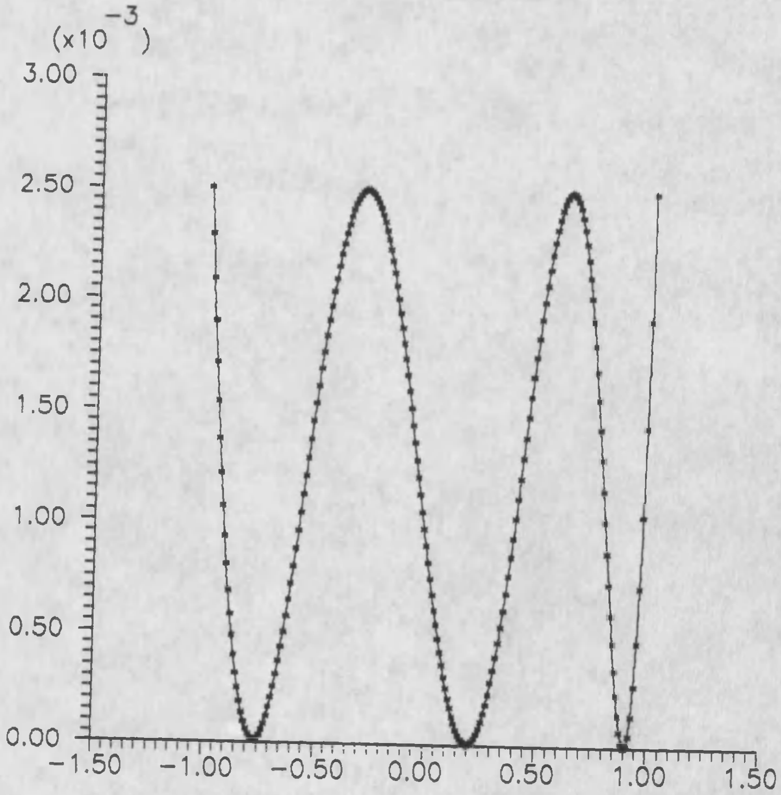


Figure 3.1.2. Derivative Function for design (3.5.2)

direct comparison between designs (3.5.1) and (3.5.2) reveals great similarity, as expected. Furthermore, the maximized values of the criteria are almost the same.

Intuitively, Bayesian optimum designs would require more support points as the uncertainty in the values of the parameters increases. A reason for this occurring is that more support points might be necessary to discriminate between the models in order to compensate for the lack of precision in the prior information. Indeed, recent results in this area appear to point in this direction. Chaloner and Larntz (1989), using uniform priors for the parameters of the logistic model for binary data, found that the number of support points of the Bayesian optimum design for parameter estimation increases if uniform priors based on wider intervals are considered. More recently, Atkinson (1992) showed an example of non–linear model discrimination in which the number of support points in the Bayesian T–optimum design varies from two for a one point prior to five for a four point equiprobable prior.

Therefore, to verify whether this phenomenon also occurs in the present example of model discrimination the sensitivity analysis initiated in the second part of Example 3.1 is carried out as follows; two dispersed priors distributed around the same point (2.0,–2.0,–1.0) in the parameter space Θ_1 are considered, the new partially T–optimum designs are determined and compared. This is the subject of Example 3.2.

EXAMPLE 3.2 (continuation of Example 3.1). Let us still assume that model 1 is true and its parameters are not precisely known. Instead, a prior distribution for their values holds as shown in Table 3.2. The aim is to search for a partially Bayesian T–optimum design, i.e. one satisfying condition (3.3.3).

As in the second part of the previous example, the last eight parameter sets of the prior displayed in Table 3.2 form a cube in Θ_1 whose centre is (2.0,–2.0,–1.0). The difference being that now the vertices lie far apart from one another so that we should expect the partially Bayesian T–optimum design to be dissimilar to the previous optimum designs (3.5.1) and (3.5.2). A rise in the number of design support points is also expected to occur. For comparison purposes, the structure of the prior probabilities of Table 3.1 is maintained, i.e. the cube centre is twice as likely as the vertices.

TABLE 3.2 – DISPERSED 9–POINT PRIOR PROBABILITY DISTRIBUTION FOR θ_1

θ_{10}	θ_{11}	θ_{12}	$p_{01}(\theta_1)$
2.0	–2.0	–1.0	0.2
–3.0	–7.0	–6.0	0.1
–3.0	–7.0	4.0	0.1
–3.0	3.0	–6.0	0.1
–3.0	3.0	4.0	0.1
7.0	–7.0	–6.0	0.1
7.0	–7.0	4.0	0.1
7.0	3.0	–6.0	0.1
7.0	3.0	4.0	0.1

The resulting design is shown below for which the maximized value of the criterion function $\gamma_2(\xi)$ is $\gamma_2(\xi^*) = 8.063 \times 10^{-2}$.

$$\xi^* = \begin{Bmatrix} -1.0 & -0.5014 & 0.5109 & 1.0 \\ 0.1659 & 0.3298 & 0.3341 & 0.1702 \end{Bmatrix} \qquad (3.5.3)$$

Surprisingly, there is no rise in the number of support points of design (3.5.3) as expected. The pattern displayed by designs (3.5.1) and (3.5.2) that the first and third points share half of the weight is again exhibited here. Another feature of design (3.5.3) is its quasi–symmetry in the design region considered. This is complemented by the almost symmetric distribution of the design weights. Each extreme point has approximately one sixth of the weight whereas the middle points have almost one third each.

Figure 3.2.1, showing the derivative function of design (3.5.3), reveals not only the above mentioned symmetry but also that the most non–informative points no longer have null derivative functions. This is a consequence of different parameter sets in Θ_1 having different sets of non–informative points. Thus, for a point in the design region to have a null derivative function the expected responses of model 1 would have to coincide with the predicted values of model 2 at such point, for all parameter sets of the prior distribution. As this is unlikely to occur the appearance of such points in Bayesian designs is rare.

Finally, a comparison between the values of the criterion functions for the partially T–optimum designs (3.5.2) and (3.5.3) shows that for the more dispersed prior the value rose approximately by a factor of 32, indicating a much greater average lack of fit for model 2 than for the more concentrated prior. This is probably due to the polynomial regression model (model 2) being "not so close" to the double exponencial model in some of the regions of Θ_1 considered in Table 3.2.

The above remark gives rise to further investigation. We now consider a third prior distribution for the parameters of model 1, in which there are twenty–five equiprobable points, distributed in the form of three cubes in Θ_1, all centered on the point (2.0,–2.0,–1.0), so that the cube of largest volume contains that of second largest volume, which in turn contains that of smallest volume. Table 3.3 shows the values of θ_1 for this prior. All three cubes are not only larger in volume than that of Table 3.1 and smaller than that of Table 3.2 but also they contain and are contained, respectively, by those.

The purpose in extending Example 3.2 in this way is to investigate both whether there will be a rise in the number of support points because of the increasing number of points in the prior and whether the value of the criterion function will inflate as a consequence of the polynomial model being "more distant" from the double exponential model in different regions of Θ_1.

TABLE 3.3 – EQUIPROBABLE DISPERSED 25–POINT PRIOR PROBABILITY DISTRIBUTION FOR $\boldsymbol{\theta}_1$

	$\boldsymbol{\theta}_1 = (\theta_{10},\theta_{11},\theta_{12})$	
	(2.0,–2.0,–1.0)	
(1.0,–3.0,–2.0)	(0.0,–4.0,–3.0)	(–1.0,–5.0,–4.0)
(1.0,–3.0, 0.0)	(0.0,–4.0, 1.0)	(–1.0,–5.0, 2.0)
(1.0,–1.0,–2.0)	(0.0, 0.0,–3.0)	(–1.0, 1.0,–4.0)
(1.0,–1.0, 0.0)	(0.0, 0.0, 1.0)	(–1.0, 1.0, 2.0)
(3.0,–3.0,–2.0)	(4.0,–4.0,–3.0)	(5.0,–5.0,–4.0)
(3.0,–3.0, 0.0)	(4.0,–4.0, 1.0)	(5.0,–5.0, 2.0)
(3.0,–1.0,–2.0)	(4.0, 0.0,–3.0)	(5.0, 1.0,–4.0)
(3.0,–1.0, 0.0)	(4.0, 0.0, 1.0)	(5.0, 1.0, 2.0)

The value of the criterion function $\gamma_2(\xi)$ at the resulting partially Bayesian T–optimum design displayed below is $\gamma_2(\xi^*) = 2.344 \times 10^{-2}$.

$$\xi^* = \begin{Bmatrix} -1.0 & -0.4835 & 0.5269 & 1.0 \\ 0.1582 & 0.3223 & 0.3418 & 0.1777 \end{Bmatrix} \qquad (3.5.4)$$

Only four support points are required in the optimum design (3.5.4) despite the 25–point dispersed prior distribution. Additionally, design (3.5.4) shows some symmetry although not as much as exhibited by design (3.5.3). The plot of the derivative function is shown in Figure 3.2.2. The interpretation given to Figure 3.2.1 is also suitable here.

The value of $\gamma_2(\xi^*)$ is about a factor of three smaller than that for the more dispersed prior of Table 3.2 which confirms that, in the present example, the region of Θ_1 influences the value of the criterion function. If the individual noncentrality parameters for each feasible $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ were available we could locate

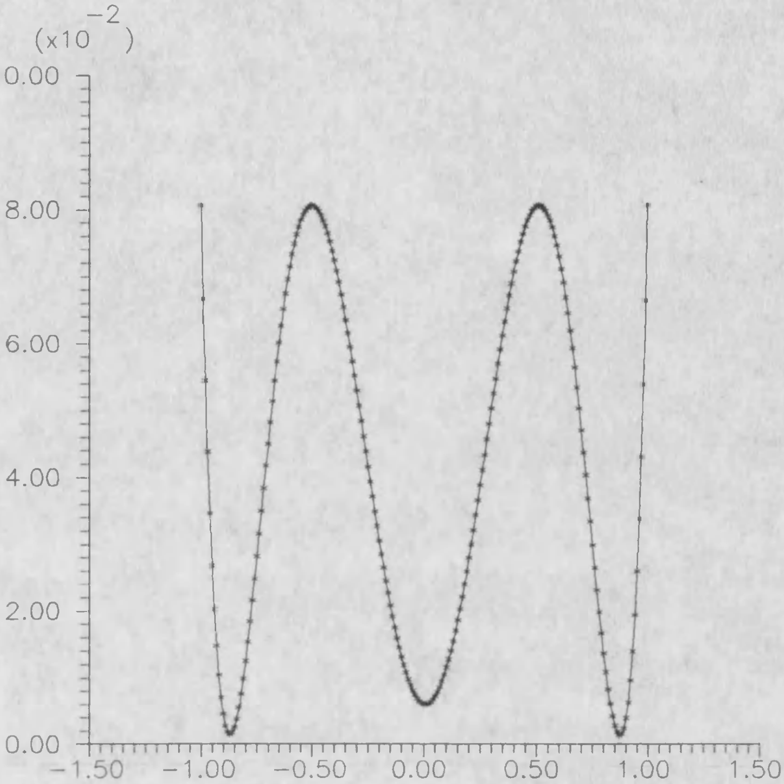


Figure 3.2.1. Derivative Function for design (3.5.3)

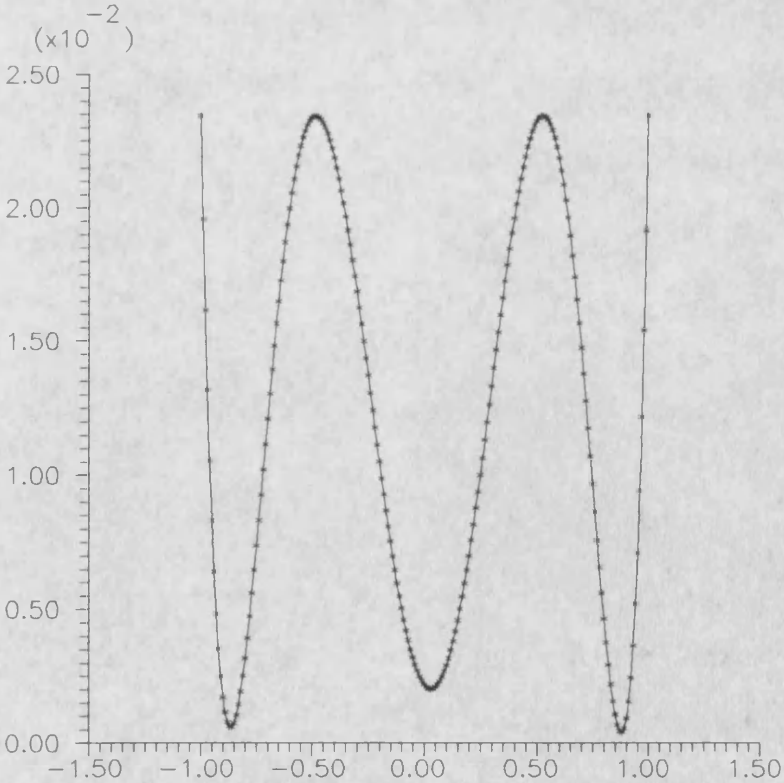


Figure 3.2.2. Derivative Function for design (3.5.4)

the parameter sets whose contribution to inflate the criterion function were larger. Such analysis would not require any extra computation, for the noncentrality parameter values are required to determine the value of the criterion function anyway.

Looking at all the results of Examples 3.1 and 3.2 a summary of the features of the locally and partially Bayesian T–optimum designs is as follows : (i) the number of support points in the optimum design is robust w.r.t. the dispersion of points in the prior; (ii) there is a pattern in the weights of all optimum designs that consists of the first and third support points, and therefore the second and the fourth, sharing 50% of the weight; (iii) there seems to be a tendency for optimum designs to be more symmetric when the underlying prior distribution is more dispersed; (iv) the value of the criterion function increases with the dispersion of the underlying prior.

Obviously, these features must not be generalized to other examples of model discrimination nor even to other regions of Θ_1 using the same models examined here. For instance, Ponce de Leon & Atkinson (1991) obtained a partially Bayesian T–optimum design with five support points using the same models, the same assumption that the first model is true, and a 10–point concentrated prior. The only difference was that a different region of Θ_1 underlined the prior distribution.

Now we present an example of a fully Bayesian T–optimum design, i.e. a design attaining condition (3.3.4).

EXAMPLE 3.3 – Suppose that the true model is unknown but there are prior probabilities for each one of the two models to be true. Conditional on these probabilities, there are prior probability distributions for the parameters of each model. Four cases, consisting of all combinations of one concentrated and one dispersed prior distribution for each set of parameters, are analysed. The prior distributions are displayed in Tables 3.4.1 and 3.4.2.

TABLE 3.4.1 – CONCENTRATED PRIOR DISTRIBUTIONS FOR θ_1 AND θ_2

θ_{10}	θ_{11}	θ_{12}	$p_{01}(\theta_1)$	θ_{20}	θ_{21}	θ_{22}	$p_{02}(\theta_2)$
5.0	–2.5	–3.0	0.2	0.0	0.5	–3.0	0.2
4.9	–2.6	–3.1	0.1	–0.1	0.4	–3.1	0.1
4.9	–2.6	–2.9	0.1	–0.1	0.4	–2.9	0.1
4.9	–2.4	–3.1	0.1	–0.1	0.6	–3.1	0.1
4.9	–2.4	–2.9	0.1	–0.1	0.6	–2.9	0.1
5.1	–2.6	–3.1	0.1	0.1	0.4	–3.1	0.1
5.1	–2.6	–2.9	0.1	0.1	0.4	–2.9	0.1
5.1	–2.4	–3.1	0.1	0.1	0.6	–3.1	0.1
5.1	–2.4	–2.9	0.1	0.1	0.6	–2.9	0.1

TABLE 3.4.2 – DISPERSED PRIOR DISTRIBUTIONS FOR θ_1 AND θ_2

θ_{10}	θ_{11}	θ_{12}	$p_{01}(\theta_1)$	θ_{20}	θ_{21}	θ_{22}	$p_{02}(\theta_2)$
5.0	–2.5	–3.0	0.2	0.0	0.5	–3.0	0.2
4.0	–3.5	–4.0	0.1	–1.0	–0.5	–4.0	0.1
4.0	–3.5	–2.0	0.1	–1.0	–0.5	–2.0	0.1
4.0	–1.5	–4.0	0.1	–1.0	1.5	–4.0	0.1
4.0	–1.5	–2.0	0.1	–1.0	1.5	–2.0	0.1
6.0	–3.5	–4.0	0.1	1.0	–0.5	–4.0	0.1
6.0	–3.5	–2.0	0.1	1.0	–0.5	–2.0	0.1
6.0	–1.5	–4.0	0.1	1.0	1.5	–4.0	0.1
6.0	–1.5	–2.0	0.1	1.0	1.5	–2.0	0.1

Analogously to Examples 3.1 and 3.2, each one of the four above prior distributions consists of parameter sets θ_1 and θ_2 forming cubes around two central points, namely $\theta_1 = (5.0,-2.5,-3.0)$ and $\theta_2 = (0.0,0.5,-3.0)$. As before, the vertices are equiprobable whereas the centre is twice as likely. For comparative purposes, the models have equal prior probabilities $\pi_{01} = \pi_{02} = 0.5$ of being the true model for all four combinations of priors. Each cube has either a small or a large volume reflecting a concentrated or a dispersed prior. A similar sensitivity analysis is carried out in Ponce de Leon & Atkinson (1991).

The aim of this example is to illustrate the effects on the structure of fully Bayesian T–optimum designs of considering mixtures of concentrated and dispersed prior distributions for θ_1 and θ_2. The results are shown in Table 3.5.

TABLE 3.5 – FULLY BAYESIAN T–OPTIMUM DESIGNS FOR FOUR DIFFERENT SETS OF PRIOR DISTRIBUTIONS FOR θ_1 AND θ_2

DESIGNS (3.5.5), (3.5.6), (3.5.7), AND (3.5.8)

PRIORS	VALUE OF $\Gamma(\xi^*)$	BAYESIAN T–OPTIMUM DESIGN
(a) both concentrated	1.787×10^{-3}	$\left\{\begin{matrix} -1.0 & -0.6849 & 0.0886 & 0.8624 \\ 0.2468 & 0.4314 & 0.2532 & 0.0686 \end{matrix}\right\}$
(b) θ_1 concentrated θ_2 dispersed	2.107×10^{-3}	$\left\{\begin{matrix} -1.0 & -0.673 & 0.1298 & 0.873 & 1.0 \\ 0.2433 & 0.4293 & 0.2545 & 0.0456 & 0.0273 \end{matrix}\right\}$
(c) θ_1 dispersed θ_2 concentrated	2.702×10^{-3}	$\left\{\begin{matrix} -1.0 & -0.6476 & 0.2182 & 1.0 \\ 0.2411 & 0.421 & 0.2589 & 0.079 \end{matrix}\right\}$
(d) both dispersed	3.169×10^{-3}	$\left\{\begin{matrix} -1.0 & -0.6329 & 0.2664 & 1.0 \\ 0.2315 & 0.4119 & 0.2685 & 0.0881 \end{matrix}\right\}$

All designs have similarities. For instance, all of them contain the support point –1.0 with approximately the same weight. In addition, the second and third support point weights are similar in all designs, the third support point varies significantly over designs whereas there is a reasonably small variation on the second. Most importantly, however, is that design (3.5.6), corresponding to a concentrated prior for θ_1 and a dispersed prior for θ_2 , has one more support point than the other designs even though it should be noted that its last two support

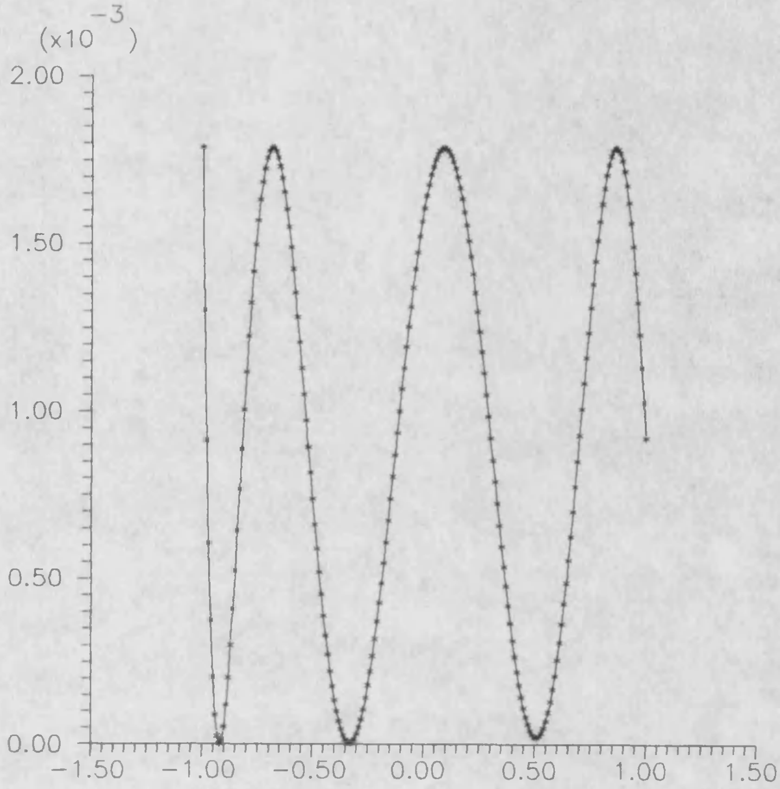


Figure 3.3.1. Derivative Function for design (3.5.5)

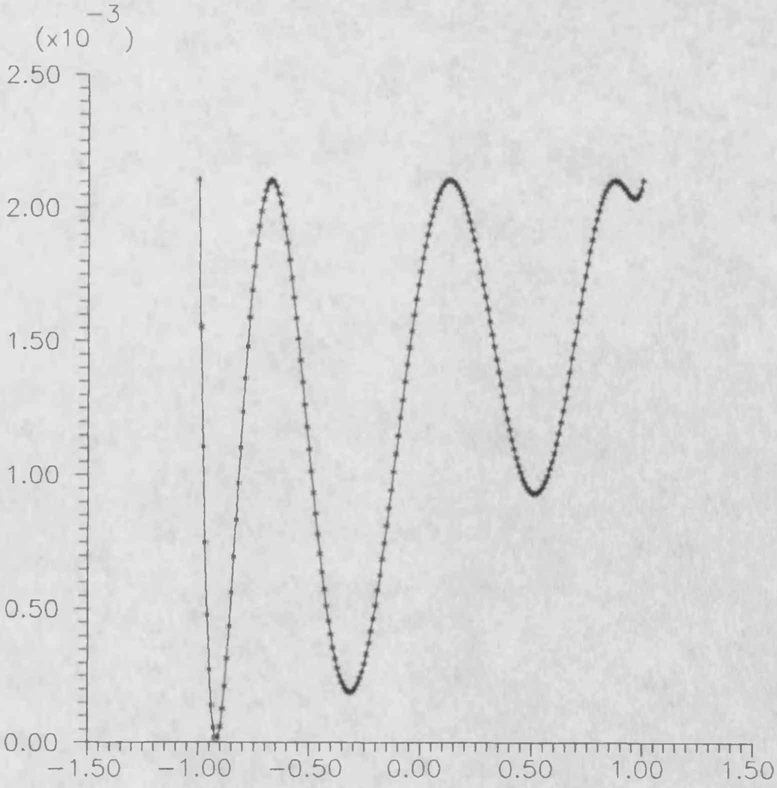


Figure 3.3.2. Derivative Function for design (3.5.6)

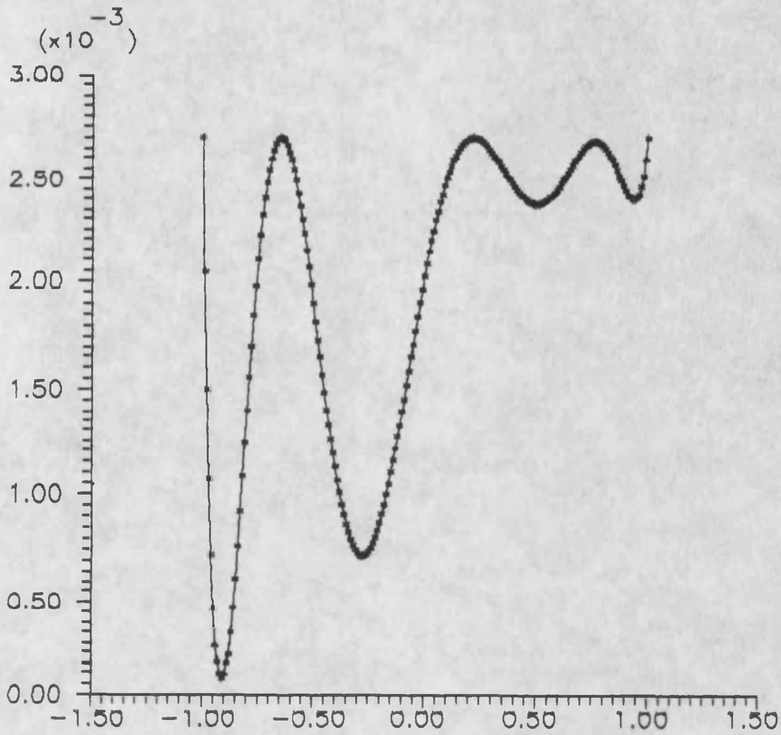


Figure 3.3.3. Derivative Function for design (3.5.7)

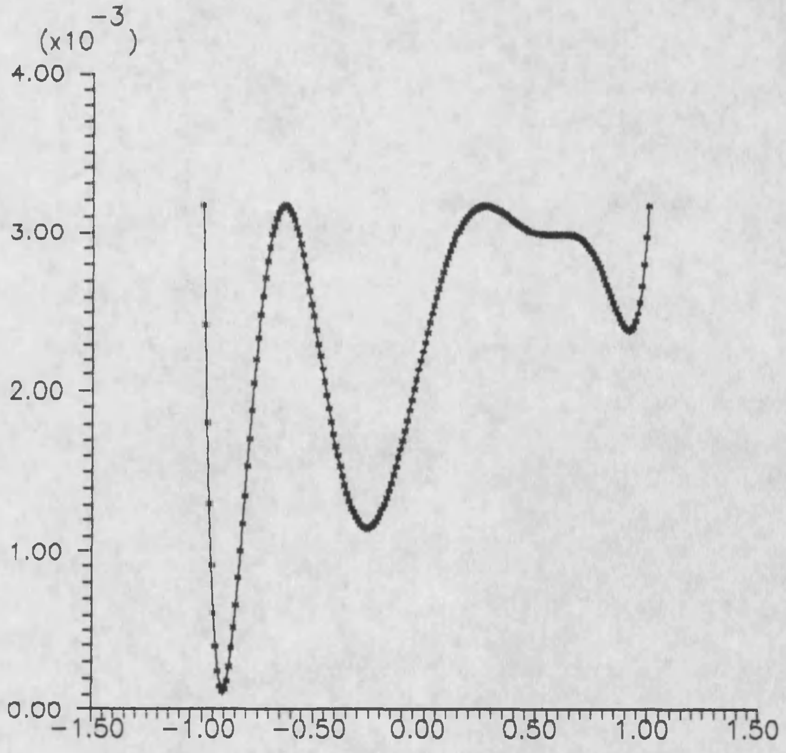


Figure 3.3.4. Derivative Function for design (3.5.8)

points weigh almost as much as the fourth support point (7.29% against 6.86%, 7.9% and 8.81%) of the remaining designs. All four–point designs in Table 3.5 present the feature that the first and third support points share 50% of the weight.

Nevertheless, the fact that design (3.5.8), corresponding to both dispersed priors, has four support points rather than five leads to the conclusion that the number of support points in fully Bayesian T–optimum designs is not influenced only by the dispersion of the underlying priors. Otherwise design (3.5.8) would not have fewer points than design (3.5.6). Perhaps, the factor that has greater influence in determining the number of support points in the optimum design is the region of the parameter spaces Θ_1 and Θ_2 in which the priors are defined.

A final remark about the designs of Table 3.5 concerns the maximized value of the criterion function. Here again, the tendency observed in partially Bayesian T–optimum designs appears to occur, that is the more dispersed the prior distributions the larger the maximized value of the criterion function. Again, an analysis based on the noncentrality parameters can be carried out to locate the most influencial parameter sets.

Figures 3.3.1, 3.3.2, 3.3.3 and 3.3.4 show the derivative functions for the designs from Table 3.5. The shape of Figure 3.3.1 resembles that of a locally T–optimum design in the sense that the derivative functions for the noninformative points are zero or approximately zero. This is due to both priors for θ_1 and θ_2 being concentrated. Figure 3.3.2 shows the only five–point design, two points of which are very close with a shallow valley in between. One might say that all points in this subregion are informative as they almost achieve the value of the maximized criterion function. Figure 3.3.3 is slightly misleading since it looks as though there are five peaks achieving the upper limit, but a close examination of the fourth value of x reveals that its derivative function does not reach the maximized value $\Gamma(\xi^*) = 2.702 \times 10^{-3}$. Finally, the shapes of Figures 3.3.3 and 3.3.4 show more markedly the effect of dispersed priors, i.e. in the presence of prior variability no single point will yield uninformative experiments. As a result the derivative function will be appreciably flatter. Chaloner & Larntz (1989) & Atkinson (1992) also show similarly shaped plots for the related derivative functions.

3.6. DISCUSSION

Carathéodory's Theorem provides an upper bound for the number of support points in the optimum continuous design for a single model (Silvey, 1980, Chapter 3 and Appendix 2). However there is no such limit for Bayesian optimum designs for model discrimination. Lack of this limit complicates the search for optimum designs although, in the example studied here, the number of support points is only one or two more than the number of parameters in the individual models. However in non–linear model discrimination between two one–parameter models with $\mathcal{X} = (0,\infty)$, Atkinson (1992) shows an example in which the number of support points increases from two to five as the prior distribution becomes more dispersed. The relative stability of the design in our example is caused by the bounded experimental region.

An alternative to the designs of this chapter would be sequential designs in which observations from earlier experiments are used to update the prior probabilities of the parameters and models. After each updating, the optimum setting for the next experiment would be found by numerical search. Such an approach would have particular advantages when there was great uncertainty in the prior information.

We have assumed independence of the prior distributions between models. It might be more realistic, in some examples, to consider priors which give equal weights to parameter values yielding similarly shaped response curves under the two models.

The examples in this chapter have assumed discrete joint prior distributions within models. The case of continuous prior distributions would involve no new ideas, but would have to be solved using numerical integration techniques, in themselves a form of discretization. Another question of interest about prior distributions is the assumption of independence among prior distributions of individual parameters within models, an assumption made, for example, by Chaloner & Larntz (1989) in their Bayesian designs for single models.

The results in this chapter are for discriminating between two models. An obvious extension is to designs for discriminating between three or more models. Atkinson & Fedorov (1975b) present one solution for locally optimum designs.

APPENDIX A

The first requirement to prove Theorem 3.1 is to verify whether criterion function (3.3.4) is a concave function on $\mathscr{H}$, that is

$$\Gamma(\xi) \geq (1-\alpha)\,\Gamma(\xi_1) + \alpha\,\Gamma(\xi_2) \tag{A.1}$$

where $\alpha \in \mathbb{R}$, $0 \leq \alpha \leq 1$ and $\xi = (1-\alpha)\,\xi_1 + \alpha\,\xi_2$; $\xi_1, \xi_2 \in \mathscr{H}$.

Due to the linearity of $\Gamma(\xi)$ with respect to the noncentrality parameters $\Delta_1(\xi,\theta_2)$ and $\Delta_2(\xi,\theta_1)$ given by (3.3.1.a) and (3.3.1.b) respectively, and their concavity on $\mathscr{H}$ it follows that (A.1) is true as demonstrated below.

Proof of (A.1)

By definition $\Gamma(\xi) = E_{\pi_0} E_{\theta_j} \left\{\Delta_{\bar{j}}(\xi,\theta_j)\right\}$, where

E_{π_0} denotes expectation over prior probabilities for the model truth given by the vector $\pi_0 = (\pi_{01},\pi_{02})$, $\pi_{02} = 1 - \pi_{01}$;

E_{θ_j} denotes expectation over the prior distribution for θ_j, given by $p_{0j}(\theta_j)$; and

$\Delta_{\bar{j}}(\xi,\theta_j) = \inf\limits_{\theta_{\bar{j}} \in \Theta_{\bar{j}}} \int_{\mathscr{X}} \left\{\eta_j(x,\theta_j) - \eta_{\bar{j}}(x,\theta_{\bar{j}})\right\}^2 \xi(dx)$; j=1,2 and $\bar{j}$ = not j.

To simplify the notation let us represent the sum of squares (integral) shown above by a function S, i.e. for j = 1,2 ; $\bar{j}$ = not j ; $\xi \in \mathscr{H}$; $\theta_j \in \Theta_j$; and $\theta_{\bar{j}} \in \Theta_{\bar{j}}$ denote

$$S(\xi,\theta_j,\theta_{\bar{j}}) = \int_{\mathscr{X}} \left\{\eta_j(x,\theta_j) - \eta_{\bar{j}}(x,\theta_{\bar{j}})\right\}^2 \xi(dx) \tag{A.2}$$

Therefore, given a design ξ and the value of θ_j we can combine equations (3.4.1) and (A.2) so as to rewrite the related noncentrality parameter $\Delta_{\bar{j}}(\xi,\theta_j)$ as

$$\Delta_{\bar{j}}(\xi,\theta_j) = S(\xi,\theta_j,\theta_{\bar{j}}^{\dagger}) = \inf_{\theta_{\bar{j}} \in \Theta_{\bar{j}}} S(\xi,\theta_j,\theta_{\bar{j}}).$$

Now, using this notation let us first prove concavity of $\Delta_{\bar{j}}(\xi,\theta_j)$ on $\mathscr{H}$. For any $\alpha \in \mathbb{R}$, $0 \le \alpha \le 1$; $\xi_1,\xi_2 \in \mathscr{H}$; let $\xi = (1-\alpha)\,\xi_1 + \alpha\,\xi_2$. Then

$$\begin{aligned}
\Delta_{\bar{j}}(\xi,\theta_j) &= \inf_{\theta_{\bar{j}}\in\Theta_{\bar{j}}} S(\xi,\theta_j,\theta_{\bar{j}}) = \inf_{\theta_{\bar{j}}\in\Theta_{\bar{j}}} S((1-\alpha)\xi_1+\alpha\xi_2,\theta_j,\theta_{\bar{j}}) \\
&= \inf_{\theta_{\bar{j}}\in\Theta_{\bar{j}}} \left\{(1-\alpha)\, S(\xi_1,\theta_j,\theta_{\bar{j}}) + \alpha\, S(\xi_2,\theta_j,\theta_{\bar{j}})\right\} \\
&\ge \inf_{\theta_{\bar{j}}\in\Theta_{\bar{j}}} \left\{(1-\alpha)\, S(\xi_1,\theta_j,\theta_{\bar{j}})\right\} + \inf_{\theta_{\bar{j}}\in\Theta_{\bar{j}}} \left\{\alpha\, S(\xi_2,\theta_j,\theta_{\bar{j}})\right\} \\
&= (1-\alpha)\,\Delta_{\bar{j}}(\xi_1,\theta_j) + \alpha\,\Delta_{\bar{j}}(\xi_2,\theta_j)
\end{aligned} \tag{A.3}$$

Thus, according to (A.3) $\Delta_{\bar{j}}(\xi,\theta_j)$ is concave on $\mathscr{H}$. Now, taking expectations over the prior probability distributions π_0 and $\{p_{0j}(\theta_j),\ j=1,2\}$ in both sides of inequality (A.3), we find

$$E_{\pi_0}E_{\theta_j}\left\{\Delta_{\bar{j}}(\xi,\theta_j)\right\} \ge E_{\pi_0}E_{\theta_j}\left\{(1-\alpha)\,\Delta_{\bar{j}}(\xi_1,\theta_j) + \alpha\,\Delta_{\bar{j}}(\xi_2,\theta_j)\right\}$$

$$\Gamma(\xi) \ge (1-\alpha)\,\Gamma(\xi_1) + \alpha\,\Gamma(\xi_2).$$

Hence, (A.1) is true; $\Gamma(\xi)$ is a concave function on $\mathscr{H}$.

The second essential requirement is to find the Fréchet derivative of $\Gamma(\xi)$, which is straightforwardly obtained due to the linearity of $\Gamma(\xi)$ w.r.t. the noncentrality parameters $\Delta_1(\xi,\theta_2)$ and $\Delta_2(\xi,\theta_1)$. It is given by

$$\frac{\partial\Gamma(\xi)}{\partial\alpha} = \sum_j \pi_{0j}\, E_{\theta_j}\left[\frac{\partial\Delta_{\bar{j}}(\xi,\theta_j)}{\partial\alpha}\right] \tag{A.4}$$

where again $\xi = (1-\alpha)\,\xi_1 + \alpha\,\xi_2$; $0 \le \alpha \le 1$; and for $j = 1,2$; $\bar{j}$ = not j

$$\frac{\partial\Delta_{\bar{j}}(\xi,\theta_j)}{\partial\alpha} = \inf_{\theta_{\bar{j}}\in\Theta^{\dagger}_{\bar{j}}(\alpha)} \left\{S(\xi_2,\theta_j,\theta_{\bar{j}}) - S(\xi_1,\theta_j,\theta_{\bar{j}})\right\} \tag{A.5}$$

where for a given $\alpha \in [0,1]$ $\Theta_{j}^{\dagger}(\alpha)$ denotes the solution set of equation (3.4.1).

Proof of (A.4)

Again, the notation is simplified as in (A.2). Before determining the Fréchet derivative of $\Gamma(\xi)$ it is convenient to find that of $\Delta_{\bar{j}}(\xi,\theta_j)$, i.e. we shall prove (A.5) and then (A.4).

For any $\alpha \in [0,1]$; $\xi_1,\xi_2 \in \mathscr{H}$; and $\xi = (1-\alpha)\,\xi_1 + \alpha\,\xi_2$ let $\Theta^{\dagger}_{\bar{j}}(\alpha)$ be the solution set of equation (3.4.1). Following the notation introduced above $\Theta^{\dagger}_{\bar{j}}(\alpha)$ is the solution set of the equation

$$S(\xi,\theta_j,\theta^{\dagger}_{\bar{j}}) = \inf_{\theta_{\bar{j}} \in \Theta_{\bar{j}}} S(\xi,\theta_j,\theta_{\bar{j}})$$

Denote the Fréchet derivative of $\Delta_{\bar{j}}(\xi,\theta_j)$, usually represented by $F_{\Delta_{\bar{j}}}(\xi_1,\xi_2)$, by $\dfrac{\partial \Delta_{\bar{j}}(\xi,\theta_j)}{\partial\alpha}$. Then

$$\frac{\partial \Delta_{\bar{j}}(\xi,\theta_j)}{\partial\alpha} = \frac{\partial}{\partial\alpha}\left\{\inf_{\theta_{\bar{j}} \in \Theta^{\dagger}_{\bar{j}}(\alpha)} S(\xi,\theta_j,\theta_{\bar{j}})\right\} = \inf_{\theta_{\bar{j}} \in \Theta^{\dagger}_{\bar{j}}(\alpha)} \left\{\frac{\partial}{\partial\alpha} S(\xi,\theta_j,\theta_{\bar{j}})\right\}$$

$$= \inf_{\theta_{\bar{j}} \in \Theta^{\dagger}_{\bar{j}}(\alpha)} \left[\lim_{\alpha \to 0^+} \frac{1}{\alpha}\left\{S((1-\alpha)\xi_1 + \alpha\xi_2,\theta_j,\theta_{\bar{j}}) - S(\xi_1,\theta_j,\theta_{\bar{j}})\right\}\right]$$

$$= \inf_{\theta_{\bar{j}} \in \Theta^{\dagger}_{\bar{j}}(\alpha)} \left\{S(\xi_2,\theta_j,\theta_{\bar{j}}) - S(\xi_1,\theta_j,\theta_{\bar{j}})\right\}.$$

where the first equality holds by applying Pshenichnyi (1971, Theorem 3.2, pp. 75). Hence, (A.5) is true. Now, suppose that the optimum design ξ^* is such that (3.4.1) has a unique solution, denoted by $\theta^*_{\bar{j}}$, when $\xi = \xi^*$ ($\xi \to \xi_1$ when $\alpha \to 0$). Then (A.5) becomes

$$F_{\Delta_{\bar{j}}}(\xi^*,\xi_2) = S(\xi_2,\theta_j,\theta^*_{\bar{j}}) - S(\xi^*,\theta_j,\theta^*_{\bar{j}})$$

$$= S(\xi_2,\theta_j,\theta^*_{\bar{j}}) - \Delta_{\bar{j}}(\xi^*,\theta_j)$$

where ξ_2 represents any design other than ξ^*. In particular when $\xi_2 = \xi_x$, where ξ_x represents the design putting all mass at the design region point $x \in \mathcal{X}$, the following result holds. For any $x \in \mathcal{X}$

$$F_{\Delta_j}(\xi^*,\xi_x) = \left\{\eta_j(x,\theta_j) - \eta_{\bar{j}}(x,\theta^*_{\bar{j}})\right\}^2 - \Delta_{\bar{j}}(\xi^*,\theta_j) \qquad \text{(A.6)}$$

Finally, applying Theorems 2.1 and 2.2, we find

$$\begin{aligned} F_{\Delta_j}(\xi^*,\xi_x) &= \psi(x,\xi^*,\theta_j) - \Delta_{\bar{j}}(\xi^*,\theta_j) \leq 0 \\ &\Longrightarrow \psi(x,\xi^*,\theta_j) \leq \Delta_{\bar{j}}(\xi^*,\theta_j)\text{ , any } x \in \mathcal{X} \end{aligned} \qquad \text{(A.7)}$$

where $\psi(x,\xi^*,\theta_j) = \left\{\eta_j(x,\theta_j) - \eta_{\bar{j}}(x,\theta^*_{\bar{j}})\right\}^2$.

However, (A.7) is condition (i) of Theorem 3.1 for a local T–optimum design. Condition (ii) follows by applying Corollary 2.1. Finally conditions (iii) and (iv) are now straightfoward.

Finally, to prove Theorem 3.1 for fully Bayesian T–optimum designs, we have to assume that at the optimum design ξ^*, (3.4.1) has unique solutions $\theta^*_{\bar{j}}$ for every $\theta_j \in \Theta_j$, relevant to the prior $p_{0j}(\theta_j)$, $\{j=1,2\}$. Now, taking expectations of (A.6) over the priors π_0 and $p_{0j}(\theta_j)$, j=1,2 and applying Theorems 2.1 and 2.2 again, we find condition (i) for this more general situation, i.e.

$$\psi(x,\xi^*) \leq \Gamma(\xi^*)\text{ , any } x \in \mathcal{X}$$

where $\psi(x,\xi^*) = E_{\pi_0}E_{\theta_j}\left\{\eta_j(x,\theta_j) - \eta_j(x,\theta^*_{\bar{j}})\right\}^2$.

Now, the other conditions follow easily.

CHAPTER 4. OPTIMUM EXPERIMENTAL DESIGN FOR DISCRIMINATING BETWEEN TWO RIVAL BINARY DATA MODELS

4.1. INTRODUCTION

In this chapter we develop methodology for finding optimum experimental designs to discriminate between two rival binary data models (also known in the literature as binary response or quantal models). The class of binary response models forms a well known subclass of generalized linear models. Therefore, the framework of generalized linear models is used to introduce the relevant concepts needed for our methodology to model discrimination, such as the criterion of optimality, which is based on the same principle as the criterion of T–optimality for regression model discrimination, and the model structural specifications. Generally speaking, our purpose is to find a design for which the false model fits the true model expected responses as badly as possible. Optimum design theory is again applied so as to derive important properties of the optimum designs yielded by this criterion. Later, we explain how they can be used as powerful tools for the numerical procedures leading to optimality. The designs which maximize the expected deviance of the false model are similar in structure to the local, partially Bayesian and fully Bayesian T–optimum designs for regression models of chapter 3. Initially, we concentrate on optimal designs for the relatively simple problem when the true model and its parameter values are known. The extension to prior distributions of parameter values and of the truth of each model is straightforward.

Although such an extension introduces numerical difficulties the design criterion remains concave and the results of optimum design theory apply. In practice, due to the complexity involved in the optimization of the criterion functions, numerical procedures must be adopted to find the optimum. The difficulties encountered at this stage are the main challenge of the whole problem.

4.2. BACKGROUND

Chambers and Cox (1967) use asymptotic arguments to find optimum designs for discriminating between binary models which require only three points of support. Their results are based on the problem of discriminating between the logistic and integrated normal (probit) binary response curves. They consider significance tests in which one of the two models is taken to be the null hypothesis and a design maximizing power when the alternative is the other model is found. Yanagisawa (1988) extended their method not only to designs requiring any number of support points but also to a range of different binary response curves. Optimal Bayesian designs for estimating functions of the parameters, as well as the parameters themselves, of logistic binary response models are described by Chaloner and Larntz (1989).

Optimum experimental designs for discriminating between two rival regression models were developed by Atkinson and Fedorov (1975a). When one model is true and there are prior probabilities for its parameters, their T–optimality criterion can be extended, as shown by Ponce de Leon and Atkinson (1991), with further details in chapter 3 of this thesis, to experimental designs that maximize the expected value of the residual sum of squares of the false model. Designs attaining this criterion are called partially Bayesian T–optimum. As well as T–optimality Bayesian T–optimality is also a concave design criterion and the standard results of optimum design theory apply. However, a partially Bayesian T–optimum design still depends upon which is the true model, therefore, is still only locally optimum. The dependence on specific prior values of parameters can be removed by taking the expectation of the criterion over the prior distribution of the parameter values. Similarly, one can take the expectation over the prior probabilities for the model truth, thus replacing dependence on the model truth by dependence on its prior probability distribution. In chapter 3 it is shown that the resulting fully Bayesian T–optimum designs are again such that optimum design theory applies.

Here we extend all these results to designs for discriminating between models for binary data by using the correspondence that exists between the two problems. Preliminary results on this subject are presented in Ponce de Leon and Atkinson (1992a) together with some examples of local T–optimum designs to discriminate the complementary log–log model against the logistic and the probit model against the logistic. These models are presented in Section 4.3. In generalized linear models the analogue of the residual sum of squares is the deviance: an optimality criterion for binary data model discrimination based upon this concept is introduced in Section 4.4. In Chapter 3 the notions of locally, partially Bayesian and fully Bayesian optimum designs were introduced. In Section 4.4 they are adapted to the present problem and used throughout Chapter 4. A Theorem derived from the General Equivalence Theorem is presented in Section 4.5 along with a discussion about its proof. Numerical examples of the optimum designs are given in Section 4.6. A brief discussion is presented in Section 4.7.

4.3. BASIC FRAMEWORK

Generalized Linear Models for binary data are described by McCullagh and Nelder (1989, Chapter 4). In this section, we give a brief description of this important class of models. Suppose that for any combination of levels of p explanatory variables a response Y_i is observed, where the index i denotes the particular combination of levels. The response Y_i can take one of two values 0 or 1, corresponding to "failure" or "success", with the probability of success $P(Y_i = 1) = \pi_i$ where $0 \leq \pi_i \leq 1$. The levels of the p explanatory variables at which the response is observed are represented by the vector $\mathbf{x}_i = (x_{i1},\ldots,x_{ip})^t$. Interest is in the relationship between π_i and $\mathbf{x}_i$. In the generalized linear model this relationship depends upon the linear predictor η_i, where $\eta_i = \eta(\mathbf{x}_i) = \sum_j x_{ij}\theta_j$ and $\{\theta_j\}$ are p unknown parameters. The probability π_i depends on η_i, and so on $\mathbf{x}_i$, through the link function g(.) such that $\eta_i = g(\pi_i)$. In the specific case of binary response models, g(.) must be a function that maps the interval [0,1] onto the real line. Three widely used link functions for binary data are given in Table 4.1. Examples of designs discriminating between pairs of these three link functions are given in Section 4.6.

TABLE 4.1 – EXAMPLES OF LINK FUNCTIONS FOR BINARY RESPONSE MODELS

	TERMINOLOGY	LINK FUNCTION
(a)	logit or logistic	$g(\pi_i) = \log\{\pi_i / (1 - \pi_i)\}$
(b)	probit or inverse normal	$g(\pi_i) = \Phi^{-1}(\pi_i)$
(c)	complementary log-log	$g(\pi_i) = \log\{-\log(1 - \pi_i)\}$

Note : $\Phi(.)$ denotes the standard normal cumulative distribution function.

An excellent feature of the problem of designing optimum experiments for discriminating between two binary response models concerns the manner in which the frameworks of optimal design theory and generalized linear models combine in these conditions.

Recall that a typical experimental design consists of n weighted support points belonging to the design region $\mathcal{X}$. For binary response models, each of the n support points $\{x_1,...,x_n\}$ is independently replicated m_i times, giving rise to a set of n binomial responses $\{Y_1,...,Y_n\}$. Let $N = \sum_i m_i$ be the sample size. Then the weight at the ith support point of the design is $p_i = m_i/N$ so that the design can be represented by a discrete measure ξ_N over the experimental region $\mathcal{X}$. The set of responses $\{Y_1,...,Y_n\}$ forms a random sample of n independent Binomials (m_i,π_i), $\{i=1,\cdots,n\}$. Removal of the restriction that m_i be an integer, as usual, yields the continuous or approximate design measure ξ.

The purpose of the experiments derived in this chapter is to determine which of the link functions provides an adequate model. To be more precise, our interest lies in finding an optimal design to discriminate between two binary data models whatever their linear predictor specifications are. Here models may have different link functions and either the same or different linear predictor structure or the same link function with different linear predictor structure. For nested models the values of the linear predictor parameters must be restricted.

Chambers and Cox (1967) proposed a criterion of optimality to discriminate between link functions (i) and (ii). Yanagisawa (1988) extended Chambers and Cox's discrimination criterion to all categories of binary response models. The method is illustrated with designs discriminating between pairs of a set of five different link functions, namely logit, probit, Aranda–Ordaz (1981), exponential, and gamma models. An alternative for optimum designs to discriminate between two or more binary response models is to consider families of link functions with an extra set of parameters to provide member identification and then search for a design that estimates the extra parameter(s) as accurately as possible. In chapter 7, one of these families of link functions is introduced and optimal designs are found for estimating the only extra parameter it contains.

In the next section a criterion of optimality to design experiments for discriminating between two binary response models in various situations, as far as prior information is concerned, is introduced.

4.4. OPTIMALITY CRITERIA

The deviance for a fitted binary model (McCullagh and Nelder, 1989, p. 118) depends on the observations y_i and on the estimated probabilities $\hat{\pi}_i$. Its purpose is to provide a measure of goodness of fit for specific models to the data and its interpretation is similar to the residual sum of squares in regression theory. Indeed, for Gaussian models, a subclass of generalized linear models, the deviance reduces to the residual sum of squares. Therefore, we can use the analogy between deviance and residual sum of squares to define a criterion of optimality for designing experiments to discriminate between pairs of binary response models. The required steps for this are described next.

Without loss of generality, take the first model as true. Since the link function and linear predictor structure of both models are supposed to be known, the expected responses for the first model, π_1, are functions of its linear predictor parameters θ_1 and of the design ξ. Hence, for a given set of linear predictor parameters θ_1 one can generate the expected responses under the first model and subsequently fit the second model to this artificial "data" set. The estimates of the second model linear predictor parameters θ_2, and consequently the probabilities π_2, are then functions of the design ξ.

The estimates $\hat{\theta}_2$ can be determined by iterative weighted least squares, a method described by McCullagh and Nelder (1989). For a given θ_1 the estimates $\hat{\theta}_2$ and $\hat{\pi}_2$ depend solely on the design ξ which generated the first model expected responses π_1 . Clearly some designs will provide less adequate "fits" than others. This leads to the idea of using the concept of deviance, a measure of goodness of fit in generalized linear models, to define a criterion of optimality for binary response model discrimination. First an important concept for this criterion is introduced. For binary data models the analogues of the non–centrality parameters are given by Definition 4.1.

Definition 4.1. Let π_1 and π_2 denote the expected responses, or probabilities, yielded by two rival binary response models. For the ith model suppose that the probability $\pi_i = h_i(x,\theta_i) = g_i^{-1}(f_i(x,\theta_i))$, where $f_i(.,.)$ and $g_i(.)$ are known functions and the inverse of the latter exists so that $h_i(.)$ is well defined. Then the quantities

$$\Delta_1(\xi,\theta_2) = \inf_{\theta_1 \in \Theta_1} \int_{\mathcal{X}} \left\{ \pi_2 \log\left[\frac{\pi_2}{\pi_1}\right] + (1-\pi_2)\log\left[\frac{1-\pi_2}{1-\pi_1}\right]\right\} \xi(dx) \qquad (4.4.1.a)$$

and

$$\Delta_2(\xi,\theta_1) = \inf_{\theta_2 \in \Theta_2} \int_{\mathcal{X}} \left\{ \pi_1 \log\left[\frac{\pi_1}{\pi_2}\right] + (1-\pi_1)\log\left[\frac{1-\pi_1}{1-\pi_2}\right]\right\} \xi(dx) \qquad (4.4.1.b)$$

are the binary discrimination totals for the first model when the second is true, and vice versa.

The function $g_i(.)$, as in generalized linear models, should not only map the interval (0,1) onto the whole real line but also be monotone and differentiable. The function $f_i(.,.)$ is assumed to be linear in the parameters θ_i even though the models that can be discriminated using the proposed methodology, could in principle have nonlinear predictors. Linearity is assumed due to lack of efficient methods to fit models whose predictors η_i are nonlinear functions of θ_i. Thus, we assume that the models to be discriminated are restricted to the class of generalized linear models for binary data.

The analogy between a noncentrality parameter in the context of regression models and a binary discrimination total is obvious. Both depend on the design ξ and on the unknown values of the respective set of parameters. The former can be described as the residual sum of squares for the false model in the absence of experimental error whereas the latter is the deviance for the false model in the absence of randomness. Both can be interpreted as a measure of the "distance" between two rival models.

In practice, however, they cannot be computed unless the true model and the values of its parameters are known, which seldom is the case. Nevertheless, if such strong assumptions are made we can solve the rather simple resulting problem and use the methodology as a starting point for the understanding of the complexities of the full problem. For instance, in the simplified problem the "distance" between two rival binary response models is only a function of the design ξ so that it is straightforward to define a criterion of discrimination based on the binary discrimination totals. Later, of course, these assumptions are replaced by knowledge of prior probability distributions. This strategy for developing the main ideas of binary data model discrimination is used to define all the criteria in this section.

We start with the simplest situation, although with the largest number of assumptions, to discriminate between two binary data models.

Definition 4.2. Consider two rival binary response models. Assume that the jth model is true. Furthermore, assume that its set of linear predictor parameters θ_j is known. Then, a *local optimum design* ξ^* to discriminate between the rival models is such that

$$\Delta_{\bar{j}}(\xi^*) = \sup_{\xi \in \mathscr{H}} \Delta_{\bar{j}}(\xi) \tag{4.4.2}$$

where

$$\Delta_{\bar{j}}(\xi) = \Delta_{\bar{j}}(\xi, \theta_j),\ j = 1,2 \text{ and } \bar{j} = \text{not } j.$$

Local optimum designs for binary response model discrimination are not as easy to obtain as are the locally T–optimum designs of Chapter 3. For the widely used binary response curves the cause for numerical difficulties can be explained by the curves being similarly shaped and also agreeing closely, especially for values of the probability of success far from the extremes 0 and 1. Therefore, the actual observed values of the criterion function are rather small, requiring highly accurate search procedures. For instance, as Chambers and Cox (1967) pointed out, the logistic and integrated normal binary response curves agree closely except in the tails making the models virtually indistinguishible. Cox (1966b) gives a numerical comparison between these curves. There are other numerical problems in finding locally optimum designs that will be commented later. For the moment, it is important to point out that locally optimum designs are limited because in practice the true model and its parameters are invariably unknown.

A more general criterion can be defined by modifying the set of hypotheses to knowledge of the true model and a prior distribution for its set of parameters.

Definition 4.3. Consider two rival binary data models. Assume that the jth model is true and that the only information about θ_j is the prior distribution $p_{0j}(\theta_j)$. Then, a *partially Bayesian optimum design* ξ^* to discriminate between the models is such that

$$\gamma_{\bar{j}}(\xi^*) = \sup_{\xi \in \mathscr{H}} \gamma_{\bar{j}}(\xi) \qquad (4.4.3)$$

where

$$\gamma_{\bar{j}}(\xi) = E_{\theta_j}\left\{\Delta_{\bar{j}}(\xi,\theta_j)\right\}, \ j = 1,2 \text{ and } \bar{j} = \text{not } j.$$

Partially Bayesian optimum designs represent a significant improvement with respect to the previous class of locally optimum designs. Generally speaking, a partially Bayesian optimum design is the result of a compromise among all plausible parameter values. Among other consequences of definition 4.3, one is that the more likely a region of the parameter space the more influential it will be in the resulting optimum design. Therefore, one should expect that an optimum design produced by (4.4.3) will discriminate between the models more adequately for parameter values in more likely regions of the parameter space. Similarly, for parameter values in less likely regions a partially Bayesian optimum design might be misleading as far as model discrimination is concerned. Based on these remarks, we can draw the additional conclusion that misspecification of the prior distribution might cause serious consequences to the actual optimality of the resulting design.

Likewise, the region itself plays a vital role in this process since optimum designs might be more sensitive to changes in the parameter values in certain regions of the parameter space than in others. Hence, it is reasonable to say that the more dispersed the prior distribution the more likely a partially Bayesian optimum design will be misleading for parameter values in more unlikely regions of the parameter space. Certainly the complexity of these relationships will vary from one particular problem to another and therefore can not be modelled entirely.

The motivation for the Bayesian extension is that in practice it will not be known which model is true, nor will the parameter values of the true model be known precisely. It is however likely that information is available which can be incorporated into prior probability distributions. Hence, suppose that there are prior probability distributions for each set of linear predictor parameters and for the truth of the model. These are denoted by

π_{0j} → probability of truth of the jth model; j=1,2.

$p_{0j}(\theta_j)$ → joint probability distribution for the jth set of linear predictor parameters.

Definition 4.4. Consider two rival binary data models. Let π_{0j} be as above and conditioned on the event that the jth model is true, let $p_{0j}(\theta_j)$ be the prior distribution for the parameters of the jth model. Then, a *fully Bayesian optimum design* ξ^* is such that

$$\Gamma(\xi^*) = \sup_{\xi \in \mathscr{H}} \Gamma(\xi) \tag{4.4.4}$$

where

$$\Gamma(\xi) = \sum_j \pi_{0j} E_{\theta_j}\left\{\Delta_j(\xi,\theta_j)\right\} = \pi_{01} E_{\theta_1}\left\{\Delta_2(\xi,\theta_1)\right\} + \pi_{02} E_{\theta_2}\left\{\Delta_1(\xi,\theta_2)\right\}.$$

It is now straightforward to prove a theorem analogous to that of Chapter 3, Section 3.4, using the same arguments. The search for fully Bayesian optimal designs to discriminate between binary data models is then possible, despite the increased amount of computing which may well raise numerical problems. These are discussed in the next sections.

4.5. CONDITIONS FOR OPTIMALITY

According to the results originated from optimal design theory not until have we proven that all the criterion functions introduced in the last section are concave on the class of design measures $\mathscr{H}$, and derived expressions for their Fréchet derivatives, we will be in a position to state a theorem characterizing optimal designs. It is easy to prove that analogous results to those found for the problem of regression model discrimination in Chapter 3 also hold in the present context.

Proof of the first result, concavity of the criterion functions, is simplified by the fact that criteria (4.4.2) and (4.4.3) are particular cases of criterion (4.4.4) so that it is enough to prove concavity of the criterion function $\Gamma(\xi)$ and consequently, concavity of the other criterion functions will follow. Further, $\Gamma(\xi)$ is a linear function of the binary discrimination totals (4.4.1.a) and (4.4.1.b). Indeed, it is the result of applying the linear operator E, expected value, over the prior distributions for parameters and model truth. Therefore, to prove concavity of $\Gamma(\xi)$ on $\mathscr{H}$ it is enough to prove concavity of the binary discrimination totals, for concavity is inherited by linear combinations of concave functions, as shown in Appendix A. Details of the proof for binary models are presented in Appendix B. The second requirement, expressions for the directional derivatives or Fréchet derivatives of the criterion functions, is also derived in Appendix B.

Similarly to the notation used in Chapter 3 let us consider the following optimization problem whose solutions are given by the binary discrimination totals (4.4.1.a) and (4.4.1.b). For any $j = 1,2$; $\bar{j}$ = not j ; $\xi \in \mathscr{H}$; and $\theta_j \in \Theta_j$ denote

$$\int_{\mathfrak{X}} \left\{ \pi_j \log\left[\frac{\pi_j}{\pi^{\dagger}_{\bar{j}}}\right] + (1-\pi_j)\log\left[\frac{1-\pi_j}{1-\pi^{\dagger}_{\bar{j}}}\right]\right\} \xi(dx)$$
$$= \inf_{\theta_{\bar{j}} \in \Theta_{\bar{j}}} \int_{\mathfrak{X}} \left\{ \pi_j \log\left[\frac{\pi_j}{\pi_{\bar{j}}}\right] + (1-\pi_j)\log\left[\frac{1-\pi_j}{1-\pi_{\bar{j}}}\right]\right\} \xi(dx) \qquad (4.5.1)$$

where $\pi_j = \pi_j(\mathbf{x},\theta_j)$, $\mathbf{x} \in \mathfrak{X}$.

The additional assumption that the optimum design is regular in the sense that when $\xi = \xi^*$ equation (4.5.1) has unique solutions, denoted by $\theta^*_{\bar{j}}$, over all $\theta_j \in \Theta_j$ which are relevant to the prior $p_{0j}(\theta_j)$; $\{j=1,2\}$, makes each result proved in Section 3.4 for discriminating between regression models have its analogue in the case of binary response models. Given the preceding conditions one can prove the following theorem.

THEOREM 4.1

(i) a necessary and sufficient condition for a design ξ^* to be optimum w.r.t. criterion (4.4.4) is fulfilment of the inequality

$$\psi(\mathbf{x},\xi^*) \leq \Gamma(\xi^*), \text{ for all } \mathbf{x} \in \mathfrak{X}$$

where $\psi(\mathbf{x},\xi^*) = \sum_j \pi_{0j} E_{\theta_j} \left\{ \pi_j \log\left[\frac{\pi_j}{\pi^*_{\bar{j}}}\right] + (1-\pi_j)\log\left[\frac{1-\pi_j}{1-\pi^*_{\bar{j}}}\right]\right\}$;

(ii) at the points of the optimum design $\psi(\mathbf{x},\xi^*)$ achieves its upper bound;

(iii) for any nonoptimum design, that is a design for which $\Gamma(\xi) < \Gamma(\xi^*)$,

$$\sup_{\mathbf{x}\in\mathfrak{X}} \psi(\mathbf{x},\xi) > \Gamma(\xi^*);$$

(iv) the set of optimum designs w.r.t (4.4.4) is convex.

In (i), the estimates $\pi^*_{\bar{j}}$ denote the predicted values yielded by the $\bar{j}$th model to estimate the expected responses under the jth model, based on the support points of the optimum design ξ^*. These estimates are obtained by iterative weighted least squares. For more transparency in the notation recall that π_j denotes $\pi_j(\mathbf{x},\theta_j)$ and $\pi^*_{\bar{j}}$ denotes $\pi_{\bar{j}}(\mathbf{x},\theta^*_{\bar{j}})$, where $\theta^*_{\bar{j}}$ is the solution of (4.5.1), when $\xi = \xi^*$.

Conditions (i) and (ii) lead to the customary check for the optimality of a candidate design. If a design ξ^0 is optimum w.r.t. (4.4.4) the plot of $\psi(\mathbf{x},\xi^0)$ over the design region $\mathcal{X}$ will reveal that $\sup_{\mathbf{x}\in\mathcal{X}} \psi(\mathbf{x},\xi^0) = \Gamma(\xi^0)$, equality ocurring at all $\mathbf{x}\in\mathcal{X}$ which are the support points of the design. If $\sup_{\mathbf{x}\in\mathcal{X}} \psi(\mathbf{x},\xi^0) > \Gamma(\xi^0)$, the design ξ^0 is not optimal w.r.t. (4.4.4).

In the next section several examples of local, partially, and fully Bayesian T–optimum designs are presented for the problem of discriminating between pairs of widely used and well known binary response models such as the probit, the logistic, and the complementary log–log models.

4.6. EXAMPLES

In this section optimum designs are found to illustrate the methods described in the previous sections of this chapter. Two features of the search method deserve mention. The first concerns the ability of checking optimality provided by the conditions of Theorem 4.1. On the other hand, lack of theoretical results concerning the number of support points in the optimum design complicates the search process because most available algorithms require that this number be specified to carry out the search. An alternative to the latter is to try different numbers of design support points and learn as much as possible from the analysis of the derivative functions corresponding to the suboptimal designs for each different number of support points. In this process, some empirical methods may be helpful such as increasing the number of support points by adding to the design those points corresponding to high peaks in the plot of the derivative function. This situation is typical of designs with less support points than required so that influential points which have been excluded might be noticed through the plot of the derivative function. At the other extreme, if a design has more support points than necessary, the resulting suboptimal design will quite often contain duplicated support points that can be combined to form the actual optimum design.

We consider three examples for discriminating between pairs of binary response models. For each pair of models locally, partially Bayesian and fully Bayesian optimum designs are found and plots of the derivative functions displayed to ensure their optimality. Selections of prior distributions are made with the purpose of illustrating the proposed methods. Only discrete prior distributions are considered in this series of examples.

EXAMPLE 4.1. Discriminating between the complementary log–log and the normal integrated (probit) curves. Initially, suppose that the true model has the complementary log–log link function. Furthermore, assume that both models have linear predictors given by a straight line $\theta_{j0} + \theta_{j1}x$,{j=1,2} and that the design region is the real line. For the true model, the parameters θ_{10} and θ_{11} are known a priori to be equal to 0.5 and 1.0, respectively. The aim is to find an optimum design that discriminates against the rival model whose link function is probit.

The local optimum design ξ^* is given below. Criterion function (4.4.2) is maximized at the value $\Delta_2(\xi^*,\theta_1) = 1.72 \times 10^{-2}$, where $\boldsymbol{\theta}_1 = (0.5,1.0)$.

$$\xi^* = \begin{Bmatrix} -5.048 & -1.076 & 1.171 \\ 0.3851 & 0.2396 & 0.3753 \end{Bmatrix} \tag{4.6.1}$$

Only three points are required in the optimum design to discriminate between these two link functions. All associated weights are significantly large with the two extreme points being almost equally important. The middle point counts for less than 25% of the total weight. Figure (4.1.1) shows the derivative function corresponding to design (4.6.1). The pattern of the derivative function plots for locally optimum designs, first observed in the case of regression model discrimination, is repeated in Figure (4.1.1), i.e. for the most noninformative points of the design region the value of the derivative is zero. A simple examination of the expression for the derivative function shows that a zero will occur if and only if the predicted value for the false model, say $\hat{\pi}_j^-$, equals the expected response for the true model, say π_j, making the resulting curves touch at such a point.

Proceeding with example 4.1 we now investigate the effects of taking prior distributions for the parameters of the true model. Table 4.2 shows a prior distribution for the set of parameters $\boldsymbol{\theta}_1$. The bidimensional points of this prior form a square on the plane, centered on $\boldsymbol{\theta}_1 = (0.5,1.0)$, the true value of $\boldsymbol{\theta}_1$ in the first part of the example. The probabilities of these points are distributed such that the center of the square is twice as likely as the other points.

TABLE 4.2 – CONCENTRATED 9–POINT PRIOR PROBABILITY DISTRIBUTION FOR $\boldsymbol{\theta}_1$

θ_{10}	θ_{11}	$p_{01}(\boldsymbol{\theta}_1)$
0.5	1.0	0.2
0.5	0.9	0.1
0.5	1.1	0.1
0.4	1.0	0.1
0.4	0.9	0.1
0.4	1.1	0.1
0.6	1.0	0.1
0.6	0.9	0.1
0.6	1.1	0.1

Now, the appropriate criterion function to be maximized is (4.4.3). The resulting partially Bayesian optimum design is shown below. The maximized value of (4.4.3), achieved for this design is equal to 1.646×10^{-2}.

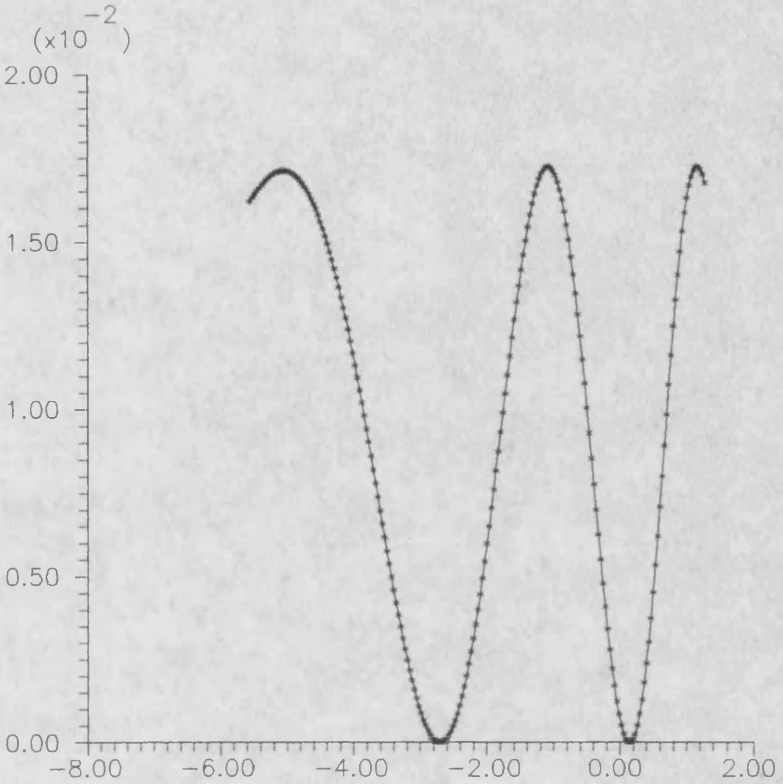


Figure 4.1.1. Derivative Function for design (4.6.1)

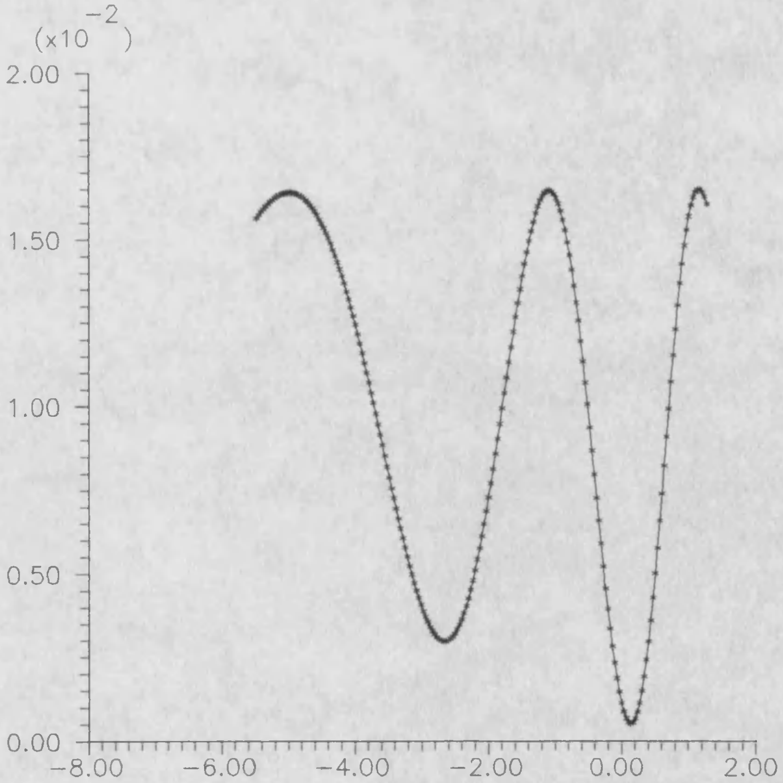


Figure 4.1.2. Derivative Function for design (4.6.2)

$$\xi^* = \begin{Bmatrix} -4.979 & -1.085 & 1.172 \\ 0.3828 & 0.2432 & 0.3741 \end{Bmatrix} \qquad (4.6.2)$$

Not surprisingly, designs (4.6.1) and (4.6.2) are very similar in structure. The support points and weights have not changed significantly from the locally optimum design to the partially Bayesian optimum one. This is due to the prior distribution for θ_1 being very concentrated around the value assumed as true in the previous example. Also expected is the shape of the derivative function for this design, shown in Figure 4.1.2. As was observed in the linear regression case, for partially Bayesian designs the most noninformative points of the design region are more informative than their counterparts of locally optimum designs such as that of Figure 4.1.1. The reason is as explained in Chapter 3, Section 3.5.

TABLE 4.3 – SLIGHTLY DISPERSED 9–POINT PRIOR PROBABILITY DISTRIBUTION FOR θ_2

θ_{20}	θ_{21}	$p_{02}(\theta_2)$
1.0	1.0	0.2
1.0	0.8	0.1
1.0	1.2	0.1
0.8	1.0	0.1
0.8	0.8	0.1
0.8	1.2	0.1
1.2	1.0	0.1
1.2	0.8	0.1
1.2	1.2	0.1

Proceeding even further with this example, let us also assume that, based on previous experiments or any other relevant source of information, prior probabilities for the models are available and, conditional on these, prior distributions for each set of parameters θ_j , {j=1,2} are also available. Tables 4.2 and 4.3 display the priors for θ_1 and θ_2 , respectively. In the prior distribution for θ_2 the bidimensional points are again arranged so that they form a square centered at the point θ_2 = (1.0,1.0). The choice of this centre point is due to the similarity between the complementary log–log and probit curves corresponding to sets of parameters θ_1 and θ_2 equal to the centre points specified by the first rows of Tables 4.2 and 4.3, respectively. The models, or links, are assumed to be true with equal probabilities.

With the amount of prior information available here, now criterion function (4.4.4) is to be maximized. The value at which the optimum design is found is $\Gamma(\xi^*) = 1.292 \times 10^{-2}$ and the fully Bayesian optimum design is given below.

$$\xi^* = \begin{Bmatrix} -4.015 & -0.896 & 1.248 \\ 0.3868 & 0.266 & 0.3472 \end{Bmatrix} \qquad (4.6.3)$$

Although the support points and weights of design (4.6.3) show some resemblance to both previous designs (4.6.1) and (4.6.2) the shape of the derivative function for this design, shown in Figure 4.1.3, is distinct from those in Figures 4.1.1 and 4.1.2. This is mainly due to the way criterion function (4.4.4) is defined, that is based on a mixture of prior distributions for the parameters of the two rival models. It is also interesting to notice that the maximized value of the criterion function for this extension of the problem is smaller than the previous observed values. This is probably due to occurrence of small values for the "deviances" when the complementary log–log model is fitted to the expected responses of the probit model.

EXAMPLE 4.2. Discriminating between the complementary log–log and the logistic (logit) curves. Suppose that the complementary log–log link is true and we wish to find an optimum design to discriminate against the logistic link based model. In addition, assume that both model linear predictors are simply described by a

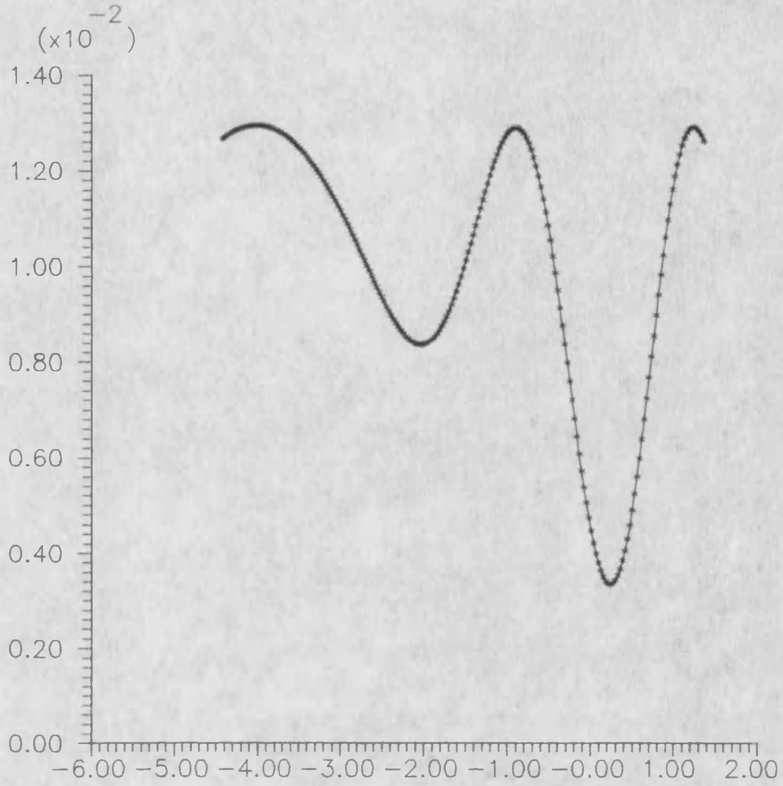


Figure 4.1.3. Derivative Function for design (4.6.3)

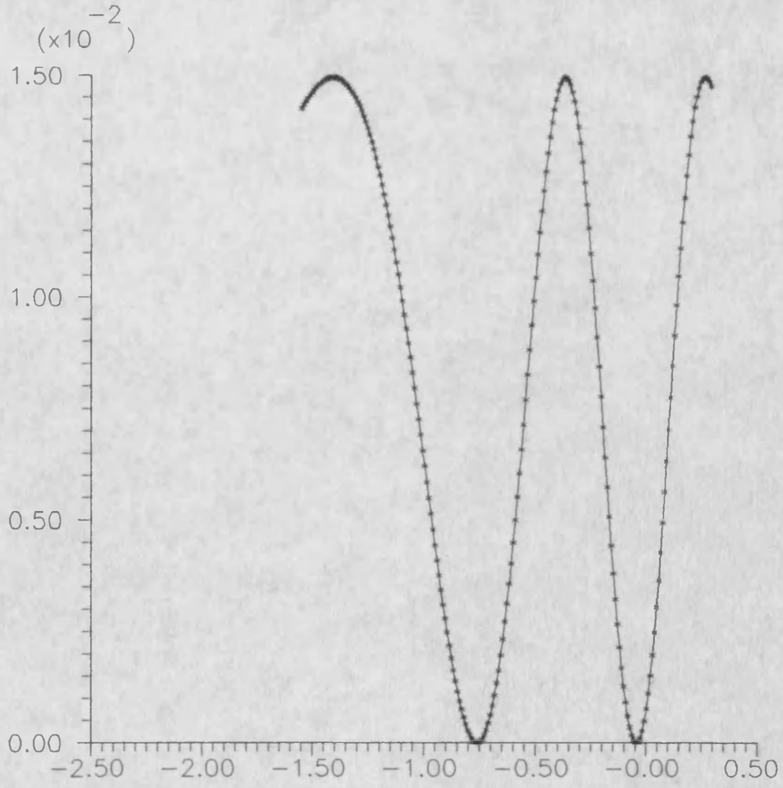


Figure 4.2.1. Derivative Function for design (4.6.4)

straight line. For j=1,2 , let $\theta_{j0} + \theta_{j1}x$ denote the linear predictors. Let the design region be the real line. Finally assume that the values of θ_{10} and θ_{11} are known to be equal to 1.0 and 3.0 respectively.

The level and amount of prior information assumed in this example are very rare in practice, but it should not be ruled out that such a situation can occur in a real problem. The criterion of optimization adopted here is criterion (4.4.2). The optimum is achieved at the value of the criterion function equal to 1.496×10^{-2}. This corresponds to the following design.

$$\xi^* = \begin{Bmatrix} -1.406 & -0.3635 & 0.2687 \\ 0.2588 & 0.2078 & 0.5334 \end{Bmatrix} \tag{4.6.4}$$

Three support points prove to be enough in order to discriminate between these two rival models. A special feature of this optimum design is that the only positive support point is assigned more than 50% of the whole weight. Figure 4.2.1 shows the derivative function for design (4.6.4). It can be seen that the most noninformative points correspond to values of the derivative equal to zero.

Continuing the present example of discriminating between the complementary log–log and logit link functions, consider the same assumptions as at the beginning of the example except knowledge of the true values of the parameters. Instead, suppose that now there is a prior probability for the parameters of the complementary log–log link based model. This prior is displayed in Table 4.4.

The bidimensional points of the above prior for θ_1 are arranged in the form of a square whose centre has coordinates $\theta_1 = (1.0,3.0)$, the true value of the parameters for the first part of this example. Again, probabilities are distributed almost uniformly, that is the centre is twice as likely as the other points in the prior. The distances from the other eight points to the centre (1.0,3.0) show a relatively large dispersion in this prior distribution. If the shapes of the curves corresponding to these nine combinations of the parameters θ_{10} and θ_{11} were to be compared, the discrepancies would be enourmous. In practice, prior distributions like that of Table 4.4 are rare.

TABLE 4.4 – DISPERSED 9–POINT PRIOR PROBABILITY DISTRIBUTION FOR θ_1

θ_{10}	θ_{11}	$p_{01}(\theta_1)$
1.0	3.0	0.2
1.0	2.5	0.1
1.0	3.5	0.1
0.5	3.0	0.1
0.5	2.5	0.1
0.5	3.5	0.1
1.5	3.0	0.1
1.5	2.5	0.1
1.5	3.5	0.1

Due to the form of the available prior information, criterion function (4.4.3) applies to this problem. Its maximized value is equal to 1.077×10^{-2} for the following optimum design.

$$\xi^* = \begin{Bmatrix} -1.373 & -0.3859 & 0.2827 \\ 0.2808 & 0.209 & 0.5103 \end{Bmatrix} \tag{4.6.5}$$

The structure of design (4.6.5) resembles that of design (4.6.4) in the sense that the first two support points in ascending order of magnitude are negative and if taken together share less than 50% of the total weight whereas the third and last point is positive and alone is responsible for the remaining 51.02%. For the purpose of checking optimality, Figure 4.2.2 shows the derivative function corresponding to design (4.6.5).

Proceeding with Example 4.2, which illustrates the search for optimum designs to discriminate between the complementary log–log and the logit links, we now consider that the true model is unknown, but that there is a prior probability for each model to be true. Conditional on these prior probabilitites, there are prior probability distributions for each of the parameter sets θ_1 and θ_2 . Under this more general situation criterion function (4.4.4) applies. The prior probabilities for θ_1 and θ_2 are displayed in Table 4.5. Again, we take the models as equiprobable.

TABLE 4.5 – 9–POINT PRIOR PROBABILITY DISTRIBUTIONS FOR PARAMETER SETS θ_1 AND θ_2

θ_{10}	θ_{11}	$p_{01}(\theta_1)$	θ_{20}	θ_{21}	$p_{02}(\theta_2)$
1.0	3.0	0.2	2.0	4.0	0.2
1.0	2.75	0.1	2.0	3.75	0.1
1.0	3.25	0.1	2.0	4.25	0.1
0.75	3.0	0.1	1.75	4.0	0.1
0.75	2.75	0.1	1.75	3.75	0.1
0.75	3.25	0.1	1.75	4.25	0.1
1.25	3.0	0.1	2.25	4.0	0.1
1.25	2.75	0.1	2.25	3.75	0.1
1.25	3.25	0.1	2.25	4.25	0.1

Each prior distribution consists of nine bidimensional points forming a square centered in $\theta_1 = (1.0,3.0)$ for the complementary log–log link and centered in $\theta_2 = (2.0,4.0)$ for the logit link. Like the previous priors, the centre is twice as likely as the other eight points. The shapes of the curves for the two parameter sets, taken as centres of the priors, are very similar.

The maximum value achieved by criterion function (4.4.4) in this example is equal to 1.29×10^{-2}. The optimum design is as follows.

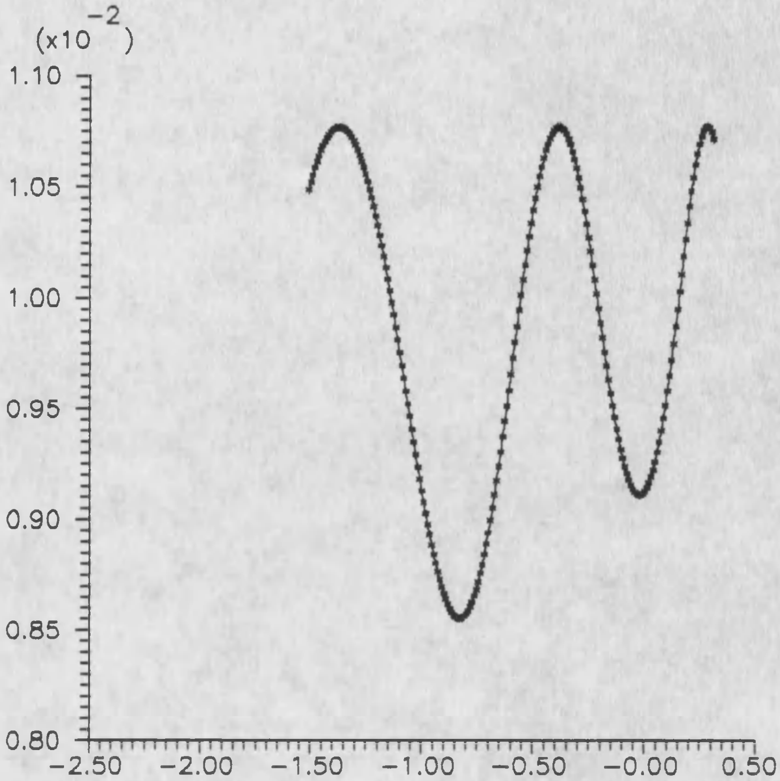


Figure 4.2.2. Derivative Function for design (4.6.5)

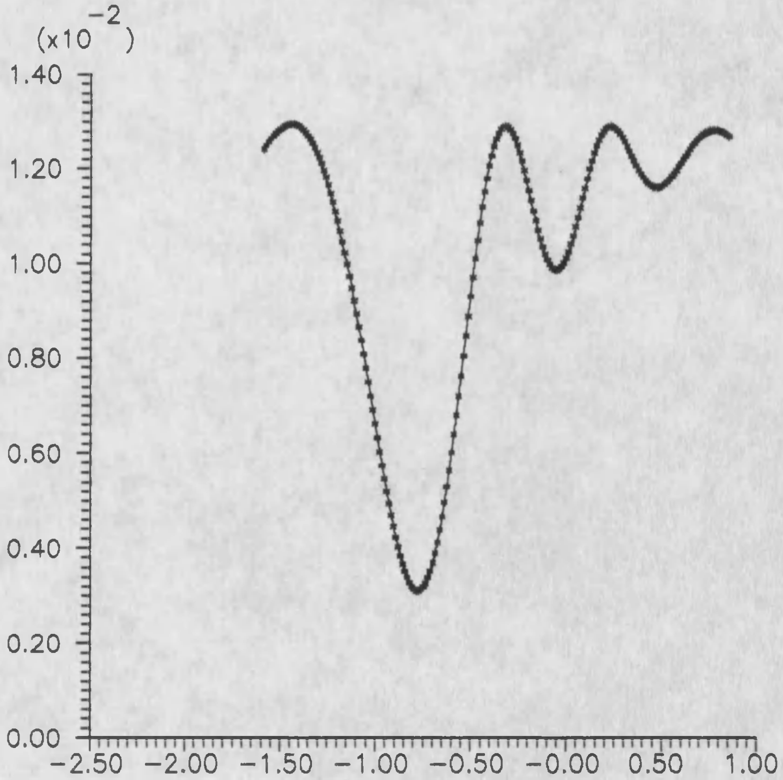


Figure 4.2.3. Derivative Function for design (4.6.6)

$$\xi^* = \begin{Bmatrix} -1.443 & -0.319 & 0.2337 & 0.7766 \\ 0.3262 & 0.2151 & 0.1645 & 0.2942 \end{Bmatrix} \qquad (4.6.6)$$

Four support points are now required in the optimum design to discriminate between these two models in the presence of reasonable uncertainty with respect to the model truth and the values of the parameters. There are two positive and two negative support points in design (4.6.6) with the first and third point weights added up counting for 49.07% of the total weight. Figure 4.2.3 shows the maximum value of criterion function (4.4.4) being achieved at all four support points of design (4.4.6). The region between the first and the second support points is noninformative regarding model discrimination whereas points in the design region between the second and fourth support points are all highly informative with respect to the same purpose.

EXAMPLE 4.3. Discriminating between the normal integrated (probit) and the logistic (logit) curves. This is the subject of the pioneering work on binary response model discrimination described by Chambers and Cox (1967). Although the present approach and theirs are based on different ideas a comparison can be made between optimum designs yielded by these different criteria. We start by assuming that the logistic link based model is true. Moreover, we assume that both model linear predictors are described by a straight line $\theta_{j0} + \theta_{j1}x$, $\{j=1,2\}$ and that $\boldsymbol{\theta}_1 = (2.0,2.0)$ are the true values of the linear predictor parameters for the true model. The purpose is to discriminate against the rival model whose link function is probit.

The maximizing value for criterion function (4.4.2) is given by $\Delta_2(\xi^*,\theta_1) = 1.853 \times 10^{-3}$ and the locally optimum design is given below.

$$\xi^* = \begin{Bmatrix} -1.761 & -0.2386 & 1.972 \\ 0.1475 & 0.2491 & 0.6034 \end{Bmatrix} \qquad (4.6.7)$$

Under the true model the probabilities of success corresponding to the three support points of design (4.6.7), in ascending order of the latter, are 0.1791, 0.8209 and 0.9974 respectively, whereas under the false model they are equal to 0.2032, 0.7968 and 0.9994 as opposed to 0.215, 0.785 and 0.9964 given by the optimum design found by Chambers and Cox (1967) under the false model.

However, the most interesting result that comes from comparing design (4.6.7) with Chambers and Cox's optimum design concerns the weights. In the former, the design weights increase as do the support points, with 60.34% of the total weight being assigned to the largest support point. Similarly, in the latter the design weights are 0.117, 0.214 and 0.669 in ascending order of their standardized support points. Their interpretation of the results was based on regions for which the curves are distinguishable or indistinguishable. Acccording to this, the first two points of the design, which lie in a region where the curves are almost indistinguishable, are probably required to estimate the linear predictor parameters whereas the third point, which lies in a region where the curves disagree, is needed to discriminate between the models. This also explains the excessive weight assigned to the only point that supposedly discriminates between the models.

Figure 4.3.1 displays the derivative function for design (4.6.7) where again the pattern of locally optimum designs is apparent. As in previous examples we proceed with the discrimination between the logit and probit links by investigating the effects of replacing the hypotheses of knowledge of the true values of parameters by prior probabilities. Suppose that the true model is now the inverse normal or probit and that there is a prior distribution for its linear predictor parameters θ_1 . Both linear predictor structures are again given by a first degree polinomial. Table 4.6.2 shows the prior for θ_1.

Criterion function (4.4.3) is appropriate in the current situation. The resulting partially Bayesian optimum design is shown below. Its maximized value is $\gamma_2(\xi^*) = 7.819 \times 10^{-4}$.

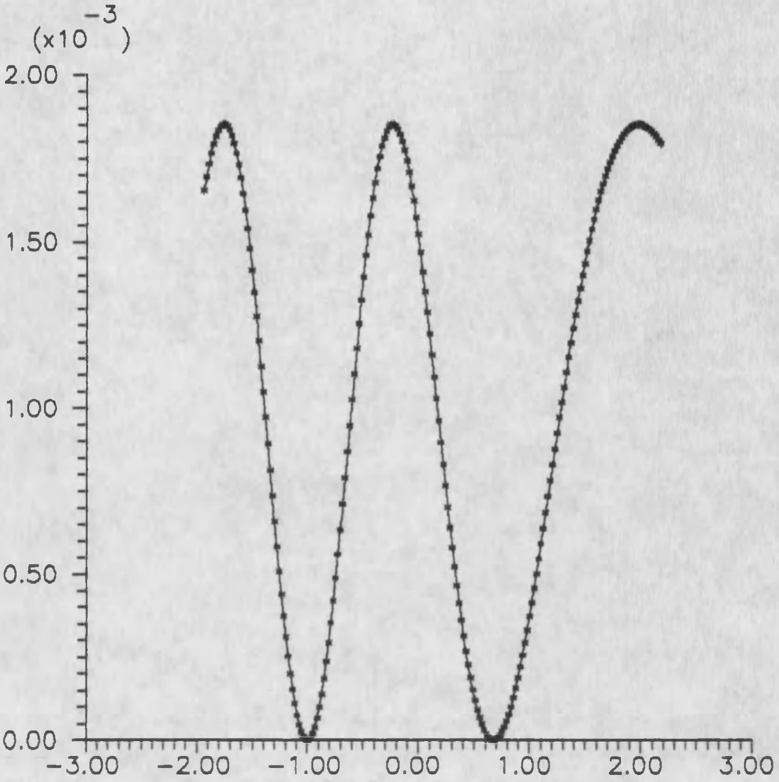


Figure 4.3.1. Derivative Function for design (4.6.7)

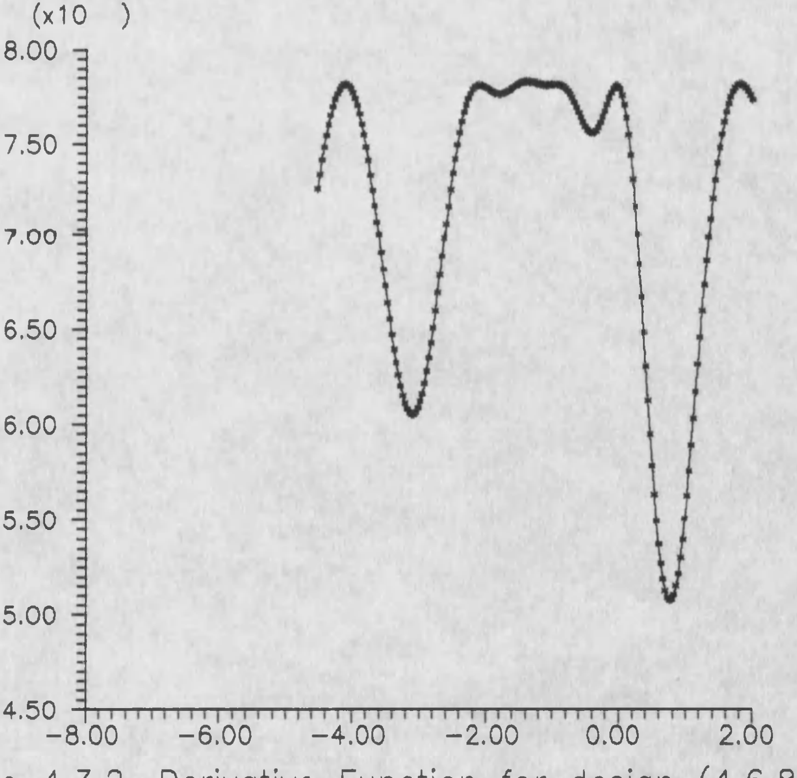


Figure 4.3.2. Derivative Function for design (4.6.8)

$$\xi^* = \begin{Bmatrix} -4.102 & -2.099 & -1.072 & -0.0384 & 1.815 \\ 0.0799 & 0.0932 & 0.1078 & 0.1866 & 0.5325 \end{Bmatrix} \tag{4.6.8}$$

Five support points are necessary to discriminate against the logistic link based model. As for design (4.6.7), the weights in design (4.6.8) increase with the values of the support points. The range of the design weights varies from 7.99% for the most negative support point to 53.25% for the largest one, the only positive number. This excessive weight for the last point suggests a large disagreement between the curves at this point making it very important to discriminate between the models. This vital support point which peaks at the maximum of the criterion function appears very close to the most noninformative point corresponding to the deepest trough in Figure 4.3.2. The design region between the second and the fourth support points seems to be very informative with respect to criterion (4.4.3).

In the last part of our investigation of the problem of searching for optimum designs to discriminate between the logistic and the inverse normal (probit) link based models, we carry out an informal and simple sensitivity analysis of fully Bayesian optimum designs to discriminate between these rival models. Basically, our interest lies in assessing the sensitivity of the optimum design to lack of accuracy in the prior distributions to the linear predictor parameter sets $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. For this purpose, four prior distributions are considered, two for each parameter set, one concentrated and the other relatively dispersed. We then combine the prior probabilities, one for $\boldsymbol{\theta}_1$ and another for $\boldsymbol{\theta}_2$, exhaustively and find the related fully Bayesian optimum designs. For each of the four combinations of priors the models are assumed equiprobable, i.e. $\pi_0 = (0.5,0.5)$. The priors are listed in Tables 4.6.1 and 4.6.2. The former contains the concentrated priors for both parameter sets whereas the latter provides the dispersed priors.

As usual the bidimensional points in the priors are arranged so that they form squares which are centered at either $\boldsymbol{\theta}_1 = (2.0,2.0)$ or $\boldsymbol{\theta}_2 = (1.0,1.0)$. The areas of the squares in Table 4.6.1 are both equal to 0.04 whereas those of Table

4.6.2 are equal to 4 and 1, respectively a factor of 100 and 25 larger than the previous ones. As probabilities are assigned to points in a similar manner for all four priors, the above figures illustrate the difference in accuracy between these priors.

TABLE 4.6.1 – CONCENTRATED PRIOR DISTRIBUTIONS FOR θ_1 AND θ_2

θ_{10}	θ_{11}	$p_{01}(\theta_1)$	θ_{20}	θ_{21}	$p_{02}(\theta_2)$
2.0	2.0	0.2	1.0	1.0	0.2
2.0	1.9	0.1	1.0	0.9	0.1
2.0	2.1	0.1	1.0	1.1	0.1
1.9	2.0	0.1	0.9	1.0	0.1
1.9	1.9	0.1	0.9	0.9	0.1
1.9	2.1	0.1	0.9	1.1	0.1
2.1	2.0	0.1	1.1	1.0	0.1
2.1	1.9	0.1	1.1	0.9	0.1
2.1	2.1	0.1	1.1	1.1	0.1

TABLE 4.6.2 – DISPERSED PRIOR DISTRIBUITONS FOR θ_1 AND θ_2

θ_{10}	θ_{11}	$p_{01}(\theta_1)$	θ_{20}	θ_{21}	$p_{02}(\theta_2)$
2.0	2.0	0.2	1.0	1.0	0.2
2.0	1.0	0.1	1.0	0.5	0.1
2.0	3.0	0.1	1.0	1.5	0.1
1.0	2.0	0.1	0.5	1.0	0.1
1.0	1.0	0.1	0.5	0.5	0.1
1.0	3.0	0.1	0.5	1.5	0.1
3.0	2.0	0.1	1.5	1.0	0.1
3.0	1.0	0.1	1.5	0.5	0.1
3.0	3.0	0.1	1.5	1.5	0.1

According to previous discussions and conclusions, one should expect that the number of support points in the optimum design increases as the combination of priors varies from both concentrated to both dispersed. The four resulting fully Bayesian optimum designs are listed in Table 4.7.

TABLE 4.7 – FULLY BAYESIAN T–OPTIMUM DESIGNS FOR FOUR COMBINATIONS OF PRIOR DISTRIBUTIONS FOR θ_1 AND θ_2

VALUE OF $\Gamma(\xi^*)$	BAYESIAN OPTIMUM DESIGN
(a) 1.608×10^{-3}	$\left\{\begin{matrix} -1.779 & -0.2138 & 1.945 \\ 0.1408 & 0.2433 & 0.6159 \end{matrix}\right\}$
(b) 1.117×10^{-3}	$\left\{\begin{matrix} -1.716 & -0.1334 & 1.892 \\ 0.1334 & 0.2435 & 0.6231 \end{matrix}\right\}$
(c) 1.273×10^{-3}	$\left\{\begin{matrix} -3.991 & -1.716 & -0.2076 & 1.959 \\ 0.0978 & 0.1588 & 0.2288 & 0.5146 \end{matrix}\right\}$
(d) 9.02×10^{-4}	$\left\{\begin{matrix} -4.104 & -2.148 & -1.325 & -0.8659 & -0.0201 & 1.871 \\ 0.0897 & 0.0916 & 0.0662 & 0.0656 & 0.1846 & 0.5023 \end{matrix}\right\}$

where

(a) both concentrated, design (4.6.9) ;

(b) θ_1 dispersed and θ_2 concentrated, design (4.6.10) ;

(c) θ_1 concentrated and θ_2 dispersed, design (4.6.11) ;

(d) both dispersed, design (4.6.12).

Apart from some similarity to designs (4.6.9) and (4.6.10) and the fact that in all designs the last support point is the only positive one and also the most influential of all, with more than 50% of the total weight, there are no other apparent features that are common to the designs of Table 4.7. The number of

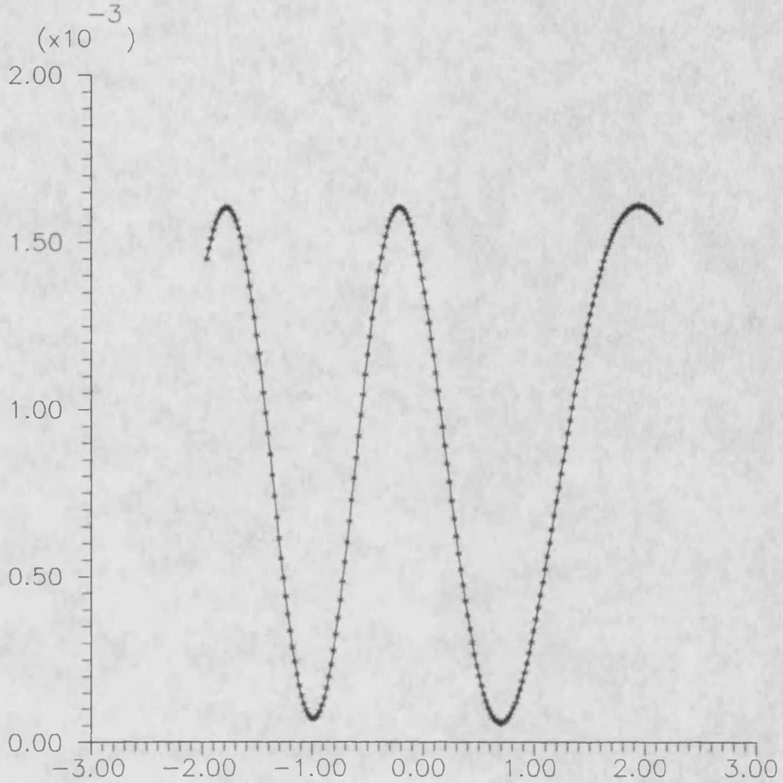


Figure 4.3.3. Derivative Function for design (4.6.9)

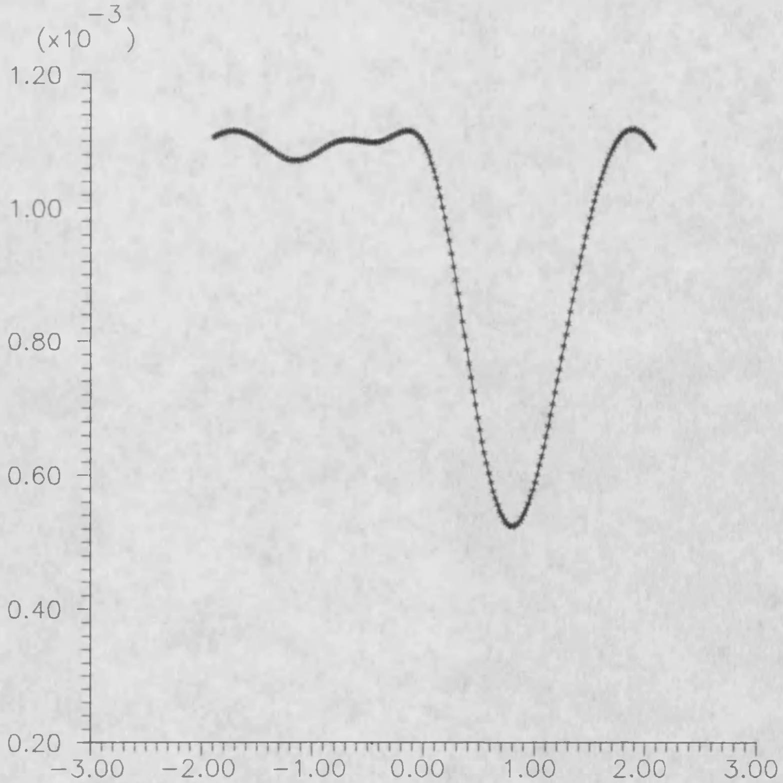


Figure 4.3.4. Derivative Function for design (4.6.10)

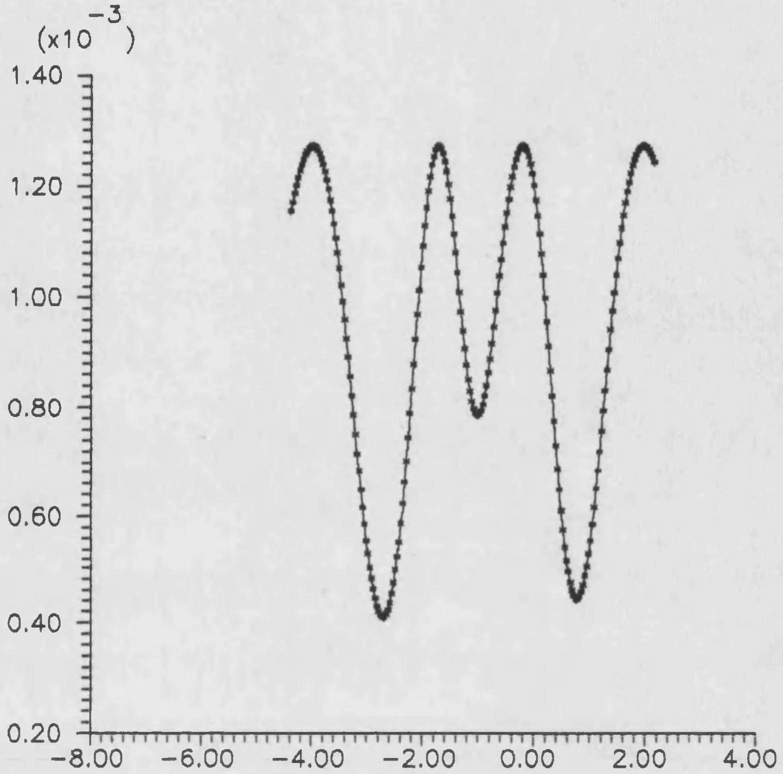


Figure 4.3.5. Derivative Function for design (4.6.11)

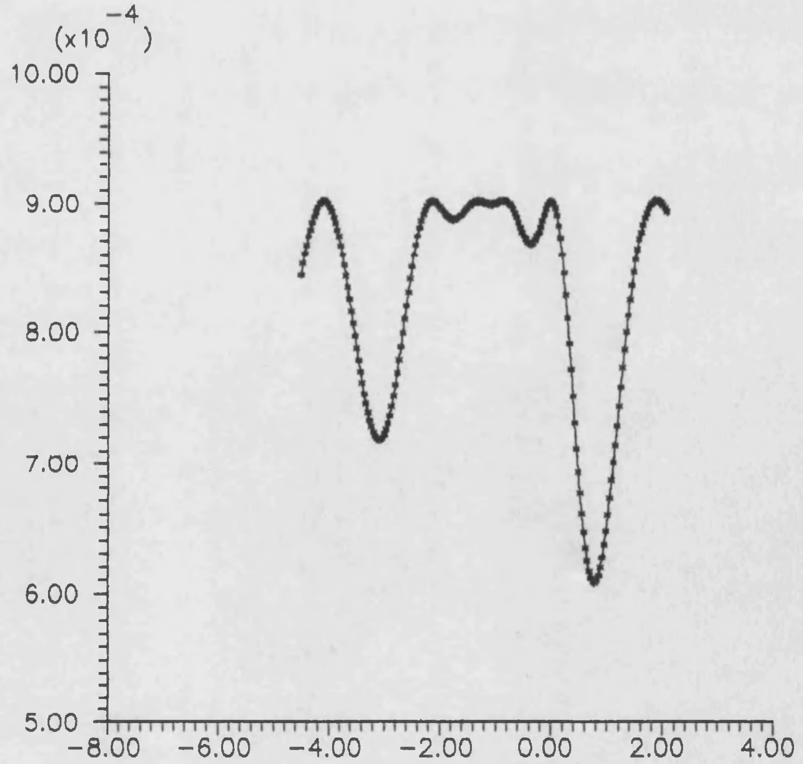


Figure 4.3.6. Derivative Function for design (4.6.12)

support points in these designs varies from three, for both concentrated priors to six, for both dispersed priors, confirming what was expected. The fact that only three support points were required to discriminate between the models in the presence of a dispersed prior for θ_1 and a concentrated one for θ_2 as opposed to four support points when there was a concentrated prior for θ_1 and a dispersed one for θ_2 seems to suggest that there is more sensitivity towards dispersion in the prior for θ_2 than for θ_1 as far as the number of support points is concerned.

Analyses of the plots for the derivative functions reveal further relevant features pertaining to these designs. Figure 4.3.3, related to design (4.6.9), shows a pattern of very well marked peaks and troughs which is typical of locally optimum designs. This is due to both priors being concentrated. In Figure 4.3.4, corresponding to design (4.6.10), there are only three peaks although in the design region between the first and second support points the shape of the derivative function is almost flat with values just below the maximized value of the criterion function, indicating that these points are very informative.

In Figure 4.3.5, related to design (4.6.11), the shape of the derivative function looks symmetric. However, the range of the design weights is very wide leading to the conclusion that, in fact, design (4.6.11) is unbalanced as opposed to the suggestion of Figure 4.3.5. Finally, Figure 4.3.6 shows the plot of the derivative function related to design (4.6.12). As in Figure (4.6.10) there is a region in which the shape of the derivative funcion is almost flat. Such a region is limited by the second and the fifth support points.

4.7 DISCUSSION

The definitions of Section 4.4 and the theoretical results of Section 4.5 highlight the similarities in construction and properties of designs for discriminating between binary data models and those for discriminating between linear and/or nonlinear regression models. A natural extension is to consider designs for discriminating between any two generalized linear models, again using as a criterion function the expected deviance under the false model. Such designs will only be of interest when the responses under the two models are of the same type with the same range. This is the subject of Chapter 5.

The designs described in this paper are non–sequential. If sequential designs are possible they are to be preferred as information obtained at one stage can be used to design a more efficient experiment at the next. The combination of sequential methods with the Bayesian approach of this chapter requires updating the prior for the parameters of each model and also, for some designs of Section 4.6, a method of updating the probability that each model is true.

There are some other more practical points. The use of discrete priors for the parameters of the two models avoids some numerical problems. Does this lead to markedly less efficient designs ? Likewise we have here calculated only approximate designs ξ^*. What about discrete or exact designs ? A final extension is to consider designs for discrimination between three or more generalized linear models. Here we might follow the approach of Atkinson & Fedorov (1975b) for regression models.

APPENDIX B

To simplify the notation used in Section 4.4, let us express the sum of the discrepancies (integral) between the expected values of two binary response models w.r.t. an underlying design ξ as follows. For j =1,2 ; $\bar{j}$ = not j ; $\xi \in \mathscr{H}$; $\theta_j \in \Theta_j$; and $\theta_{\bar{j}} \in \Theta_{\bar{j}}$ denote

$$b(\xi,\theta_j,\theta_{\bar{j}}) = \int_{\mathfrak{X}} \left\{ \pi_j \log\left[\frac{\pi_j}{\pi_{\bar{j}}}\right] + (1-\pi_j)\log\left[\frac{1-\pi_j}{1-\pi_{\bar{j}}}\right] \right\} \xi(dx) \tag{B.1}$$

where $\pi_j = \pi_j(x,\theta_j)$, $x \in \mathfrak{X}$.

According to this, for any design $\xi \in \mathscr{H}$; and any $\theta_j \in \Theta_j$ the binary discrimination totals (4.4.1.a) and (4.4.1.b) may be represented by

$$\Delta_{\bar{j}}(\xi,\theta_j) = b(\xi,\theta_j,\theta_{\bar{j}}^{\dagger}) = \inf_{\theta_{\bar{j}} \in \Theta_{\bar{j}}} b(\xi,\theta_j,\theta_{\bar{j}}) \text{ , j=1,2} \tag{B.2}$$

What we want to prove is that (B.2) is a concave function on $\mathscr{H}$, i.e.

$$\Delta_{\bar{j}}(\xi,\theta_j) \geq (1-\alpha)\, \Delta_{\bar{j}}(\xi_1,\theta_j) + \alpha\, \Delta_{\bar{j}}(\xi_2,\theta_j) \tag{B.3}$$

where $0 \leq \alpha \leq 1$ and $\xi = (1-\alpha)\,\xi_1 + \alpha\,\xi_2$; ξ_1, $\xi_2 \in \mathscr{H}$. This proof, presented below, is analogous to the proof of concavity for the noncentrality parameters given in Appendix A, result (A.3).

<u>Proof of (B.3)</u>

Let $\alpha \in \mathbb{R}$, $0 \leq \alpha \leq 1$; $\xi_1,\xi_2 \in \mathscr{H}$; and $\xi = (1-\alpha)\,\xi_1 + \alpha\,\xi_2$.

Then

$$\begin{aligned}
\Delta_{\bar{j}}(\xi,\theta_j) &= \inf_{\theta_{\bar{j}} \in \Theta_{\bar{j}}} b(\xi,\theta_j,\theta_{\bar{j}}) = \inf_{\theta_{\bar{j}} \in \Theta_{\bar{j}}} \left\{ b((1-\alpha)\xi_1+\alpha\xi_2,\theta_j,\theta_{\bar{j}}) \right\} \\
&= \inf_{\theta_{\bar{j}} \in \Theta_{\bar{j}}} \left\{ (1-\alpha)\, b(\xi_1,\theta_j,\theta_{\bar{j}}) + \alpha\, b(\xi_2,\theta_j,\theta_{\bar{j}}) \right\} \\
&\geq \inf_{\theta_{\bar{j}} \in \Theta_{\bar{j}}} \left\{ (1-\alpha)\, b(\xi_1,\theta_j,\theta_{\bar{j}}) \right\} + \inf_{\theta_{\bar{j}} \in \Theta_{\bar{j}}} \left\{ \alpha\, b(\xi_2,\theta_j,\theta_{\bar{j}}) \right\} \\
&= (1-\alpha)\, \Delta_{\bar{j}}(\xi_1,\theta_j) + \alpha\, \Delta_{\bar{j}}(\xi_2,\theta_j).
\end{aligned}$$

Hence $\Delta_{\bar{j}}(\xi,\theta_j)$ is a concave function on $\mathscr{H}$. Further, concavity of criterion function (4.4.4) follows from linearity.

The steps to determine the directional derivatives of $\Gamma(\xi)$ for the problem of binary response model discrimination are also analogous to those for the equivalent of (4.4.4) in the regression case. Here, the expression for the derivative function is also given by

$$\frac{\partial\Gamma(\xi)}{\partial\alpha} = \sum_{j} \pi_{0j}\, E_{\theta_j}\left\{\frac{\partial\Delta_{\bar{j}}(\xi,\theta_j)}{\partial\alpha}\right\} \tag{B.4}$$

where $\xi = (1-\alpha)\ \xi_1 + \alpha\ \xi_2$; $0 \leq \alpha \leq 1$; and for $j = 1,2$

$$\frac{\partial\Delta_{\bar{j}}(\xi,\theta_j)}{\partial\alpha} = \inf_{\theta_{\bar{j}}\in\Theta^{\dagger}_{\bar{j}}(\alpha)}\left\{b(\xi_2,\theta_j,\theta_{\bar{j}}) - b(\xi_1,\theta_j,\theta_{\bar{j}})\right\} \tag{B.5}$$

where for a given $\alpha \in [0,1]$, $\Theta^{\dagger}_{\bar{j}}(\alpha)$ is the solution set of equation (4.5.1).

<u>Proof of (B.4)</u>

Let us first prove (B.5). For any real number $\alpha \in [0,1]$; $\xi_1,\xi_2 \in \mathscr{H}$; and $\xi = (1-\alpha)\ \xi_1 + \alpha\ \xi_2$ let $\Theta^{\dagger}_{\bar{j}}(\alpha)$ be the solution set of equation (4.5.1). In our notation $\Theta^{\dagger}_{\bar{j}}(\alpha)$ is the solution set of the following equation

$$b(\xi,\theta_j,\theta^{\dagger}_{\bar{j}}) = \inf_{\theta_{\bar{j}}\in\Theta_{\bar{j}}} b(\xi,\theta_j,\theta_{\bar{j}})$$

Then, the Fréchet derivative of $\Delta_{\bar{j}}(\xi,\theta_j)$ is derived as follows.

$$\frac{\partial\Delta_{\bar{j}}(\xi,\theta_j)}{\partial\alpha} = \frac{\partial}{\partial\alpha}\left\{\inf_{\theta_{\bar{j}}\in\Theta^{\dagger}_{\bar{j}}(\alpha)} b(\xi,\theta_j,\theta_{\bar{j}})\right\} = \inf_{\theta_{\bar{j}}\in\Theta^{\dagger}_{\bar{j}}(\alpha)}\left\{\frac{\partial}{\partial\alpha} b(\xi,\theta_j,\theta_{\bar{j}})\right\}$$

$$= \inf_{\theta_{\bar{j}}\in\Theta^{\dagger}_{\bar{j}}(\alpha)}\left[\lim_{\alpha\to 0^{+}} \frac{1}{\alpha}\left\{b((1-\alpha)\xi_1+\alpha\xi_2,\theta_j,\theta_{\bar{j}}) - b(\xi_1,\theta_j,\theta_{\bar{j}})\right\}\right]$$

$$= \inf_{\theta_{\bar{j}}\in\Theta^{\dagger}_{\bar{j}}(\alpha)}\left[\lim_{\alpha\to 0^{+}} \frac{1}{\alpha}\alpha\ \left\{b(\xi_2,\theta_j,\theta_{\bar{j}}) - b(\xi_1,\theta_j,\theta_{\bar{j}})\right\}\right]$$

$$= \inf_{\theta_{\bar{j}}\in\Theta^{\dagger}_{\bar{j}}(\alpha)}\left\{b(\xi_2,\theta_j,\theta_{\bar{j}}) - b(\xi_1,\theta_j,\theta_{\bar{j}})\right\}$$

where again the first equality holds by applying Pshenichnyi (1971, Theorem 3.2, pp. 75). Hence, (B.5) is true. Now, on the assumption that the optimum design ξ^* is such that (4.5.1) has a unique solution $\theta_{\bar{j}}^*$ when $\xi = \xi^*$ ($\xi \to \xi_1$, when $\alpha \to 0$), (B.5) becomes

$$\begin{aligned} F_{\Delta_{\bar{j}}}(\xi^*,\xi_2) &= b(\xi_2,\theta_j,\theta_{\bar{j}}^*) - b(\xi^*,\theta_j,\theta_{\bar{j}}^*) \\ &= b(\xi_2,\theta_j,\theta_{\bar{j}}^*) - \Delta_{\bar{j}}(\xi^*,\theta_j). \end{aligned}$$

where ξ_2 represents any design other than ξ^*. In particular when $\xi_2 = \xi_x$, the design putting all mass at the point $x \in \mathfrak{X}$, it follows that for any $x \in \mathfrak{X}$

$$F_{\Delta_{\bar{j}}}(\xi^*,\xi_x) = \pi_j \log\left[\frac{\pi_j}{\pi_{\bar{j}}^*}\right] + (1-\pi_j)\log\left[\frac{1-\pi_j}{1-\pi_{\bar{j}}^*}\right] - \Delta_{\bar{j}}(\xi^*,\theta_j) \tag{B.6}$$

However, according to Theorems 2.1 and 2.2

$$\begin{aligned} F_{\Delta_{\bar{j}}}(\xi^*,\xi_x) &= \psi(x,\xi^*,\theta_j) - \Delta_{\bar{j}}(\xi^*,\theta_j) \leq 0 \\ &\Longrightarrow \psi(x,\xi^*,\theta_j) \leq \Delta_{\bar{j}}(\xi^*,\theta_j) \text{, for any } x \in \mathfrak{X} \end{aligned} \tag{B.7}$$

where $\psi(x,\xi^*,\theta_j) = \pi_j \log\left[\frac{\pi_j}{\pi_{\bar{j}}^*}\right] + (1-\pi_j)\log\left[\frac{1-\pi_j}{1-\pi_{\bar{j}}^*}\right]$

Inequality (B.7) is condition (i) of Theorem 4.1 for a local optimum design. Again, Corollary 2.1 applies here so as to derive condition (ii). Conditions (iii) and (iv) are then straigthforward.

For the more general case corresponding to criterion (4.4.4) we must assume that at the optimum design ξ^* (4.5.1) has unique solutions $\theta_{\bar{j}}^*$ for every $\theta_j \in \Theta_j$ which is relevant to the prior $p_{0j}(\theta_j)$, $\{j=1,2\}$. Taking expectations of (B.6) over the priors π_0 and $p_{0j}(\theta_j)$, $\{j=1,2\}$ yields condition (i) of Theorem 4.1 in its more general version given in Section 4.4, namely

$$\psi(x,\xi^*) \leq \Gamma(\xi^*) \text{, for any } x \in \mathfrak{X}$$

where $\psi(x,\xi^*) = E_{\pi_0} E_{\theta_{\bar{j}}} \left\{ \pi_j \log\left[\frac{\pi_j}{\pi_{\bar{j}}^*}\right] + (1-\pi_j)\log\left[\frac{1-\pi_j}{1-\pi_{\bar{j}}^*}\right] \right\}.$

Conditions (ii), (iii) and (iv) can now be obtained analogously.

CHAPTER 5. THE DESIGN OF EXPERIMENTS TO DISCRIMINATE BETWEEN TWO RIVAL GENERALIZED LINEAR MODELS

5.1. INTRODUCTION

In this chapter the theoretical results presented previously in Chapters 3 and 4 are extended to the class of generalized linear models. This extension is motivated by the similarities encountered in problems of model discrimination for other categories of statistical models beyond the scope of regression, such as Gamma, Poisson, etc. The analogy between residual sum of squares in regression models and deviance in generalized linear models is again explored for the introduction of the criterion of optimality (the same argument is used in Chapter 4, although for the particular case of binary response models).

To describe the contents of this chapter more precisely we could say that it concerns an extension of the criterion of T–optimality (Atkinson and Fedorov, 1975a) to the class of generalized linear models by means of the natural generalization of the residual sum of squares into the deviance. This leads to the term Generalized T–optimality for the criterion introduced in Section 5.3. Such generalization provides a much wider framework for applying the methods to determine an optimum experimental design to discriminate between two statistical models.

In unplanned experiments the fitting of data by generalized linear models requires the specification of both the link function and the linear predictor structure. The usual solution is to try several combinations of link function and linear predictor structure until a satisfactory fit is found. In most cases, however, the search for the model that best fits the data is reduced to a few structures between which a well designed experiment could help to discriminate. Further, there are situations in which the structures and models that are more likely to provide the best fits are known a priori. In these cases the planning of the experiment with the purpose of discrimination, if possible, is appealing.

For instance, in the previous chapter the logistic, probit and complementary log–log link functions were compared and optimum designs for link discrimination were found in different situations as far as prior information on the linear predictor parameters is concerned. Then, the reason for applying methods of model discrimination was that the expected responses yielded by models with these link functions are similar. Problems of link choice as well as linear predictor structure choice can arise in all categories of generalized linear models, thereby creating a wide scope for application of the ideas of model discrimination based on T–optimality. This motivates the methodology to be introduced in this chapter.

5.2. BACKGROUND

With a few exceptions, the methods of optimal design theory have not yet been applied to GLMs, as defined by McCullagh and Nelder (1989). Although rather briefly, Chalòner (1987) approaches the problem of designing optimum experiments to estimate the linear predictor parameters of a GLM, but later emphasizes the subclass of logistic regression for binary response models. Two criteria are suggested, namely expected D–optimality and A–optimality, to deal with the problems of estimating the linear predictor parameters and relevant functions of these. Following this work, Chaloner and Larntz (1989) give further details of the problems. However, no thorough investigation has been carried out hitherto on the problem of designing experiments under the framework of GLMs.

Ponce de Leon and Atkinson (1992b) approach the problem of optimum designs for model discrimination under this framework, although only locally optimum designs are taken into consideration. In this chapter these preliminary results are extended so as Bayesian optimum designs can also be determined. Moreover, we illustrate the theory by means of an extensive investigation of the problem of discriminating between the logarithimic and the reciprocal links for a Gamma model.

In Section 5.3 we describe the general problem and define the generalized T–optimality criterion function along with specific expressions of it for five subclasses of GLMs. In Section 5.4 we show how to find the derivative functions required for the construction and checking of optimum designs. Four numerical examples are provided in Section 5.5. The chapter concludes with a brief discussion in Section 5.6.

5.3. GENERALIZED T–OPTIMALITY

Suppose that two well defined structures (models) are to be discriminated between in the early stages of designing the experiment. For reasons of simplicity assume that both structures belong to the same subclass of GLMs (this is not a crucial assumption since in some practical situations models belonging to different subclasses of GLMs might well provide good fits for the data). Further, assume that for both models the link function and linear predictor structure are known a priori. The link functions and/or linear predictor structures may or may not be the same, so the models may or may not be nested. There is no restriction on the kind of linear structure nor on the number of linear predictor parameters. Thus, let μ_j denote the vector of means corresponding to the jth model and η_j the linear predictor vector (j=1,2). Under the framework of GLMs we have

$$\eta_j = (g_j(\mu_{j1}),\ldots,g_j(\mu_{jn}))'$$

where

$g_j(.)$ is the link function ; $\eta_j = X_j\beta_j$ is the vector of linear predictors ; X_j is the design matrix, $(n \times k_j)$; β_j is the vector of unknown parameters, $(k_j \times 1)$; k_j is the number of parameters ; and n is the sample size.

Under these circumstances the problem we are concerned with can be summarized as the search for an experimental design that discriminates most efficiently between the two GLMs.

In our approach to model discrimination we assume that prior information on the linear predictor parameters for both models and on the model truth are available by means of probability distributions. Thus, let $\pi_0 = (\pi_{01},\pi_{02})$, where $\pi_{01} + \pi_{02} = 1$, denote the prior probabilities for each model to be true. Let $p_{0j}(\beta_j)$, j=1,2, denote the conditional (on the model truth) probability distributions for the parameter sets $\{\beta_j\}$.

Proceeding as in Chapters 3 and 4, we now define two variables, functions of the design measure ξ and the true values of the linear predictor parameters β_1 or β_2. They represent generalizations of the noncentrality parameters under regression theory with normality assumptions described by Definition 3.1.

Definition 5.1. The quantities

$$\Delta_1(\xi,\beta_2) = \underset{\beta_1 \in B_1}{\mathrm{Min}} \int_{\mathcal{X}} d(x,\beta_2,\beta_1)\ \xi(dx) \tag{5.3.1.a}$$

where

$$d(x,\beta_2,\beta_1) = 2w[\mu_2(x,\beta_2)\{\theta_2(x,\beta_2) - \theta_1(x,\beta_1)\} - b\{\theta_2(x,\beta_2)\} + b\{\theta_1(x,\beta_1)\}]$$

and

$$\Delta_2(\xi,\beta_1) = \underset{\beta_2 \in B_2}{\mathrm{Min}} \int_{\mathcal{X}} d(x,\beta_1,\beta_2)\ \xi(dx) \tag{5.3.1.b}$$

where

$$d(x,\beta_1,\beta_2) = 2w[\mu_1(x,\beta_1)\{\theta_1(x,\beta_1) - \theta_2(x,\beta_2)\} - b\{\theta_1(x,\beta_1)\} + b\{\theta_2(x,\beta_2)\}].$$

are the generalized discrimination totals for the first model when the second is true, and vice versa.

TABLE 5.1 – EXPRESSIONS FOR $d(x,\beta_1,\beta_2)$

Subclass	$d(x,\beta_1,\beta_2)$
Normal	$\{\mu_1(x,\beta_1) - \mu_2(x,\beta_2)\}^2$
Binomial	$2\left[\mu_1(x,\beta_1)\log\left[\frac{\mu_1(x,\beta_1)}{\mu_2(x,\beta_2)}\right] + \{1 - \mu_1(x,\beta_1)\}\log\left[\frac{1 - \mu_1(x,\beta_1)}{1 - \mu_2(x,\beta_2)}\right]\right]$
Poisson	$2\left[\mu_1(x,\beta_1)\log\left[\frac{\mu_1(x,\beta_1)}{\mu_2(x,\beta_2)}\right] - \{\mu_1(x,\beta_1) - \mu_2(x,\beta_2)\}\right]$
Gamma	$2\left[-\log\left[\frac{\mu_1(x,\beta_1)}{\mu_2(x,\beta_2)}\right] + \frac{\mu_1(x,\beta_1) - \mu_2(x,\beta_2)}{\mu_2(x,\beta_2)}\right]$
Inverse Gaussian	$\{\mu_1(x,\beta_1) - \mu_2(x,\beta_2)\}^2 / \{\mu_2(x,\beta_2)\}^2\, \mu_1(x,\beta_1)$

In both expressions $d(x,\beta_2,\beta_1)$ and $d(x,\beta_1,\beta_2)$ given above, w is a known prior weight; $\mu_2(x,\beta_2)$ and $\mu_1(x,\beta_1)$ provide the expected values under the rival models; $\theta_2(x,\beta_2)$ and $\theta_1(x,\beta_1)$ are the canonical parameters; and $b\{\theta_2(x,\beta_2)\}$ and $b\{\theta_1(x,\beta_1)\}$ are known functions of the canonical parameters. For each subclass of GLMs there are specific expressions for $\theta_j(x,\beta_j)$ and $b\{\theta_j(x,\beta_j)\}$, j=1,2. For further details about the notation and meaning of the components of a GLM, see McCullagh and Nelder (1989, Chapter 2). Particular expressions for $d(x,\beta_1,\beta_2)$ are given in Table 5.1. Analogous expressions follow for $d(x,\beta_2,\beta_1)$. In Table 5.1 the factor 2 which appears in the Poisson, Binomial and Gamma models may be discarded as it will have no effect on the optimization procedures.

From Table 5.1 and expressions (5.3.1.a) and (5.3.1.b) we can find expressions (3.3.1.a) and (3.3.1.b) as well as (4.4.1.a) and (4.4.1.b) corresponding to the Normal and Binomial subclasses of GLMs, respectively. For other subclasses it is straightforward to obtain the quantities that will be used to define the criteria of optimality for model discrimination. We start with the case in which the true model and its linear predictor parameters are known.

<u>Definition 5.2.</u> Assume that the jth model is true and its linear predictor parameters $\{\beta_{jk}\}$ are known. A design ξ^* is said to be *local generalized T–optimum* if and only if

$$\Delta_{\bar{j}}(\xi^*) = \sup_{\xi \in \mathscr{H}} \Delta_{\bar{j}}(\xi) \tag{5.3.2}$$

where

$$\Delta_{\bar{j}}(\xi) = \Delta_{\bar{j}}(\xi,\beta_j),\ j=1,2 \text{ and } \bar{j} = \text{not } j.$$

By taking the Normal subclass of GLMs, (5.3.2) becomes the original criterion of T–optimality as defined by Atkinson and Fedorov (1975a). Another possibility in practice arises when the true model is known and its linear predictor parameters are known to be distributed according to a given prior probability distribution. This corresponds to either π_{01} or π_{02} equal to one.

<u>Definition 5.3.</u> Assume that the jth model is true and that $p_{0j}(\beta_j)$ denotes the prior probability for its linear predictor parameters. Then, a design ξ^* is said to be *partially Bayesian generalized T–optimum* if and only if

$$\gamma_{\bar{j}}(\xi^*) = \sup_{\xi \in \mathscr{H}} \gamma_{\bar{j}}(\xi) \tag{5.3.3}$$

where

$$\gamma_{\bar{j}}(\xi) = E_{\beta_j}\left\{\Delta_{\bar{j}}(\xi,\beta_j)\right\},\ j=1,2 \text{ and } \bar{j} = \text{not } j.$$

In the general situation both prior probability distributions are spread around in the parameter space rather than concentrated on a single point as are (5.3.2) and (5.3.3).

Definition 5.4. Assume that there is a prior probability that each model is true, and conditionally on the event that the jth model is true, there is a prior probability distribution $p_{0j}(\beta_j)$ for its linear predictor parameters. Then, a design ξ^* is said to be *fully Bayesian generalized T–optimum* if and only if

$$\Gamma(\xi^*) = \sup_{\xi \in \mathscr{H}} \Gamma(\xi) \tag{5.3.4}$$

where

$$\Gamma(\xi) = \sum_j \pi_{0j} E_{\beta_j}\left\{\Delta_j(\xi,\beta_j)\right\} = \pi_{01} E_{\beta_1}\left\{\Delta_2(\xi,\beta_1)\right\} + \pi_{02} E_{\beta_2}\left\{\Delta_1(\xi,\beta_2)\right\}.$$

All criterion functions introduced so far in this thesis are particular cases of (5.3.4). Basically, there are two ways of making (5.3.4) collapses into these particular cases, namely by assuming that the models to be discriminated belong to a specific category of GLMs and/or by taking prior distributions with mass functions at a single point in the related parameter spaces. This covers several problems of designing experiments for model discrimination.

In the next section an extended theorem based on the General Equivalence Theorem is stated and the Fréchet derivative for the Generalized T–optimality criterion function is presented along with its specific expressions for the five subclasses of GLMs considered in this section.

5.4. A THEOREM FOR GENERALIZED T–OPTIMUM DESIGNS

Following the definition of Generalized T–optimality in the previous section, some relevant properties of the criterion functions are derived in this section with the purpose of establishing conditions of optimality. Basically, what ought to be verified is whether analogous results to those of Chapters 3 and 4, which provide important means for checking optimality and constructing optimal designs, also hold in the general situation considered in this chapter.

Based on the same rationale used in Section 4.5 it suffices to prove concavity of expressions (5.3.1.a) and/or (5.3.1.b) in order to prove concavity of the Generalized T–optimality criterion function (5.3.4). Directional differentiation of this criterion function, in the Fréchet's sense, provides the other essential result for proving a Theorem which generalizes Theorems 3.1 and 4.1. Further details of these demonstrations are presented in Appendix C.

Let us denote the solution of the optimization problem presented by the generalized discrimination totals (5.3.1.a) and (5.3.1.b) as follows. For $j = 1,2$; $\bar{j} =$ not j ; $\xi \in \mathcal{H}$; and $\beta_j \in B_j$ let

$$\int_{\mathfrak{X}} d(x,\beta_j,\beta_{\bar{j}}^{\dagger})\ \xi(dx) = \inf_{\beta_{\bar{j}} \in B_{\bar{j}}} \int_{\mathfrak{X}} d(x,\beta_j,\beta_{\bar{j}})\ \xi(dx) \qquad (5.4.1)$$

where $d(x,\beta_j,\beta_{\bar{j}})$ is as in definition 5.1 and Table 5.1.

Suppose that, when $\xi = \xi^{*}$, (5.4.1) has unique solutions denoted by $\beta_{\bar{j}}^{*}$ over all $\beta_j \in B_j$ which is relevant to the prior $p_{0j}(\beta_j)$, $\{j = 1,2\}$. Then, we may state the following general Theorem for discriminating between two generalized linear models.

THEOREM 5.1

(i) a necessary and sufficient condition for a design ξ^{*} to be Bayesian generalized T–optimum is fulfilment of the inequality

$$\psi(x,\xi^{*}) \leq \Gamma(\xi^{*})\ , \text{ for all } x \in \mathcal{H}$$

where $\psi(x,\xi^{*}) = \sum_{j} \pi_{0j} E_{\beta_j}\left\{d(x,\beta_j,\beta_{\bar{j}}^{*})\right\}$;

(ii) at the points of the Bayesian generalized T–optimum design $\psi(x,\xi^{*})$ achieves its upper bound ;

(iii) for any non–optimum design ξ, that is a design for which $\Gamma(\xi) < \Gamma(\xi^{*})$,

$$\sup_{x \in \mathfrak{X}} \psi(x,\xi) > \Gamma(\xi^{*})\ ;$$

(iv) the set of Bayesian generalized T–optimum designs is convex.

The derivatives $d(x,\beta_j,\beta^*_{\bar{j}})$ particularize into the five subclasses of GLMs in expressions which are very similar to those of Table 5.1. Indeed, the only change is due to replacing $\beta_{\bar{j}}$ by $\beta^*_{\bar{j}}$. The computation of such derivative functions is crucial at the stage of checking optimality as seen in previous chapters. In practice, the derivative function ought to be computed over a grid in the design region $\mathcal{X}$ and then plotted against the components of this grid so as to demonstrate the optimality or non–optimality of the design candidate.

In the next section we present four numerical examples to illustrate these procedures.

5.5. EXAMPLES

To illustrate the results introduced in this chapter we concentrate on a particular problem of link choice for the Gamma subclass of GLMs since the Normal and Binomial subclasses have been studied in Chapters 3 and 4, respectively. Here, two link functions are regarded, namely the reciprocal and the logarithmic. For a given range of the response variable both link functions combined with suitable linear predictor structures are known to provide reasonably good fits to data supposedly originating from a Gamma distribution. Because the expected values under these two models agree closely only a well designed experiment could provide the means for the experimenter to decide upon which link best fits the data. In what follows we show some examples of locally, partially Bayesian and fully Bayesian T–optimum designs with this purpose.

Firstly, we ought to find an expression for the criterion function specifically for the Gamma subclass of generalized linear models. From Table 5.1 we find that the criterion is defined as the search for a design ξ^* such that (5.3.4) is satisfied, where

$$\Delta_{\bar{j}}(\xi,\beta_j) = \underset{\beta_{\bar{j}} \in B_{\bar{j}}}{\mathrm{Min}} \int_{\mathfrak{X}} 2\left\{-\log\left[\frac{\mu_j(x,\beta_j)}{\mu_{\bar{j}}(x,\beta_{\bar{j}})}\right] + \frac{\mu_j(x,\beta_j)-\mu_{\bar{j}}(x,\beta_{\bar{j}})}{\mu_{\bar{j}}(x,\beta_{\bar{j}})}\right\}\xi(dx),\ j=1,2;\ \bar{j}=\text{not } j.$$

Now, depending upon the nature of prior information that is available, the expression for the criterion function will vary accordingly. In all examples of this section the models to be discriminated are described by their expected values with respect to a single observation, i.e. $\mu_1 = g_1^{-1}(\eta_1) = 1/\eta_1$ and $\mu_2 = g_2^{-1}(\eta_2) = \exp(\eta_2)$ where η_1 and η_2 are given linear structures. Due to the nature of the Gamma distribution the expected values, as well as any data generated by this distribution, must be positive. For the canonical link, the reciprocal, this constraint requires some restrictions on the values of the linear predictor parameters $\{\beta_{1k}\}$ so that the related values of η_1 are positive. However, no restrictions need to be imposed on the linear predictor set of parameters β_2. In all the following examples, we take the design region to be the interval [1.0,10.0] and the values of the linear predictor parameters $\{\beta_{1k}\}$, for the reciprocal link, to be positive so that the constraint is satisfied. The only exception to this assumption concerns Example 5.4 where the design region is the interval [0.1,10.0] although the linear predictor parameters $\{\beta_{1k}\}$ for the reciprocal link based model are again assumed to be positive.

EXAMPLE 5.1. Suppose that the first model is true with inverse link and that its linear predictor structure is $\beta_{10} + \beta_{11}x$ where $\beta_{10} = 0.0$ and $\beta_{11} = 1.0$. The aim is to find a design that discriminates against rival models with logarithmic link function and identical linear predictor structure. Criterion (5.3.2) applies.

The resulting locally T–optimum design is shown below. The optimum value of the criterion function is $\Delta_2(\xi^*,\beta_1) = 9.53 \times 10^{-2}$, where $\beta_1 = (0.0,1.0)$.

$$\xi^* = \begin{Bmatrix} 1.0 & 3.9086 & 10.0 \\ 0.3037 & 0.5513 & 0.145 \end{Bmatrix} \tag{5.5.1}$$

A remarkable feature of design (5.5.1) is that both extremes of the design region, the interval [1.0,10.0], belong to the optimum design although more than 55% of the weight is assigned to the middle support point 3.9086. The plot of the derivative function for design (5.5.1) is displayed in Figure 5.1.1.

TABLE 5.2 – CONCENTRATED 5–POINT PRIOR DISTRIBUTION FOR β_1

β_{10}	β_{11}	$p_{01}(\beta_1)$
0.0	1.0	0.2
–0.1	1.0	0.2
–0.1	0.9	0.2
0.1	1.0	0.2
0.1	0.9	0.2

To proceed with this numerical investigation suppose that both models have the same linear predictor structure as before and consider a combination of values for the linear predictor parameters for the second model, β_2 which provides similar expected values to those for the first model over the interval [1.0,10.0]. As for the first model, five equally likely combinations of values for β_1 are considered. They lie in a tiny region around the point $\beta_1 = (0.0,1.0)$ from the first part of this example. This discrete uniform prior distribution is shown in Table 5.2. Both models are taken to be equally probable to be true. Under these circumstances, criterion (5.3.4) applies.

The resulting fully Bayesian T–optimum design is shown below. The maximized value of the criterion function is $\Gamma(\xi^*) = 7.017 \times 10^{-2}$.

$$\xi^* = \left\{\begin{matrix} 1.0 & 4.0027 & 10.0 \\ 0.2401 & 0.5018 & 0.2581 \end{matrix}\right\} \tag{5.5.2}$$

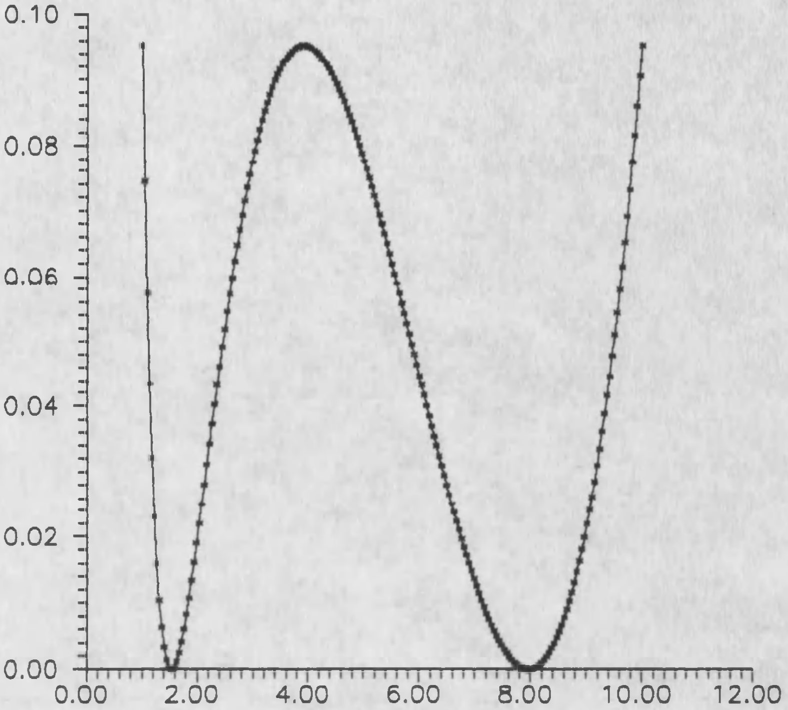


Figure 5.1.1. Derivative Function for design (5.5.1)

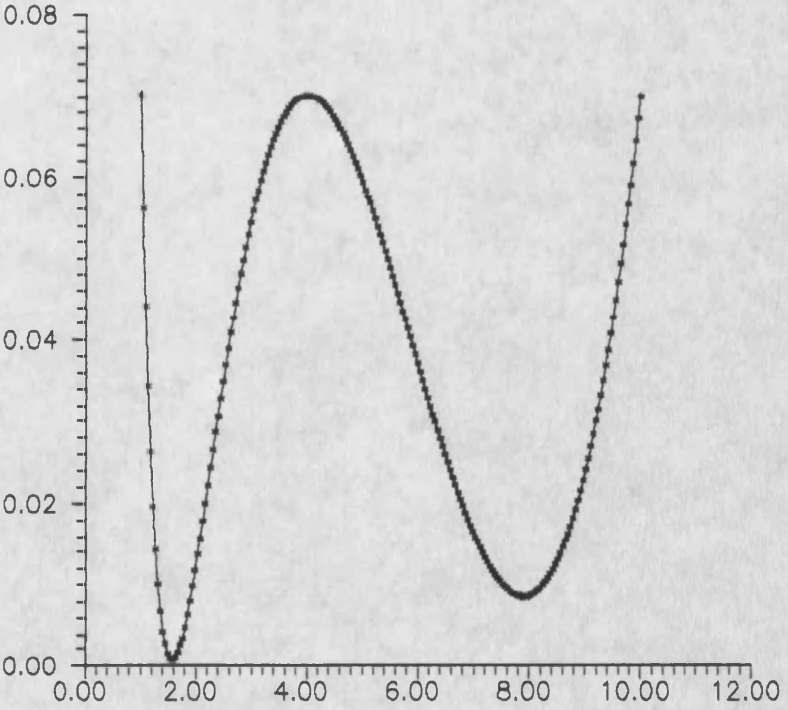


Figure 5.1.2. Derivative Function for design (5.5.2)

As expected, designs (5.5.1) and (5.5.2) show some similarities. Their support points are virtually the same, but the weights are slightly different. The derivative function for this design is displayed in Figure 5.1.2. The main difference between the plots of the derivative functions for designs (5.5.1) and (5.5.2) is the level of smoothness of their valleys. The former is spikier due to its local optimality.

TABLE 5.3 – 9–POINT CONCENTRATED PRIOR PROBABILITY DISTRIBUTIONS FOR PARAMETER SETS β_1 AND β_2

β_{10}	β_{11}	$p_{01}(\beta_1)$	β_{20}	β_{21}	$p_{02}(\beta_2)$
0.0	1.0	0.2	–0.3	–0.2	0.2
0.0	1.1	0.1	–0.3	–0.1	0.1
0.0	0.9	0.1	–0.3	–0.3	0.1
–0.1	1.0	0.1	–0.4	–0.2	0.1
–0.1	1.1	0.1	–0.4	–0.1	0.1
–0.1	0.9	0.1	–0.4	–0.3	0.1
0.1	1.0	0.1	–0.2	–0.2	0.1
0.1	1.1	0.1	–0.2	–0.1	0.1
0.1	0.9	0.1	–0.2	–0.3	0.1

Further investigation is carried out by taking two rather concentrated prior distributions for β_1 and β_2 for which the curves of the expected values in the design region considered are shaped similarly, therefore justifying the need for model discrimination. The prior probabilities for each model to be true are again supposed to be equal to 1/2 and conditioned on the event that the jth model is true the prior distribution for β_j is given in Table 5.3 (j=1,2). The bidimensional points of the parameter spaces B_1 and B_2 listed in Table 5.3 form squares around the points [0.0,1.0] in the former and around [–0.3,–0.2] in the latter, with both centres of squares being twice as likely as the other points. Again, criterion (5.3.4) applies.

The resulting fully Bayesian T–optimum design is shown below for which the maximized value of the criterion function is $\Gamma(\overset{*}{\xi}) = 7.172 \times 10^{-2}$.

$$\xi^* = \begin{Bmatrix} 1.0 & 3.95 & 10.0 \\ 0.231 & 0.4869 & 0.2821 \end{Bmatrix} \tag{5.5.3}$$

As compared to designs (5.5.1) and (5.5.2) the structure of optimum design (5.5.3) is kept although the middle point is assigned less weight than before. The plot of the derivative function for this design, shown in Figure 5.1.3, is shaped like the plots of the previous designs, displayed in Figures 5.1.1 and 5.1.2.

To summarize the results in this example we could say that for all combinations of values for the linear predictor parameters β_1 and β_2 that were considered, there always are three points in the design region which seem to contain most of the information about discrimination, two of them being the extreme points of the design region whereas the remaining is located around 4.0 and is assigned most of the weight in all optimum designs from (5.5.1) to (5.5.3).

EXAMPLE 5.2. Now, suppose that the linear predictor structures are described by $\beta_{10} + \beta_{11}x + \beta_{12}x^2$ and $\beta_{20} + \beta_{21}x$ for the reciprocal link and the logarithmic link based models, respectively. Initially, let us assume that the first model is true and its linear predictor parameters are known to be equal to $\beta_1 = (0.0,0.1,0.01)$. Thus, the aim is to find a design to maximize lack of fit for the second model, the logarithmic. Accordingly, the criterion of optimization applied is (5.3.2).

The resulting locally T–optimum design is shown below. The optimum value of the criterion is $\Delta_2(\overset{*}{\xi},\beta_1) = 1.0846 \times 10^{-1}$, where $\beta_1 = (0.0,0.1,0.01)$.

$$\xi^* = \begin{Bmatrix} 1.0 & 3.9876 & 10.0 \\ 0.2975 & 0.5547 & 0.1478 \end{Bmatrix} \tag{5.5.4}$$

According to the above design most of the trials should be assigned to the middle support point 3.9876. Again, the extremes of the design region are

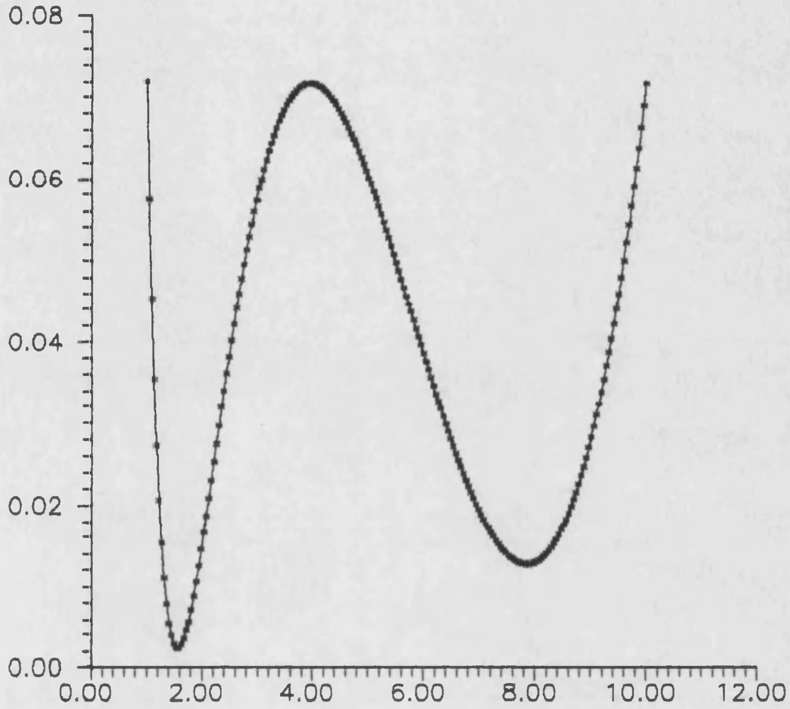


Figure 5.1.3. Derivative Function for design (5.5.3)

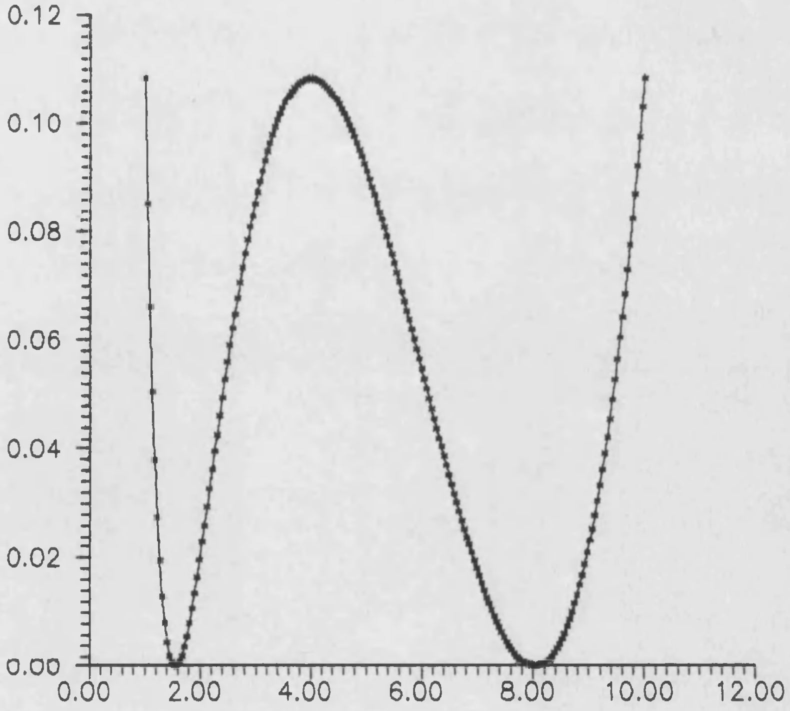


Figure 5.2.1. Derivative Function for design (5.5.4)

contained in the optimum design. The plot of the directional derivative w.r.t. design (5.5.4) is shown in Figure 5.2.1.

Now, let us extend this analysis by taking the models as equally likely to be true. Further, suppose that the linear predictor parameters $\{\beta_{jk}\}$ are known (conditionally on the event that the jth model is true). For this situation to be more realistic we proceeded as follows: a vector of expected values for the reciprocal link based model, where $\beta_1 = (0.0,0.1,0.01)$ and $x \in [1.0,10.0]$, was generated over a grid of fifty equally spaced points, starting at $x = 1.0$ and ending at $x = 10.0$. Then, the rival model (logarithmic) was fitted to these expected values. The resulting linear predictor parameter estimates were approximately equal to $\hat{\beta}_2 = (1.95,-0.29)$. Thus, by taking the values of β_2 to be equal to 1.95 and -0.29, the expected responses under each of the rival models should agree closely in most points of the design region, therefore, justifying the utilization of experimental designing methods for model discrimination.

Criterion (5.3.4) applies, with the prior distributions for the linear predictor parameters β_1 and β_2 having mass one at the points $\beta_1 = (0.0,0.1,0.01)$ and $\beta_2 = (1.95,-0.29)$, respectively, making the expression for the criterion function reduce to $\frac{1}{2}\left[\Delta_2(\xi,\beta_1) + \Delta_1(\xi,\beta_2)\right]$. In words, the present problem concerns searching for a design that maximizes an equally weighted compromise between the fitting of the second model w.r.t. the expected values under the first and vice versa.

After rounding the values of the support points and related weights to four decimal places, the resulting fully Bayesian T–optimum design, shown below, becomes identical to design (5.5.4), whilst the maximized criterion function is approximately a half of the value for the criterion function at the previous design, that is $\Gamma(\xi^*) \cong 5.423 \times 10^{-2}$ whereas $\Delta_2(\xi^*,\beta_1) \cong 1.0846 \times 10^{-1}$ and $\Delta_1(\xi^*,\beta_2) \cong 0.0$.

$$\xi^* = \begin{Bmatrix} 1.0 & 3.9876 & 10.0 \\ 0.2975 & 0.5547 & 0.1478 \end{Bmatrix} \tag{5.5.5}$$

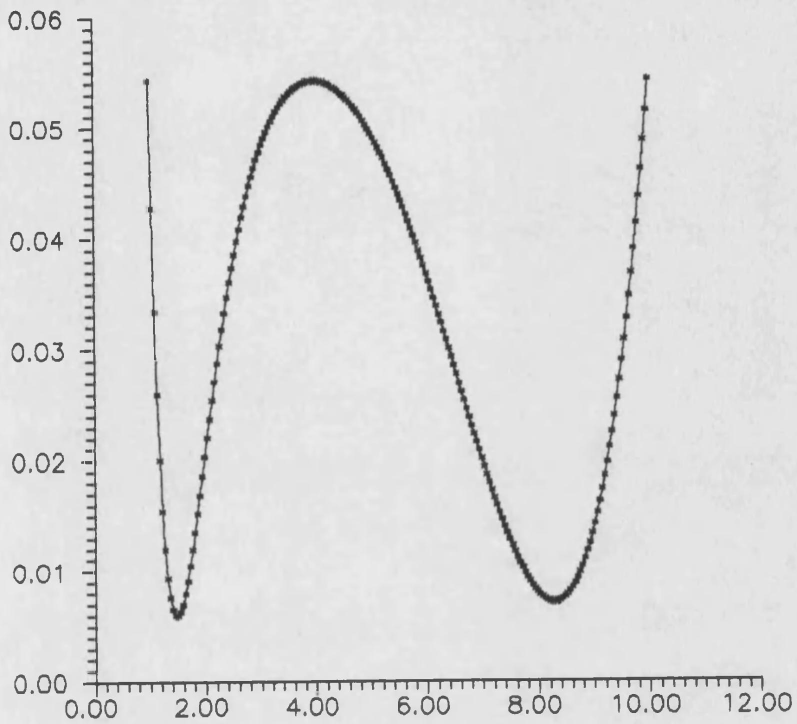


Figure 5.2.2. Derivative Function for design (5.5.5)

Figure 5.2.2 shows the plot of the directional derivative for this design. It is important to notice that although designs (5.5.4) and (5.5.5) are almost identical, the related plots of the directional derivatives are not, the former showing typical features of locally optimum designs.

These results highlight an absolute influence of the first part of the expression of criterion (5.3.4), i.e. $\frac{1}{2}\left[\Delta_2(\xi,\beta_1)\right]$, in determining the optimum design. However, the oustanding numerical result arising from this example has to do with the fact that $\Delta_1(\xi^*,\beta_2) \cong 0.0$. This means that the expected values under the second model, with $\beta_2 = (1.95,-0.29)$, are fitted almost exactly by the estimates of the expected values under the first model at the support points of design (5.5.5). By contrast, the converse process does not provide such a precise fit. Indeed, its contribution to the criterion function at the optimum counts for almost 100%.

To complement this sensitivity analysis two additional situations are considered. The first concerns the case in which the second model is known to be true but there is uncertainty w.r.t. the true values of its parameters. This case is considered with the purpose of finding out what happens to the structure of the optimum design when the influence of the first model linear predictor parameters is removed. In the second not only is either model true with probability 1/2 but also there is uncertainty about the values of both linear predictor parameter sets, reflected by conditional prior distributions on the event that the jth model is true. Here, our aim is to verify whether the dominance of the first model features over the second, as far as the optimum design is concerned, is carried over to other points in the parameter space around the original point $\beta_1 = (0.0,0.1,0.01)$.

TABLE 5.4 – 9–POINT PRIOR PROBABILITY DISTRIBUITONS FOR PARAMETER SETS β_1 AND β_2

β_{10}	β_{11}	β_{12}	$p_{01}(\beta_1)$	β_{20}	β_{21}	$p_{02}(\beta_2)$
0.0	0.10	0.010	0.2	1.95	–0.29	0.2
0.0	0.10	0.005	0.1	1.95	–0.30	0.1
0.0	0.10	0.015	0.1	1.95	–0.28	0.1
0.0	0.05	0.010	0.1	1.90	–0.29	0.1
0.0	0.05	0.005	0.1	1.90	–0.30	0.1
0.0	0.05	0.015	0.1	1.90	–0.28	0.1
0.0	0.15	0.010	0.1	2.00	–0.29	0.1
0.0	0.15	0.005	0.1	2.00	–0.30	0.1
0.0	0.15	0.015	0.1	2.00	–0.28	0.1

For comparative purposes, the prior for β_2 is the same in both situations and the values for the linear predictor parameters β_1 and β_2 in the priors lie very closely to the previous values of $\beta_1 = (0.0,0.1,0.01)$ and $\beta_2 = (1.95,-0.29)$. Table 5.4 displays these values and their prior probabilities.

The first coordinates of all nine tridimensional points for β_1 are equal to zero making the corresponding expected values be of the form $\frac{1}{\beta_{11}x + \beta_{12}x^2}$. In addition, the sets described by the second and third coordinates of β_1 form a square centered at $\beta_{11} = 0.1$ and $\beta_{12} = 0.01$. Similarly, the nine bidimensional points for β_2 form a square centered at $\beta_2 = (1.95,-0.29)$. In each of the priors the center of the square is twice as likely as the other points.

In the search corresponding to the former case, where the second model is true, criterion (5.3.3) applies, yielding the partially Bayesian T–optimum design (5.5.6) for which the maximized value of the criterion is $\gamma_1(\overset{*}{\xi}) = 7.373 \times 10^{-3}$.

$$\overset{*}{\xi} = \begin{Bmatrix} 1.0 & 2.3872 & 6.4243 & 10.0 \\ 0.0935 & 0.1982 & 0.3920 & 0.3163 \end{Bmatrix} \tag{5.5.6}$$

For the latter, criterion (5.3.4) applies, achieving the maximum value $\Gamma(\overset{*}{\xi}) = 5.635 \times 10^{-2}$ at the fully Bayesian T–optimum design (5.5.7).

$$\overset{*}{\xi} = \begin{Bmatrix} 1.0 & 3.9968 & 10.0 \\ 0.2962 & 0.5560 & 0.1478 \end{Bmatrix} \tag{5.5.7}$$

The reduction of the number of support points from four in design (5.5.6) to three in design (5.5.7) is inconsistent with the intuitive idea that the more uncertain the amount of prior information the greater the number of support points in the optimal design (recall that in the former case there was greater knowledge about the true values of the parameters involved in the problem than in the latter). However, a closer examination of the numerical results reveals that for design (5.5.7), the values of the partial quantities expressed by (5.3.1.a) are much greater than those expressed by (5.3.1.b). Indeed, the latter quantities approach zero reflecting almost perfect fits for the expected values under the second model. In other words, similarly to previous analysis, the value of $\Gamma(\overset{*}{\xi}) = 5.635 \times 10^{-2}$ arises from the first part of the criterion function (5.3.4), that is $\frac{1}{2}\left\{E_{\beta_1}\left[\Delta_2(\xi,\beta_1)\right]\right\}$ once $E_{\beta_2}\left[\Delta_1(\xi,\beta_2)\right] \cong 0.0$, where the expected values are taken over the priors of

Table 5.4. This confirms the dominance of the first model features over the second that had been verified before.

Thus, we could say that the search for the optimum design is dominated by the values of β_1, that is the power for model discrimination is much greater under the assumption that the first model is true than vice–versa. Consequently, design (5.5.7) is very similar to previous designs (5.5.5) and (5.5.4) where in the former there is a similar domination and in the latter the first model is taken to be true. More important for these similar structures, however, is that the prior values of β_1 in Table 5.4 lie closely in the parameter space to the value of β_1 taken to be true previously. Hence, only three support points are needed in design (5.5.7) as were in designs (5.5.5) and (5.5.4) even though there is a much less degree of precision under the priors of Table 5.4 than there was in the first two cases.

On the other hand, when the second model is assumed true with a prior for its linear predictor parameters, design (5.5.6) is obtained. There are four support points in this design and apart from the feature that the two extremes of the interval [1.0,10.0] belong to the optimum design there are no other resemblances to design (5.5.7). This remark seems to support the idea that it is the region of the parameter space that most influences the number of support points in the optimum design rather than lack of precision in the prior information. However, it should be pointed out that the number of support points in design (5.5.6) is one more than the number of linear predictor parameters in the rival model, a result that also appeared when the reverse situation was considered, i.e. when the reciprocal model was assumed to be true with the rival model (logarithmic) having two linear predictor parameters. This resulted in three support points for design (5.5.4).

Another interesting result is that the criterion function is maximized at levels which are smaller than those corresponding to design (5.5.7), highlighting the difference between the levels of precision in which the first model expected values fit to the second model ones and vice–versa.

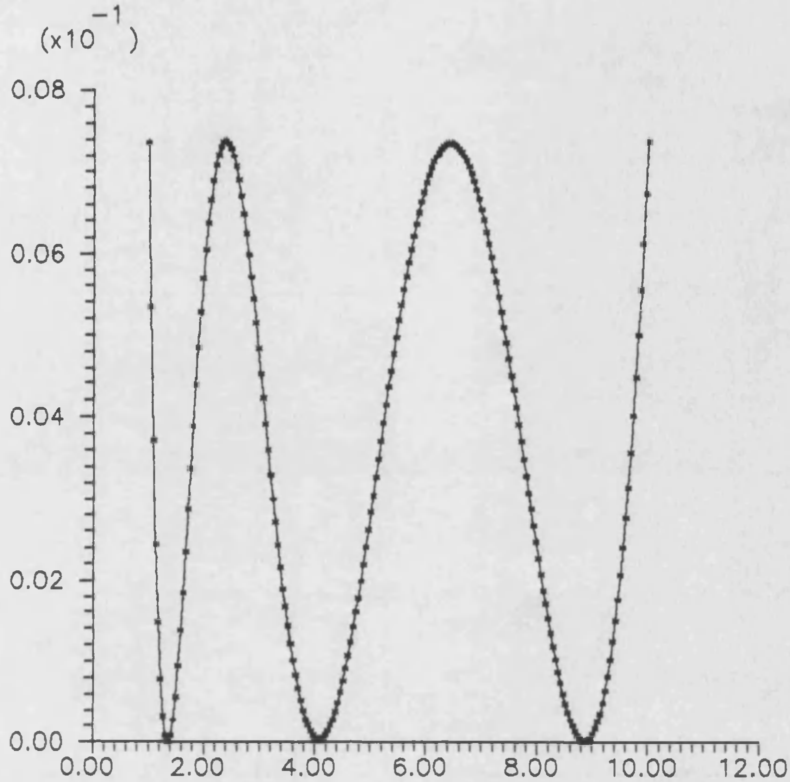


Figure 5.2.3. Derivative Function for design (5.5.6)

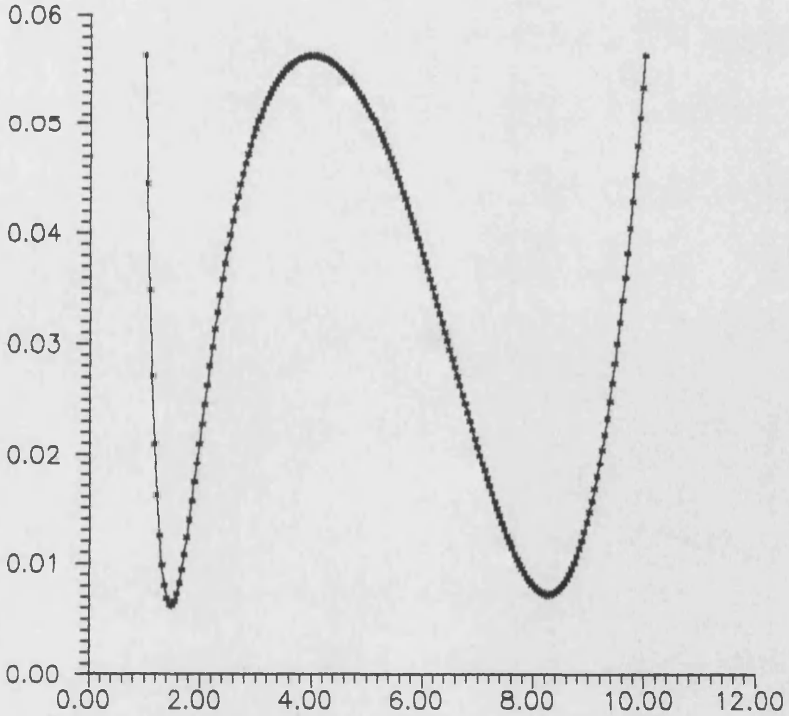


Figure 5.2.4. Derivative Function for design (5.5.7)

Figures 5.2.3 and 5.2.4, corresponding to designs (5.5.6) and (5.5.7), respectively, show the shape of the derivative function curves over the design region. The latter being shaped like that related to design (5.5.5).

EXAMPLE 5.3. With the purpose of investigating, in an informal manner, differences in the features of the optimum designs as we move from region to region in the parameter space, or in the combination of parameter spaces, we now take the logarithmic link based model to be true with a first degree polynomial linear predictor. The aim is to find an optimum design to discriminate the rival model, a reciprocal link based one with a second degree linear predictor structure. Additionally, we suppose that the values of the linear predictor parameters in the true model are known to be $\beta_{10} = 2.5$ and $\beta_{11} = -0.5$.

Criterion (5.3.2) applies where the first model is the logarithmic and the second, the reciprocal. The local T–optimum design is shown below. The maximized value of the criterion is $\Delta_2(\xi^*, \beta_1) = 1.41 \times 10^{-1}$, for $\beta_1 = (2.5, -0.5)$.

$$\xi^* = \begin{Bmatrix} 1.0 & 2.0097 & 5.3627 & 10.0 \\ 0.0709 & 0.1040 & 0.3669 & 0.4582 \end{Bmatrix} \qquad (5.5.8)$$

Design (5.5.8) is rather skewed in the sense that only 17.49% of the total weight is assigned to the first two support points, leaving the bulk of the weight to the last two points. It is important to point out that the numbers of support points in designs (5.5.8) and (5.5.6) coincide. Since the logarithmic link was assumed to be true in both cases that suggests a cause for this happening other than merely coincidence. Furthermore, the structures of these designs are alike despite the linear predictor parameters in both cases being different. It remains to investigate whether the excessive weight assigned to the last two points is needed for the discrimination process or to the estimation of the linear predictor parameters. That is, of course, if these objectives can be divided in this way. The plot of the derivative function for design (5.5.8) is shown in Figure 5.3.1.

TABLE 5.5 – 9–POINT CONCENTRATED PRIOR PROBABILITY DISTRIBUTION FOR β_1

β_{10}	β_{11}	$p_{01}(\beta_1)$
2.50	–0.50	0.2
2.50	–0.45	0.1
2.50	–0.55	0.1
2.45	–0.50	0.1
2.45	–0.45	0.1
2.45	–0.55	0.1
2.55	–0.50	0.1
2.55	–0.45	0.1
2.55	–0.55	0.1

Let us now suppose that there is a prior distribution for the linear predictor parameters $\{\beta_{1j}\}$ but we keep the assumption that the first model is true, i.e. the logarithmic. The linear predictor structures are the same as before. Table 5.5 shows the prior distribution for β_1. Criterion (5.3.3) applies.

The distribution of probabilities and points in the above prior follows our usual arrangement so far, that is the value of β_1 assumed to be true in the first part of the current example is the center of a square formed by the other eight points and the probabilities are equally distributed among the points except the center which is twice as likely. Because the area of this square is small the prior is a rather concentrated one reflecting a slight lack of precision in the prior information.

The resulting partially Bayesian T–optimum design is shown below, for which the maximized value of the criterion is $\gamma_2(\xi^*) = 1.47178 \times 10^{-1}$.

$$\xi^* = \begin{Bmatrix} 1.0 & 2.0008 & 5.3056 & 10.0 \\ 0.0698 & 0.1010 & 0.3627 & 0.4665 \end{Bmatrix} \qquad (5.5.9)$$

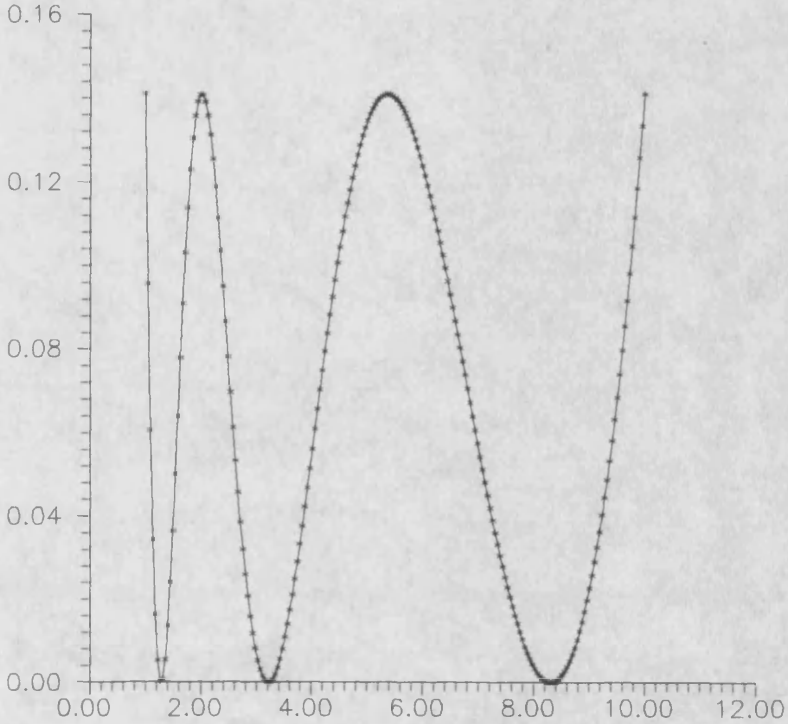


Figure 5.3.1. Derivative Function for design (5.5.8)

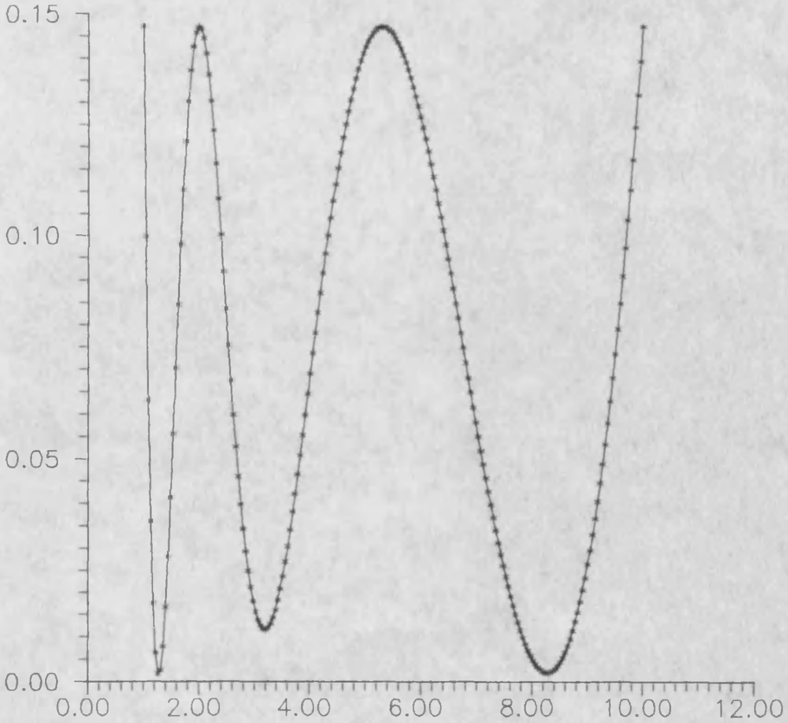


Figure 5.3.2. Derivative Function for design (5.5.9)

As it should be expected, designs (5.5.8) and (5.5.9) are very similar as were the values of the criterion functions. The plot of the derivative function corresponding to design (5.5.9) is displayed in Figure 5.3.2, where it can be noticed that the curve does not touch zero, a feature of Bayesian designs.

EXAMPLE 5.4. Here, a more thorough investigation of Bayesian designs to discriminate between the logarithmic and the reciprocal links is carried out. Suppose that the linear predictor structures for those models are given by a first and a second degree polynomials, respectively. Consider eight sets of linear predictor parameters β_1 (logarithmic link) which are listed in the first part of Table 5.6.

TABLE 5.6 – SETS OF PARAMETERS β_1 FOR THE FIRST MODEL AND ITS RELATED ESTIMATED PARAMETERS $\hat{\beta}_2$ FOR THE SECOND, BASED ON A GRID OF 50 EQUIDISTANT POINTS OVER THE INTERVAL [0.1,10.0]

β_{10}	β_{11}	$\hat{\beta}_{20}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$
2.5	–0.40	0.1198	–0.0458	0.0332
2.5	–0.45	0.1339	–0.0762	0.0487
2.5	–0.50	0.1506	–0.1150	0.0691
2.5	–0.55	0.1699	–0.1625	0.0950
2.5	–0.60	0.1916	–0.2192	0.1274
2.5	–0.65	0.2155	–0.2851	0.1670
2.5	–0.70	0.2415	–0.3605	0.2145
2.5	–0.75	0.2693	–0.4453	0.2708

Using a grid of 50 equidistant points in the interval [0.1,10.0] a curve linking the expected values of the response variable, for the logarithmic model and parameter sets as listed in Table 5.6, is shown in Figure 5.4.1. Then, the reciprocal link based model with a second degree polynomial linear predictor was fitted for each of the eight sets of expected values generated over the 50–point grid. The resulting estimates $\hat{\beta}_2$ are shown in the second part of Table 5.6 and the curves

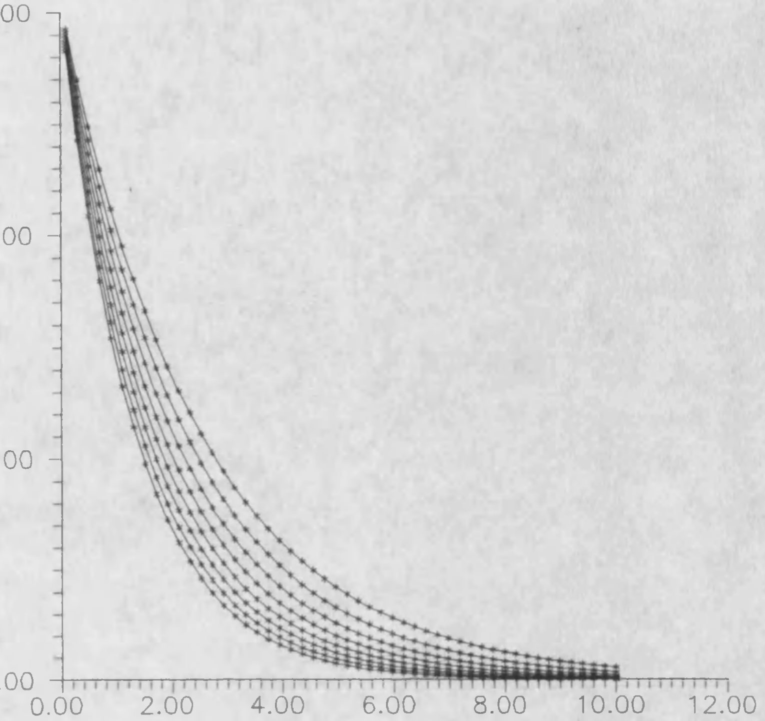


Figure 5.4.1. Curves of the expected values for 8 parameter sets. Logarithmic link.

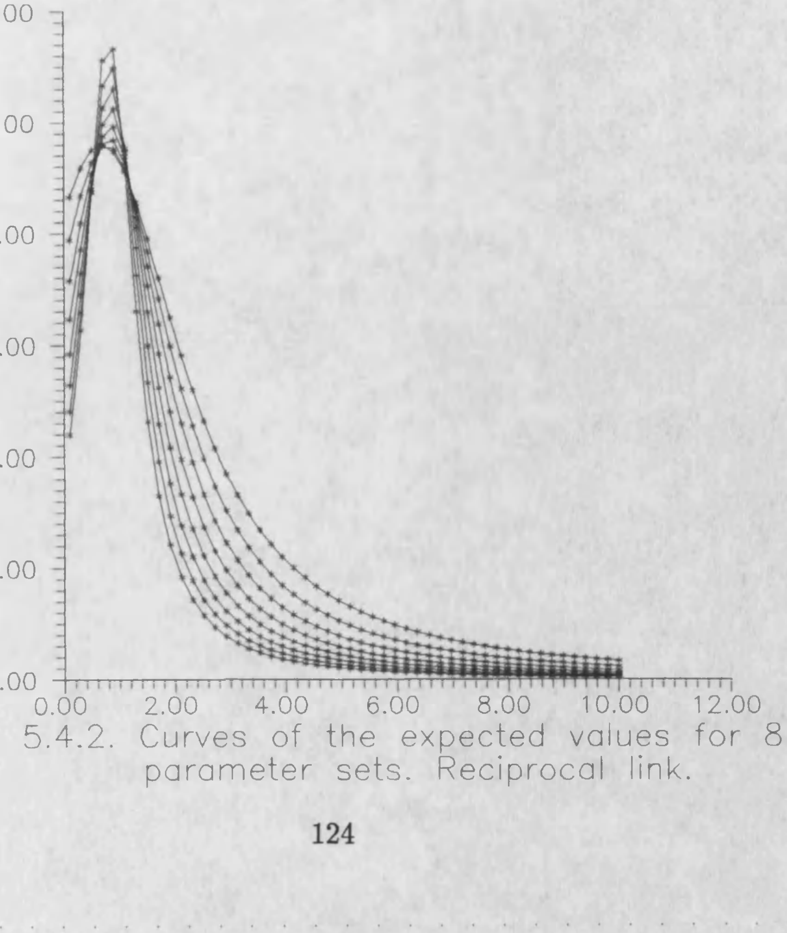


Figure 5.4.2. Curves of the expected values for 8 parameter sets. Reciprocal link.

corresponding to the estimated expected values under the reciprocal link based model are shown in Figure 5.4.2.

An overall look at Figures 5.4.1 and 5.4.2 reveals similar patterns generated by the logarithmic and reciprocal links. Indeed, the latter set of curves was determined from fitting the former. However, in a real situation each pair of curves (generated by separate sets of parameters in the same row of Table 5.6) could well describe the pattern of a given data set. Furthermore, the curves in Figures 5.4.1 and 5.4.2 might represent the plausible patterns in situations where there is uncertain prior information about the true model and its parameters.

Therefore, the following case is regarded. Both models are equally likely to be true and conditionally on the event that each model is true there is a prior distribution for the linear predictor parameters which is uniform over the corresponding set of eight points listed in Table 5.6. The aim is to find a design optimizing criterion (5.3.4).

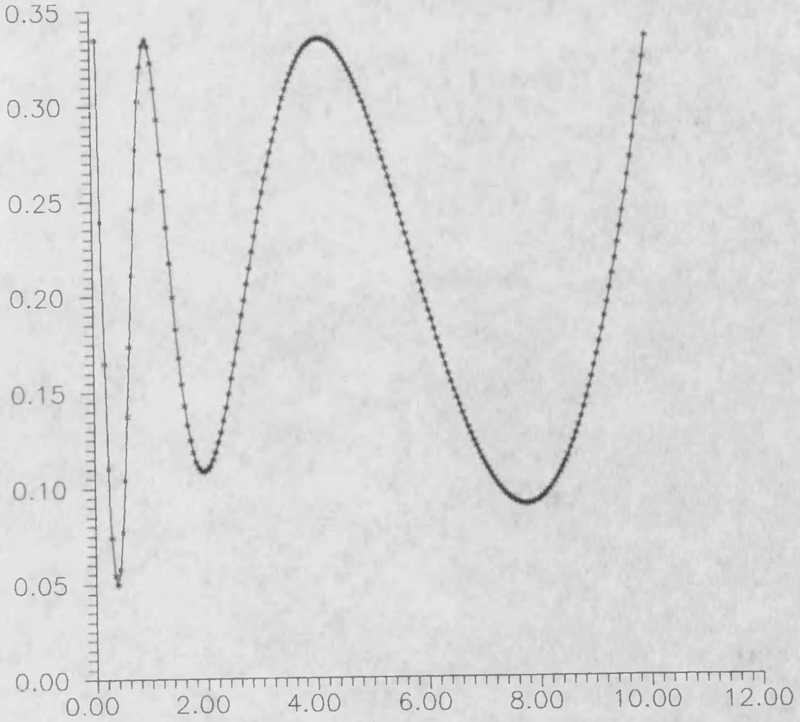


Figure 5.4.3. Derivative Function for design (5.5.10)

The resulting fully Bayesian T–optimum design is shown below for which the maximized value of the criterion function is $\Gamma(\xi^*) = 3.3477 \times 10^{-1}$.

$$\xi^* = \begin{Bmatrix} 0.1 & 0.99 & 4.1271 & 10.0 \\ 0.053 & 0.1169 & 0.3406 & 0.4895 \end{Bmatrix} \qquad (5.5.10)$$

Following the tendency of all previous optimum designs in this chapter, both extremes of the design region belong to optimum design (5.5.10). Another remarkable feature of this design is the increase of weights from the smallest support point to the largest. However, it is rather difficult to interpret the fact that almost 50% of the weight is assigned to the largest support point in design (5.5.10) since a comparison between the curves of Figures 5.4.1 and 5.4.2 suggests that points in the first quarter of the design region should be highly informative to discriminate between the models as the shapes of the patterns provided by the models in this region are very discrepant. By contrast, the patterns seem to coincide in the remaining parts of the design region, especially towards x = 10.0, the upper limit of the interval (design region).

So, why is it necessary to assign so much weight to the fourth support point x = 10.0 ? An explanation for this apparently counter–intuitive optimum design is that there would be no need to put excessive weight on points of the design region where the expected values under each model vary significantly as that would not contribute further to the maximization of the criterion function. The reason is the need for estimating the linear predictor parameters so that the patterns of the models are well determined and thus the process of discriminating between the models can be carried out reliably.

Figure 5.4.3 shows the plot of the derivative function related to design (5.5.10) where typical features of Bayesian designs can be seen. Furthermore, we can see that the two middle support points of design (5.5.10) are located in the regions where the curves of Figures 5.4.1 and 5.4.2 mostly disagree. In the intervals between the second and third support points, as well as the third and the fourth, there seem to be agreement between the curves which is reflected by the two valleys apparent in Figure 5.4.3. To conclude this chapter, further questions are discussed in the next section.

5.6. DISCUSSION

The research described in this chapter provides designs for discrimination between any two GLMs as long as they belong to the same subclass. Calculations for further examples have led to designs which, like those described here, have the number of design points one greater than the number of parameters in the rival models. Similar results have also been obtained for nonlinear regression models using a version of the T–optimality criterion.

Based on theoretical results and specific expressions for the criterion functions presented in this chapter it is feasible to obtain numerical examples of optimum designs to discriminate between two GLMs belonging to subclasses other than Normal, Binomial and Gamma. There is no reason to believe that there will be more difficulties in determining optimum designs for subclasses like Poisson, Inverse Gaussian or even Negative Binomial. Any situation in which there are two competing generalized linear models these methods can be applied.

Also based on theoretical results of this chapter, sequential methods for determining optimum designs for model discrimination can be developed. To implement this one could follow the same steps as those in Atkinson and Fedorov (1975a) and generalize their sequential method to deal with any situation that concerns discriminating between two generalized linear models. In this case, the use of simulation methods is essential to verify the behaviour, convergence and efficiency of sequentially designed experiments.

Moreover, both sequential and Bayesian approaches may be combined to provide a more powerful tool for searching optimum designs w.r.t. T–optimality or similarly defined criteria. Under this combined approach, more sophisticated concepts and theoretical results from Bayesian statistics have to be used, increasing both the theoretical and numerical complexities involved in the problem. In addition, the need for a reasonably large amount of observations rules out situations in which there are constraints in the amount of experimentation.

The problem of discriminating between two generalized nonlinear models remains to be investigated.

APPENDIX C

The results presented in this appendix generalize those of Appendices A and B. We adopt the notation corresponding to Definition 5.1 where the generalized discrimination totals are introduced. Thus, for j=1,2 ; $\bar{j}$=not j ; $\xi \in \mathfrak{X}$; and $\beta_j \in B_j$ denote

$$d(x,\beta_j,\beta_{\bar{j}}) = 2w[\mu_j(x,\beta_j)\{\theta_j(x,\beta_j) - \theta_{\bar{j}}(x,\beta_{\bar{j}})\} - b\{\theta_j(x,\beta_j)\} + b\{\theta_{\bar{j}}(x,\beta_{\bar{j}})\}] \qquad (C.1)$$

and

$$l(\xi,\beta_j,\beta_{\bar{j}}) = \int_{\mathfrak{X}} d(x,\beta_j,\beta_{\bar{j}})\ \xi(dx) \qquad (C.2)$$

According to (C.1) and (C.2), the generalized discrimination totals (5.3.1.a) and (5.3.1.b) may now be denoted, respectively, by

$$\Delta_1(\xi,\beta_2) = l(\xi,\beta_2,\beta_1^{\dagger}) = \inf_{\beta_1 \in B_1} l(\xi,\beta_2,\beta_1) \qquad (C.3.1)$$

and

$$\Delta_2(\xi,\beta_1) = l(\xi,\beta_1,\beta_2^{\dagger}) = \inf_{\beta_2 \in B_2} l(\xi,\beta_1,\beta_2) \qquad (C.3.2)$$

Let us now prove concavity of (C.3.1) and (C.3.2) on the set of measures $\mathscr{H}$, i.e. for j=1,2

$$\Delta_{\bar{j}}(\xi,\beta_j) \geq (1-\alpha)\ \Delta_{\bar{j}}(\xi_1,\beta_j) + \alpha\ \Delta_{\bar{j}}(\xi_2,\beta_j) \qquad (C.4)$$

where $0 \leq \alpha \leq 1$; $\xi = (1-\alpha)\xi_1 + \alpha\xi_2$; and $\xi_1,\xi_2 \in \mathscr{H}$.

Proof of (C.4)

For j=1,2, let $\alpha \in \mathbb{R}$, $0 \leq \alpha \leq 1$; $\xi = (1-\alpha)\xi_1 + \alpha\xi_2$; and $\xi_1,\xi_2 \in \mathscr{H}$.

Then

$$\begin{aligned}
\Delta_{\bar{j}}(\xi,\beta_j) &= \inf_{\beta_{\bar{j}} \in B_{\bar{j}}} l(\xi,\beta_j,\beta_{\bar{j}}) = \inf_{\beta_{\bar{j}} \in B_{\bar{j}}} \Big\{l((1-\alpha)\xi_1+\alpha\xi_2,\beta_j,\beta_{\bar{j}})\Big\} \\
&= \inf_{\beta_{\bar{j}} \in B_{\bar{j}}} \Big\{(1-\alpha)\ l(\xi_1,\beta_j,\beta_{\bar{j}}) + \alpha\, l(\xi_2,\beta_j,\beta_{\bar{j}})\Big\} \\
&\geq \inf_{\beta_{\bar{j}} \in B_{\bar{j}}} \Big\{(1-\alpha)\ l(\xi_1,\beta_j,\beta_{\bar{j}})\Big\} + \inf_{\beta_{\bar{j}} \in B_{\bar{j}}} \Big\{\alpha\, l(\xi_2,\beta_j,\beta_{\bar{j}})\Big\} \\
&= (1-\alpha)\ \Delta_{\bar{j}}(\xi_1,\beta_j) + \alpha\ \Delta_{\bar{j}}(\xi_2,\beta_j).
\end{aligned}$$

Thus, the generalized discrimination totals $\{\Delta_{\bar{j}}(\xi,\beta_j),\ j=1,2\}$ are concave functions on $\mathscr{H}$. Furthermore, taking expectations over the prior probability distributions π_0 and $\{p_{0j}(\beta_j),\ j=1,2\}$ in both sides of inequality (C.4), we find

$$E_{\pi_0}E_{\beta_j}\left\{\Delta_{\bar{j}}(\xi,\beta_j)\right\} \geq E_{\pi_0}E_{\beta_j}\left\{(1-\alpha)\ \Delta_{\bar{j}}(\xi_1,\beta_j) + \alpha\ \Delta_{\bar{j}}(\xi_2,\beta_j)\right\}$$

$$\Gamma(\xi) \geq (1-\alpha)\Gamma(\xi_1) + \alpha\Gamma(\xi_2).$$

Hence, the criterion function related to the Bayesian generalized T–optimality criterion (Definition 5.4) is also concave on $\mathscr{H}$.

In order to prove Theorem 5.1, the Fréchet derivative of $\Gamma(\xi)$ is required. Because of the linearity of Fréchet derivatives, we can write

$$\frac{\partial\Gamma(\xi)}{\partial\alpha} = \sum_j \pi_{0j}\ E_{\beta_j}\left[\frac{\partial\Delta_{\bar{j}}(\xi,\beta_j)}{\partial\alpha}\right] \tag{C.5}$$

where $\xi = (1-\alpha)\xi_1 + \alpha\xi_2$; $0 \leq \alpha \leq 1$; and the Fréchet derivatives of the generalized discrimination totals are given by

$$\frac{\partial\Delta_{\bar{j}}(\xi,\beta_j)}{\partial\alpha} = \inf_{\beta_{\bar{j}}\in B^{\dagger}_{\bar{j}}}\left\{l(\xi_2,\beta_j,\beta_{\bar{j}}) - l(\xi_1,\beta_j,\beta_{\bar{j}})\right\}\ (j=1,2) \tag{C.6}$$

where $B^{\dagger}_{\bar{j}}$ denotes the solution set of equation (5.4.1).

<u>Proof of (C.6)</u>

In the notation adopted here, the solution set of equation (5.4.1) is represented by

$$l(\xi,\beta_j,\beta^{\dagger}_{\bar{j}}) = \inf_{\beta_{\bar{j}}\in B_{\bar{j}}} l(\xi,\beta_j,\beta_{\bar{j}})$$

where $\xi = (1-\alpha)\xi_1 + \alpha\xi_2$; $\alpha \in [0,1]$; and $\xi_1,\xi_2 \in \mathscr{H}$. Then, for j=1,2

$$\frac{\partial\Delta_{\bar{j}}(\xi,\beta_j)}{\partial\alpha} = \frac{\partial}{\partial\alpha}\left\{\inf_{\beta_{\bar{j}}\in B^{\dagger}_{\bar{j}}(\alpha)} l(\xi,\beta_j,\beta_{\bar{j}})\right] = \inf_{\beta_{\bar{j}}\in B^{\dagger}_{\bar{j}}(\alpha)}\left[\frac{\partial}{\partial\alpha} l(\xi,\beta_j,\beta_{\bar{j}})\right\}$$

$$= \inf_{\beta_{\bar{j}}\in B^{\dagger}_{\bar{j}}(\alpha)}\left[\lim_{\alpha\to 0}\frac{1}{\alpha}\left\{l((1-\alpha)\xi_1+\alpha\xi_2,\beta_j,\beta_{\bar{j}}) - l(\xi_1,\beta_j,\beta_{\bar{j}})\right\}\right]$$

$$= \inf_{\beta_{\bar{j}} \in B^{\dagger}_{\bar{j}}(\alpha)} \left[\lim_{\alpha \to 0} \frac{1}{\alpha}\, \alpha \left\{ l(\xi_2,\beta_j,\beta_{\bar{j}}) - l(\xi_1,\beta_j,\beta_{\bar{j}}) \right\} \right]$$

$$= \inf_{\beta_{\bar{j}} \in B^{\dagger}_{\bar{j}}} \left\{ l(\xi_2,\beta_j,\beta_{\bar{j}}) - l(\xi_1,\beta_j,\beta_{\bar{j}}) \right\}.$$

where the first equality holds by applying Pshenichnyi (1971, Theorem 3.2, pp.75). Hence, (C.6) is true. Now, if we suppose that equation (5.4.1) has a unique solution $\beta^*_{\bar{j}}$ when $\xi = \xi^*$, the optimum design for the optimization problem (5.3.2), (C.6) becomes

$$F_{\Delta_{\bar{j}}}(\xi^*,\xi_2) = l(\xi_2,\beta_j,\beta^*_{\bar{j}}) - l(\xi^*,\beta_j,\beta^*_{\bar{j}})$$
$$= l(\xi_2,\beta_j,\beta^*_{\bar{j}}) - \Delta_{\bar{j}}(\xi^*,\beta_j)$$

where ξ_2 represents any design except ξ^*. Moreover, if $\xi_2 = \xi_{\mathbf{x}}$, the design putting all mass at the point $\mathbf{x} \in \mathcal{X}$, then

$$F_{\Delta_{\bar{j}}}(\xi^*,\xi_{\mathbf{x}}) = d(\mathbf{x},\beta_j,\beta^*_{\bar{j}}) - \Delta_{\bar{j}}(\xi^*,\beta_j)$$

Now, following Theorems 2.1 and 2.2, we get

$$F_{\Delta_{\bar{j}}}(\xi^*,\xi_{\mathbf{x}}) = \psi(\mathbf{x},\xi^*,\beta_j) - \Delta_{\bar{j}}(\xi^*,\beta_j) \le 0 \quad \text{, for any } \mathbf{x} \in \mathcal{X}$$

or

$$\psi(\mathbf{x},\xi^*,\beta_j) \le \Delta_{\bar{j}}(\xi^*,\beta_j) \quad \text{, for any } \mathbf{x} \in \mathcal{X} \qquad (C.7)$$

where $\psi(\mathbf{x},\xi^*,\beta_j) = d(\mathbf{x},\beta_j,\beta^*_{\bar{j}})$.

Inequality (C.7) is condition (i) of Theorem 5.1 for a local optimum design. Corollary 2.1 applies for condition (ii). Conditions (iii) and (iv) are straightforward. For the case corresponding to fully Bayesian optimum designs, the same assumptions made in Appendices A and B apply here.

CHAPTER 6. OPTIMUM EXPERIMENTAL DESIGNS FOR PARAMETER ESTIMATION IN BINARY RESPONSE MODELS

6.1. INTRODUCTION

The first ideas giving rise to the theory of optimal design originated in the context of searching for experimental designs maximizing the precision for parameter estimation in linear regression. Its main pillar, the celebrated General Equivalence Theorem, was also originally proven in this context. In the present chapter we return to this problem in a slightly different framework. Here, the interest lies in optimal designs for parameter estimation in a generalized linear model for binary responses.

Searching for optimal designs in linear regression situations is relatively simple, for the criterion functions related to existing criteria of optimization depend upon the design matrix X only, so that global optimal designs can be determined. However, global optimality cannot be achieved for nonlinear regression models under the same conditions. This is because these criterion functions not only depend upon the design matrix X but also upon the true values of the model parameters. Therefore, similarly to the problems of model discrimination dealt with in Chapters 3 to 5, only local, sequential and Bayesian optimum designs can be determined in such circumstances. An additional class of optimum designs which can be found in this context is the class of minimax designs as proposed by Atkinson and Fedorov (1975a).

Such dependence (in nonlinear regression) of optimum designs upon the true values of the parameters provides the link between the first part of this thesis and both the current and the next chapters. From the optimization point of view, and to a certain extent also from the optimal design theory point of view, both problems are essentially the same although their purposes and, therefore, the optimization criteria which are utilized are not.

6.2. BACKGROUND

A number of sensible criteria have been suggested for the problem of optimizing parameter estimation in linear regression, the most important of which are defined as functions of the Fisher information matrix corresponding to the vector of unknown parameters of the model. Here, we adopt the criterion of D–optimality, perhaps the most intuitive and used of all, and one of its extensions, Bayesian D–optimality.

An excellent reference for the study of D–optimality is Atkinson and Donev (1992) which presents the most relevant properties and the sequential construction of D–optimum designs (Chapter 11); introduces algorithms for the construction of exact D–optimum designs (Chapter 15); and also gives examples of local and Bayesian D–optimum designs (Chapters 18 and 19). In addition to this text, the pioneering books of Fedorov (1970) and Silvey (1980) cover the subject with great emphasis.

Chaloner (1987) and Chaloner and Larntz (1989) deal with the specific problem of searching for Bayesian D–optimum and Bayesian A–optimum designs to estimate the parameters of a logistic regression model for binary data. Their definition of Bayesian D–optimality is given by the expected value of the D–optimality criterion function over a prior distribution for the linear predictor parameters. Bayesian A–optimality is analogously defined. As illustrations, they provide some examples in which independent uniform distributions are taken as priors for the parameters of a first degree polynomial linear predictor. In addition to these papers, Chaloner and Larntz (1988) developed a software package, called Logit–Design, to find Bayesian D–optimum and Bayesian A–optimum designs based on independent beta prior distributions for the linear predictor parameters.

In another context, namely optimum experimental designs for estimating the parameters of a regression model containing controlled and uncontrolled factors, Fedorov and Atkinson (1988) proposed five possibilities for extending the criterion of D–optimality in order to define Bayesian D–optimality. The second of their suggestions is utilized in the work of Chaloner and Larntz. Here, this option is also regarded as the optimization criterion.

Other approaches regarding optimum experimental designs in the context of binary responses have been proposed. Most of them, however, concern the class of exact designs as opposed to that of approximate designs considered in this chapter. Some of these are mentioned below.

Tsutakawa (1972), for instance, presents a Bayesian formulation in which the underlying dose–response curve belongs to a family of distributions $\{F_\theta\}$ whose parameter θ has a given prior distribution $\lambda(\theta)$. The criterion of optimality is defined as minimizing the expected value of the approximate posterior variance of θ over the prior $\lambda(\theta)$. As illustration, a one–parameter logistic distribution is regarded as the family $\{F_\theta\}$ whereas a Normal $(0,\tau^2)$ is taken as the prior $\lambda(\theta)$. Two further features of this approach are the addition of a constant in the posterior variance of θ to avoid discrepancies and the consideration of two–stage designs. Later, Tsutakawa (1980) extended his method for finding optimum designs to estimate a given percentile of a logistic distribution when the scale parameter is known.

Abdelbasit and Plackett (1983) studied the efficiency (robustness) of optimum designs arising both from the fiducial method proposed by Finney (1971) and from local D–optimality. In the context of logistic regression, Minkin (1987) considered two–stage designs based on confidence regions of minimum area.

For the problem of estimating the LD_p, $0 < p < 1$, for logit, probit, and extreme value models when the scale parameter is known, Khan (1988) showed that the related Fisher information function is unimodal as a function of the unknown location parameter, say θ, as well as a single design level. Then, the optimality criterion for finding a one point Bayesian optimum design to estimate the LD_p is defined as the expectation of the Fisher information function over the posterior distribution of θ, given a sample of n independent Binomial $\{Y_1,\cdots,Y_n\}$ based on $\{N_1,\cdots,N_n\}$ replications of a set of design levels $\{x_1,\cdots,x_n\}$. For parameter estimation, Khan and Yazdi (1988) adopted the criterion of D–optimality.

The main results contained in this chapter concern two extensions of Chaloner and Larntz's results. Firstly, the linear predictor is assumed to have any structure as long as it is, of course, linear in the parameters. Secondly, we extend the available methodology for logit models to probit, complementary log–log, and any other model whose link function complies with the restrictions imposed by the nature of binary data under the framework of generalized linear models (see McCullagh and Nelder(1989), Section 4.3 for further details).

Our notation, definitions, and other relevant results concerned with Bayesian D–optimality are presented in Section 6.3. Firstly, the criterion of D–optimality is recalled and then extended to the criterion of Bayesian D–optimality. Also, the corresponding version of the General Equivalence Theorem, is presented. In Section 6.4, several examples that illustrate not only how to obtain a Bayesian D–optimum design in this context but also how to check whether or not it is optimum are shown. The first three examples also illustrate how the number of support points in Bayesian D–optimum designs fluctuates as a function of the nature of the prior information. Finally, further extensions related to this problem are discussed in Section 6.5.

6.3. BAYESIAN D–OPTIMALITY

Throughout this chapter and Chapter 7, the binary data models considered are a subclass of generalized linear models. Thus, we adopt the terminology and notation of McCullagh & Nelder (1989) for modelling, but use optimal design theory notation elsewhere.

Suppose that a random sample $Y_1,...,Y_n$ is to be observed, where Y_i follows a Binomial (m_i,π_i) distribution. The interest lies in the relationship between π_i and the covariates $\{x_{i1},...,x_{ip}\}$. According to generalized linear model assumptions this relationship is described by the linear predictor $\eta_i = \sum_{j=1}^{p} x_{ij}\,\beta_j$, and a monotonic and differentiable function g(.) such that $\eta_i = g(\pi_i)$. The set of parameters $\{\beta_j\}$ is supposed to be unknown. Then, the log likelihood function is

$$l(\pi,m;y) = \sum_{i=1}^{n} \left[y_i \log\left[\frac{\pi_i}{1-\pi_i}\right] + m_i \log(1-\pi_i)\right] \qquad (6.3.1)$$

where $\pi_i = g^{-1}(\eta_i)$. The term that does not depend on π can be neglected.

For defining D–optimality, and Bayesian D–optimality, the Fisher information matrix, say M, for the vector of parameters β must be obtained. This is done, for instance, in McCullagh and Nelder (1989, Section 4.4) where the following matrix is determined.

$$M_{rs} = -E\left[\frac{\partial^2 l}{\partial\beta_r \partial\beta_s}\right] = \left\{X^t W\, X\right\}_{rs} \qquad (6.3.2)$$

where X is the associated design matrix; and W is a diagonal matrix given by

$$W = \text{diag}\left\{m_i\left[\frac{d\pi_i}{d\eta_i}\right]^2 / \pi_i(1-\pi_i)\right\} \qquad (6.3.3)$$

where m_i is interpreted as the number of replications of the point $\mathbf{x}_i$ in the design, or equivalently $m_i/(\sum_j m_j)$ is the weight of $\mathbf{x}_i$.

Because W is diagonal, (6.3.2) can be simplified to $M_{rs} = \sum_i w_i x_{ri} x_{si}$, where w_i is the ith component of W, and $\mathbf{x}_r = \{x_{ri}\}$ and $\mathbf{x}_s = \{x_{si}\}$ are respectively the rth and sth columns of X. As (6.3.3), and consequently (6.3.2), depend upon the design ξ and the values of the unknown parameters β we denote M as $M(\xi,\beta)$, and for a specific β this matrix will be called a local Fisher information matrix.

Table 6.1 shows specific expressions of W for logit, complementary log–log, and probit models corresponding to a single design point $\mathbf{x}$ replicated just once (for n design points $\{x_1,\cdots,x_n\}$, each x_i being replicated m_i times, the general form of W is easily obtainable from Table 6.1). The notation used in Table 6.1 shows explicitly the dependence of the Fisher information matrix M on ξ and β as both the expected response $\pi(\mathbf{x})$ and the linear predictor $\eta(\mathbf{x})$ not only depend upon the specific design point $\mathbf{x}$ but also upon the vector of parameters β.

TABLE 6.1 – SOME EXPRESSIONS FOR (6.3.3)

LINK	DIAGONAL MATRIX W
Logit	$\pi(\mathbf{x},\beta)\left[1-\pi(\mathbf{x},\beta)\right]$
Complementary log–log	$\dfrac{\pi(\mathbf{x},\beta)}{1-\pi(\mathbf{x},\beta)}\left[\ln\,\pi(\mathbf{x},\beta)\right]^2$
Probit	$\dfrac{\exp(-\eta^2(\mathbf{x},\beta)/2)}{(2\pi)\;\pi(\mathbf{x},\beta)\{1-\pi(\mathbf{x},\beta)\}}$

For any link function $\eta = g(\pi)$ such that g(.) maps the interval [0,1] onto the whole real line, the corresponding expression of (6.3.3) can be easily obtained and subsequently the local Fisher information matrix $M(\xi,\beta)$ is straightforward to determine. For example, the inverse of any distribution function F(.) related to a continuous r.v. attains the necessary conditions set above.

When there is a single explanatory variable the calculations required to obtain (6.3.2) reduce significantly. For instance, Chaloner and Larntz (1989) take the linear predictor as $\eta = -\beta(x-\mu)$ and the canonical link function, logit, so that under the framework of a dose–response relationship problem, μ can be interpreted as the dose x at which the expected response is a half (such a value of x is known in the literature as the LD50) and β is the slope on the logit scale (in fact, μ can be equally interpreted for models with the probit link function, but not for the complementary log–log). Because of this useful interpretation such parameterization in logistic regression is quite convenient. Further calculations lead to the 2×2 local Fisher information matrix whose determinant is given by a rather simple function of the design points and the expected response π which in turn is a function of the parameters μ and β as well as the design points. More generally, however, (6.3.3) and consequently (6.3.2) will be very complex matrices.

Prior to defining the criterion of Bayesian D–optimality, it is important to point out, as mentioned above, that there are five different possibilities for doing so; for further details and a numerical example comparing these five possibilities, see Atkinson and Donev (1992, Sections 19.2 and 19.3). To be consistent with other definitions in previous chapters, our definition of Bayesian D–optimality is based on the expected value over the prior probability distribution for β of the non–Bayesian D–optimality criterion function.

Definition 6.1. For a specific vector of linear predictor parameters β, denote the determinant of the local Fisher's information matrix by $|M(\xi,\beta)|$. Let $p_0(\beta)$ be a prior distribution for the vector of linear predictor parameters β. Then, the Bayesian D–optimality criterion function is defined as

$$\Delta(\xi) = \begin{cases} E_\beta \log |M(\xi,\beta)|, & \text{if } M(\xi,\beta) \text{ is nonsingular for all } \beta \text{ relevant to the prior } p_0(\beta) \\ -\infty, & \text{otherwise} \end{cases} \tag{6.3.4}$$

Definition 6.2. A Bayesian D–optimum design ξ^* w.r.t. $p_0(\beta)$ is such that

$$\Delta(\xi^*) = \sup_{\xi \in \mathscr{H}} \Delta(\xi) \tag{6.3.5}$$

Concavity of (6.3.4) in the set of design measures $\mathscr{H}$ is ensured by linearity, and by concavity of the logarithmic function. Thus, with the purpose of applying results from optimal design theory the Fréchet directional derivative of criterion function (6.3.4) remains to be determined. This is derived in Appendix D. Given these conditions and provided that $|M(\xi,\beta)| \neq 0$ for all β relevant to the prior $p_0(\beta)$ we can now state the following Theorem for Bayesian D–optimum designs.

THEOREM 6.1

(i) a necessary and sufficient condition for a design ξ^* to be Bayesian D–optimum is fulfilment of the inequality

$$\psi(\mathbf{x},\xi^*) \leq p \text{ , for all } \mathbf{x} \in \mathcal{X}$$

where

$$\psi(\mathbf{x},\xi^*) = E_\beta\left\{w(\mathbf{x},\beta)\, f(\mathbf{x})^t\, [M(\xi^*,\beta)]^{-1} f(\mathbf{x})\right\};$$

$$w(\mathbf{x},\beta) = \frac{1}{\pi(\mathbf{x},\beta)[1 - \pi(\mathbf{x},\beta)]}\left[\frac{d\pi}{d\eta}\right]^2_{\pi=\pi(\mathbf{x},\beta)};$$

$f(\mathbf{x})^t = (f_1(\mathbf{x}),\cdots,f_p(\mathbf{x}))$; and

p is the number of linear predictor parameters.

(ii) at the points of the Bayesian D–optimum design $\psi(\mathbf{x},\xi^*)$ achieves its upper bound;

(iii) for any nonoptimum design ξ, that is a design for which $\Delta(\xi) < \Delta(\xi^*)$

$$\sup_{\mathbf{x}\in\mathcal{X}} \psi(\mathbf{x},\xi) > p \,;$$

(iv) the set of Bayesian D–optimum designs is convex.

An extra condition for the criterion of G–optimality can be added to this theorem. A powerful consequence of Theorem 6.1 is that the useful technique for checking optimality can be applied. In order to prove the conditions of Theorem 6.1 one should apply the general results of Chapter 2. In the next section some examples are presented to illustrate the theory of Bayesian D–optimality applied to binary response models.

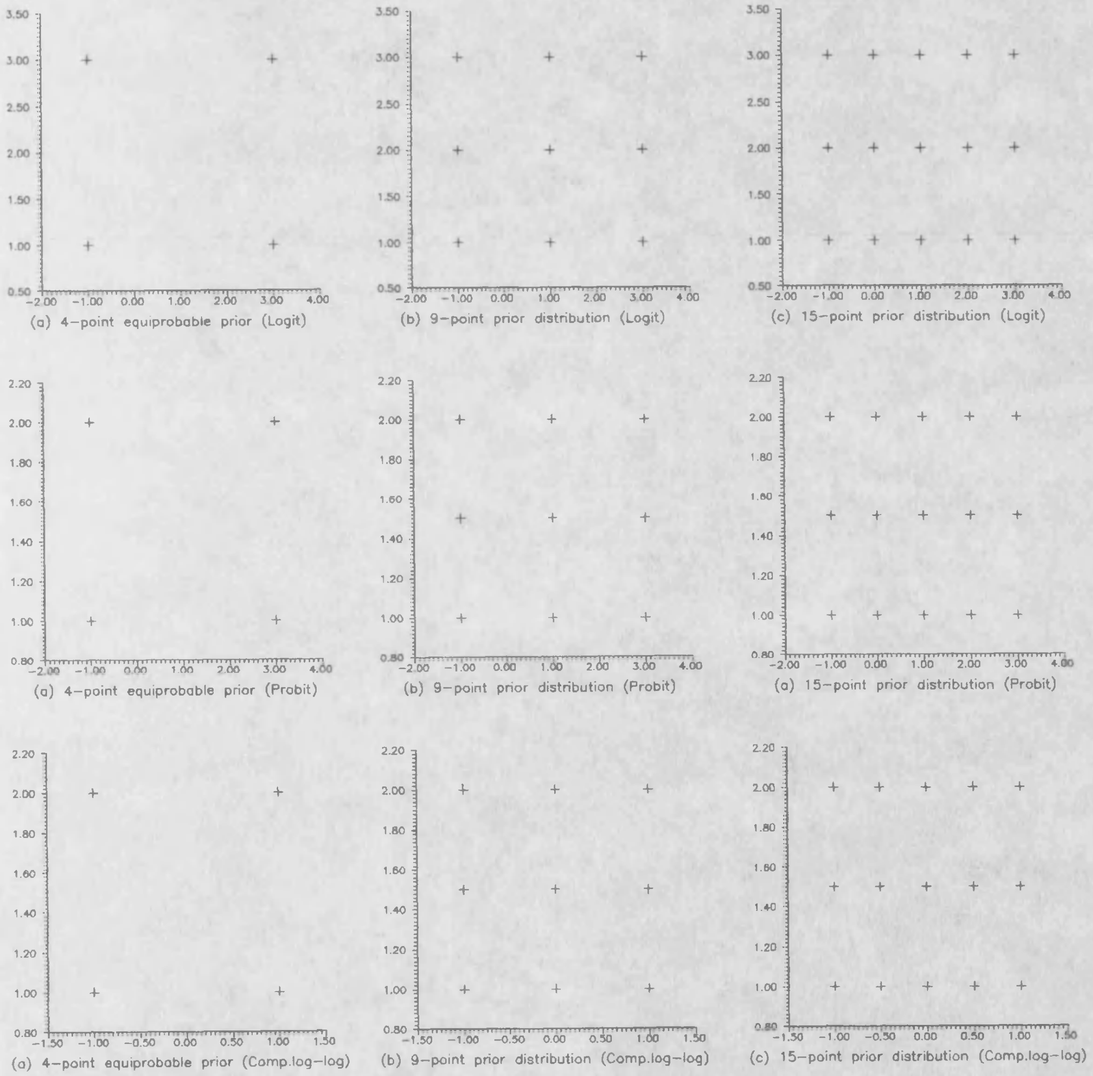
6.4. EXAMPLES

In this section five examples are presented. The main purpose in all of them is to investigate the effects of distinct prior distributions on the resulting Bayesian D–optimum designs. The common interest in these examples lies in estimating both parameters of the linear predictor $\eta = \alpha + \beta x$ when prior information is available by means of probability distributions, either discrete or continuous. For all examples rectangular subregions of the whole parameter spaces are taken as the regions within which the underlying prior distributions are considered. The first part of the investigation concerns Examples 6.1, 6.2, and 6.3 whereas the second consists of Examples 6.4 and 6.5.

In the first stage of the former, three discrete prior distributions are examined. They contain four, nine, and fifteen parameter values reflecting increasingly dense prior information, although limited by the same rectangular region of the parameter space. In the second stage two independent uniform prior distributions, again defined in the same rectangular region, are regarded as priors for the linear predictor parameters α and β. This strategy is adopted not only with the purpose of assessing the fluctuation of the number of support points in the optimum designs but also to compare some structural aspects of the resulting Bayesian D–optimum designs.

For each of the first three examples one of the link functions mentioned in Table 6.1 is examined. No attention is paid to how the expected values under the distinct models agree, for in this chapter our interest is restricted to parameter estimation rather than model discrimination. Equiprobable or quasi–equiprobable discrete prior distributions are assumed in all examples. Whenever a discrete prior consists of an odd number of parameter values the centre point of the corresponding rectangular region is assigned slightly larger probabilitites than the region vertices. Figure 6.0.1 displays the parameter values constituting all discrete priors regarded in the first three examples. Each sequence of three graphs in the same row shows the priors utilized under a specific link function.

The following conventions hold for the priors in Figure 6.0.1 : (i) four–point priors are equiprobable ; (ii) nine–point priors assign probabilities equal to 0.11 for all parameter values, with the exception of the centre of the region which is assigned probability 0.12 ; (iii) fifteen–point priors assign probabilities equal to 0.066 for all parameter values with the exception of the centre of the region which is assigned probability 0.076.



(a) 4–point equiprobable prior (Logit)

(b) 9–point prior distribution (Logit)

(c) 15–point prior distribution (Logit)

(a) 4–point equiprobable prior (Probit)

(b) 9–point prior distribution (Probit)

(a) 15–point prior distribution (Probit)

(a) 4–point equiprobable prior (Comp.log–log)

(b) 9–point prior distribution (Comp.log–log)

(c) 15–point prior distribution (Comp.log–log)

Figure 6.0.1. Prior distributions for Examples 6.1 to 6.3.

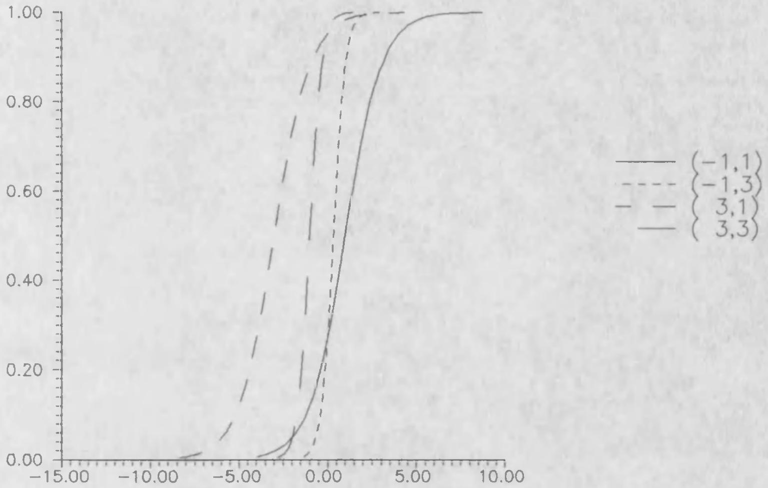


Figure 6.0.2. (Logit) Patterns of expected responses

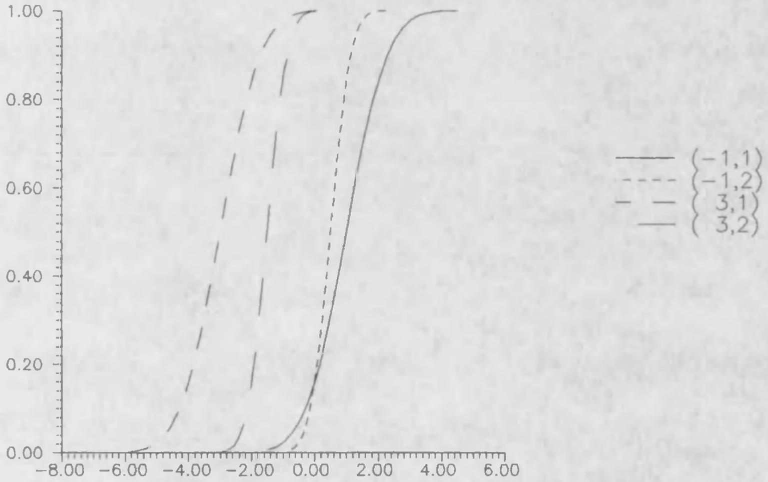


Figure 6.0.2. (Probit) Patterns of expected responses

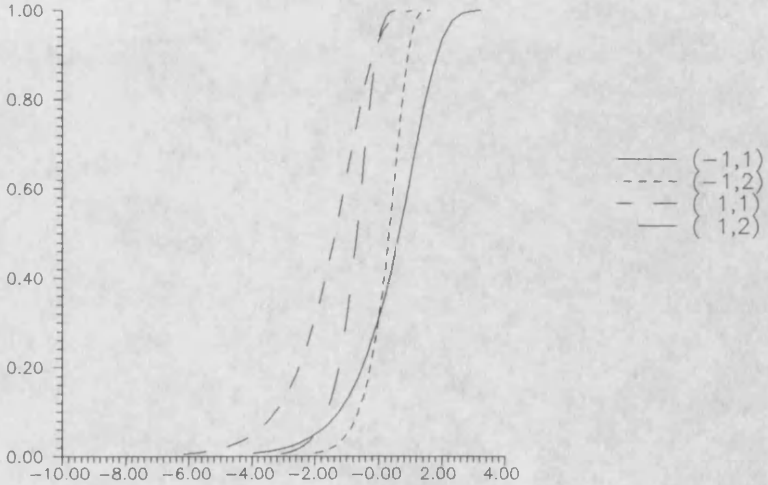


Figure 6.0.2. (Cplog) Patterns of expected responses

For the discrete priors in Figure 6.0.1, the variances of the marginal prior distributions for α and β decrease with the increasing number of relevant parameter values. This is particularly the case for the marginal distributions for α, which although being defined in the same regions, have two, three, and five relevant values for α in the four, nine, and fifteen–point priors, respectively. The aims of the investigation are to find out whether, and how, these variations on the precision of the available prior information will be reflected in the resulting Bayesian D–optimum designs. Here, it is important to notice that the more dispersed the priors of Figure 6.0.1 the smaller the variance; dispersion meaning a larger number of parameter values spread over the rectangular regions.

The three graphs of Figure 6.0.2 show the curves of expected responses generated by pairs (α,β) belonging to the four–point priors of Figure 6.0.1. The legends on the graphs display the pairs (α,β) generating the specific curves whereas the link (logit, probit, or complementary log–log) is specified in the title. As the curves are generated by the pairs (α,β) which are the vertices of the rectangular regions they establish limiting patterns for those being generated by other pairs in the same region. Therefore, for the discrete priors consisting of nine and fifteen pairs (α,β), the curves of expected responses determine intermediate patterns w.r.t. those of Figure 6.0.2. Thus, if we plot the additional patterns for the remaining pairs (α,β) belonging to the nine and fifteen–point priors they will lie between the boundaries established by those of Figure 6.0.2.

EXAMPLE 6.1. Suppose that a binary data model has the canonical link function (logit). The rectangular region in which the prior distributions, three discrete and one continuous, are considered is $\Theta_0 = \{(\alpha,\beta)\colon \alpha \in [-1,3] \text{ and } \beta \in [1,3]\}$. The discrete priors are specified by four, nine, and fifteen points as indicated in Figure 6.0.1 and the three rules described above whereas the continuous prior is given by two independent uniform distributions each over one of the intervals defining Θ_0. The resulting Bayesian D–optimum designs are displayed in Table 6.2.

TABLE 6.2 – BAYESIAN D–OPTIMUM DESIGNS FOR THE CANONICAL LINK FUNCTION (LOGIT)

VALUE OF $\Delta(\xi^*)$	BAYESIAN D–OPTIMUM DESIGN
(a) -5.0837	$\left\{\begin{matrix} -4.1097 & -1.4327 & -0.3363 & 0.9438 \\ 0.1255 & 0.2595 & 0.3387 & 0.2763 \end{matrix}\right\}$
(b) -5.0173	$\left\{\begin{matrix} -3.6640 & -1.6280 & -0.6407 & -0.1210 & 0.8347 \\ 0.0603 & 0.2778 & 0.1972 & 0.1814 & 0.2833 \end{matrix}\right\}$
(c) -4.9287	$\left\{\begin{matrix} -3.2979 & -1.4872 & -0.3961 & 0.6830 \\ 0.0661 & 0.3121 & 0.2955 & 0.3263 \end{matrix}\right\}$
(d) -4.7886	$\left\{\begin{matrix} -1.6052 & -0.5134 & 0.5541 \\ 0.3563 & 0.2679 & 0.3758 \end{matrix}\right\}$

where
(a) design (6.4.1) related to the four–point underlying prior;
(b) design (6.4.2) related to the nine–point prior;
(c) design (6.4.3) related to the fifteen–point prior; and
(d) design (6.4.4) related to the bivariate uniform prior.

The first prominent feature of the four resulting Bayesian D–optimum designs is that the range in which the optimum design support points lie gradually shrinks as the number of parameter values in the prior increases (smaller variances for the marginal priors). This shrinkage is substantial when designs (6.4.1), (6.4.2), and (6.4.3) corresponding to the four, nine, and fifteen–point priors, respectively, are compared. This phenomenon is even more evident if designs (6.4.1) and (6.4.4), corresponding respectively to a four–point prior and a bivariate uniform prior are compared separately. Within the region Θ_0, the former is as dispersed a four–point prior as possible whereas the latter could be regarded as a limiting equiprobable discrete distribution, so that they represent two extreme possibilities for prior distributions as far as equiprobability and dispersion are concerned.

This finding can be summarized as follows: in the subregion Θ_0 of the parameter space the vaguer a prior distribution for (α,β) the narrower the interval containing the support points for the corresponding Bayesian D–optimum design will be. In other words, the points in the design region containing the largest amounts of information for estimation of the parameters α and β seem to be more concentrated as the prior information about α and β is more spread out in Θ_0.

Recall that the pairs (α,β) being added to the discrete priors in Figure 6.0.2 generate intermediate patterns w.r.t. those of the four–point priors. Thus, the actual information about α and β becomes more concentrated over the design region $\mathcal{X}$ as more points in the region Θ_0 are added to the priors. Therefore, this explains the shrinkage of the range of support points in the Bayesian D–optimum designs.

Another feature of the four Bayesian D–optimum designs regarded in the order presented in Table 6.2 is the increasing value of the maximum achieved by criterion function (6.3.3). This and the shrinkage phenomenon mentioned above occur concomitantly and can be explained similarly.

However, unlike the increasing value of the optimized criterion function or the gradual shrinkage of the region containing the optimum design support points, the number of support points in the Bayesian D–optimum designs does not have a consistent pattern as a function of the increasing vagueness of the prior information in Θ_0. Firstly there are four points in the optimum design for the four–point equiprobable prior. This number increases to five, then decreases to four again and finally decreases to three for the independent uniform priors.

As in previous chapters, these results do not give any support to the intuitive idea that the number of support points in the optimum design would increase when the prior information was more dispersed. On the contrary, based on the evidence shown in this example, this number seems to be smaller for denser prior parameter values as is the case of uniform priors.

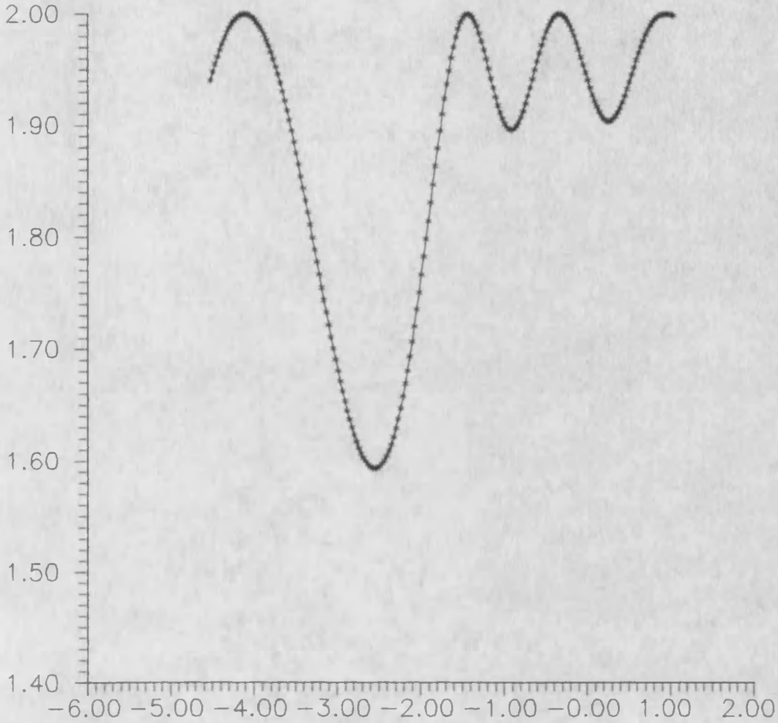


Figure 6.1.1. Derivative Function for design (6.4.1)

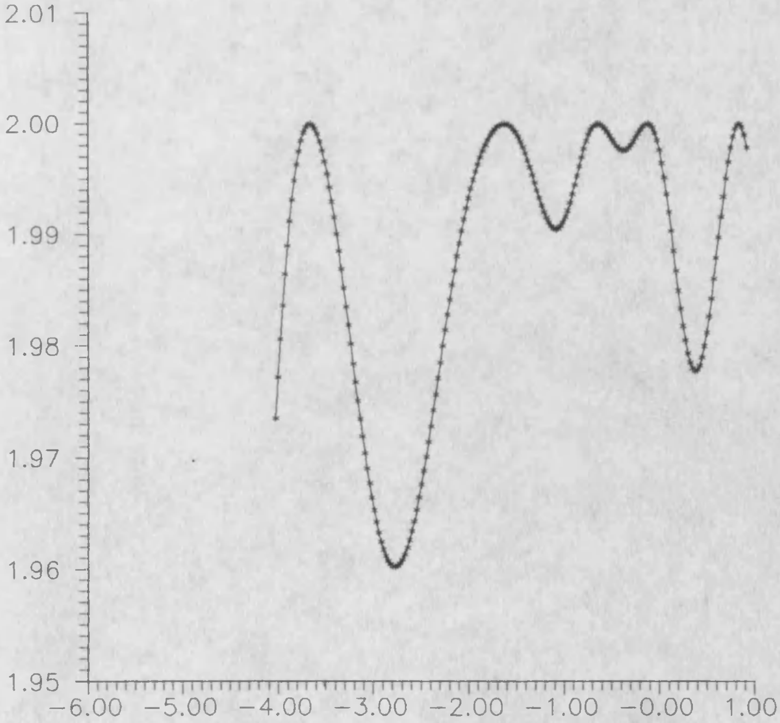


Figure 6.1.2. Derivative Function for design (6.4.2)

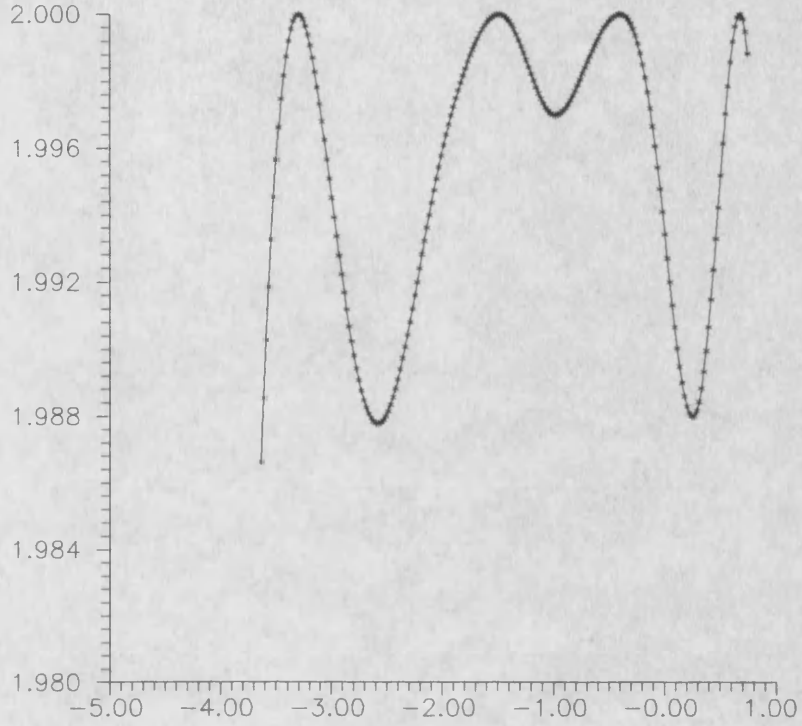


Figure 6.1.3. Derivative Function for design (6.4.3)

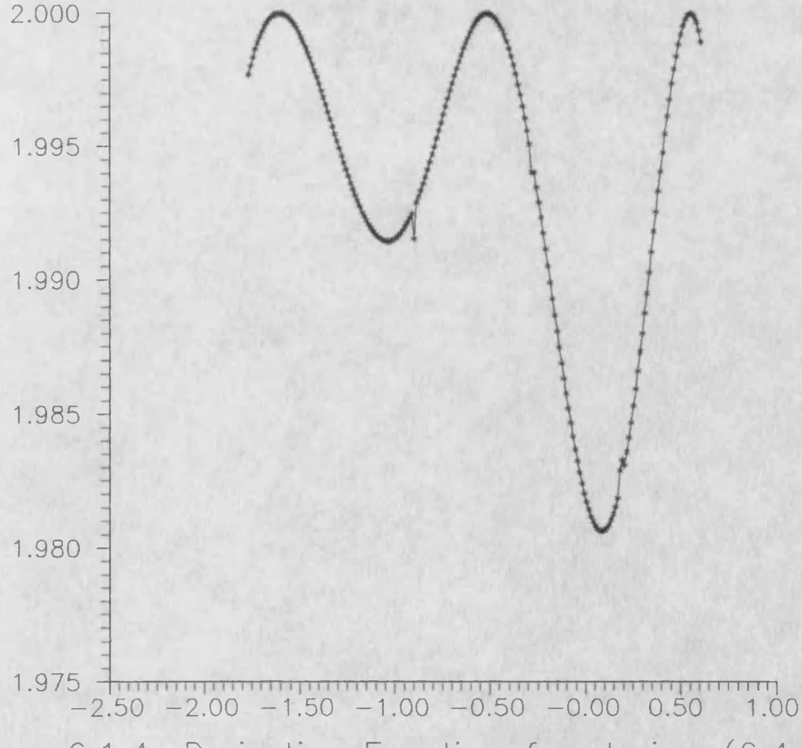


Figure 6.1.4. Derivative Function for design (6.4.4)

Further conclusions can be drawn from Figures 6.1.1, 6.1.2, 6.1.3, and 6.1.4. They show the derivative functions for the Bayesian D–optimal designs of Table 6.2. Here, it is important to recall that the upper limit for the derivative function corresponding to the optimum design equals the number of parameters to be estimated.

EXAMPLE 6.2. Now, suppose that the link function is the inverse normal or probit. We take the region $\Theta_0 = \{(\alpha,\beta)\text{: } \alpha \in [-1,3] \text{ and } \beta \in [1,2]\}$. The linear predictor structure is maintained. The sequence of three graphs in the second row of Figure 6.0.1 displays the priors to be investigated in this example. In addition, conventions (i) to (iii) apply to the priors. The resulting Bayesian D–optimum designs are displayed in Table 6.3.

TABLE 6.3 – BAYESIAN D–OPTIMUM DESIGNS FOR THE INVERSE NORMAL LINK FUNCTION (PROBIT)

	VALUE OF $\Delta(\xi^*)$	BAYESIAN D–OPTIMUM DESIGN
(a)	–3.7563	$\left\{\begin{matrix} -4.0415 & -1.9538 & -0.5392 & 1.1743 \\ 0.1102 & 0.2186 & 0.4224 & 0.2488 \end{matrix}\right\}$
(b)	–3.635	$\left\{\begin{matrix} -3.6172 & -2.2382 & -1.0503 & -0.0634 & 1.1724 \\ 0.0399 & 0.1907 & 0.3055 & 0.2826 & 0.1813 \end{matrix}\right\}$
(c)	–3.4956	$\left\{\begin{matrix} -3.1680 & -1.9778 & -0.9935 & -0.2668 & 0.8574 \\ 0.0501 & 0.2034 & 0.2468 & 0.2828 & 0.2169 \end{matrix}\right\}$
(d)	–3.266	$\left\{\begin{matrix} -1.9261 & -0.6484 & 0.5827 \\ 0.2841 & 0.4337 & 0.2822 \end{matrix}\right\}$

where

(a) design (6.4.5) related to the four–point underlying prior;
(b) design (6.4.6) related to the nine–point prior;
(c) design (6.4.7) related to the fifteen–point prior; and
(d) design (6.4.8) related to the bivariate uniform prior.

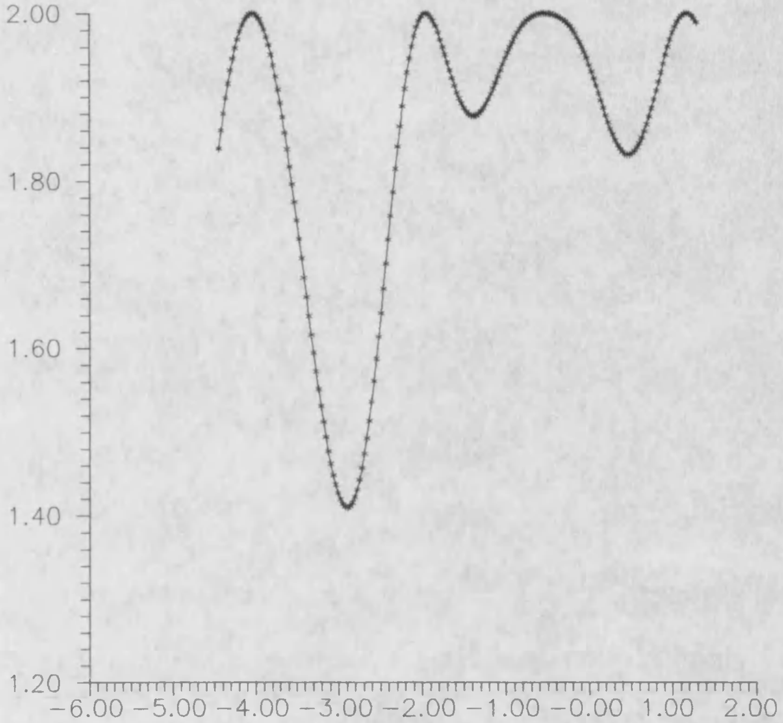


Figure 6.2.1. Derivative Function for design (6.4.5)

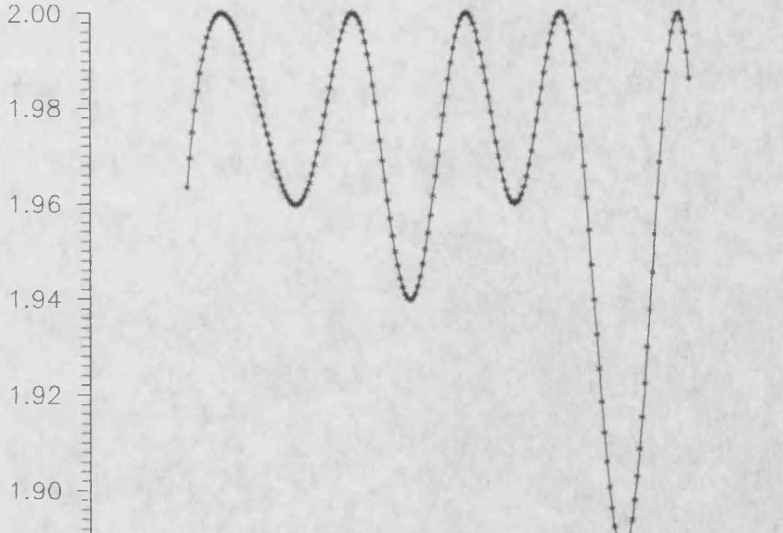


Figure 6.2.2. Derivative Function for design (6.4.6)

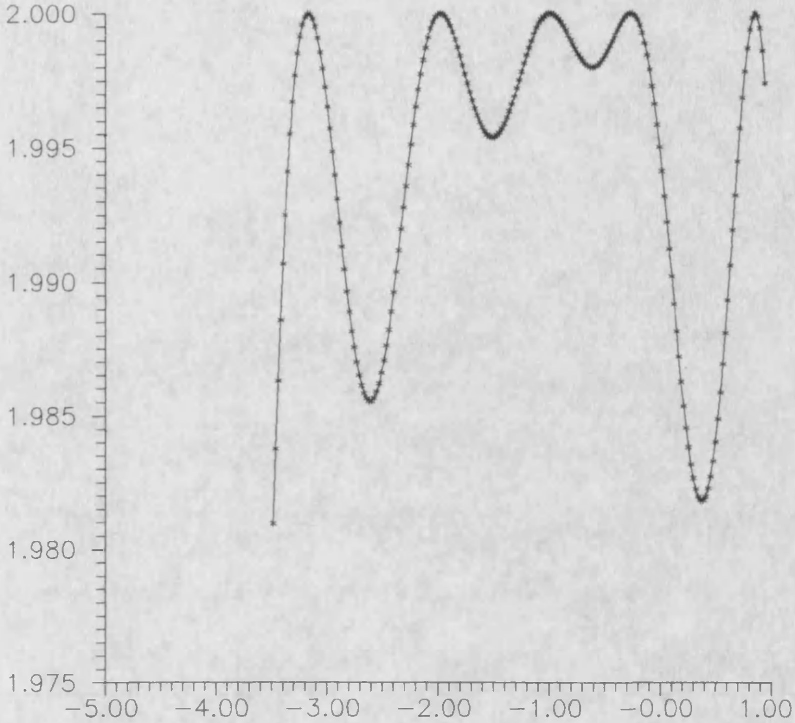


Figure 6.2.3. Derivative Function for design (6.4.7)

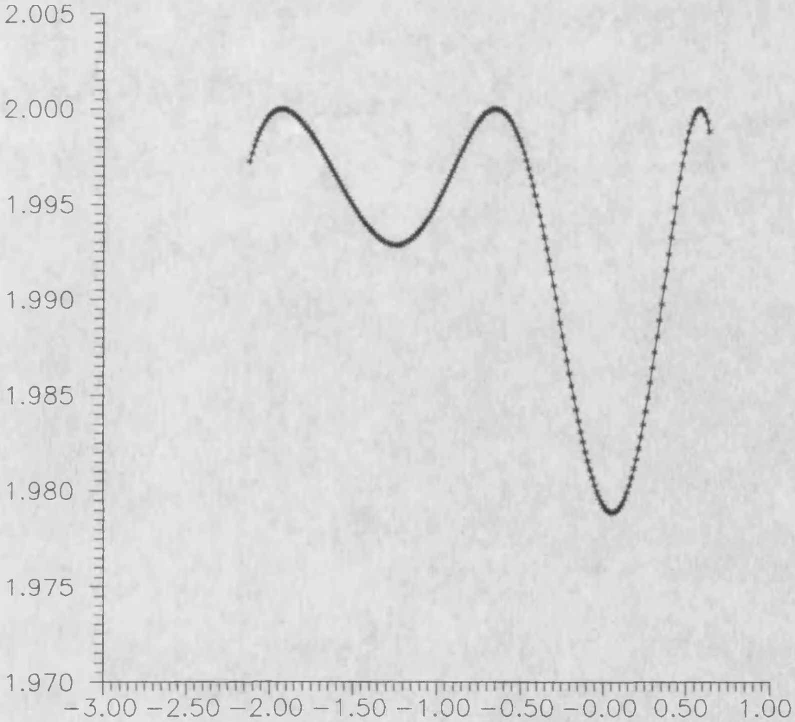


Figure 6.2.4. Derivative Function for design (6.4.8)

Similar comments to those of Example 6.1 can also be made here: namely, the range specified by the largest and the smallest support points in the optimum design shrinks with the increasing vagueness of the prior distributions, and the optimized values of criterion function (6.3.3) increase as a function of this vagueness. Further the number of support points in the optimum design fluctuates from four for the basic four–point prior to three for the bivariate uniform one. To be more precise, it starts as four, increases to five, continues to be five, and at last decreases to three for the limiting uniform prior.

Figures 6.2.1 to 6.2.4 show the derivative functions corresponding to Bayesian D–optimum designs (6.4.5) to (6.4.8), respectively. It is important to notice that the scales for the derivative functions are different in all four plots. For instance, Figure 6.2.1, related to design (6.4.5) shows the deepest valley whereas the shallowest one occurs in Figure 6.2.3 corresponding to design (6.4.7). In fact, Figures 6.2.3 and 6.2.4 show that all points in the ranges displayed are highly informative w.r.t. estimating parameters α and β since the values of the derivative functions are not smaller than 1.98. On the other hand, Figures 6.2.1 and 6.2.2 do not show such high values, indicating that if points other than the optimum are assigned to the experimental design there will be an appreciable loss of precision in the estimation of α and β.

EXAMPLE 6.3. In the last of this series of examples we suppose that the underlying link is now the complementary log–log. The linear predictor structure is assumed to be the same while the region Θ_0 is $\{(\alpha,\beta)$: $\alpha \in [-1,1]$ and $\beta \in [1,2]\}$. As in the previous examples, a sequence of priors is investigated which is shown in the three graphs in the third row of Figure 6.0.1 with conventions (i) to (iii) holding once again. To complement this sequence, we take a prior consisting of two independent uniform distributions for α and β each defined over one of the intervals making the region Θ_0. The resulting Bayesian D–optimum designs are shown in Table 6.4.

TABLE 6.4 – BAYESIAN D–OPTIMUM DESIGNS FOR THE COMPLEMENTARY LOG–LOG LINK FUNCTION

VALUE OF $\Delta(\xi^*)$	BAYESIAN D–OPTIMUM DESIGN
(a) –3.2712	$\left\{\begin{matrix} -1.3979 & 0.0066 & 1.0795 \\ 0.3699 & 0.3990 & 0.2311 \end{matrix}\right\}$
(b) –3.2143	$\left\{\begin{matrix} -1.2276 & 0.1207 & 0.8017 & 1.0073 \\ 0.4109 & 0.3601 & 0.0629 & 0.1661 \end{matrix}\right\}$
(c) –3.1381	$\left\{\begin{matrix} -1.1562 & 0.1950 & 0.8857 \\ 0.4335 & 0.3432 & 0.2233 \end{matrix}\right\}$
(d) –3.0208	$\left\{\begin{matrix} -1.0588 & 0.3554 & 0.8178 \\ 0.4712 & 0.3518 & 0.1770 \end{matrix}\right\}$

where
(a) design (6.4.9) related to the four–point underlying prior;
(b) design (6.4.10) related to the nine–point prior;
(c) design (6.4.11) related to the fifteen–point prior; and
(d) design (6.4.12) related to the bivariate uniform prior.

Again, very similar phenomena to those occurring in the previous examples are present here although less dramatically. Likewise, the number of support points does not show a great variation. For example, there are three support points in the Bayesian D–optimum design based on the equiprobable four–point prior, four for the nine–point prior, and three for both the fifteen–point and the bivariate uniform priors.

The same pattern is observed in Examples 6.1, 6.2, and 6.3. For instance, in all examples Bayesian D–optimum designs corresponding to the bivariate uniform priors have three support points. Also, the number of support points in the optimum design is increased by one when a nine–point prior is used

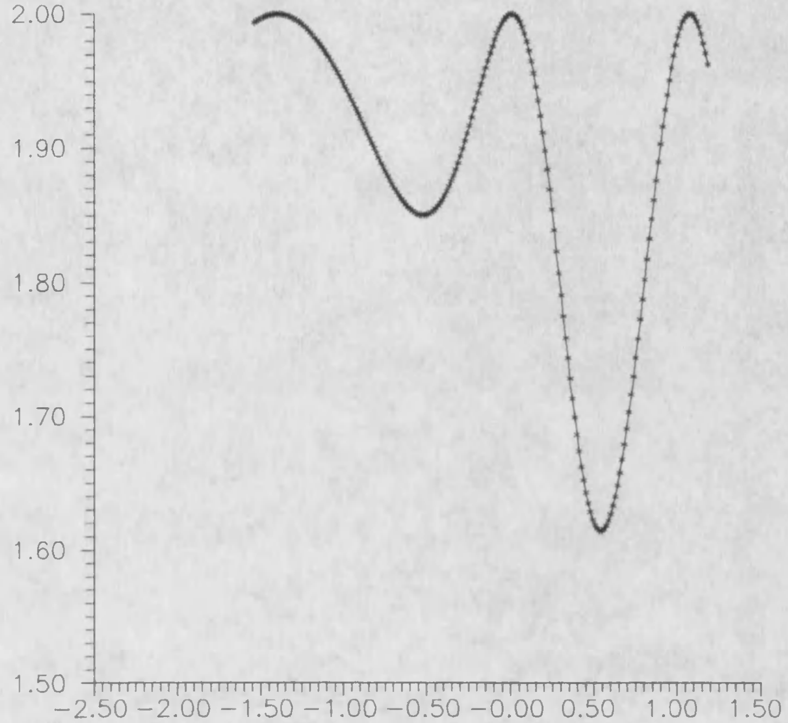


Figure 6.3.1. Derivative Function for design (6.4.9)

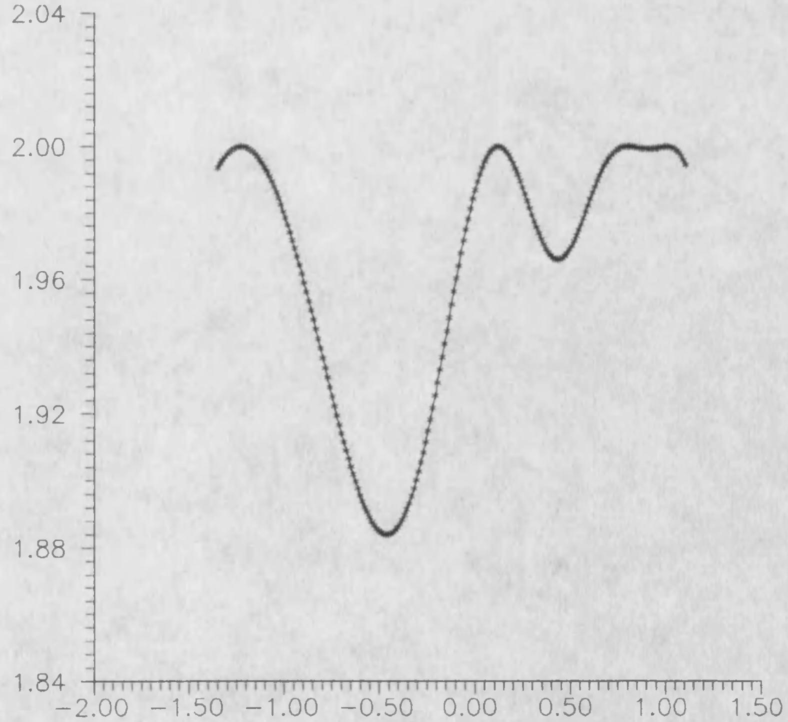


Figure 6.3.2. Derivative Function for design (6.4.10)

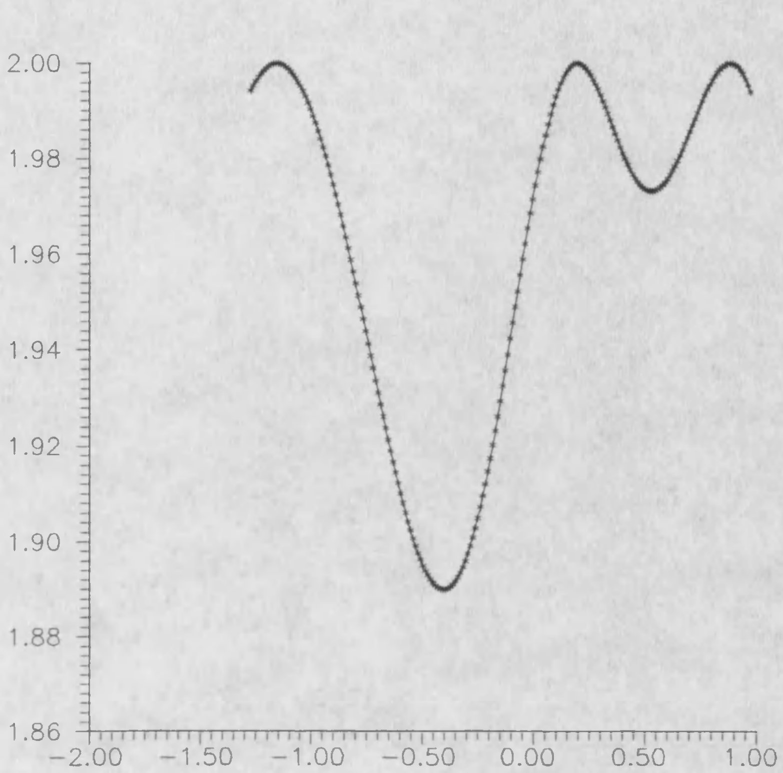


Figure 6.3.3. Derivative Function for design (6.4.11)

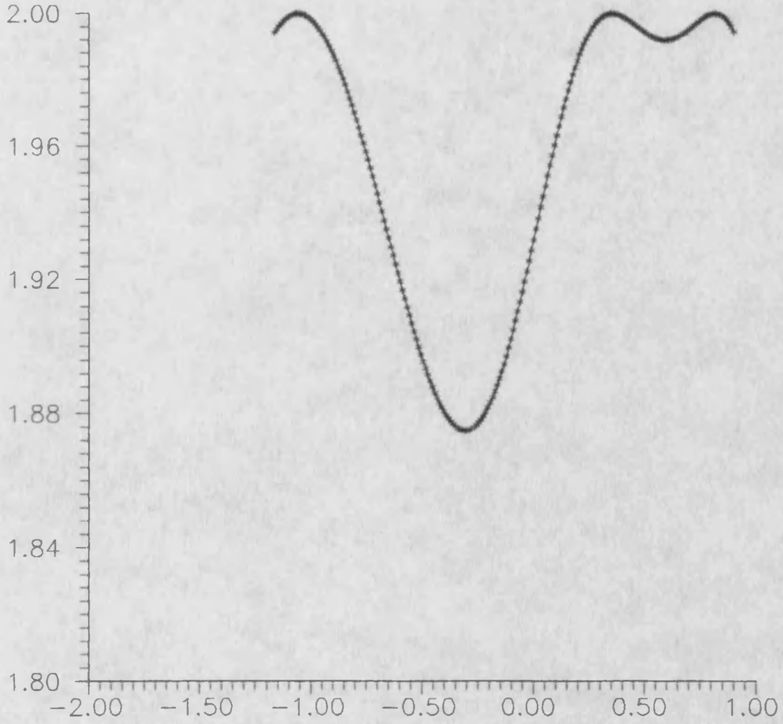


Figure 6.3.4. Derivative Function for design (6.4.12)

rather than a four–point one. A plausible interpretation for this change is that by increasing the uncertainty about the true values of the parameters, although limiting it to the region Θ_0, one more support point in the optimum design becomes necessary. However, this does not occur when a fifteen–point prior is used rather than a nine–point one, which contradicts the previous explanation.

Figures 6.3.1 to 6.3.4 confirm that designs (6.4.9) to (6.4.12) are indeed optimum w.r.t. criterion (6.3.4). They also show that some points in the corresponding ranges of the explanatory variables considered are relatively noninformative. However, as in the previous examples, the values of the derivative functions are large, frequently approaching the upper limit, two.

We now present two examples with the purpose of illustrating with more details the process of searching for the optimum design. Because of lack of results concerning the number of support points in Bayesian optimum designs on one hand and the need to specify this number for the numerical search procedures on the other, the actual search must involve an element of trial and error. This is described in the following examples.

EXAMPLE 6.4. The link function is logit and the linear predictor structure is simply $\alpha + \beta x$ again. The region Θ_0 is given by $\{(\alpha,\beta)\colon \alpha \in [-1,3] \text{ and } \beta \in [1,3]\}$. Consider a ten–point prior distribution for (α,β), which is similar to the nine–point prior shown in Figure 6.0.1. The only differences are that the point $(\alpha{=}2,\beta{=}1.5)$ is added to the prior relevant parameter values and that the distribution is now equiprobable.

Our first attempt is a design with five support points. Using a starting design whose support points are spread out in the range [–2,1] of the design region $\mathcal{X}$ the Bayesian D–optimum five–point design resulting from a numerical search is given below.

$$\xi = \begin{Bmatrix} -3.3447 & -1.6782 & -0.4252 & 0.7511 & 0.7511 \\ 0.0654 & 0.3029 & 0.3284 & 0.0570 & 0.2463 \end{Bmatrix} \qquad (6.4.13)$$

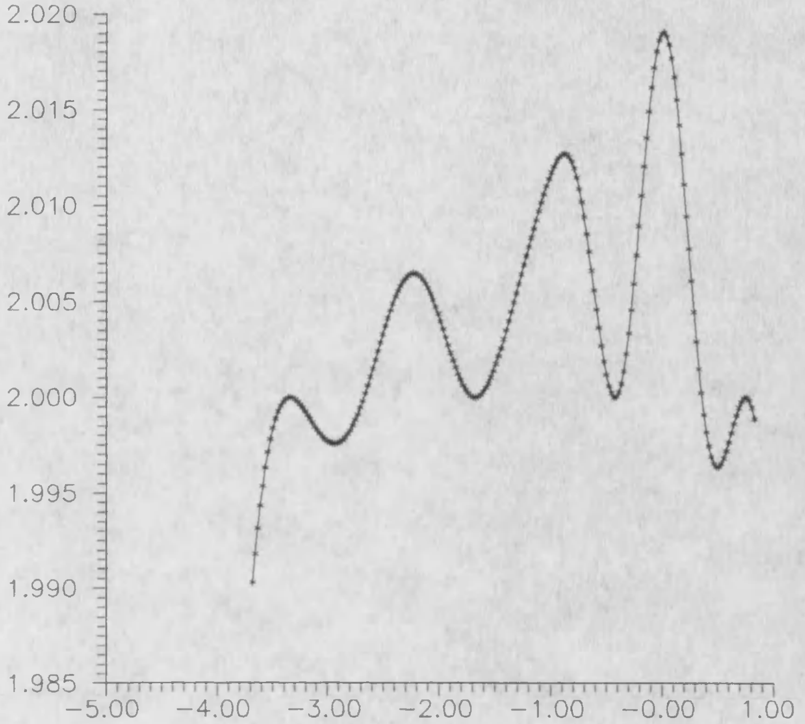


Figure 6.4.1. Derivative Function for design (6.4.13)

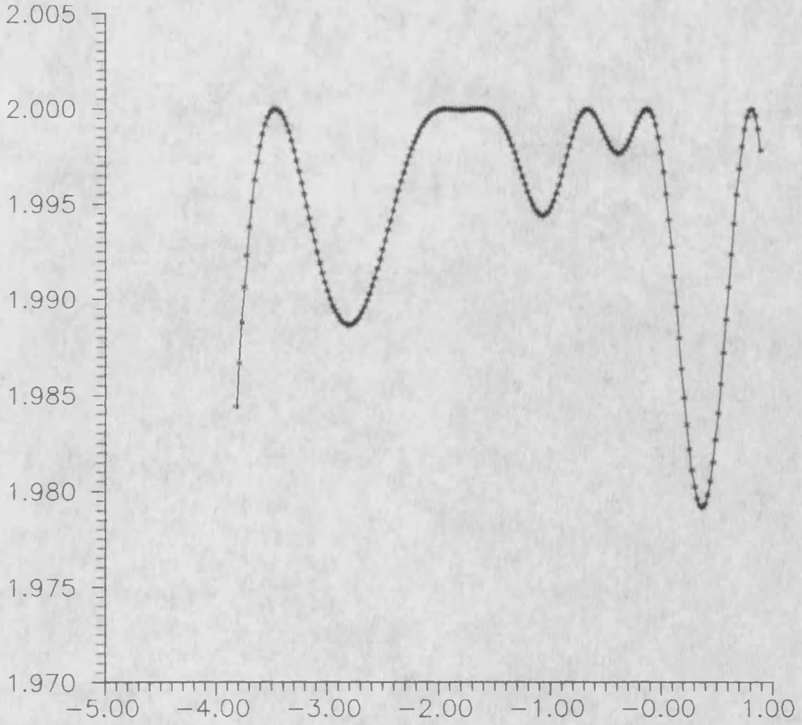


Figure 6.4.2. Derivative Function for design (6.4.14)

In fact, design (6.4.13) has four support points as the last two points are the same. Prior to drawing any conclusion about the optimality of this design we must examine the plot of its derivative function over a range in the design region including the support points of (6.4.13). This is shown in Figure 6.4.1 where it becomes obvious that design (6.4.13) is not optimal w.r.t. criterion (6.3.4). Indeed, the upper limit of its derivative function is greater than two and there are three peaks above this value which may be interpreted as corresponding to potential support points of the actual Bayesian D–optimum design. Thus, in an attempt to include all the potential informative points to estimate (α,β) it is sensible to increase the number of support points in the starting design. Therefore, in the second attempt we will proceed searching for the optimum design over the class of, for example, six–point designs.

$$\xi^* = \begin{Bmatrix} -3.4691 & -1.9192 & -1.6289 & -0.6660 & -0.1293 & 0.8083 \\ 0.0536 & 0.1325 & 0.1598 & 0.1995 & 0.1812 & 0.2734 \end{Bmatrix} \qquad (6.4.14)$$

After repeating the numerical search procedure with a six–point starting design, design (6.4.14) is obtained. To check optimality, its derivative function is shown in Figure 6.4.2 confirming that this is indeed a Bayesian D–optimum design, for the upper limit for the derivative function is now not only obeyed at all points in the design region but also equalled at the support points of design (6.4.14). Figure 6.4.2 also shows that the derivative function values corresponding to design region points lying between the second and third support points of the optimum design are virtually equal to two, indicating that all values of x in this interval are highly informative.

EXAMPLE 6.5. Let the link function be the integrated normal, or probit. Suppose that the region Θ_0 is given by $\{(\alpha,\beta)$: $\alpha \in [-1,3]$ and $\beta \in [0.5,3]\}$ and that the prior distribution for α and β, $p_0(\alpha,\beta)$, consists of twenty equiprobable pairs (α,β) in the region Θ_0 as shown in Table 6.5. Finally, let the linear predictor be $\eta = \alpha + \beta x$..

TABLE 6.5 – EQUIPROBABLE TWENTY–POINT PRIOR.

PAIRS OF PARAMETERS (α,β). (PROBIT LINK)

Pair (α,β)			
(–1,0.5)	(–1,1)	(–1,2)	(–1,3)
(0,0.5)	(0,1)	(0,2)	(0,3)
(1,0.5)	(1,1)	(1,2)	(1,3)
(2,0.5)	(2,1)	(2,2)	(2,3)
(3,0.5)	(3,1)	(3,2)	(3,3)

Again, our first guess is to search for an optimum design over the class of five–point designs. The arbitrarily chosen starting design consists of support points {–3,–2.5,–1.5,–1,0} equally weighed. After the first numerical search, the following Bayesian D–optimum design candidate is obtained.

$$\xi = \begin{Bmatrix} -7.0163 & -3.7452 & -1.3273 & -0.3971 & 0.5469 \\ 0.0278 & 0.1055 & 0.2705 & 0.3446 & 0.2516 \end{Bmatrix} \qquad (6.4.15)$$

Figure 6.5.1 displays the plot of the derivative function corresponding to design (6.4.15). In the first half of the design region that is plotted, the derivative function attains the specifications under optimality, but in the second half it does not so that design (6.4.15) is not optimum and the search must be restarted. Again the three peaks above the value of two observed in Figure 6.5.1 can be interpreted as indicative of values x carrying valuable information about α and β that are not included in the optimum design. Therefore, in the next attempt the number of support points in the design should be increased.

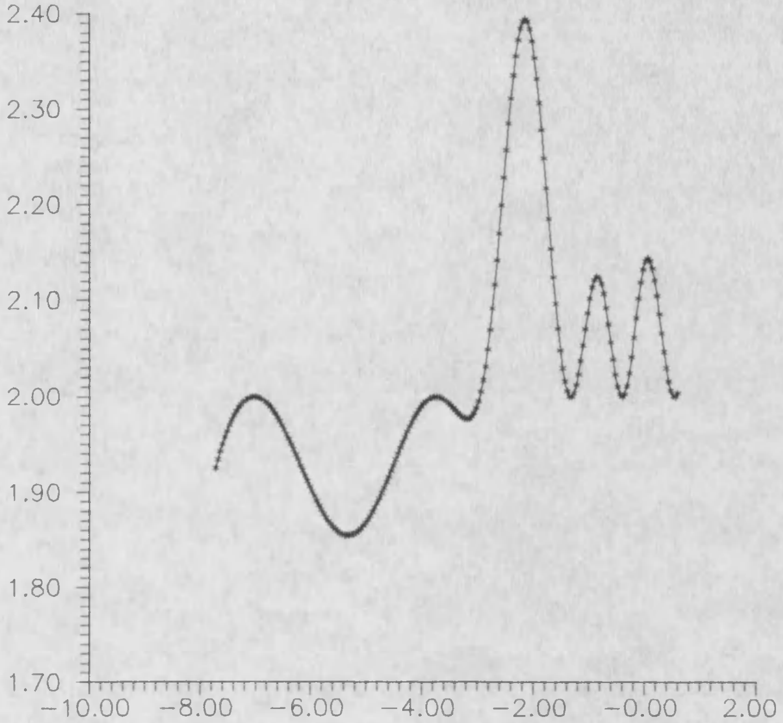


Figure 6.5.1. Derivative Function for design (6.4.15)

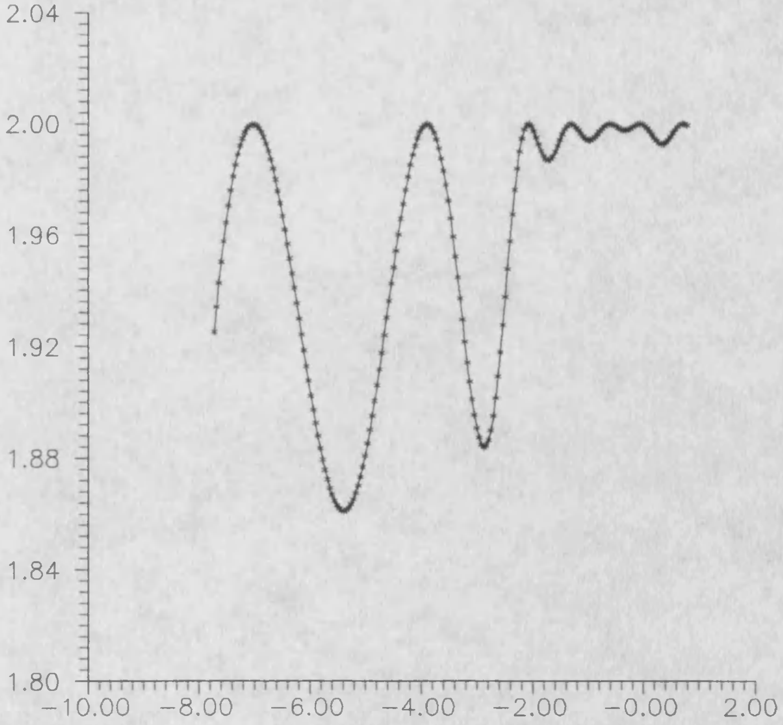


Figure 6.5.2. Derivative Function for design (6.4.16)

For instance, let the numerical search be carried out over the class of seven–point supported designs since the optimum design seems to require at least two more support points as suggested by Figure 6.5.1. Then, the following Bayesian D–optimal design candidate is determined.

$$\xi^* = \begin{Bmatrix} -7.0186 & -3.8917 & -2.0738 & -1.3106 & -0.6035 & -0.0612 & 0.7156 \\ 0.0274 & 0.0889 & 0.0645 & 0.1967 & 0.2303 & 0.1982 & 0.1940 \end{Bmatrix} \qquad (6.4.16)$$

Design (6.4.16) turns out to be Bayesian D–optimum as shown by the shape of its derivative function, displayed in Figure 6.5.2. Further, in the region where the derivative function of design (6.4.15) peaked above the upper limit, now the derivative function related to design (6.4.16) is almost flat, and most importantly, below or equal to the upper limit. A comparison of designs (6.4.15) and (6.4.16) shows that the extra two support points belonging to the latter were inserted in this region, confirming what was suggested before, i.e. high peaks in the derivative function of nonoptimum designs usually correspond to points in the design region that ought to belong to the optimum design.

Hence, this example illustrates very well the sort of situations that can occur in practice when searching for optimality of experimental designs without any help from theoretical results establishing boundaries for the number of support points in the optimum design. Due to this difficulty as well as an inadequate starting design sometimes it can be very hard to achieve optimality.

6.5. DISCUSSION

The numerical results obtained in this chapter provide good examples of how to obtain informative experimental designs for binary response models in practical situations. To avoid excess of computational effort, in all the examples we considered linear predictors consisting of only one explanatory variable. This has enabled us to do all the calculations on a PC 486. All the periods of time elapsed to determine optimum designs and derivative functions were quite short, always less than a minute, making the use of these techniques of searching for optimum experimental designs a simple task to carry out with modern PCs.

A general NAG routine was used as the optimization procedure in all examples. However, more specific methods of numerical search are required for models with several explanatory variables for which the necessary calculations to obtain the optimum designs will be much more complex and take much longer. Furthermore, as the dimension of the problem increases the occurrence of local optima might be more frequent so that search procedures which are able to avoid local optima should be preferred. For instance, when using NAG routine E04JAF the search for an optimum five–point design in a model with a single independent variable like those in this chapter is carried out in nine dimensions. A similar search in a model with multiple independent variables would have to be performed in a much larger number of dimensions. However, once these numerical problems are successfully dealt with the techniques described in this chapter can be very useful for researchers in apllied statistics.

In the Bayesian approach for determining experimental designs optimizing parameter estimation, a great amount of input is expected from the experimenter such as the specifications of the binary response model (link function and linear predictor structure), and a prior distribution for the linear predictor parameters. However, in practical situations subjective choices of "optimum" experimental designs can also be made by trying out different model specifications and/or prior distributions. This alternative gives another dimension for the Bayesian techniques described here, i.e. they can also be useful as exploratory tools in the planning of an experiment.

In this chapter the purpose of the experiment is restricted to estimation of the parameters of a binary response model as precisely as possible

(D–optimality). For other important criteria such as D_s–optimality, D_a–optimality, etc, similar methodology can be developed so as to extend the existing results to their Bayesian generalizations.

APPENDIX D

In this appendix the aim is to determine the Fréchet derivative of criterion function (6.3.4). Recall its definition; $\Delta(\xi) = E_{\beta} \log|M(\xi,\beta)|$, if $M(\xi,\beta)$ is nonsingular for all β relevant to the prior distribution $p_0(\beta)$. Suppose that this is the case, i.e. $|M(\xi,\beta)| \neq 0$, for all β relevant to $p_0(\beta)$.

Then, let us denote $M_1 = M(\xi_1,\beta)$; $M_2 = M(\xi_2,\beta)$, $\xi_1,\xi_2 \in \mathscr{H}$ and let p be the dimension of the vector of parameters β. Based on the linearity of the expectation operator E, it is straightforward to extend to Bayesian D–optimality the result proven, for instance, in Silvey (1980, Chapter 3, pp. 21) for non–Bayesian D–optimality. That is, the Fréchet derivative function of $\Delta(\xi)$ at M_1 in the direction of M_2 is given by

$$F_{\Delta}(M_1,M_2) = E_{\beta}\left\{\mathrm{tr}\left[M(\xi_2,\beta)\, M^{-1}(\xi_1,\beta)\right]\right\} - p \tag{D.1}$$

Now, let us take ξ_2 to be a design measure putting all mass at a specific point $\mathbf{x}$ in the design region $\mathfrak{X}$ and denote such a measure by $\xi_{\mathbf{x}}$. Then, according to (6.3.3) the diagonal matrix W related to $\xi_{\mathbf{x}}$ is unidimensional whereas the design matrix X corresponding to $\xi_{\mathbf{x}}$ reduces to a row vector so that according to (6.3.2) the local Fisher information matrix for β based on design $\xi_{\mathbf{x}}$ is given by

$$M(\xi_{\mathbf{x}},\beta) = w(\mathbf{x},\beta)\, f(\mathbf{x})^t\, f(\mathbf{x}) \tag{D.2}$$

where

$f(\mathbf{x}) = (f_1(\mathbf{x}),\cdots,f_p(\mathbf{x}))$;

$w(\mathbf{x},\beta) = \dfrac{1}{\pi(\mathbf{x},\beta)\{1 - \pi(\mathbf{x},\beta)\}} \left[\dfrac{d\pi}{d\eta}\right]^2_{\pi=\pi(\mathbf{x},\beta)}$;

$\pi(\mathbf{x},\beta) = g^{-1}(\eta(\mathbf{x},\beta))$; g(.) is the link function ;

and $\eta(\mathbf{x},\beta) = \sum_j f_j(\mathbf{x})\beta_j$.

Hence, replacing ξ_2 by $\xi_{\mathbf{x}}$, or alternatively, replacing $M(\xi_2,\beta)$ as in (D.1) by $M(\xi_{\mathbf{x}},\beta)$ given by (D.2) we obtain

$$F_{\Delta}(M_1,M_{\mathbf{x}}) = E_{\beta}\left\{\mathrm{tr}\left[w(\mathbf{x},\beta)\ f(\mathbf{x})^t\ f(\mathbf{x})\ M^{-1}(\xi_1,\beta)\right]\right\} - p$$

$$= E_{\beta}\left\{w(\mathbf{x},\beta)\ \mathrm{tr}\left[f(\mathbf{x})\ M^{-1}(\xi_1,\beta)\ f(\mathbf{x})^t\right]\right\} - p$$

$$= E_{\beta}\left[w(\mathbf{x},\beta)\ f(\mathbf{x})\ M^{-1}(\xi_1,\beta)\ f(\mathbf{x})^t\right] - p. \qquad \text{(D.3)}$$

That completes the derivation of the Frechet derivative of criterion function (6.3.4). Now, considering that (6.3.4) is a concave function on the set of design measures $\mathscr{H}$, a necessary condition for applying optimal design theory results, and applying Theorems 2.1, 2.2, and Corollary 2.1, the conditions of Theorem 6.1 follow. That is, when ξ_1 is replaced by ξ^*, an optimum design with respect to (6.3.5), in (D.3), then for all $\mathbf{x} \in \mathcal{X}$

$$E_{\beta}\left[w(\mathbf{x},\beta)\ f(\mathbf{x})\ M^{-1}(\xi^*,\beta)\ f(\mathbf{x})^t\right] \leq p.$$

CHAPTER 7. OPTIMUM EXPERIMENTAL DESIGNS FOR THE CHOICE OF LINK FUNCTION FOR A BINARY DATA MODEL

7.1. INTRODUCTION

In this chapter we investigate some consequences of including an extra parameter to extend the link function on the modelling of binary data. Specifically, we address the problem of designing optimum experiments in this framework. The extended link includes the logistic and complementary log–log as special cases. An important feature of this approach concerns the wide choice of criteria that can be considered, namely optimum designs to estimate the link function parameter, the linear predictor parameters, or the whole set of unknown parameters. In the latter case two possibilities are taken into consideration. Firstly, the parameters are estimated regardless of their importance, and secondly complementary weights are assigned to the estimation of the link function and the linear predictor parameters. Different criteria can be used to assign these weights. But, whichever criterion is used, the values of the weights should reflect any specific priority of the experimenter. In each case a different criterion function is defined. As in the previous chapters, prior information is incorporated in the design criteria as the related criterion functions depend upon the parameters being estimated.

In the next section a brief review of the literature concerning families of link functions for modelling binary data is presented. We also introduce our choice of link function family together with two numerical examples which illustrate the practical applications of this approach. In Section 7.3 the Fisher information matrix for the extended set of parameters is derived. Then, a selection of criteria reflecting different optimization aims is defined. As an illustration for the proposed methodology some examples of local optimum as well as Bayesian optimum designs are provided in Section 7.4. Finally, Section 7.5 contains a discussion about the methods and numerical results obtained in this chapter as well as alternative approaches.

7.2. BACKGROUND AND THE GENERALIZED LINK FUNCTION

In the modelling of binary data the choice of link function plays an important role, as illustrated later in this section by Examples 7.1 and 7.2. However, only three functions are widely used in most applications. They are the logistic (logit), the inverse normal (probit) and the complementary log–log link functions. In practice, by increasing the number of link function candidates one should expect to obtain better results as far as the fitting of binary data is concerned.

In the previous chapter, the purpose was to determine optimum designs to estimate the linear predictor parameters given that the model was specified by one of the above mentioned link functions. This problem, nevertheless, becomes much more general, as well as interesting, when the choice (or estimation) of the link function is also taken into consideration. Recent developments in the modelling of binary data have made this possible.

Several alternative link functions, and link function families have been proposed hitherto making the choice of models for binary data widen substantially. However, an inevitable consequence of extending the choice of link function in this manner is the inclusion of additional parameters in the underlying binary data model as the family membership must be specified, and therefore, the extra parameter(s) must be estimated. A simple solution for the problem of estimating the extra parameter(s) consists of taking nested likelihood estimators. This process of estimation is known as the likelihood or deviance profile.

Prentice (1976) suggested a four–parameter link function family which includes the inverse normal (probit) and the logit links among other important link functions. In addition to the location and scale parameters the model considers two other parameters related to the shape of the dose–response curve. Based on a similar approach, Pregibon (1980) proposed another four–parameter link function family in which the logit link is a particular case.

Aranda–Ordaz (1981) considers two families of power transformations (link functions) for the probability of success. The first adds one extra parameter to the model and satisfies a condition of symmetry for the probabilities of failure and success. The second consists of assymetric transformations in the above sense whilst adding also only one parameter to the model. The logistic and linear link functions are particular cases of the former. So is the complementary log–log link for the latter.

More recently, Rocke (1993) suggested a transformation family which is based on the incomplete beta function. It includes the logistic, arcsin square–root, and the identity link functions among others. Further, other transformation families are discussed and compared to the beta transformation family.

Czado (1992) pointed out that the estimated variances of the parameter estimates inflate as a consequence of estimating the additional link parameter(s). Then, she proposed a unified approach for choosing parametric link function families, for generalized linear models, that reduces the variance inflation.

The generalization to be proposed in this section consists of extending the class of link functions by adding only one extra parameter to the model. Since this extra parameter (the link function parameter) is estimable its inclusion offers the experimenter a wide choice of estimation purposes.

However, if priorities are to be assigned it is more sensible to allocate a greater amount of effort to estimate the link function parameter, for the estimation of the linear predictor parameters clearly depends upon the link function specification. Having said that, there might be situations in which two or more link functions fit the data rather well, even though the linear predictor parameter estimates differ significantly for each link, thus, making the estimation of the latter more crucial.

The basic assumptions made for the models of Chapter 6 are again supposed for the models of this chapter. Now, we introduce the generalized link function for binary data models that is mentioned by both Pregibon (1980) and Aranda–Ordaz (1981), and reproduced in McCullagh and Nelder (1989, p. 378).

$$g(\pi_i,\lambda) = \log\left[\left\{\left[\frac{1}{1-\pi_i}\right]^{\lambda} - 1\right\} / \lambda\right];\ \lambda \in \mathbb{R}^{+} \qquad (7.2.1)$$

where $g(\pi_i,\lambda) = \eta_i$, the linear predictor given by $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$.

When $\lambda = 1$ (7.2.1) reduces to the logistic link. Furthermore, it is straightforward to prove that $\lim_{\lambda\to 0} g(\pi_i,\lambda) = \log\{-\log(1-\pi_i)\}$, the complementary log–log link function. Therefore, the value of λ restricted to the interval (0,1), may be interpreted as a measure of the distance between the logistic and the complementrary log–log models.

Nevertheless, there is no reason whatsoever for restricting attention to models in which λ belongs to this interval, for any nonnegative value of λ is a potential candidate to provide a reasonable fit for binary data. Negative values of λ are not considered as numerical problems may arise in the computation of the link function inverse.

To illustrate how the maximum likelihood estimator for λ, say $\hat{\lambda}$, can be obtained we present two numerical examples using real data sets. Note that in both examples, $\hat{\lambda}$ lies outside the interval [0,1].

EXAMPLE 7.1. Toxicity of Rotenone to *Macrosiphoniella sanborni.* Finney (1947, Ex.1, p.26, Table 2) gives the details. To investigate how the value of λ affects the goodness of fit for a binary data set, the criterion adopted was the deviance. The estimation of λ and β was carried out in two steps. Firstly, we may suppose that the value of λ is known and proceed to estimate β. Then, the log likelihood with respect to λ is maximized, or equivalently the deviance is minimized. Such a procedure is analogous to the so–called nested least squares. Analytical solutions for this problem appear to be very complicated. However, a numerical search for the minimal deviance over a grid for λ is effective and easy to implement. The model fitted to the rotenone data had the linear predictor $\eta_i = \beta_0 + \beta_1 x_i$, where $\{x_i\}$ are values of log concentration of rotenone. For each value of λ, the parameters β_0 and β_1 were estimated by iterative weighted least squares as described in McCullagh & Nelder (1989, pp. 40–43).

Figure 7.1 shows the deviance profile, i.e. the deviances as a function of λ over a grid in the interval [0.0,5.0]. Although the change in deviance is not large, values of λ in the interval (1.25,1.5) yield smaller deviances, the optimum lying near $\lambda = 1.3$. Obviously, further analysis ought to be carried out before regarding any particular model as suitable.

The linear predictor parameter estimates vary with the value of λ. For example, when $\lambda = 0$ (complementary log–log), $\hat{\beta}_0 = -3.512$ and $\hat{\beta}_1 = 4.426$ whereas for $\lambda = 1$ (logistic), $\hat{\beta}_0 = -4.839$ and $\hat{\beta}_1 = 7.068$. These results are later referred to in order to justify the assumption of prior distributions for the linear predictor parameters being conditioned on the value of λ. In fact, there seems to be a trend in the behaviour of $\hat{\beta}_0$ and $\hat{\beta}_1$ that could be investigated more carefully.

EXAMPLE 7.2. Milicer & Szczotka (1966) give the number of schoolgirls having menstruated as a function of age. The data are reproduced by Aranda–Ordaz (1981, Table 2). As in the previous example the linear predictor was simply $\eta_i = \beta_0 + \beta_1 x_i$. The procedure described in Example 7.1, to estimate λ, β_0 and β_1 was again applied.

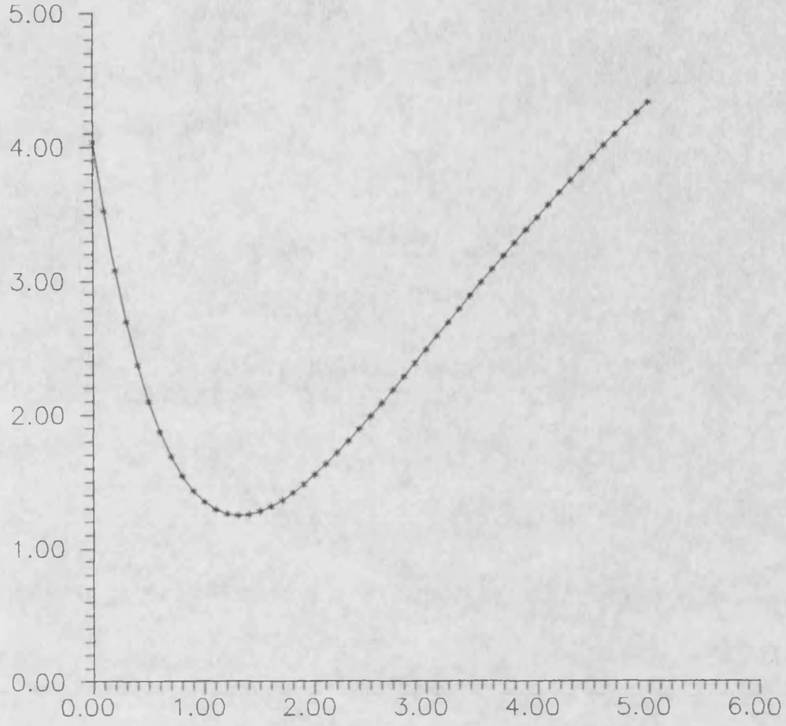


Figure 7.1. Deviance (lambda) for Example 7.1

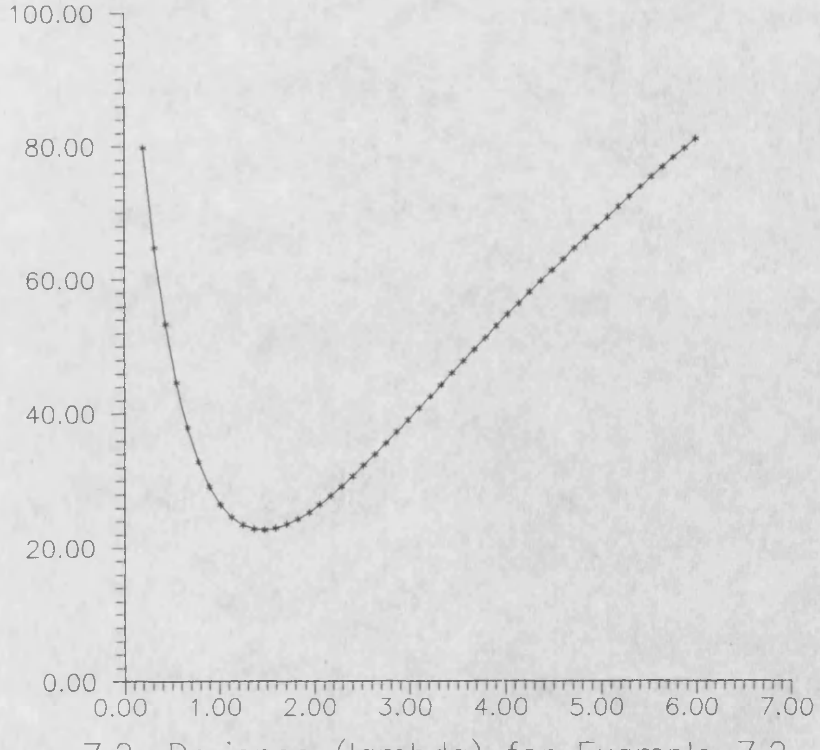


Figure 7.2. Deviance (lambda) for Example 7.2

The deviances are shown in Figure 7.2. The search was carried out over a grid in the interval [0.2,6.0]. As can be seen, the value of λ that provides the smallest deviance lies in the interval (1.25,1.7), more precisely close to 1.475.

In the next section we obtain the Fisher information matrix relative to the full problem, i.e. when λ is unknown. In addition, a number of design criteria are suggested and discussed.

7.3. A SELECTION OF OPTIMALITY CRITERION

We can now address the problem of designing optimal experiments to estimate λ and/or β. The main conclusion that can be drawn from the above examples is that a substantial importance must be given to the estimation of λ, whenever (7.2.1) is taken as the link function. Thus, from the optimal design theory point of view, it is sensible to focus the optimization on the estimation of λ rather than on the estimation of the linear predictor parameters β. A third possibility is to find the right balance between the two purposes.

7.3.1. THE FISHER INFORMATION MATRIX

Prior to defining the criterion functions, the Fisher information matrix for the set of parameters (λ,β) must be obtained. This requires a sequence of differentiations of the log likelihood function (6.3.1) with respect to λ and β. The resulting Fisher information matrix for (λ,β) is the following.

$$\frac{1}{N} M(\lambda,\beta,\xi_N) = \begin{bmatrix} \sum_1^n w_i \left[\frac{d\pi_i}{d\lambda}\right]^2 & \left[\sum_1^n w_i \frac{d\pi_i}{d\lambda} \frac{d\pi_i}{d\eta_i} x_{ir}\right]^t \\ \left[\sum_1^n w_i \frac{d\pi_i}{d\lambda} \frac{d\pi_i}{d\eta_i} x_{ir}\right] & \left[\sum_1^n w_i \left[\frac{d\pi_i}{d\eta_i}\right]^2 x_{ir} x_{is}\right] \end{bmatrix} \qquad (7.3.1)$$

where

$$N = \sum_1^n m_i; \quad w_i = \frac{p_i}{\pi_i(1-\pi_i)}; \quad p_i = \frac{m_i}{N}; \quad \xi_N = \begin{Bmatrix} x_1 \dots x_n \\ p_1 \dots p_n \end{Bmatrix};$$

$$\frac{d\pi_i}{d\lambda} = \frac{-\left[\frac{1}{\lambda^2}\right] \log(\lambda e^{\eta_i} + 1) + \frac{(1/\lambda) e^{\eta_i}}{\lambda e^{\eta_i} + 1}}{(\lambda e^{\eta_i} + 1)^{1/\lambda}};$$

$$\frac{d\pi_i}{d\eta_i} = \frac{\lambda e^{\eta_i}}{(\lambda e^{\eta_i} + 1)^{1/\lambda}}; \quad r,s = 1,\dots,p.$$

The dimensions of the submatrices of matrix (7.3.1) are respectively by row, 1×1, 1×p, p×1, and p×p. Alternatively, (7.3.1) can be written in a compact manner, given by

$$M(\lambda,\beta,\xi_N) = \begin{bmatrix} (1^t W_{\lambda\lambda} 1) & (1^t W_{\lambda\eta} X) \\ (X^t W_{\lambda\eta} 1) & (X^t W_{\eta\eta} X) \end{bmatrix} \tag{7.3.2}$$

where

$$W_{\lambda\lambda} = \text{diag}\left[m_i \left[\frac{d\pi_i}{d\lambda}\right]^2 / \pi_i(1-\pi_i)\right]; \quad W_{\lambda\eta} = \text{diag}\left[m_i \frac{d\pi_i}{d\lambda} \frac{d\pi_i}{d\eta_i} / \pi_i(1-\pi_i)\right];$$

$W_{\eta\eta} = \text{diag}\left[m_i \left[\frac{d\pi_i}{d\eta_i}\right]^2 / \pi_i(1-\pi_i)\right]$; 1 is a column vector containing only ones ; and X is the associated design matrix.

It is interesting to notice that the lower right submatrix of matrix (7.3.2) is identical to the Fisher information matrix for a binary data model in the absence of the link function parameter λ (as derived in Section 6.3 of Chapter 6). Here, however, the expression for $\left[\frac{d\pi_i}{d\eta_i}\right]$ includes λ.

Here, as in all the previous chapters, we are only concerned with the approximate theory of optimal design so that matrix (7.3.1) must be adapted to this framework. This means to write down the Fisher information matrix corresponding to a design measure ξ which is given below.

$$M(\lambda,\beta,\xi) = \begin{bmatrix} \int w\left[\frac{d\pi}{d\lambda}\right]^2 \xi(dx) & \left[\int w\, \frac{d\pi}{d\lambda}\frac{d\pi}{d\eta}\, x_r\, \xi(dx)\right]^t \\ \left[\int w\, \frac{d\pi}{d\lambda}\frac{d\pi}{d\eta}\, x_r\, \xi(dx)\right] & \left[\int w\left[\frac{d\pi}{d\eta}\right]^2 x_r x_s\, \xi(dx)\right] \end{bmatrix} \qquad (7.3.3)$$

All the integrals in (7.3.3) are taken over the design region $\mathfrak{X}$ and all the variables have the same definition as before, but with no index. The exception is the weight w which is given by $w = [\pi(1-\pi)]^{-1}$. In order to shorten the notation, from now on $M(\lambda,\beta,\xi)$ will be denoted simply by M, and its submatrices by

$$M_{11} = \int_{\mathfrak{X}} w\left[\frac{d\pi}{d\lambda}\right]^2 \xi(dx); \qquad M_{12} = \left[\int_{\mathfrak{X}} w\, \frac{d\pi}{d\lambda}\frac{d\pi}{d\eta}\, x_r\, \xi(dx)\right]^t;$$

$$M_{22} = \left[\int_{\mathfrak{X}} w\left[\frac{d\pi}{d\eta}\right]^2 x_r x_s\, \xi(dx)\right]; \text{ and } \qquad M_{21} = \left[\int_{\mathfrak{X}} w\, \frac{d\pi}{d\lambda}\frac{d\pi}{d\eta}\, x_r\, \xi(dx)\right].$$

The reason for partitioning matrix (7.3.3) in this way is that M as well as M_{11}, M_{12}, M_{21}, and M_{22} are essential for defining the optimization criterion functions in the next subsection.

7.3.2. OPTIMIZATION CRITERIA

In this subsection, we consider some possibilities for the choice of criterion function according to the aim of the experiment. There are several purposes of interest in a binary data experiment, such as estimating a subset of parameters, estimating the LD50, or any other percentile, or estimating a certain function of the model parameters that might be meaningful. Obviously, the choice of criterion function needs to reflect the particular purpose. Here, we are concerned with the estimation of parameters and particular subsets of them. To be more specific, our main interest lies in four distinct but interrelated purposes, namely

(i) estimate the link function parameter λ ;

(ii) estimate the vector of linear predictor parameters $\boldsymbol{\beta}$;

(iii) estimate both λ and $\boldsymbol{\beta}$;

(iv) estimate λ and $\boldsymbol{\beta}$ with complementary weights (relative importance).

Due to the features of the first two problems the criterion adopted in (i) and (ii) is that of D_s–optimality. D–optimality is suitable for the full problem (iii) whereas for problem (iv) a linear combination of criterion functions related to problems (i) and (ii) is adopted. For convenience, we split the formulation of the criteria and related derivative functions into these four cases. The criterion and criterion function for each case of interest are given below. The criterion functions for cases (i) and (ii) are presented in two equivalent expressions. For further details the reader should refer to Appendix E. Here, $|.|$ denotes the determinant operator.

(i) Estimating $\boldsymbol{\lambda}$

$$\underset{\xi \in \mathcal{H}}{\text{Maximize}}\ \Psi_1(M), \text{ where}$$

$$\Psi_1(M) = \log \left| M_{11} - M_{12} M_{22}^{-1} M_{21} \right| = \log \left\{ \frac{|M|}{|M_{22}|} \right\} \tag{7.3.4}$$

(ii) Estimating $\boldsymbol{\beta}$

$$\underset{\xi \in \mathcal{H}}{\text{Maximize}}\ \Psi_p(M), \text{ where}$$

$$\Psi_p(M) = \log \left| M_{22} - M_{22} M_{11}^{-1} M_{12} \right| = \log \left\{ \frac{|M|}{|M_{11}|} \right\} \tag{7.3.5}$$

(iii) Estimating $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$

$$\underset{\xi \in \mathcal{H}}{\text{Maximize}}\ \Psi(M), \text{ where}$$

$$\Psi(M) = \log |M| \tag{7.3.6}$$

(iv) Mixture criterion

$$\underset{\xi \in \mathcal{H}}{\text{Maximize}}\ \Psi_m(M), \text{ where for } 0 \leq \alpha \leq 1$$

$$\Psi_m(M) = \alpha\, \Psi_1(M) + (1-\alpha)\, \Psi_p(M) = \log \left\{ \frac{|M|}{|M_{11}|^{1-\alpha} |M_{22}|^{\alpha}} \right\} \tag{7.3.7}$$

The inclusion of a pair of complementary weights, representing the relative importance of estimating λ and β, in the mixture criterion (iv) is an alternative formulation for (iii). The weights can either be specified by the experimenter, based on some subjective criterion, or can be determined through the optimization procedure, by finding the value of α corresponding to the mixture criterion optimum design achieving maximum efficiency with respect to both optimum designs resulting from applying criteria (i) and (ii). Although only the former method is used in the illustrations of Section 7.4, further details concerning the latter are presented in Section 7.5. Criterion functions (i) and (ii) are particular cases of criterion function (iv) when either purpose is allocated weight one.

As (7.3.4), (7.3.5), and (7.3.6) are concave functions on the set of design measures $\mathscr{H}$, a property required to apply optimal design theory, a theorem similar to that of the previous chapter can be proven. Further, it is straightforward to prove that (7.3.7) is concave on $\mathscr{H}$ so that a similar theorem also holds for this case.

Optimum designs for all the above criteria depend on the parameters λ and β. If λ were known, say $\lambda = 1$, the problem would reduce to designing for estimation of β. The optimum design would still require knowledge of β. The reverse problem of known β with λ unknown does not make sense in practice, unless λ and β are orthogonal as defined by Cox and Reid (1987). Examples 7.1 and 7.2 show that this is not necessarily the case. We are thus left with the dependence of the optimum designs on the parameter values.

7.3.3. DERIVATIVE FUNCTIONS

To check the optimality of any optimum design candidate we require the derivative function of the design criterion. Suppose that for fixed λ and β, $M(\xi^*)$ and $M(\xi_x)$ denote the normalized information matrices at the optimum design ξ^*

and at the design ξ_x, respectively, where ξ_x is the measure assigning mass one to the point $x \in \mathcal{X}$. By replacing ξ_x in (7.3.3) we obtain $M(\xi_x)$ which is given by

$$M(\xi_x) = \begin{bmatrix} w_{\lambda\lambda} & w_{\lambda\eta} f(x)^t \\ w_{\lambda\eta} f(x) & w_{\eta\eta} f(x) f(x)^t \end{bmatrix} \tag{7.3.8}$$

where

$$w_{\lambda\lambda} = \left[\frac{d\pi}{d\lambda}\right]^2_{\pi=\pi(x)} / \pi(x)(1-\pi(x)) \; ; \; w_{\lambda\eta} = \left[\frac{d\pi}{d\lambda}\right]_{\pi=\pi(x)} \left[\frac{d\pi}{d\eta}\right]_{\pi=\pi(x)} / \pi(x)(1-\pi(x)) \; ;$$

$$w_{\eta\eta} = \left[\frac{d\pi}{d\eta}\right]^2_{\pi=\pi(x)} / \pi(x)(1-\pi(x)) \; ; \text{ and } f(x)^t = (f_1(x), \cdots, f_p(x)).$$

Then, the Frechet derivative of the criterion function at $M(\xi^*)$ in the direction of $M(\xi_x)$, or the derivative function, provides the following bounds which are essential for checking the optimality of a design candidate (see Appendix E, for details).

(i) **Estimating λ** – For all $x \in \mathcal{X}$,

$$\psi_1(x) = \mathrm{tr}\left[M(\xi_x)\left\{M(\xi^*)\right\}^{-1}\right] - \mathrm{tr}\left[M_{22}(\xi_x)\left\{M_{22}(\xi^*)\right\}^{-1}\right] \le 1 \tag{7.3.9}$$

(ii) **Estimating β** – For all $x \in \mathcal{X}$,

$$\psi_p(x) = \mathrm{tr}\left[M(\xi_x)\left\{M(\xi^*)\right\}^{-1}\right] - \mathrm{tr}\left[M_{11}(\xi_x)\left\{M_{11}(\xi^*)\right\}^{-1}\right] \le p \tag{7.3.10}$$

(iii) **Estimating λ and β** – For all $x \in \mathcal{X}$,

$$\psi(x) = \mathrm{tr}\left[M(\xi_x)\left\{M(\xi^*)\right\}^{-1}\right] \le p + 1 \tag{7.3.11}$$

(iv) **Mixture criterion** – For all $x \in \mathcal{X}$,

$$\begin{aligned}\psi_m(x) = \mathrm{tr}\left[M(\xi_x)\left\{M(\xi^*)\right\}^{-1}\right] &- \alpha\, \mathrm{tr}\left[M_{22}(\xi_x)\left\{M_{22}(\xi^*)\right\}^{-1}\right] \\ &- (1-\alpha)\, \mathrm{tr}\left[M_{11}(\xi_x)\left\{M_{11}(\xi^*)\right\}^{-1}\right] \le \alpha + (1-\alpha)p \end{aligned} \tag{7.3.12}$$

Because of the dependence of the information matrix on the unknowm parameter values, the results obtained so far yield local optimum designs for estimating the parameters λ and/or β. However, by incorporating prior distributions into the model, Bayesian optimum designs are also obtainable. This is developed in the next subsection.

7.3.4. THE BAYESIAN APPROACH

It is straightforward to prove that criterion functions defined as expectations of (7.3.4), (7.3.5), (7.3.6) and (7.3.7) are still concave functions on $\mathscr{H}$. Consequently all results from optimal design theory continue to hold.

As suggested by Examples 7.1 and 7.2 it is reasonable to take priors for the parameters $\{\beta_i\}$ conditional on the value of λ. In fact, the assumption of a conditional probability distribution, say $p_0(\beta|\Lambda=\lambda)$, is general in the sense that if values of $\{\beta_i\}$ are independent of λ all results to be developed here will still hold.

Taking the priors $p_0(\beta|\Lambda=\lambda)$ and $p_0(\lambda)$ into consideration the criteria are defined as the expected values of criterion functions (7.3.4), (7.3.5), (7.3.6) and (7.3.7), expectations being taken in two steps, first over $p_0(\beta|\Lambda=\lambda)$ and then over $p_0(\lambda)$. Analogously, the derivative functions are defined as the expected values of expressions (7.3.9), (7.3.10), (7.3.11), and (7.3.12).

In the next section we show how to obtain local and Bayesian optimum designs through a series of examples illustrating the use of each criterion and the procedures for checking optimality.

7.4. EXAMPLES

The first two examples concern local optimum designs, that is designs which are calculated for a single assumed value of λ and β. In Example 7.3, the interest lies in the estimation of the link function parameter λ, whereas in Example 7.4 the aim is to determine local optimum designs for all four different purposes, using the same exact prior information on the parameters. In both examples we assume that the linear predictor structure is of the kind $\eta = \beta_0 + \beta_1 x$. Hence, the dimension of the problem is three.

EXAMPLE 7.3. The values of the parameters are $\lambda = 1.4$, $\beta_0 = 0.5$, and $\beta_1 = 1.0$. The aim is to estimate λ. Therefore, criterion function (7.3.4) is applied, yielding the following local optimum design.

$$\xi^* = \begin{Bmatrix} -2.384 & 0.9266 & 3.335 \\ 0.6954 & 0.2691 & 0.0355 \end{Bmatrix} \tag{7.4.1}$$

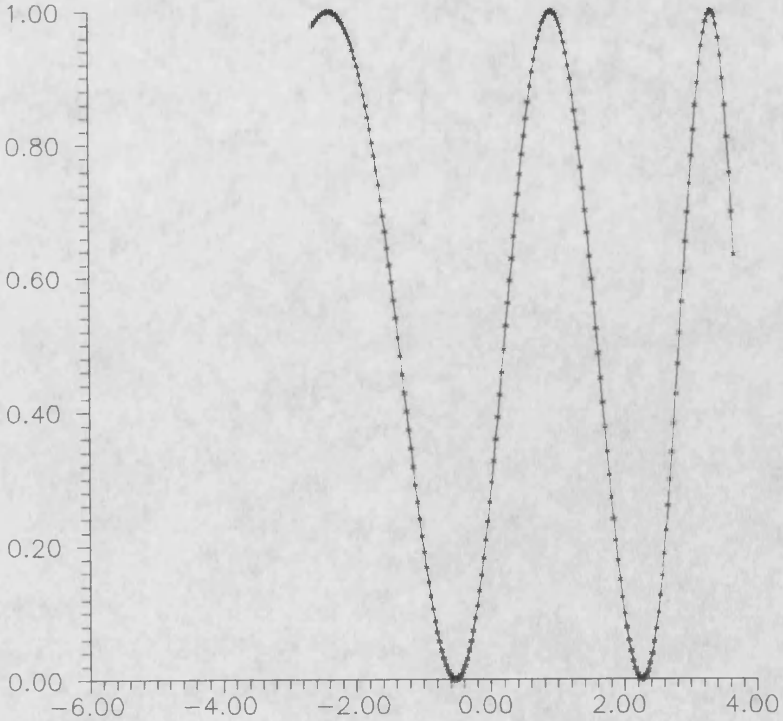


Figure 7.3. Derivative Function for design (7.4.1)

The value of the criterion function at design (7.4.1) is $\Psi_1(\overset{*}{\xi}) = -5.766$. Most of the total weight is assigned to x = –2.384 reflecting the importance of such a point in estimating the link function parameter λ. To make sure that the above design is indeed optimum, Figure 7.3 shows the derivative function (7.3.9) in which the upper limit, $\psi_1(x) = 1$ is achieved only at the support points for the optimum design. In addition, the two points in the design region $\mathcal{X}$, at which the derivative function is zero, can be considered as noninformative about λ.

EXAMPLE 7.4. All criteria introduced in Section 7.3 are utilized. The parameter values are $\lambda = 0.001$, $\beta_0 = 0.5$, and $\beta_1 = 1.0$. Table 7.1 shows the resulting local optimum designs and respective optimum values for the related criteria.

In spite of using the same exact prior information about the parameters λ, β_0 and β_1, we obtain significantly distinct local optimum designs. For instance, more than 50% of the weight is allocated to the only positive support point 1.376 in the optimum design to estimate λ, whereas to estimate β_0 and β_1, nearly 50% of the weight is allocated to the most negative support point of the design, –2.841. The unbalanced allocation of weights in both designs suggests that the greatest part of the information about λ is concentrated in positive values of the covariate x, as opposed to negative values of x, which appear to contain most part of the information about β_0 and β_1. To reinforce this interpretation design (7.4.4), the optimum design to estimate λ and β, allocates equal weights to the support points as though it were a combination between estimating λ and β, neither being emphasized.

Intermediate situations arise when the mixture criterion function (7.3.7) is used. For instance, the less importance is assigned by the mixture criterion to estimate λ the more weighty the most negative support point of the optimum design becomes. This result, together with a similar occurrence of larger weights for the only positive support point as the relative importance of estimating λ increases, also agrees with the above interpretation that the information about λ seems to be concentrated in the positive support point. It is also interesting to notice that the range corresponding to the optimum design support points for situations in which estimating λ is given more importance is wider, although the narrowest range corresponds to criterion function (7.3.6).

TABLE 7.1 – OPTIMUM DESIGNS FOR ALL CRITERIA (SECTION 7.3)

PURPOSE	UPPER LIMIT	OPTIMUM VALUE OF CRITERION	OPTIMUM DESIGN
(a) Estimate λ	$\psi_1(x) \leq 1$	–2.526	$\begin{Bmatrix} -3.633 & -0.1416 & 1.376 \\ 0.2681 & 0.2276 & 0.5043 \end{Bmatrix}$
(b) Estimate β	$\psi_2(x) \leq 2$	–31.62	$\begin{Bmatrix} -2.841 & -0.4758 & 1.296 \\ 0.4997 & 0.3477 & 0.1526 \end{Bmatrix}$
(c) Estimate both	$\psi(x) \leq 3$	–32.77	$\begin{Bmatrix} -2.645 & -0.1452 & 1.117 \\ 0.3333 & 0.3333 & 0.3333 \end{Bmatrix}$
(d) Mixture ($\alpha=0.5$)	$\psi_m(x) \leq 1.5$	–17.24	$\begin{Bmatrix} -3.028 & -0.3692 & 1.309 \\ 0.3969 & 0.3231 & 0.28 \end{Bmatrix}$
(e) Mixture ($\alpha=0.25$)	$\psi_m(x) \leq 1.75$	–24.46	$\begin{Bmatrix} -2.905 & -0.4257 & 1.296 \\ 0.4505 & 0.3412 & 0.2082 \end{Bmatrix}$
(f) Mixture ($\alpha=0.8$)	$\psi_m(x) \leq 1.2$	–8.466	$\begin{Bmatrix} -3.301 & -0.2646 & 1.341 \\ 0.3245 & 0.2782 & 0.3974 \end{Bmatrix}$

where
(a) design (7.4.2); (b) design (7.4.3); (c) design (7.4.4);
(d) design (7.4.5); (e) design (7.4.6); and (f) design (7.4.7).

Another feature of the designs in Table 7.1 is that they all are supported on three points, regardless the number of parameters they are meant to estimate. This can be explained by the fact that the number of support points in the

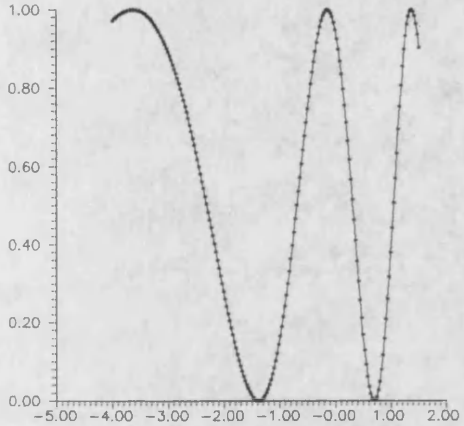


Figure 7.4.1. Derivative Function for design (7.4.2)

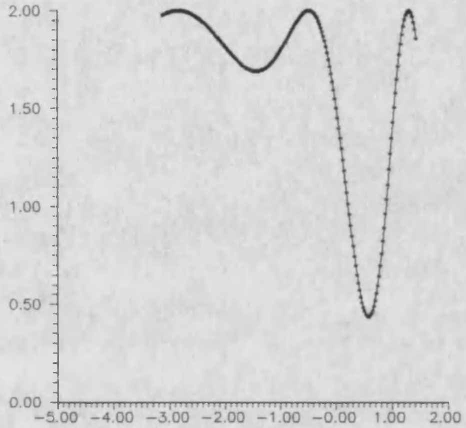


Figure 7.4.2. Derivative Function for design (7.4.3)

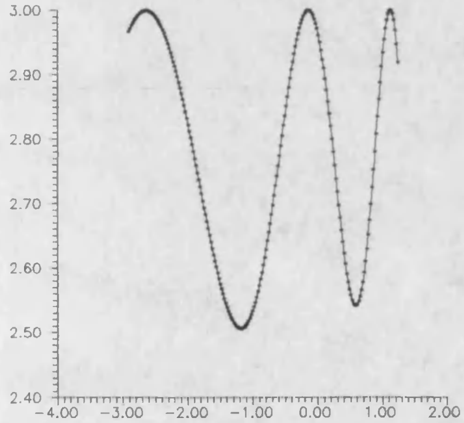


Figure 7.4.3. Derivative Function for design (7.4.4)

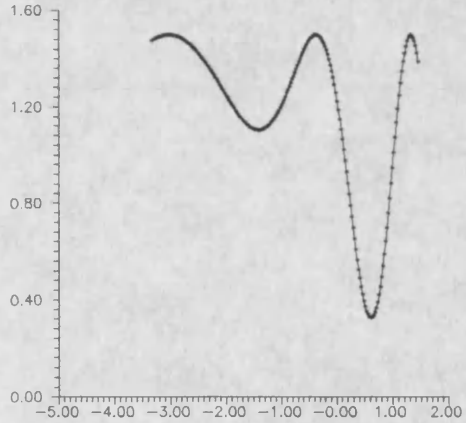


Figure 7.4.4. Derivative Function for design (7.4.5)

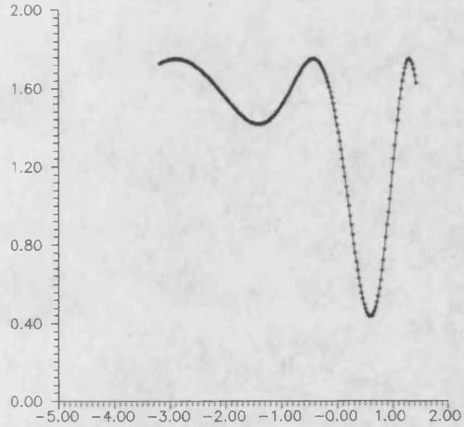


Figure 7.4.5. Derivative Function for design (7.4.6)

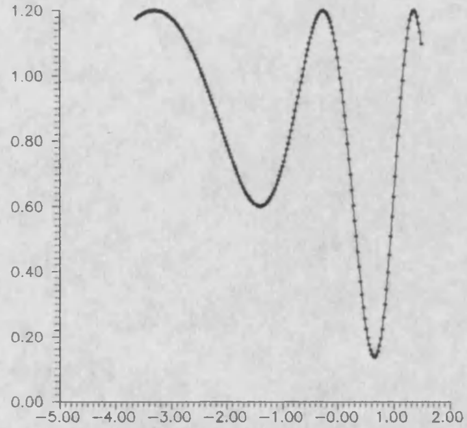


Figure 7.4.6. Derivative Function for design (7.4.7)

optimum design depends strongly on the dimension of the problem, or the number of unknown parameters, rather than on the number of parameters to be estimated in the criterion adopted. Figures 7.4.1 to 7.4.6 show the derivative functions corresponding to the designs of Table 7.1. Note the upper limit varies from one to three, depending not only upon the number of parameters to be estimated but also upon the importance assigned to each purpose.

We now present two further examples to illustrate the use of Bayesian optimum designs. With the purpose of comparing the results the same prior is considered in Example 7.6. Example 7.5 focuses on the specific problem of estimating λ whereas all four criteria are regarded in Example 7.6.

EXAMPLE 7.5. Suppose the interest of the experiment lies in estimating λ and that prior information about its value is available. Two cases are considered. In the first, information about λ is relatively accurate, whereas in the second it is rather dispersed. Moreover, we suppose, in the first case, that the distribution of $\beta|\Lambda=\lambda$ is slightly inaccurate whilst independent of λ but, in the second, it is precise although dependent on λ. Prior distributions are shown below in Table 7.2. In both cases we assume that the linear predictor structure is simply $\beta_0 + \beta_1 x$.

TABLE 7.2 – PRIOR DISTRIBUTIONS FOR λ AND $\beta|\lambda$ (TWO CASES)

CASE	λ	Prob(λ)	$\beta\|\lambda$	Prob{$\beta\|\lambda$}
Accurate Information	0.001	0.5	(0.5,1.0)	0.5
			(0.5,1.5)	0.5
	0.002	0.5	(0.5,1.0)	0.5
			(0.5,1.5)	0.5
Dispersed Information	0.001	0.5	(0.5,1.0)	1.0
	1.0	0.5	(0.0,2.0)	1.0

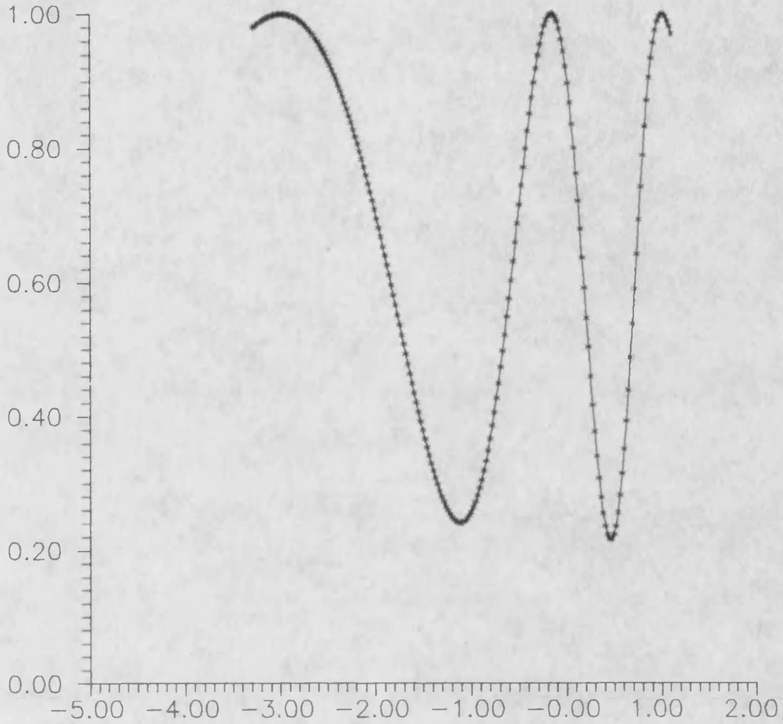


Figure 7.5.1. Derivative Function for design (7.4.8)

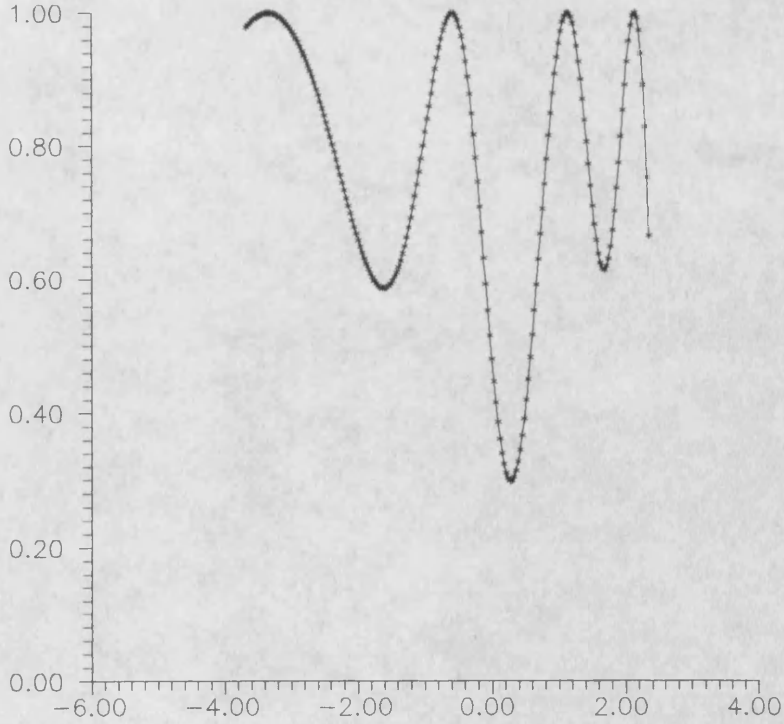


Figure 7.5.2. Derivative Function for design (7.4.9)

The resulting Bayesian optimum design for the first case has only three support points as opposed to four in the second. This difference is not surprising as the number of support points in the Bayesian optimum design is likely to increase as the accuracy of the prior information about λ decreases. Both the designs and the derivative functions are shown in Table 7.3 and Figures 7.5.1 and 7.5.2, respectively.

TABLE 7.3 – OPTIMUM BAYESIAN DESIGNS (EXAMPLE 7.5)

CASE	OPTIMUM VALUE OF CRITERION	BAYESIAN OPTIMUM DESIGN
(a) Accurate Information	–2.784	$\begin{Bmatrix} -2.985 & -0.1544 & 1.01 \\ 0.2781 & 0.2595 & 0.4624 \end{Bmatrix}$
(b) Dispersed Information	–4.153	$\begin{Bmatrix} -3.344 & -0.5939 & 1.127 & 2.137 \\ 0.2418 & 0.4247 & 0.2902 & 0.0433 \end{Bmatrix}$

where (a) design (7.4.8) ; and (b) design (7.4.9)

A noticeable distinction between designs (7.4.8) and (7.4.9) concerns the widenning of the range containing the support points when the information about λ becomes more dispersed. The distributions of weights for these designs are also rather dissimilar. Another distinction is that the largest support point of design (7.4.9), $x = 2.137$, is far away from its equivalent in design (7.4.8), $x = 1.01$. Although with a relatively small importance (4.33% of weight), this point may be necessary in the optimum design for discriminating between the two values of λ.

EXAMPLE 7.6. Another interesting situation arises when all four criteria are used with the same prior information on λ and $\beta|\lambda$. Here, we consider prior distributions that are quite concentrated around specific values for both λ and $\beta|\Lambda=\lambda$. Table 7.4 shows the priors. Again, the linear predictor is assumed to be given by $\eta = \beta_0 + \beta_1 x$. The Bayesian optimum designs are displayed in Table 7.5.

TABLE 7.4 – PRIOR DISTRIBUTIONS FOR λ AND $(\beta_0,\beta_1)|\Lambda=\lambda$

Lambda	Prob(λ)	$(\beta_0,\beta_1)|\lambda$	Prob$\{(\beta_0,\beta_1)|\lambda\}$
0.001	0.25	(0.0,1.0)	0.25
		(0.0,1.5)	0.25
		(0.5,1.0)	0.25
		(0.5,1.5)	0.25
0.002	0.25	(0.25,1.5)	0.2
		(0.25,1.0)	0.3
		(0.50,1.0)	0.2
		(0.50,1.5)	0.3
0.003	0.25	(0.8,1.5)	0.3
		(0.4,2.0)	0.2
		(1.0,2.0)	0.2
		(0.5,1.5)	0.3
0.004	0.25	(1.0,1.00)	0.4
		(0.8,0.90)	0.2
		(1.0,1.50)	0.2
		(0.7,0.95)	0.2

All values of λ in the prior of Table 7.4 yield models which are similar to that related to the complementary log–log link ($\lambda\to0$). However, the prior information about $\beta|\lambda$ is variable so that the curves of expected responses under each of the above sixteen models do not agree closely. Such a dispersion on the prior distributions of the underlying parameters may require a large number of support points in the resulting Bayesian optimum designs.

Regardless of the specific purpose all Bayesian optimum designs in Table 7.5 have five support points. As expected designs (7.4.11) and (7.4.15), corresponding to estimating β with weights respectively 1.0 and 0.9, are very similar. The same applies for designs (7.4.10) and (7.4.14), corresponding to estimating λ with weights respectively 1.0 and 0.8, although the similarities are less marked. Following a similar reasoning, one should expect that designs (7.4.12) and

(7.4.13) were more alike than they actually are since they result from criteria which supposedly do not give priority to any single purpose of estimation. But, in fact, while criterion function (7.3.7), for $\alpha=0.5$, explicitly assigns equal weights for estimating λ as well as β, criterion function (7.3.6) fails to specify any numerically explicit priority. In fact, it is difficult to interpret the manner which the latter operates to allocate such priorities.

TABLE 7.5 – BAYESIAN OPTIMUM DESIGNS (EXAMPLE 7.6)

	UPPER LIMIT	OPTIMUM VALUE OF CRITERION	BAYESIAN OPTIMUM DESIGN
(a)	$\psi_1(x) \leq 1$	–3.136	$\left\{\begin{matrix} -2.842 & -0.2885 & 0.372 & 1.24 & 0.8549 \\ 0.2678 & 0.2045 & 0.0925 & 0.1788 & 0.2564 \end{matrix}\right\}$
(b)	$\psi_2(x) \leq 2$	–29.39	$\left\{\begin{matrix} -2.323 & -0.5246 & 0.3792 & 0.8493 & 1.156 \\ 0.4861 & 0.3392 & 0.0477 & 0.0771 & 0.0499 \end{matrix}\right\}$
(c)	$\psi(x) \leq 3$	–30.65	$\left\{\begin{matrix} -2.149 & -0.2971 & 0.3721 & 0.8277 & 1.058 \\ 0.3224 & 0.2836 & 0.1438 & 0.2102 & 0.04 \end{matrix}\right\}$
(d)	$\psi_m(x) \leq 1.5$	–16.4	$\left\{\begin{matrix} -2.463 & -0.4535 & 0.3712 & 0.8419 & 1.187 \\ 0.3949 & 0.3055 & 0.0667 & 0.0888 & 0.1441 \end{matrix}\right\}$
(e)	$\psi_m(x) \leq 1.2$	–8.487	$\left\{\begin{matrix} -2.641 & -0.3762 & 0.3694 & 0.8466 & 1.214 \\ 0.326 & 0.2572 & 0.0809 & 0.2042 & 0.1317 \end{matrix}\right\}$
(f)	$\psi_m(x) \leq 1.9$	–26.8	$\left\{\begin{matrix} -2.34 & -0.5129 & 0.3776 & 0.8462 & 1.161 \\ 0.4695 & 0.3355 & 0.0512 & 0.088 & 0.0558 \end{matrix}\right\}$

where

(a) Estimating λ – design (7.4.10); (b) Estimating β – design (7.4.11)

(c) Estimating λ and β – design (7.4.12); (d) Mixture (α=0.5) – design (7.4.13)

(e) Mixture (α=0.8) – design (7.4.14); and (f) Mixture (α=0.1) – design (7.4.15).

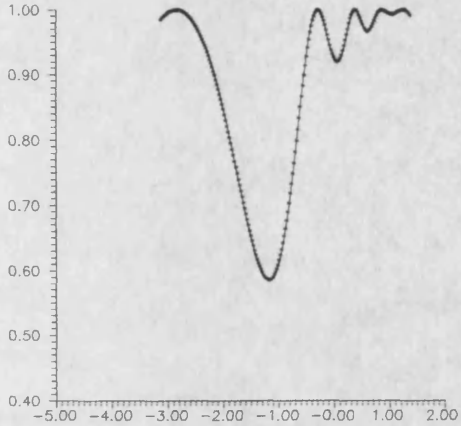


Figure 7.6.1. Derivative Function for design (7.4.10)

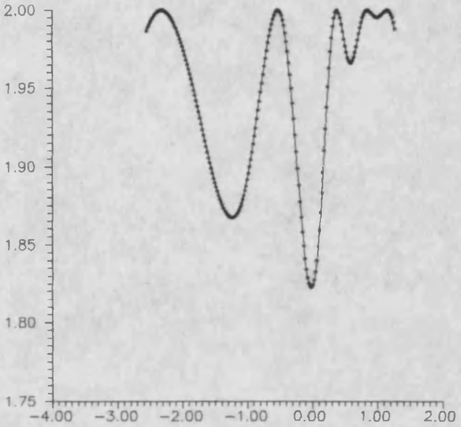


Figure 7.6.2. Derivative Function for design (7.4.11)

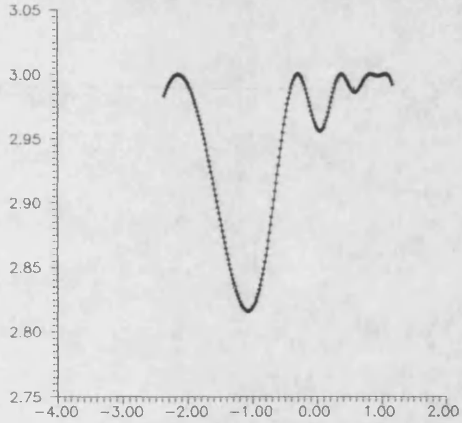


Figure 7.6.3. Derivative Function for design (7.4.12)

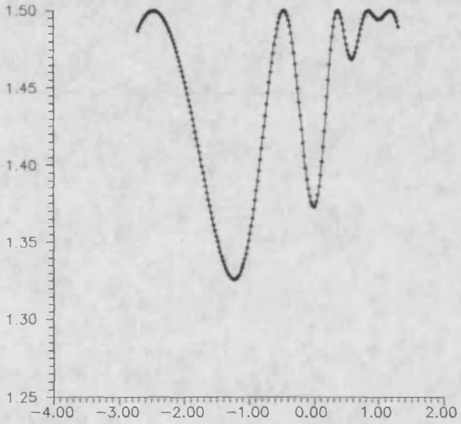


Figure 7.6.4. Derivative Function for design (7.4.13)

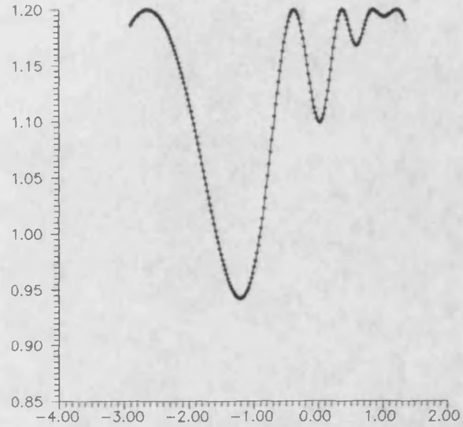


Figure 7.6.5. Derivative Function for design (7.4.14)

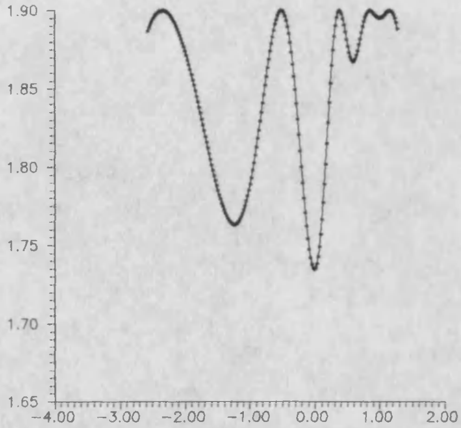


Figure 7.6.6. Derivative Function for design (7.4.15)

Another marked feature of designs (7.4.11) and (7.4.15), both resulting from criteria stressing the estimation of β, is that the first two support points in numerical order have, respectively, 82.53% and 80.5% of the total weight making this region crucial for the related purpose. On the other hand, designs (7.4.10) and (7.4.14), resulting from stressing the estimation of λ, hold weights which are reasonably less concentrated on single support points. Further, in both the lattter designs, the last support point in numerical order is given substantial importance, especially in design (7.4.10). This seems to suggest that points in this subregion of the design region $\mathcal{X}$ contain more information about λ since only a small weight is allocated to such points when the purpose is to estimate β.

In order to check optimality of the designs in Table 7.5 all the related derivative functions are plotted in Figures 7.6.1 to 7.6.6. The shapes of all plots are similar, especially in the regions containing the last three support points. Nevertheless, there is some variation on how informative is the region between the second and third support points compared to other similar regions in $\mathcal{X}$. The other common factor among the derivative functions corresponding to the Bayesian optimum designs in Table 7.5 is the marked valley between the first and second support points. This suggests that points in this region should be avoided as they provide little information about the related parameters.

7.5. DISCUSSION

Local and Bayesian optimum designs can be obtained for each of the four problems considered in this chapter. Optimum designs to estimate λ, the link function parameter, are particularly useful, for other models than the logistic, probit and complementary log–log may also yield good fits. Comparison with designs to discriminate between models can be made for the case of designing to estimate λ. Simulation methods might be used to evaluate how informative the local and

Bayesian optimum designs are compared to other designs.

In practice, a good strategy to obtain sensible experimental designs is to utilize the mixture criterion function (7.3.7) with a number of different values of α. By doing so, we can extract some important features related to the particular situation such as which regions seem to contain more information about the link function parameter and/or the linear predictor parameters. Another possibility is to find out how the optimum designs, resulting from the mixture criterion, perform as opposed to optimum designs for specific purposes. As a measure of efficiency we suggest to calculate the ratio between the values for the specific criterion function at the mixture criterion optimum design and the specific optimum design.

This also suggests adding another criterion of optimality to those of Subsection 7.3.2. Instead of maximizing the mixture criterion function (7.3.7) for a fixed value of α, a further step of optimization could be introduced. For example, we could search over values of α in the interval [0,1] to find the best compromise optimum design, that is a design which maximizes a certain function of the efficiencies of the mixture criterion optimum design relative to both specific optimum designs, for estimating λ and β. This function could be defined, for example, as the product of these efficiencies.

Finally, all the ideas developed in this chapter can also be implemented for other families of link functions not only for binary data models but also for the other classes of generalized linear models. For instance, similar methods to those developed here can be used to estimate the parameter related to the Box–Cox power transformation in regression models.

APPENDIX E

E.1. D_A, D_s–OPTIMALITY AND THE MIXTURE CRITERION

We follow the definition of D_s–optimality given in Silvey (1980, pp.45). First, D_A–optimality must be introduced.

Definition E.1.1. To simplify the notation denote the local normalized Fisher information matrix $M(\lambda,\beta,\xi)$ just by M. Suppose that the main interest of the experiment is to estimate s linear combinations of the k unknown parameters, $s < k$. Let A^t be an s×k matrix of rank s, whose rows contain the coefficients of the s linear combinations of interest. Then, with the same motivation as for D–optimality, the criterion of D_A–optimality is defined as the maximization over the set of design measures $\mathscr{H}$ of the following function.

$$\Psi_A(M) = \begin{cases} -\log\ |A^t M^{-1} A| & \text{, if M is nonsingular} \\ -\infty & \text{, otherwise} \end{cases}$$

For simplicity, only the case of M nonsingular is regarded here. Now, the criterion of D_s–optimality can be introduced as a particular case of D_A–optimality. Suppose that the aim of the experiment is to estimate the first s parameters of the whole vector of unknown parameters. This situation can be considered by taking the matrix of coefficients $A^t = (I_s\ 0)$.

Then, a simplified expression for the D_s–optimality criterion function can be derived. Let the normalized Fisher information matrix and its inverse be partitioned as follows

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}; \qquad M^{-1}(\xi) = \begin{bmatrix} M^{11} & M^{12} \\ M^{21} & M^{22} \end{bmatrix}$$

where M_{11} and M^{11} have orer s×s, and M_{22} and M^{22} (k–s)×(k–s).

Then, it is straightforward to show that $(A^t M^{-1} A) = M^{11}$. Further, it is a well known result that M^{11} can be expressed in terms of the partition of M as $(M_{11} - M_{12} M_{22}^{-1} M_{21})^{-1}$. Hence, the D_s–optimality criterion function is as follows.

$$\Psi_s(M) = \begin{cases} \log |M_{11} - M_{12} M_{22}^{-1} M_{21}| & \text{, if M is nonsingular} \\ -\infty & \text{, otherwise} \end{cases}$$

To derive an equivalent and sometimes more useful way of expressing the D_s–optimality criterion function, the following Lemmas are required.

<u>LEMMA E.1.2.</u> – Let A and B be square matrices such that $|A| \neq 0$ and $|B| \neq 0$. Let C be a matrix whose dimensions are compatible with those of A and B according to the matrix operations shown below. Similarly, let 0 denote the null matrix. Then the following results hold.

(i) $\begin{vmatrix} A & 0 \\ 0^t & B \end{vmatrix} = |A|\ |B|$;

(ii) $\begin{vmatrix} A & C \\ 0^t & B \end{vmatrix} = |A|\ |B|$; and

(iii) $\begin{vmatrix} A & 0 \\ C^t & B \end{vmatrix} = |A|\ |B|$.

PROOF

(i) $\begin{vmatrix} A & 0 \\ 0^t & B \end{vmatrix} = \left| \begin{bmatrix} A & 0 \\ 0^t & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0^t & B \end{bmatrix} \right| = \begin{vmatrix} A & 0 \\ 0^t & I \end{vmatrix} \begin{vmatrix} I & 0 \\ 0^t & B \end{vmatrix} = |A|\ |B|$.

(ii) $\begin{vmatrix} A & C \\ 0^t & B \end{vmatrix} = \left| \begin{bmatrix} A & 0 \\ 0^t & B \end{bmatrix} \begin{bmatrix} I & A^{-1}C \\ 0^t & I \end{bmatrix} \right| = \begin{vmatrix} A & 0 \\ 0^t & B \end{vmatrix} \begin{vmatrix} I & A^{-1}C \\ 0^t & I \end{vmatrix} = |A|\ |B|$.

(iii) $\begin{vmatrix} A & 0 \\ C^t & B \end{vmatrix} = \left| \begin{bmatrix} A & 0 \\ 0^t & B \end{bmatrix} \begin{bmatrix} I & 0 \\ B^{-1}C^t & I \end{bmatrix} \right| = \begin{vmatrix} A & 0 \\ 0^t & B \end{vmatrix} \begin{vmatrix} I & 0 \\ B^{-1}C^t & I \end{vmatrix} = |A| \; |B|.$

<u>LEMMA E.1.3.</u> – Let A be a square matrix such that $|A| \neq 0$. Let A be partitioned as follows

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

Then the following results hold

(i) $|A| = |A_{22}| \; |A_{11} - A_{12} A_{22}^{-1} A_{21}|$

(ii) $|A| = |A_{11}| \; |A_{22} - A_{21} A_{11}^{-1} A_{12}|$

PROOF

(i) First premultiply and postmultiply A as follows

$$\begin{bmatrix} I & -A_{12} A_{22}^{-1} \\ 0^t & I \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ -A_{22} A_{21} & I \end{bmatrix} = \begin{bmatrix} A_{11} - A_{12} A_{22}^{-1} A_{21} & 0 \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ -A_{22}^{-1} A_{21} & I \end{bmatrix}$$

$$= \begin{bmatrix} A_{11} - A_{12} A_{22}^{-1} A_{21} & 0 \\ 0^t & A_{22} \end{bmatrix}.$$

Now, taking determinants on both sides of the equality and using the results of Lemma E.1.2. we obtain result (i). For proving (ii), we similarly premultiply and postmultiply A as follows

$$\begin{bmatrix} I & 0 \\ -A_{21} A_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} I & -A_{11}^{-1} A_{12} \\ 0^t & I \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ 0^t & A_{22} - A_{21} A_{11}^{-1} A_{12} \end{bmatrix} \begin{bmatrix} I & -A_{11}^{-1} A_{12} \\ 0^t & I \end{bmatrix}$$

$$= \begin{bmatrix} A_{11} & 0 \\ 0^t & A_{22} - A_{21} A_{11}^{-1} A_{12} \end{bmatrix}.$$

Again, taking determinants on both sides of the above equality and using the results of Lemma E.1.2. we obtain result (ii). Thus, results (i) and (ii) of Lemma E.1.3. prove the equivalence of the expressions used for criterion functions (7.3.4) and (7.3.5).

Let us now show how the expression presented for the mixture criterion function (7.3.7) is derived. Recall that criterion function (7.3.7) is defined as a linear combination of criterion functions (7.3.4) and (7.3.5), that is for $0 \leq \alpha \leq 1$

$$\Psi_m(M) = \alpha\, \Psi_1(M) + (1-\alpha)\, \Psi_p(M)$$

Then

$$\Psi_m(M) = \alpha \log\left\{|M| \,/\, |M_{22}|\right\} + (1-\alpha) \log\left\{|M| \,/\, |M_{11}|\right\}$$

$$= \log\left\{\frac{|M|^{\alpha} \quad |M|^{1-\alpha}}{|M_{22}|^{\alpha} \quad |M_{11}|^{1-\alpha}}\right\} = \log\left\{\frac{|M|}{|M_{22}|^{\alpha} \quad |M_{11}|^{1-\alpha}}\right\}.$$

E.2. THE DERIVATIVE FUNCTIONS

To find the Fréchet derivatives related to criterion functions (7.3.4) to (7.3.7) we refer to the derivation of the Fréchet derivative for the D–optimality criterion function. Silvey (1980, Chapter 3, pp. 21), for example, derived such results. Let the D–optimality criterion function be defined as

$$\Psi(M) = \begin{cases} -\log\ |M| & \text{, if } M \text{ is nonsingular} \\ -\infty & \text{, otherwise} \end{cases}$$

Then, the Fréchet derivative of $\Psi(.)$ at $M(\lambda,\beta,\xi_1)$, say M_1 in the direction of $M(\lambda,\beta,\xi_2)$, say M_2, denoted by $F_\Psi(M_1,M_2)$, is

$$F_\Psi(M_1,M_2) = \text{tr}\left[M_2\, M_1^{-1}\right] - k \qquad \text{(E.2.1)}$$

where k is the dimension of the matrix M. In our context, this corresponds to the Fréchet derivative of criterion function (7.3.6).

Result (E.2.1) is also used to derive the Fréchet derivative of criterion functions (7.3.4) and (7.3.5). The main argument is based on the fact that the D_s–optimality criterion function can be written as

$$\Psi_s(M) = \log \left\{ \frac{|M|}{|M_{22}|} \right\} = \log |M| - \log |M_{22}|$$

Therefore, by first applying the property of additivity for Fréchet derivatives and then result (E.2.1) for both terms of the above expression, we find that the Fréchet derivative of $\Psi_s(.)$ at M_1 in the direction of M_2, $F_{\Psi_s}(M_1,M_2)$, is

$$F_{\Psi_s}(M_1,M_2) = \operatorname{tr}\left[M_2\, M_1^{-1}\right] - \operatorname{tr}\left[M_{2,22}\, M_{1,22}^{-1}\right] - s \qquad \text{(E.2.2)}$$

where $M_{2,22}$ and $M_{1,22}^{-1}$ are the lower right submatrices for the corresponding partitions of matrices M_2 and M_1, respectively. Now, applying result (E.2.2) to both criterion functions (7.3.4) and (7.3.5) we obtain their Fréchet derivatives, respectively given by

$$F_{\Psi_1}(M_1,M_2) = \operatorname{tr}\left[M_2\, M_1^{-1}\right] - \operatorname{tr}\left[M_{2,22}\, M_{1,22}^{-1}\right] - 1 \qquad \text{(E.2.3)}$$

$$F_{\Psi_p}(M_1,M_2) = \operatorname{tr}\left[M_2\, M_1^{-1}\right] - \operatorname{tr}\left[M_{2,11}\, M_{1,11}^{-1}\right] - p \qquad \text{(E.2.4)}$$

To determine the Fréchet derivative of criterion function (7.3.7), it is necessary to apply the property of additivity for Fréchet derivatives, and thereafter substitute (E.2.3) and (E.2.4) into the resulting expression, as shown below.

$$F_{\Psi_m}(M_1,M_2) = \alpha\, F_{\Psi_1}(M_1,M_2) + (1-\alpha)\, F_{\Psi_p}(M_1,M_2)$$

$$= \alpha \left\{ \mathrm{tr}\left[M_2\, M_1^{-1}\right] - \mathrm{tr}\left[M_{2,22}\, M_{1,22}^{-1}\right] - 1 \right\} + (1-\alpha) \left\{ \mathrm{tr}\left[M_2\, M_1^{-1}\right] - \mathrm{tr}\left[M_{2,11}\, M_{1,11}^{-1}\right] - p \right\}$$

$$= \mathrm{tr}\left[M_2\, M_1^{-1}\right] - \alpha\, \mathrm{tr}\left[M_{2,22}\, M_{1,22}^{-1}\right] - (1-\alpha)\, \mathrm{tr}\left[M_{2,11}\, M_{1,11}^{-1}\right] - \alpha - (1-\alpha)p \tag{E.2.5}$$

Finally, replacing ξ_1 by ξ^*, an optimum design with respect to one of the criteria defined in Subection 7.3.2 ; ξ_2 by ξ_x, the design measure putting mass one at the point $x \in \mathcal{X}$; and applying Theorem 2.2 respectively to results (E.2.1), (E.2.3), (E.2.4), and (E.2.5), we obtain the derivative functions (7.3.9) to (7.3.12).

CHAPTER 8. CONCLUSIONS

In this chapter general and specific conclusions are drawn based upon the results obtained in the thesis. A number of comments apply to all chapters such as the consequences of the lack of theoretical results concerning the number of support points in the optimum design, the numerical complications caused by linear predictors with several explanatory variables, etc. Other comments are more specific such as the relevance of considering link function families for binary data models.

Undoubtedly, the major contribution of this thesis is to show that some of the main results of optimum design theory can also be applied to situations in which prior information is incorporated into existing optimality criteria such as D–optimality, D_s–optimality, and T–optimality. This extension could play an important role in practical situations where the classical optimum designs depend upon the true values of the parameters. The advantages are particularly relevant when the prior knowledge of the experimenter reflects the behaviour of the underlying parameters of interest. Moreover, the techniques described in this thesis can also be used as tools for explanatory analysis since they require no observations.

Another contribution of this thesis concerns the extensions of the T–optimality criterion for discriminating between two deterministic structures to other classes of statistical models, such as binary data models, Poisson models, etc.

A combination of sequential and Bayesian techniques seems to be the best approach for both problems considered. The advantages of this methodology are obvious. Firstly, a starting optimum design could be obtained as described in the previous chapters. The next stage would involve sampling from the population of interest. The posterior distribution for the parameters, based on the observations from the sampling stage, could then be used as an updated prior distribution for the next stage. This process could continue until a reasonable stability of the optimum design is reached.

Another possibility for obtaining efficient designs is to adopt a pure sequential approach. For example, in the context of model discrimination for linear regression, Atkinson and Fedorov (1975a) proposed a procedure in which at each step one further design point is included in the design to give the highest possible efficiency. The criterion for inclusion of the new design point is to choose the point in the design region maximizing the square of the difference between the predicted values under the two models. This method can be extended to the problem of model discrimination for generalized linear models. More sophisticated techniques based upon multiple inclusion, or exchange, of points may accelerate the process of obtaining more efficient designs.

In both cases of sequential designing mentioned above, standard simulation techniques could be used to evaluate convergence, speed, and other related features of the procedures. As an illustration of the kind of comparisons that can be made, suppose that the true values of the set(s) of the parameters are known (of course, this is only required for the simulation process). Thus, a locally optimum design can be determined as illustrated in this thesis. Now, given an initial prior distribution for the set(s) of parameters, a Bayesian optimum design can be obtained and its efficiency, compared to that of the locally optimum design. The next steps follow those described above. At each stage of the sequential procedure, the whole sequence of relative efficiencies can be analysed with the purpose of making the decision to stop. This can be implemented for either general problem considered in the thesis.

Several examples are used as illustrations throughout the thesis. One of the main interests in these examples concerns the number of support points in the optimum design. The lack of theoretical results about this, as opposed to the existence of an upper limit for the case of parameter estimation in regression models, leads to considerable difficulty in the determination of optimum designs (particularly, Bayesian optimum designs). Sensitivity analyses carried out in different chapters appear to suggest that both the precision of the prior distributions and the region of the parameter space in which these priors are defined may be influential. This is certainly an important point for future investigation, both analytical and numerical.

With the exception of the examples of Chapter 6, where uniform prior distributions are considered, all the other priors in the illustrations are discrete probability distributions. The case of continuous priors requires the use of numerical integration techniques. According to our experience in the examples of Chapter 6, this slowed down the computations significantly, although it does not seem to be extremely computer time–consuming. Further computations for the case of continuous priors may be carried out, especially for the problem of model discrimination, to evaluate the numerical complexity involved compared to the case of discrete distributions.

As mentioned above, all the computations were carried out on a PC 486 by means of Fortran 77 programs combined with NAG routines, for the optimization steps. Not one optimum design, or candidate for optimality, took more than a minute to obtain. In fact, most of them were determined in less than 20 seconds. As expected, Bayesian optimum designs often took longer than locally optimum designs. However, these rather fast results can be explained by the fact that the design regions regarded in all examples were either the real line or an interval of the real line. For problems of higher dimension, the convergence times will certainly be greater.

Further, more specific optimization techniques could be considered as the NAG routines might not be suitable for higher dimension problems. This is because NAG routine E04JAF, used in all optimization programs, had to be adapted for our purposes. In this adaptation, the design support points together with the respective weights are regarded as elements of a single vector. Because of the restriction that the weights sum to one, the dimension of this vector is equal to twice the number of support points minus one. Some examples in which the search was performed with seven support points, hence thirteen dimensions, are among those that took longest to converge. Therefore, further adaptations of this or other NAG routines, implementing the searches described in this thesis for problems with more than one explanatory variable, would cause the convergence times increase drastically. To be more specific, let p be the number of explanatory variables in the model and let n be the number of design points in a particular search. Then, using an adaptation of the NAG routine EO4JAF, the single vector storing the design points and weights would have dimension equal to $(n\times p)+(n-1)$.

REFERENCES

Abdelbasit, K.M. and Plackett, R.L. (1983). Experimental Design for Binary Data. J. Am. Statist. Assoc. 78, 381, 90–98.

Andrews, D.F. (1971). Sequentially designed experiments for screening out bad models with F tests. Biometrika 58, 3, 427–432.

Aranda–Ordaz, F.J. (1981). On two families of transformations to additivity for binary response data. Biometrika 68, 2, 357–363.

Atkinson, A.C. (1972). Planning experiments to detect inadequate regression models. Biometrika 59, 2, 275–293.

Atkinson, A.C. (1981). A Comparison of Two Criteria for the Design of Experiments for Discriminating Between Models. Technometrics 23, 3, 301–305.

Atkinson, A.C. (1982). Developments in the Design of Experiments. International Statistical Review 50, 161–177.

Atkinson, A.C. (1988). Recent Developments in the Methods of Optimum and Related Experimental Designs. International Statistical Review 56, 2, 99–115.

Atkinson, A.C. (1992). Optimum Experimental Designs for Parameter Estimation and for Discrimination between Models in the Presence of Prior Information. Model Oriented Data–Analysis, 3–30. Physica–Verlag.

Atkinson, A.C. and Cox, D.R. (1974). Planning experiments for discriminating between models. J. R. Statist. Soc. B 36, 321–348.

Atkinson, A.C. and Donev, A.N. (1992). Optimum Experimental Designs. Oxford Statistical Science Series. Clarendon Press, Oxford.

Atkinson, A.C. and Fedorov, V.V. (1975a). The design of experiments for discriminating between two rival models. Biometrika 62, 57–70.

Atkinson, A.C. and Fedorov, V.V. (1975b). Optimal design: Experiments for discriminating between several models. Biometrika 62, 2, 289–303.

Bliss, C.I. (1935). The calculation of the dosage–mortality curve. Annals of Applied Biology 22, 134–167.

Borth, D.M. (1975). A Total Entropy Criterion for the Dual Problem of Model Discrimination and Parameter Estimation. J. R. Statist. Soc. B 37, 77–87.

Box, G.E.P. and Hill, W.J. (1967). Discrimination Among Mechanistic Models. Technometrics 9, 1, 57–71.

Box, G.E.P. and Lucas, H.L. (1959). Design of experiments in nonlinear situations. Biometrika 46, 77–90.

Chaloner, K. (1987). An approach to experimental design for generalized linear models. Lecture Notes in Economics and Mathematical Systems 297, 3–12. Model–Oriented Data Analysis, Proceedings, Eisenach, GDR. Springer–Verlag.

Chaloner, K. and Larntz, K. (1988). Software for logistic regression experiment design. Optimal Design and Analysis of Experiments (Ed. Dodge, Fedorov and Wynn), 207–211. North Holland.

Chaloner, K. and Larntz, K. (1989). Optimal Bayesian design applied to logistic regression experiments. J. Statist. Planning Inf. 21, 191–208.

Chambers, E.A. and Cox, D.R. (1967). Discrimination between alternative binary response models. Biometrika 54, 3, 573–578.

Chernoff, H. (1953). Locally optimal designs for estimating parameters. The Annals of Mathematical Statistics 24, 586–602.

Chernoff, H. (1959). Sequential design of experiments. Ann. Math. Statist. 30, 755–770.

Cox, D.R. (1966b). Some procedures connected with the logistic qualitative response curve. Research Papers in Statistics: Essays in Honour of J. Neyman's

70th Birthday (ed. F.N. David), Wiley, London, 55–71.

Cox, D.R. and Reid, N. (1987). Parameter Othogonality and Approximate Conditional Inference. J. Roy. Statist. Soc. B 49, 1–39.

Crump, K.S. (1979). Dose Response Problems in Carcinogenesis. Biometrics 35, 157–167.

Czado, C. (1992). On Link Selection in Generalized Linear Models. Advances in GLIM and Statistical Modelling, Lecture Notes in Statistics 78, 60–65. Springer–Verlag.

Fedorov, V.V. (1972). Theory of Optimal Experiments. Academic Press, New York.

Fedorov, V.V. (1978). Some extremal problems in designing discriminating experiments. Commun. Statist. Theor. Meth. A7, 14, 1339–1345.

Fedorov, V.V. and Atkinson, A.C. (1988). The optimum design of experiments in the presence of uncontrolled variability and prior information. Optimal Design and Analysis of Experiments (Ed. Dodge, Fedorov and Wynn), 327–344. North Holland.

Fedorov, V.V. and Khabarov, V. (1986). Duality of optimal designs for model discrimination and parameter estimation. Biometrika 73, 1, 183–190.

Finney, D.J. (1971). Probit Analysis (3rd Edition). Cambridge University Press, Cambridge.

Hill, P.D. (1978a). A Review of Experimental Design Procedures for Regression Model Discrimination. Technometrics 20, 1, 15–21.

Hill, P.D. (1978b). A note on the equivalence of D–optimal design measures for three rival linear models. Biometrika 65, 3, 666–667.

Hill, W.J. and Hunter, W.G. (1969). A Note on Designs for Model Discrimination: Variance Unknown Case. Technometrics 11, 2, 396–400.

Hill, W.J., Hunter, W.G. and Wichern, D.W. (1968). A Joint Design Criterion for the Dual Problem of Model Discrimination and Parameter Estimation. Technometrics 10, 1, 145–160.

Huang, C.Y. (1991). Planification d'expériences pour la discrimination entre structures de modeles. These en Sciences. Université de Paris–Sud Centre D'Orsay.

Hunter, W.G. and Reiner, A.M. (1965). Designs for Discriminating Between Two Rival Models. Technometrics 7, 3, 307–323.

Jones, E.R. and Mitchell, T.J. (1978). Design criteria for detecting model inadequacy. Biometrika 65, 3, 541–551.

Khan, M.K. (1988). Optimal Bayesian estimation of the median effective dose. J. Statist. Planning Inf. 18, 69–81.

Khan, M.K. and Yazdi, A.A. (1988). On D–optimal designs for binary data. J. Statist. Planning Inf. 18, 83–91.

Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). Ann. Statist. 2, 849–879.

Kiefer, J. and Wolfowitz, J. (1960). The Equivalence of two extremum problems. Can. J. Math. 12, 363–366.

McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models, 2nd Edition. Chapman and Hall, London New York.

Meeter, D., Pirie, W. and Blot, W. (1970). A Comparison of Two Model–Discrimination Criteria. Technometrics 12, 3, 457–470.

Milicer, H. and Szczotka, F. (1966). Age at menarche in Warsaw girls in 1965. Human Biology 38, 199–203.

Minkin, S. (1987). Optimal Designs for Binary Data. J. Am. Statist. Assoc. 82, 1098–1103.

Morgan, B.J.T. (1992). Analysis of Quantal Response Data. Chapman and Hall.

Pereira, B.B. (1977). Discriminating among Separate Models: a Bibliography. International Statistical Review 45 163–172.

Ponce de Leon, A.C. and Atkinson, A.C. (1991). Optimum experimental design for discriminating between two rival models in the presence of prior information. Biometrika 78, 3, 601–608.

Ponce de Leon, A.C. and Atkinson, A.C. (1992a). Optimal design for discriminating between two rival binary data models in the presence of prior information. PROBASTAT '91, 123–129. Proceedings of the 10th Conference on Probability and Mathematical Statistics, Bratislava, Czechoslovakia.

Ponce de Leon, A.C. and Atkinson, A.C. (1992b). The Design of Experiments to Discriminate Between Two Rival Generalized Linear Models. Advances in GLIM and Statistical Modelling, Lecture Notes in Statistics 78, 159–164. Springer–Verlag.

Pregibon, D. (1980). Goodness of Link Tests for Generalized Linear Models. Applied Statistics 29, 15–24.

Prentice, R.L. (1976). A Generalization of the Probit and Logit Methods for Dose Response Curves. Biometrics 32, 761–768.

Pronzato, L., Huang, C.Y. and Walter, E. (1991). Nonsequential T–optimal design for model discrimination: New Algorithms. PROBASTAT '91, 130–136. Proceedings of the 10th Conference on Probability and Mathematical Statistics, Bratislava, Czechoslovakia.

Pshenichnyi, B.N. (1971). Necessary Conditions for an Extremum. Marcel Dekker. New York.

Ramsey, B. and Chesher, A. (1976). Some measures of the "Difference" Between Regression models. J. Am. Statist. Assoc. 71, 356, 972–976.

Rocke, D.M. (1993). On the Beta Transformation Family. Technometrics 35, 1, 72–81.

Tsutakawa, R.K. (1972). Design of Experiment for Bioassay. J. Am. Statist. Assoc. 67, 584–590.

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

Wijesinha, M.C. and Khuri, A.I. (1987). Construction of optimal designs to increase the power of the multiresponse lack of fit test. J. Statist. Planning Inf. 16, 179–192.

Yanagisawa, Y. (1988). Designs for discriminating between binary response models. J. Statist. Planning Inf. 19, 31–41.