Optimal Design of Experiments for the Quadratic Logistic Model

Ellinor Fackle Fornius

Stockholm
University

Doctoral Dissertation
Department of Statistics
Stockholm University
S-106 91 Stockholm
Sweden

Abstract
Optimal design of experiments for binary data is the topic of this thesis. A particular logistic model including a quadratic term in the linear predictor is considered. Determining an optimal design for this model is complicated by the fact that the optimal design is dependent on the unknown true parameters. Methods to obtain locally c- and D-optimal designs are illustrated. c-optimal designs are derived via the canonical design space. This space offers an useful geometric interpretation of the design problem. Using the canonical design space it is shown how the number of design points in a c-optimal design varies depending on the parameter being estimated. Furthermore, formulae for finding the design points along with the corresponding design weights are derived. The small sample performance of the locally optimal designs is compared to the performances of some non-optimal designs in a simulation study. The evaluations are made in terms of mean squared error of the maximum likelihood estimator. The small sample distribution of the maximum likelihood estimator is demonstrated to be quite different from the asymptotic distribution. It was also concluded that non-existence of the maximum likelihood estimator is a critical problem for the quadratic logistic model. The designs differed considerably in this respect and this problem also turned out to be parameter dependent. As a solution to this problem another type of parameter estimator is suggested, which is also evaluated in the simulation study. It performs better in this respect, but not completely satisfactory because it fails in other respects. Two kinds of sequential design approaches are proposed for the purpose of finding the point of optimum response. One is a parametric optimal design approach where c-optimal designs are updated sequentially. The other one is a nonparametric stochastic approximation approach. The suggested designs are evaluated and compared via simulations. Based on the simulation results the c-optimal design approach was consistently favored. Sequential estimation proved to be an effective way to handle the parameter dependence issue.

Key words: Logistic regression, D-optimality, c-optimality, Canonical design space, Maximum likelihood estimation, Sequential design, Stochastic approximation.

ISBN 978-91-7155-644-8
Printed in Sweden by ABA Intercopy, Stockholm 2008
Distributor: Department of Statistics, Stockholm University

TACK

Honestly, I am not sure I understand how it all happened. I am absolutely sure, though, that nothing would have happened if it were not for my supervisor Professor Hans Nyquist. First of all, your exceptional scientific skills and ability to make the most complicated matters seem so easy have been invaluable to me. Thank you for answering all my questions. Moreover, you always knew what was in my best interest long before I did, which I am most grateful for.

I am also sure that these years had not been so enjoyable if it were not for all my friends/colleagues at the department. In particular I want to thank my fellow doctoral students (DB, SB, MC, JF, JK, BL, JL, NP, JSS, KS, PS, TS, CT, MT, BW, LW) for sharing this experience.

Special thanks also to Gösta Hägglund for making me start my Ph.D. studies, to Professors Daniel Thorburn, Gebre Ghilagaber and Jan Wretman for guidance and inspiration, to Karin and Per Dahmström for careful proofreading, to Håkan for always being helpful (with the computers and more), to Bibbi and Elisabet for friendliness and solving practical issues.

Jessica, Daniel, Bertil, Pär and the rest of the VVL-gang, thank you for many exciting lunch trips.

Thank you Daniel/Gunde for numerous discussions on important subjects like running, optimal design theory and programming. I wish I could run as fast as you.

Karin, thank you for always being cheerful, for taking coffee breaks with me and for teaching me dalmål.

I have gained a true friend in my roommate Jessica. You have helped me in so many ways, with statistics and far beyond. Thank you for excellent teamwork when taking courses, shopping and drinking wine.

My dear friends Marie and Johanna, thank you for all the stimulating discussions on the mysteries of life.

My precious family - mormor Gertrud, mamma Eva, lillebror Mikael and lillasyster Lill, thank you for your unconditional love.

Ola, I could not have made it without your constant intellectual and emotional support. Thank you for making me happy every day. I love you!

Stockholm, April 2008 Ellinor Fackle Fornius

Contents

Chapter 1

Introduction

Experimentation constitutes a cornerstone of the empirical sciences. With the help of experiments it is possible to answer questions, to test hypotheses and ultimately to either confirm or refute theories. An experiment is a controlled study in which observations are made and data collected that forms the basis for analysis and subsequent conclusions. The quality of the analysis depends directly on the experimental design. A well-designed experiment allows valid conclusions to be drawn. Statistical methods are of vital importance to achieve this ambition. Available as an alternative is to perform an observational study in which the researcher does not make any interventions but merely observes existing states. The advantage with experiments over observational studies is the opportunity to control the experimental conditions and to determine which variables to include. Observational studies are vulnerable in the sense that the interpretations may be distorted by important variables that are not measured or even unknown.

A common objective is to gain knowledge about a process or system, such as a manufacturing process or a biomedical system, that is affected by one or more controllable variables and possibly a number of uncontrollable variables. A response variable reflects some observable aspect of the output that is of interest, like the yield of a chemical process or whether a product is defect or not. The person conducting the experiment, the experimenter, can vary the levels of the control variable(s) systematically in order to investigate how it influences the response variable(s). Response Surface Methodology (RSM) treats the statistical methods for design and analysis of such experiments. The search for optimum operating conditions in a

manufacturing process is a frequent application.

RSM rests on the notion that there are many equally satisfactory paths that lead to correct conclusions and that the learning process is iterative. It is often wise to split the resources into several stages instead of devoting everything to one large experiment. For instance, it may very well happen that a screening experiment points in a different direction than anticipated beforehand. For example moving from current towards optimum operating conditions is usually accomplished in more than one step.

The theory of optimal design provides an approach that enables the experimental design to be customized to a specific inferential goal. Planning and performing experiments requires resources, optimal experimental design is about getting good value for the time and money invested. By carefully considering certain choices before the experiment is conducted the information obtained can be maximized (given a cost constraint) or the expenses can be minimized (for a desired precision). The choices that have to be made include deciding on which variables to examine, at what levels to make observations and the corresponding proportion of observations as well as how the resources should be divided between different stages of the experiment.

A pioneer in the design of experiments area was Sir R. A. Fisher. He introduced statistical principles to the experimental design in the studies of agricultural systems. The work of Fisher starting in the 1930's laid the foundation to statistical experimental design, see for example Fisher (1935). The applications were mainly in the agricultural, biological and medical fields. In the 1950's the development of RSM caused statistical design of experiments to enter the industrial field of application. RSM originated with the paper by Box and Wilson (1951) on which an extensive growth followed during the next decades. The book by Box and Draper (1987) gives a comprehensive treatment of the RSM techniques.

It was also in the late 1950's that the optimal design theory was initiated where many contributions are attributable to J. Kiefer. For instance the alphabetical terminology referring to the optimality criteria was introduced by Kiefer (1959) and the General Equivalence Theorem is owing to Kiefer and Wolfowitz (1959) and Kiefer (1961). Fedorov (1972), Silvey (1980), Pazman (1986), Atkinson and Donev (1992), Pukelsheim (1993) and Fe-

dorov and Hackl (1997) are all classic books that cover optimal design of experiments.

Research was predominantly concentrated to linear models at first but during the last decades it has been extended to concern non-linear models including Generalized Linear Models, see e.g. Ford et al. (1989). The crux of the problem of finding optimal designs for the latter cases is that the optimal design generally depends on the unknown model parameters. It may seem discouraging that the construction of an optimal design to estimate the parameters requires the very same parameters to be known prior to the experiment. However, there exist several approaches to solve this problem. The most straightforward solution is to base the construction on a best guess, obtained from earlier experiments or based on expert knowledge, which obviously risks being poor if the guess is poor. Such a design is called a locally optimal design and was introduced by Chernoff (1953). A natural development of this approach is the so called optimum on average designs (also known as Bayesian designs) which assume a prior distribution on the parameters instead of focusing on just one guess, see e.g. Fedorov and Hackl (1997). Another alternative is the sequential construction of designs, as in Silvey (1980). The idea is that the parameter estimates and the optimal design are updated stepwise. A design is derived at each stage assuming the parameter estimates obtained at the previous stage to be true. The advantage with a sequential strategy is that modifications are allowed as experimentation proceeds and more information is gained. Sequential designs are particularly effective when the response values can be obtained in a short time.

Much of the research concerning optimal experimental design for GLMs has been devoted to the logistic two-parameter model. A common application are the dose-response models that relates the control variable, e.g. the dose of a drug, to the probability of response, e.g. that a test subject is cured. The interest is often to find the dose associated with some specific response rate, referred to as the effective dose. An overview of optimal designs for the two-parameter logistic model can be found in Myers et al. (1994). Wu (1985) presents various sequential designs for binary data, which for instance are suitable to find the effective dose.

This thesis treats optimal design of experiments for a particular logistic model, the quadratic logistic model. In addition to the usual two parameters

it includes a third parameter related to a quadratic term. The overall ambition is to find optimal designs for estimation of the associated quadratic response curve. The main topics of the thesis are:

- to illustrate methods to obtain locally optimal designs for the quadratic logistic model, including the use of the canonical design space (Chapter 3)
- to explore the performance of optimal as well as some non-optimal designs in small samples (Chapter 4)
- to evaluate the performance of two kinds of sequential designs through a simulation study (Chapter 6)

The outline of the thesis is as follows. Chapter 2 introduces the concept of generalized linear models and reviews the theory of optimal design of experiments. It also treats Response Surface Methodology and gives the definition of the quadratic logistic model. The subsequent chapter includes a detailed discussion about how to find locally optimal designs for this model. Chapter 4 contains a simulation study with the aim to examine the performance of different designs in small samples. Comparisons are made between optimal and non-optimal designs and between two types of estimators. In Chapter 5 two sequential approaches are proposed as a solution to the parameter dependence issue. The performances of the sequential designs and their abilities to find the optimum operating conditions are evaluated through simulations in Chapter 6. Conclusions together with some suggestions for future research areas appear in the last chapter.

Chapter 2

Theory

2.1 Generalized Linear Models

The concept of a generalized linear model (GLM) unifies the analysis of a wide variety of statistical models. The standard linear regression model with the responses being continuous and normally distributed can be viewed as a special case. However, both continuous and discrete data as well as several other probability distributions fit into this framework. Consider for example experiments where the outcome is one out of two possibilities, e.g. success or failure, explode or not explode, resulting in a binary response variable. Logit and probit regression are commonly used models for such situations. Generalized Linear Models were introduced in Nelder and Wedderburn (1972), a comprehensive book-length treatment is provided by McCullagh and Nelder (1989) and an introduction is given by Dobson (2002). The class of GLMs shares the following characteristics.

- *The random component* specifies the distribution of the independent response variables $Y_1, \ldots, Y_N$. The distribution can be any member of the exponential family and the response variable may be either discrete or continuous.

- *The linear predictor* defined as

$$\eta_i = \mathbf{x}_i'\boldsymbol{\theta}, \quad i = 1, ..., N$$

where $\mathbf{x}_i$ is a $p \times 1$ vector containing the k control variables $(x_1, .., x_k)$ and $\boldsymbol{\theta}$ is a $p \times 1$ parameter vector. The vector $\mathbf{x}_i$ may include non-linear elements, in contrast to the parameter vector $\boldsymbol{\theta}$. For instance, a model with an intercept, a quadratic term and a cubic term yields $\mathbf{x}_i = \begin{bmatrix} 1 & x_i & x_i^2 & x_i^3 \end{bmatrix}'$.

- *The link function*

$$g(\mu_i) = \eta_i \quad i = 1, ..., N$$

defines the connecting link between the mean response, $E(Y_i) = \mu_i$, and the linear predictor, η_i. g is assumed to be a differentiable and monotonic function. Examples on standard GLMs and corresponding link functions are shown in Table 2.1.

TABLE 2.1: Some GLM examples.

REGRESSION MODEL	RESPONSE	DISTRIBUTION	LINK FUNCTION	INVERSE LINK FUNCTION
Linear	Continuous	Normal	identity: $\eta = \mu$	$\mu = \eta$
Logistic	Binary	Binomial	logit: $\eta = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp\{\eta\}}{1+\exp\{\eta\}}$
Probit	Binary	Binomial	probit: $\eta = \Phi^{-1}(\mu)$	$\mu = \Phi(\eta)$
Poisson	Counts	Poisson	log: $\eta = \ln \mu$	$\mu = \exp\{\eta\}$

The mean response μ_i is a function of the linear predictor, see Table 2.1, which in turn is a function of the control variables and the model parameters. Furthermore, the variance of Y_i generally depends on the mean response μ_i. For example, in the case of a binomial response variable $V(Y_i) = r_i \mu_i (1 - \mu_i)$, where r_i is the number of observations at $\mathbf{x}_i$, and for a Poisson response variable $V(Y_i) = \mu_i$, whereas in the standard linear regression model $V(Y_i) = \sigma^2$. The function

$$v(\mathbf{x}_i) = \frac{1}{V(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

is called the GLM weight. For binomial Y_i with logit link the GLM weight becomes $v(\mathbf{x}_i) = r_i \mu_i (1 - \mu_i)$, for a Poisson response and log link $v(\mathbf{x}_i) = \mu_i$ and for a normally distributed response it is constant, $v(\mathbf{x}_i) = \sigma^{-2}$. This

is what makes the construction of optimal designs for GLMs generally more complicated than for linear models. An optimal design often attempts to maximize the precision of the parameter estimates, or equivalently minimize the variance, e.g. to make good predictions of the response. The GLM weight enters as a term into the variance associated with estimating the model parameters. Since $v(\mathbf{x}_i)$ depends on the true parameters through μ_i the optimal design will also depend on the true parameters.

2.2 Optimal Design of Experiments

The experimenter has to decide on what levels on the control variable should be used in the experiment, such as the amounts of a certain fertilizer in an agricultural experiment, the quantities of a new drug in a biomedical experiment or the temperatures in a manufacturing experiment. A design point is a particular level on the control variable, like 150 degrees or 5 ml of the drug. The locations of the design points have a direct influence on the amount of information that can be extracted from the experiment. Other issues related to the experimental design refers to the number of design points and the allocation of observations to the points. In the case of sequential experimentation additional questions arise, for example are a few large batches preferred over many batches of smaller size?

The purpose of conducting experiments is to increase knowledge, to gain as much information as possible. The core of the theory of optimal design is therefore the information matrix which mathematically summarizes the amount of information. Optimal designs are derived by optimizing some carefully chosen function of the information matrix.

2.2.1 The information matrix

Let ξ denote a design formulated according to

$$\xi = \left\{ \begin{array}{cccc} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_n \\ r_1 & r_2 & \dots & r_n \end{array} \right\}, \quad \sum_{i=1}^{n} r_i = N, \quad \mathbf{x} \in \chi \subseteq \mathbb{R}^k,$$

where r_i is the number of observations taken at the design point $\mathbf{x}_i$. The design space χ is the set of possible $\mathbf{x}$-values. It can be either restricted or unrestricted. Some applications imply a restricted design space, e.g.

a dose of a drug is restricted to be above zero and below a safety level. Alternatively, a design can be formulated as

$$\xi = \left\{ \begin{array}{cccc} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_n \\ w_1 & w_2 & \dots & w_n \end{array} \right\}, \quad 0 \leq w_i \leq 1, \quad \sum_{i=1}^{n} w_i = 1,$$

where the design weights $w_i = r_i/N$, specify the allocation to the design points. The restriction that r_i should be integer needs to be imposed for the design to be realizable. The fulfilment of such condition is what distinguishes an exact design from a continuous design. It is typically more practical to handle continuous designs calculation-wise although such designs are not guaranteed to be feasible. However, an exact design can often serve as a satisfactory approximation, provided that N is not too small. Let the $N \times p$ design matrix be given by

$$\mathbf{X}_N = \left[\begin{array}{l} \left. \begin{array}{c} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_1' \end{array} \right\} r_1 \text{ times} \\ \vdots \\ \left. \begin{array}{c} \mathbf{x}_i' \\ \vdots \\ \mathbf{x}_i' \end{array} \right\} r_i \text{ times} \\ \vdots \\ \left. \begin{array}{c} \mathbf{x}_n' \\ \vdots \\ \mathbf{x}_n' \end{array} \right\} r_n \text{ times} \end{array} \right]$$

where the exact appearance of the $1 \times p$ vector $\mathbf{x}_i'$ is determined by the model. Each of the N observations made according to the design ξ enters as a row in $\mathbf{X}_N$, one time each. If there are any replications some of the rows will be exactly the same. The design matrix contains information about the location of the design points in the design space χ. The information matrix for a GLM

$$\mathbf{I}(\xi, \boldsymbol{\theta}) = \mathbf{X}_N' \mathbf{V}_N \mathbf{X}_N$$

is composed of the design matrix and the $N \times N$ GLM weight matrix

$$\mathbf{V}_N = diag \left[\underbrace{v(\mathbf{x}_1) \dots v(\mathbf{x}_1)}_{r_1 \text{ times}} \quad \dots \quad \underbrace{v(\mathbf{x}_i) \dots v(\mathbf{x}_i)}_{r_i \text{ times}} \quad \dots \quad \underbrace{v(\mathbf{x}_n) \dots v(\mathbf{x}_n)}_{r_n \text{ times}} \right].$$

That is, smaller weights are assigned to less informative locations/points and vice versa. The weights are determined by the model parameters $\boldsymbol{\theta}$, as argued in the previous section. Again, there may be elements in $\mathbf{V}_N$ that are identical if replicates are made at any design point. The information matrix reduces to

$$\mathbf{I}(\xi) = \frac{1}{\sigma^2}\mathbf{X}_N'\mathbf{X}_N$$

for the linear regression model, in which case it suffices to consider the design matrix when deriving optimal designs (not the model parameters). Furthermore, it is customary to deal with the standardized information matrix that reveals the average information per observation. Adding more points to a design will never decrease the total information, worst case scenario is that it remains the same. Unless a candidate point causes the average information to increase it would pay off more to make replications at an already existing point. Seeking to maximize average information assures the inclusion of only highly informative design points. The standardized information matrix for a design is given by

$$\mathbf{M}(\xi, \boldsymbol{\theta}) = N^{-1}\mathbf{I}(\xi, \boldsymbol{\theta}) = N^{-1}\mathbf{X}_N'\mathbf{V}_N\mathbf{X}_N = \mathbf{X}_n'\mathbf{V}_n\mathbf{W}_n\mathbf{X}_n.$$

The matrix $\mathbf{X}_n$ of dimension $n \times p$ is given by

$$\mathbf{X}_n = \begin{bmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_i' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix}$$

and the GLM weight matrix $\mathbf{V}_n$ of dimension $n \times n$ is accordingly

$$\mathbf{V}_n = diag\begin{bmatrix} v(\mathbf{x}_1) & \dots & v(\mathbf{x}_i) & \dots & v(\mathbf{x}_n) \end{bmatrix}.$$

The $n \times n$ design weight matrix $\mathbf{W}_n$ contains the proportions of observations on the diagonal according to

$$\mathbf{W}_n = diag\begin{bmatrix} w_1 & \dots & w_i & \dots & w_n \end{bmatrix}.$$

The standardized information matrix can also be expressed as the weighted sum of the information obtained from the individual design points

$$\mathbf{M}(\xi, \boldsymbol{\theta}) = \sum_{i=1}^{n} w_i v(\mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i' = \sum_{i=1}^{n} w_i \mathbf{m}(\boldsymbol{\theta}, \mathbf{x}_i).$$

$\mathbf{m}(\boldsymbol{\theta}, \mathbf{x}_i)$ denotes the contribution of the design point $\mathbf{x}_i$ to the total information. Moreover, the standardized variance associated with making predictions of the response at $\mathbf{x}$ is defined as

$$d(\mathbf{x}, \xi) = v(\mathbf{x}) \mathbf{x}' \mathbf{M}^{-1}(\xi, \boldsymbol{\theta}) \mathbf{x} = tr\left[\mathbf{m}(\boldsymbol{\theta}, \mathbf{x}) \mathbf{M}^{-1}(\xi, \boldsymbol{\theta})\right].$$

$d(\mathbf{x}, \xi)$, known as the standardized predictor variance, plays an important role in the construction of D-optimal designs.

2.2.2 Maximum Likelihood Estimation

The inferential goal of an experiment is often to achieve high precision/low variance when estimating the model parameters. Maximum likelihood estimates for a GLM is obtained by solving the score equations

$$\frac{\partial l}{\partial \theta_j} = U_j = \sum_{i=1}^{N} \frac{(y_i - \mu_i) x_{ij}}{V(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right) = 0 \qquad j = 1, 2, ..., p$$

where l is the logarithm of the likelihood function $L(\boldsymbol{\theta}; \mathbf{y})$. The solutions to these equations generally have to be found numerically. The method of Fisher scoring (see e.g. Dobson, 2002) is an iterative method that is useful to compute the estimates. Given a guess $\boldsymbol{\theta}^{(m-1)}$ of the parameter vector $\boldsymbol{\theta}$ a new guess $\boldsymbol{\theta}^{(m)}$ is obtained by

$$\boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^{(m-1)} + \left[\mathbf{I}^{(m-1)}\right]^{-1} \mathbf{U}^{(m-1)}.$$

$\mathbf{I}^{(m-1)}$ is the information matrix and $\mathbf{U}^{(m-1)}$ is the vector of scores (U_j), both evaluated at $\boldsymbol{\theta}^{(m-1)}$. This expression can be rewritten as

$$\boldsymbol{\theta}^{(m)} = \left(\mathbf{X}_N' \mathbf{V}_N \mathbf{X}_N\right)^{-1} \mathbf{X}_N' \mathbf{V}_N \mathbf{z},$$

where $\mathbf{X}_N$ is the $N \times p$ design matrix, $\mathbf{V}_N$ is the $N \times N$ GLM weight matrix and $\mathbf{z}$ is a $N \times 1$ vector with elements

$$z_i = \mathbf{x}_i' \boldsymbol{\theta}^{(m-1)} + (y_i - \mu_i) \left(\frac{\partial \mu_i}{\partial \eta_i}\right).$$

$\mathbf{V}_N$ and $\mathbf{z}$ are evaluated at $\boldsymbol{\theta}^{(m-1)}$. Iterations are continued until a termination criterion is reached, e.g. when the relative difference between $\boldsymbol{\theta}^{(m)}$ and $\boldsymbol{\theta}^{(m-1)}$ is less than a predetermined small number. $\boldsymbol{\theta}^{(m)}$ is then taken as the ML estimate of $\boldsymbol{\theta}$ denoted by $\widehat{\boldsymbol{\theta}}$. The asymptotic sampling distribution of $\sqrt{N}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)$ is normal with covariance matrix

$$V\left(\widehat{\boldsymbol{\theta}}\right)=\left(\mathbf{X}_n'\mathbf{V}_n\mathbf{W}_n\mathbf{X}_n\right)^{-1}.$$

Unfortunately, for certain samples there exist no maximum likelihood estimates. The data pattern of a sample determines whether the maximum likelihood estimate exists or not. These issues are treated in Albert and Anderson (1984). As an extreme example consider a binary response experiment in which only zeros (nonresponse) are observed at all design points, in such case it is intuitive that no parameter estimates can be obtained.

2.2.3 Optimality Criteria

Ideally a design that maximizes the entire standardized information matrix $\mathbf{M}(\xi,\boldsymbol{\theta})$ would be chosen as the optimal design. Unfortunately, such an optimization task is generally not doable because it is not possible to rank matrices. Some appropriate function of $\mathbf{M}(\xi,\boldsymbol{\theta})$ will instead be the subject of optimization. Let $\Psi\{\mathbf{M}(\xi,\boldsymbol{\theta})\}$ denote a criterion function that agrees well with the inferential goal of the experiment. Generally a design ξ^* is said to be Ψ-optimal if

$$\xi^*=\arg\min_{x_i,w_i,n}\Psi\{\mathbf{M}(\xi,\boldsymbol{\theta})\}.$$

Convexity of the criterion function Ψ is assumed. Let $\bar{\xi}$ be a design with design weight 1 at the design point $\mathbf{x}$ and let ξ' be defined as

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

for $0\leq\alpha\leq 1$. The directional derivative of $\Psi(\xi,\boldsymbol{\theta})$ in the direction $\bar{\xi}$ is given by

$$\phi(\mathbf{x},\xi)=\lim_{\alpha\to 0^+}\frac{1}{\alpha}\left[\Psi\{\mathbf{M}(\xi',\boldsymbol{\theta})\}-\Psi\{\mathbf{M}(\xi,\boldsymbol{\theta})\}\right]$$

A design is Ψ-optimal if and only if the minimum of the directional derivative $\phi(\mathbf{x},\xi^*)\geq 0$ for all $\mathbf{x}\in\chi$. This result is stated in The General

Equivalence Theorem (Kiefer and Wolfowitz (1959) and Kiefer, (1961)) together with two equivalent conditions on ξ^*:

1. The design ξ^* minimizes $\Psi\{\mathbf{M}(\xi, \boldsymbol{\theta})\}$
2. The minimum of $\phi(\mathbf{x}, \xi^*) \geq 0$
3. The derivative $\phi(\mathbf{x}, \xi^*)$ achieves its minima at the points of the design,

see e.g. Atkinson and Donev (1992). Various optimality criteria have been proposed, some popular examples are listed below.

D-optimality The perhaps most widely used optimality criteria is that of D-optimality, where the optimal design is found by minimizing

$$\Psi\{\mathbf{M}(\xi, \boldsymbol{\theta})\} = \ln\left|\mathbf{M}^{-1}(\xi, \boldsymbol{\theta})\right|.$$

Equivalently, a D-optimal design is found by maximizing

$$\Psi^*\{\mathbf{M}(\xi, \boldsymbol{\theta})\} = \ln|\mathbf{M}(\xi, \boldsymbol{\theta})|$$

because $|\mathbf{M}^{-1}| = |\mathbf{M}|^{-1}$. Numerical methods are usually required to solve this task. The rationale of this criterion function is that $|\mathbf{M}(\xi, \boldsymbol{\theta})|^{-1/2}$ is proportional to the volume of the asymptotic confidence region for the parameters and it is desirable to have this region as small as possible. When $\Psi\{\mathbf{M}(\xi, \boldsymbol{\theta})\} = \ln|\mathbf{M}^{-1}(\xi, \boldsymbol{\theta})|$ it can be shown that

$$\phi(\mathbf{x}, \xi) = p - d(\mathbf{x}, \xi),$$

see for example Silvey (1980), with the resulting alternative formulation of the conditions on ξ^* in The General Equivalence Theorem

1. The design ξ^* minimizes $\Psi\{\mathbf{M}(\xi, \boldsymbol{\theta})\}$
2. $d(\mathbf{x}, \xi^*) \leq p$
3. $d(\mathbf{x}, \xi^*)$ achieves its maxima at the points of the design.

The practical implication of this is that the optimality of a suggested design can easily be verified or disproved. A graphical examination of a plot of $d(\mathbf{x}, \xi)$ reveals whether a design is optimal or not. In the case of a non-optimal design the appearance of the curve can give a clue about the optimal number of design points.

Sometimes interest is in s linear combinations of the p parameters, a situation that often arises in medical experiments when interest is in making comparisons between a control group and one or more treatment groups. The asymptotic covariance matrix for the linear combination $\mathbf{A}'\boldsymbol{\theta}$ is given by

$$\mathbf{A}'\mathbf{M}^{-1}(\xi, \boldsymbol{\theta})\,\mathbf{A}$$

where $\mathbf{A}$ is a $p \times s$ matrix. A special case of D-optimality, referred to as D_A-optimality occurs when

$$\Psi\left\{\mathbf{M}(\xi, \boldsymbol{\theta})\right\} = \ln\left|\mathbf{A}'\mathbf{M}^{-1}(\xi, \boldsymbol{\theta})\,\mathbf{A}\right|$$

is minimized.

Another special case, D_s-optimality, comes about in situations where there are s parameters of interest and $p-s$ nuisance parameters or when interest is in model checking. Consider for example the parameter vector partitioned as $\boldsymbol{\theta} = \begin{bmatrix}\boldsymbol{\theta}_s & \boldsymbol{\theta}_{p-s}\end{bmatrix}$ and that the aim is to find an optimal design to estimate $\boldsymbol{\theta}_s$. It is analogous to D_A-optimality if $\mathbf{A}$ is set equal to $\begin{bmatrix}\mathbf{I}_s & \mathbf{0}\end{bmatrix}$, $\mathbf{I}_s$ being the $s \times s$ identity matrix and $\mathbf{0}$ a $(p-s) \times s$ matrix with zeros.

c-optimality

c-optimality is an appropriate criterion when the aim is to estimate some function $h(\boldsymbol{\theta})$ of the model parameters with minimum variance. The criterion function to be minimized is then

$$\Psi\left\{\mathbf{M}(\xi, \boldsymbol{\theta})\right\} = \mathbf{c}'\mathbf{M}^{-1}(\xi, \boldsymbol{\theta})\,\mathbf{c}$$

where $\mathbf{c}$ is a $p \times 1$ vector. Specifically, when the target function is nonlinear in $\boldsymbol{\theta}$, the asymptotic variance of $h\left(\widehat{\boldsymbol{\theta}}\right)$ is given by

$$\begin{aligned} V\left(h\left(\widehat{\boldsymbol{\theta}}\right)\right) &= \mathbf{c}'\mathbf{V}\left(\widehat{\boldsymbol{\theta}}\right)\mathbf{c}, \\ \mathbf{c} &= \frac{\partial h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \end{aligned}$$

where $h(\boldsymbol{\theta})$ is assumed to be differentiable in a neighborhood of $\boldsymbol{\theta}$. Hence, it follows that c-optimality is a natural criterion when the purpose of the experiment is accurate estimation of $h(\boldsymbol{\theta})$.

c-optimality will also be the choice when the objective is to estimate one single parameter. Such situations often arise in the RSM context when interest is in the parameter that determines the point of optimum response. Consider for example manufacturing a food product that needs to be prepared in oven, an experiment might then be conducted to find out the baking time that maximizes the probability of a good product.

A-optimality

Even if the volume of a confidence ellipsoid is small (as strived for using the D-optimality criterion) all variances of the parameter estimates are not necessarily small. As an alternative, A-optimality strives for minimizing the sum of the variances of the parameter estimators. The diagonal elements of the inverse of the standardized information matrix are proportional to the asymptotic variance of the ML estimator of $\boldsymbol{\theta}$. The design that minimizes

$$\Psi\left\{\mathbf{M}\left(\xi,\boldsymbol{\theta}\right)\right\} = tr\left[\mathbf{M}^{-1}(\xi,\boldsymbol{\theta})\right]$$

is called A-optimal.

It needs to be pointed out that these optimal design criteria rely on asymptotics for the GLM models. In practice all designs are limited regarding the number of observations. As a consequence the success of a particular design depends on the degree of agreement between the asymptotic sampling distribution and the sampling distribution for the current finite sample size.

2.2.4 The Canonical Design Space

An analysis of the geometry of the canonical design space can provide useful insights into the construction of optimal designs. It can also result in the formulation of some geometric rules. In an important paper, Elfving (1952) used geometrical arguments to derive c-optimal design points along with optimal weights for the linear regression model. Kitsos, Titterington and Torsney (1988) applied these principles for a specific nonlinear design problem. Ford, Torsney and Wu (1992) used the canonical form to construct locally optimal designs for generalized linear models with one control variable. Sitter and Torsney (1995) used the same method for binary response models with two control variables. Biedermann, Dette, and Zhu (2005) explored designs for binary response models with a single response variable

and using a general class of criterion functions as well as a restricted range of the control variable.

The canonical design space $Z \subseteq \mathbb{R}^p$ arises from a transformation of the design points defined by the GLM weight function as

$$\mathbf{z} = \sqrt{v(\mathbf{x})}\mathbf{x}.$$

The transformation leads to a formulation of the design problem that is independent of the unknown parameters. The problem can thus be solved independent of the parameters. When the dimensionality allows, i.e. for $p \leq 3$, a plot of the space Z can be very useful. The geometry of Z reveals both the number of design points and their (approximate) locations together with the design weights. More details are given in Chapter 3 where it is applied as a method to find c-optimal designs.

2.3 Response Surface Methodology

The statistical methods for designing and analyzing the outcome of experiments where interest is in a response variable that is affected by one or several variables are known as *Response Surface Methodology* (RSM). A thorough exposition of these techniques is given in the book by Box and Draper (1987). RSM has traditionally been used for finding optimum operating conditions in the industry. It is now common in many different fields like physical, chemical, biological, clinical and social sciences.

The principal objective of RSM is to explore the unknown relationship between the response/output variable and the control/input variables. Should the exact nature of this relationship be known, which is rarely the case, a mechanistic model could be formulated. Mostly, RSM deals with approximate empirical models. As an example, if the response variable is the yield of a chemical process, the control variables might be temperature and pressure. If the response variable instead is the reaction time of an individual, possible control variables are dose of alcohol and amount of sleep. RSM attempts to answer questions about how the response variable is affected when the levels of the control variables are changed, like what happens to the yield when the temperature and pressure levels are varied. A common application for RSM is to find the optimum operating conditions, e.g. to find out for what levels of temperature and pressure the yield is maximized.

RSM is essentially a sequential process where the experimental design is gradually updated as investigation proceeds. Initially, choices have to be made regarding the model, the number of replications, the levels on the control variables and the size and location of the region of interest. The aim is to have a procedure such that the right conclusions can be drawn even if the initial experimental design is poor and that the path to arrive there is as short as possible, for details see Box and Draper (1987).

In general a response variable Y is observed in an experiment and the relationship between Y and the control variables $\mathbf{x} = \begin{bmatrix} x_1 & \dots & x_k \end{bmatrix}'$ is assumed to have the functional form

$$y = f(\mathbf{x}, \boldsymbol{\theta}) + \varepsilon.$$

Observing the response may be associated with measurement errors and there may be errors due to variations in the experimental setting. These kinds of errors as well as natural variation around the mean response are all captured by the term ε. Hence, the mean response, expressed as

$$E(Y) = f(\mathbf{x}, \boldsymbol{\theta}),$$

forms the response surface. Because the exact true functional form $f(\mathbf{x}, \boldsymbol{\theta})$ generally is unknown it needs to be approximated. A function $g(\mathbf{x}, \boldsymbol{\theta})$ is taken as a local approximation to $f(\mathbf{x}, \boldsymbol{\theta})$. The approximation is local because it is restricted to be valid only in a limited region of interest, R. The function $g(\mathbf{x}, \boldsymbol{\theta})$ is usually in the form of a polynomial. The use of a polynomial stems from making a Taylor expansion of $f(\mathbf{x}, \boldsymbol{\theta})$, a polynomial of degree d corresponds to truncating the Taylor's series after the $d : th$ order term. The higher degree of complexity of the polynomial, the better the approximation. Given a certain degree d a smaller R will increase the closeness of the approximating function to the true function. Of course there has to be a balance between the complexity and a reasonable dimension, increasing the complexity increases the number of parameters to be estimated.

An RSM experiment is often performed stepwise. One important application is to determine the conditions on the control variables that maximize or minimize the response function, for example to find the optimum operating conditions for an industrial process. Starting at current operating

conditions it may be adequate to fit a lower order model at the first stage in order to find the direction towards the optimum operating conditions. The method of steepest ascent (descent) is a sequential procedure where successive steps are taken towards the maximum increase (decrease) in the response. Assume that two control variables are believed to influence the response variable and a first order model is fitted to start with. A contour plot of the first order response surface may look like Figure 2.1, the steepest ascent is then found in the direction of the arrow. The steps are proportional to the parameter estimates, the actual step length is obtained by normalizing the vector $\widehat{\boldsymbol{\theta}}$ and then multiplying it with an appropriate step size. Observations on the response variable are made along this path until the response decreases, when a new first order model is fitted, a new direction of steepest ascent is located and a new path is followed. Experimentation proceeds in this manner until a lack of fit test indicates that a first order approximation no longer is suitable. This usually happens when the maximum of the response surface is nearby. A higher order model is then fitted and analyzed.

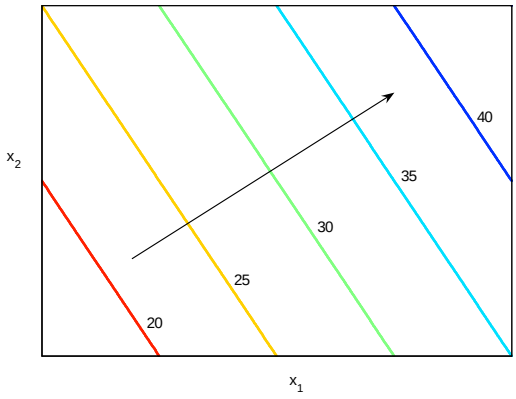


FIGURE 2.1: A contour plot of a first order response surface, the arrow points in the direction of steepest ascent.

The applications of RSM were primarily concentrated to linear regression models. However, a response surface may just as well be fitted for a GLM

model, in which case the mean response

$$E(Y) = \mu$$

that forms the response surface is connected to the linear predictor η via the inverse link function. Some examples on inverse link functions can be found in Table 2.1 in Section 2.1. In the special case of a linear model the connection between the mean response and the linear predictor is direct, that is $\mu = \eta$. This thesis deals with a second order logistic regression model for binary data which is presented in detail in the next section.

2.4 The Quadratic Logistic Model

The response variable in the logistic model is binary. Success/failure, broken/not broken and pass/fail a test are examples on outcomes of a binary response variable. The control variable(s) can be either continuos, e.g. temperature, or categorical, such as treatment group. Many times the purpose of the analysis is to explore the relationship between the probability of response (e.g. a success) and the control variable(s). The response surface (or response curve in the case of two dimensions) that approximates this relationship takes on different shapes depending on the linear predictor. For example, the logistic two-parameter model yields a sigmoidally shaped response curve. Adding a quadratic term yields a quadratic response curve, consequently named the quadratic logistic model.

The quadratic logistic model belongs to the class of GLMs. The binary response variable Y is assumed to be Bernoulli distributed with response probability $E(Y) = P(Y = 1) = \pi(x)$ given by

$$\pi(x) = \frac{\exp\{\eta(x)\}}{1 + \exp\{\eta(x)\}},$$

where the linear predictor is defined as

$$\eta(x) = \alpha + \beta(x - \mu)^2.$$

The logit link

$$\ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \eta(x)$$

specifies the connection between $\pi(x)$ and $\eta(x)$. Let $\boldsymbol{\theta} = (\alpha, \beta, \mu)'$. The parameter μ defines the point of optimum response, that is the value of x at which $\pi(x)$ attains its maximum or minimum. α controls the height of the response curve at the point of optimum response. β is related to the width of the curve and the kind of optimum. Throughout the thesis $\beta < 0$ is assumed, i.e. the response curve is assumed to have a maximum. The choice is arbitrary, analogous methods and results apply to $\beta > 0$ in which case the response curve has a minimum.

The response curve is symmetric around μ, the point of maximum response. The parameter μ can be thought of as a location parameter for the response curve. Changing μ shifts the response curve along the x-axis. Furthermore, the parameter β can be thought of as a scale parameter because the curve is stretched or contracted when β is changed. Define $\pi^*(x)$ as

$$\pi^*(x) = \frac{\exp\{\alpha - x^2\}}{1 + \exp\{\alpha - x^2\}}, \tag{2.1}$$

i.e. $\pi^*(x)$ is the same as $\pi(x)$ with the parameters β and μ set to -1 and 0, respectively. Thus, $\pi^*(x)$ corresponds to $\pi(x)$ based on the parameter set $\boldsymbol{\theta}^* = (\alpha, -1, 0)'$. The response curve $\pi^*(x)$ is standardized in the sense of being a curve with location at $x = 0$ and standardized width, given a certain height of the response curve at the point of maximum response (which is determined by the parameter α). Taking $\pi^*(x)$ as the starting point the response curve $\pi(x)$ with parameters $\boldsymbol{\theta} = (\alpha, \beta, \mu)'$ can be obtained as $\pi^*\left[(x-\mu)\sqrt{|\beta|}\right]$:

$$\begin{aligned}
\pi^*\left[(x-\mu)\sqrt{|\beta|}\right] &= \frac{\exp\left\{\alpha - \left[(x-\mu)\sqrt{|\beta|}\right]^2\right\}}{1 + \exp\left\{\alpha - \left[(x-\mu)\sqrt{|\beta|}\right]^2\right\}} \\
&= \frac{\exp\left\{\alpha + \beta(x-\mu)^2\right\}}{1 + \exp\left\{\alpha + \beta(x-\mu)^2\right\}} = \pi(x).
\end{aligned}$$

The information matrix under the quadratic logistic model, given an observation at x, is

$$\mathbf{I}(\boldsymbol{\theta}, x) = v(x)\left(\frac{\partial \eta(x)}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \eta(x)}{\partial \boldsymbol{\theta}}\right)',$$

where

$$\left(\frac{\partial \eta(x)}{\partial \boldsymbol{\theta}}\right) = \begin{pmatrix} 1 \\ (x-\mu)^2 \\ -2\beta(x-\mu) \end{pmatrix}$$

and

$$v(x) = \frac{1}{V(Y)}\left(\frac{\partial \pi(x)}{\partial \eta(x)}\right)^2 = \pi(x)(1-\pi(x)).$$

An important application for this model is to find the point of maximum or minimum response, i.e. to estimate the parameter μ. As an example consider manufacturing a product where defect items have to be discarded and the probability of a good product is maximized for the right adjustment of machine speed. As another example, consider a pharmaceutical manufacturer who produces boxes containing asthma medicine. To make the boxes waterproof a plastic wrap is inserted. The quality of this wrap is affected by the variable pressure. The control variable can be varied in an experiment to determine the level of pressure for which the probability of a leaking box is minimized. The applications are not limited to maximizing the probability of response $\pi(x)$, optimizing any general function $h(x)$ may as well be the objective. For instance, let $c(x)$ be the production cost and $h(x) = c(x)/\pi(x)$ be the cost per produced unit that is saleable. Interest may then be in finding the value of x that minimizes $h(x)$.

Chapter 3

Locally Optimal Designs for the Quadratic Logistic Model

Before an optimal design can be determined it needs to be established what sense of optimality is intended. Therefore, a decision on an appropriate criterion function has to be made. It should be chosen to match the aim of the experiment. The locally optimal design is then the choice of points and corresponding weights that optimize the selected criterion function. The criterion function is commonly related in some way to the precision of the parameter estimates, such as the size of a confidence region or the sum of the variances of the parameter estimators. It can concern estimation of all parameters, a subset, or some combination of the parameters.

In any case, the criterion function involves the standardized information matrix. The standardized information matrix given a particular design ξ is the weighted sum of the contributions from each of the n design points. For the quadratic logistic model it is given by the 3×3 matrix

$$\begin{aligned}
\mathbf{M}(\boldsymbol{\theta},\xi) &= \sum_{i=1}^{n} w_i v(x_i) \left(\frac{\partial \eta(x_i)}{\partial \boldsymbol{\theta}}\right) \left(\frac{\partial \eta(x_i)}{\partial \boldsymbol{\theta}}\right)' \\
&= \sum_{i=1}^{n} w_i \pi(x_i)(1-\pi(x_i)) \begin{pmatrix} 1 & (x_i-\mu)^2 & -2\beta(x_i-\mu) \\ (x_i-\mu)^2 & (x_i-\mu)^4 & -2\beta(x_i-\mu)^3 \\ -2\beta(x_i-\mu) & -2\beta(x_i-\mu)^3 & 4\beta^2(x_i-\mu)^2 \end{pmatrix}.
\end{aligned}$$

The notation $\mathbf{M}(\boldsymbol{\theta},\xi)$ is used to stress the fact that the information matrix depends on the unknown model parameters.

Locally optimal designs may not be that usable per se, but sequential procedures are greatly facilitated if there are easy ways to find locally optimal designs. They are also important as benchmarks, e.g. when making efficiency comparisons. In this section the procedures to obtain locally D- and c-optimal designs are demonstrated and illustrated with some examples. When deriving c-optimal designs particularly, it is shown how the geometry of the canonical design space can be made to use.

3.1 D-optimality

There are several methods at hand when it comes to the practise of determining the optimal design. These include algorithms, analytical, numerical and graphical methods, used separately or in combinations. For a review of the available methods, see e.g. Atkinson and Donev (1992). There is no method that is generally favorable, it depends on the problem at hand. The method selected to derive D-optimal designs for the quadratic logistic model is described below.

D-optimality is achieved by minimizing $|\mathbf{M}^{-1}(\xi, \boldsymbol{\theta})|$, or equivalently maximizing $|\mathbf{M}(\xi, \boldsymbol{\theta})|$. To start with the number of design points n is not known. Though, it is known that there exists a D-optimal design with $p \leq n \leq \frac{p(p+1)}{2}$ design points. A plot of the standardized predictor variance, $d(x, \xi)$ serves as a useful tool, it helps to reveal whether a suggested design is optimal or not. If the design is optimal the maximum of $d(x, \xi)$ should be equal to the number of parameters in the model. The maxima will also appear at the design points. In the non-optimal case the plot can give a hint of the optimal number of design points by looking at the number of peaks of the function $d(x, \xi)$. The D-optimal design is essentially obtained according to the following steps.

1. Begin to assume a p-point design.

2. Minimize $|\mathbf{M}^{-1}(\xi, \boldsymbol{\theta})|$ yielding the best possible p-point design.

3. Plot the standardized predictor variance, $d(x, \xi)$. If a visual inspection indicates that the design clearly is non-optimal return to step 2 and try a design with $p + 1$ points.

4. Verify optimality of the suggested design by the General Equivalence Theorem, see Section 2.2.3, by assuring that the maxima of $d(x, \xi)$ are attained at the candidate design points (either analytically or numerically). If the design cannot be verified to be optimal, go back to step 2 and evaluate a $(p+1)$-point design.

etc.

The procedure to find D-optimal designs for the quadratic logistic model is facilitated by seeking an optimal design for the standardized response curve $\pi^*(x)$ given in (2.1). Using $\pi^*(x)$ as the basis corresponds to seeking a locally D-optimal design for $\boldsymbol{\theta}^* = (\alpha, -1, 0)'$. A design point x for the response curve $\pi^*(x)$ corresponds to the point

$$\frac{x}{\sqrt{|\beta|}} + \mu \tag{3.1}$$

for the response curve $\pi(x)$, see Figure 3.1. Once a locally D-optimal design is found for $\boldsymbol{\theta}^*$, the locally D-optimal design points for estimation of $\boldsymbol{\theta}$ can easily be obtained by the transformation (3.1). It is motivated by comparing the determinants of the two information matrices $\mathbf{M}(\xi, \boldsymbol{\theta}^*)$ and $\mathbf{M}(\xi, \boldsymbol{\theta})$, the latter evaluated at $x/\sqrt{|\beta|} + \mu$. First, the expression for the determinant of the standardized information matrix based on $\boldsymbol{\theta}^*$ is

$$|\mathbf{M}(\xi, \boldsymbol{\theta}^*)| = \left| \sum w_i v^*(x_i) \begin{pmatrix} 1 & x_i^2 & 2x_i \\ x_i^2 & x_i^4 & 2x_i^3 \\ 2x_i & 2x_i^3 & 4x_i^2 \end{pmatrix} \right|.$$

Secondly, the determinant based on $\boldsymbol{\theta}$ evaluated at $x/\sqrt{|\beta|} + \mu$ is given by

$$|\mathbf{M}(\xi, \boldsymbol{\theta})|_{\frac{x}{\sqrt{|\beta|}}+\mu} = \left| \sum w_i v\left(\frac{x_i}{\sqrt{|\beta|}} + \mu\right) \begin{pmatrix} 1 & \frac{x_i^2}{|\beta|} & -2\sqrt{|\beta|}x_i \\ \frac{x_i^2}{|\beta|} & \frac{x_i^4}{\beta^2} & \frac{-2x_i^3}{\sqrt{|\beta|}} \\ -2\sqrt{|\beta|}x_i & \frac{-2x_i^3}{\sqrt{|\beta|}} & 4|\beta| x_i^2 \end{pmatrix} \right|.$$

Then, noting that

$$v^*(x) = v\left(\frac{x}{\sqrt{|\beta|}} + \mu\right),$$

the following equality holds

$$\left|\mathbf{M}\left(\xi,\boldsymbol{\theta}\right)\right|_{\frac{x}{\sqrt{|\beta|}}+\mu}=\frac{|\beta|}{4}\left|\mathbf{M}\left(\xi,\boldsymbol{\theta}^{*}\right)\right|.$$

The fact that $\left|\mathbf{M}\left(\xi,\boldsymbol{\theta}^{*}\right)\right| \propto \left|\mathbf{M}\left(\xi,\boldsymbol{\theta}\right)\right|_{\frac{x}{\sqrt{|\beta|}}+\mu}$ shows that the locally D-optimal depends only on the parameter α since $\boldsymbol{\theta}^{*}$ only includes α.

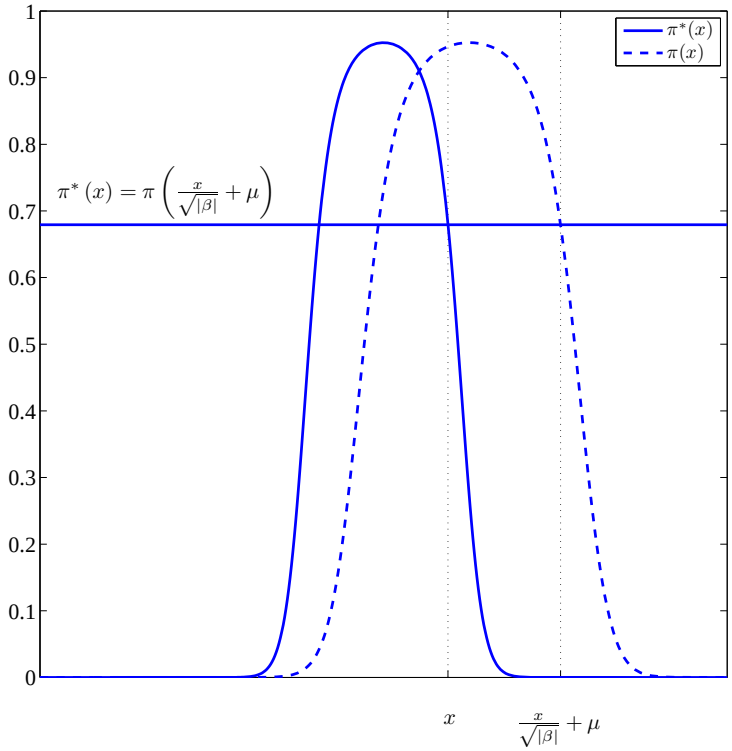


FIGURE 3.1: The response curves $\pi^{*}\left(x\right)$ and $\pi\left(x\right)$ for $\alpha=3$.

The derivation of D-optimal designs will be illustrated for three parameter sets: $\boldsymbol{\theta}_{A}=\left(3,-1,0\right)'$, $\boldsymbol{\theta}_{B}=\left(0,-1,0\right)'$ and $\boldsymbol{\theta}_{C}=\left(-3,-1,0\right)'$, according to the steps given above.

Step 1. Due to the symmetry property of the response curve that $\pi^{*}\left(x\right)=\pi^{*}\left(-x\right)$ one might expect that the optimal design is symmetric as well. To begin with the following symmetric design consisting of $p=3$ points with equal design weights is assumed.

$$\xi_3 = \left\{ \begin{array}{ccc} -x & 0 & x \\ 1/3 & 1/3 & 1/3 \end{array} \right\}.$$

Step 2. The standardized information matrix for ξ_3 is

$$\begin{aligned} \mathbf{M}(\xi_3, \boldsymbol{\theta}^*) &= \frac{1}{3}\left[2v^*(x) \begin{pmatrix} 1 & x^2 & 0 \\ x^2 & x^4 & 0 \\ 0 & 0 & 4x^2 \end{pmatrix} + v^*(0) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \right] \\ &= \frac{2}{3} \begin{pmatrix} v^*(x) + \frac{v^*(0)}{2} & v^*(x)\, x^2 & 0 \\ v^*(x)\, x^2 & v^*(x)\, x^4 & 0 \\ 0 & 0 & v^*(x)\, 4x^2 \end{pmatrix} \end{aligned}$$

where $v^*(x)$ is the GLM weight function based on $\pi^*(x)$, i.e. $v^*(x) = \pi^*(x)[1 - \pi^*(x)]$,

$$\begin{aligned} v^*(x) &= v^*(-x) = \frac{\exp\{\alpha - x^2\}}{(1 + \exp\{\alpha - x^2\})^2}, \\ v^*(0) &= \frac{\exp\{\alpha\}}{(1 + \exp\{\alpha\})^2}. \end{aligned}$$

Now, the determinant of the standardized information matrix is

$$|\mathbf{M}(\xi_3, \boldsymbol{\theta}^*)| = \left(\frac{2}{3}\right)^3 \left(2x^6 v^*(x)^2 v^*(0)\right) \propto x^6 v^*(x)^2$$

and maximizing yields the following expression

$$x = \sqrt{\frac{3}{2 - 4\pi^*(x)}}. \tag{3.2}$$

The solutions to formula (3.2) can be found numerically, using Mathcad 12.0 for instance. The resulting 3-point designs are given in Table 3.1.

TABLE 3.1: Suggested 3-point designs.

TRUE PARAMETERS	DESIGN
$\boldsymbol{\theta}_A = (3, -1, 0)'$	$\xi_{3A} = \begin{Bmatrix} -1.957 & 0 & 1.957 \\ 1/3 & 1/3 & 1/3 \end{Bmatrix}$
$\boldsymbol{\theta}_B = (0, -1, 0)'$	$\xi_{3B} = \begin{Bmatrix} -1.407 & 0 & 1.407 \\ 1/3 & 1/3 & 1/3 \end{Bmatrix}$
$\boldsymbol{\theta}_C = (-3, -1, 0)'$	$\xi_{3C} = \begin{Bmatrix} -1.238 & 0 & 1.238 \\ 1/3 & 1/3 & 1/3 \end{Bmatrix}$

Step 3. Plots of the standardized predictor variance, $d(x, \xi)$ are given in Figure 3.2. These plots suggest that the 3-point designs are D-optimal for the two parameter sets $\boldsymbol{\theta}_B$ and $\boldsymbol{\theta}_C$. $d(x, \xi) \leq 3$ and the maxima are attained at the design points for these two models which is in line with the General Equivalence Theorem. However, $d(x, \xi) > 3$ for model A and the design ξ_{3A} is clearly non-optimal. The plot indicates that the D-optimal design could consist of 4 symmetric points. Therefore, the next step is to test a 4-point design. Let ξ_{4A} denote a design with 4 points symmetric around μ specified as

$$\xi_{4A} = \begin{Bmatrix} -x_2 & -x_1 & x_1 & x_2 \\ w & 0.5 - w & 0.5 - w & w \end{Bmatrix}.$$

The standardized information matrix for the design ξ_{4A} is then

$$\mathbf{M}(\xi_{4A}, \boldsymbol{\theta}^*) = 2wv^*(x_2) \begin{pmatrix} 1 & x_2^2 & 0 \\ x_2^2 & x_2^4 & 0 \\ 0 & 0 & 4x_2^2 \end{pmatrix} + 2(0.5 - w) v^*(x_1) \begin{pmatrix} 1 & x_1^2 & 0 \\ x_1^2 & x_1^4 & 0 \\ 0 & 0 & 4x_1^2 \end{pmatrix}.$$

In this case the determinant of the standardized information matrix

$$\begin{aligned} & |\mathbf{M}(\xi_{4A}, \boldsymbol{\theta}^*)| \\ = \; & 8w(1 - 2w) v^*(x_1) v^*(x_2) \left(x_2^2 - x_1^2\right)^2 \left[2wv^*(x_2) x_2^2 + (1 - 2w) v^*(x_1) x_1^2\right] \end{aligned}$$

is a function of three variables: $f(x_1, x_2, w)$. The D-optimal design can now be found as the solutions to the set of equations

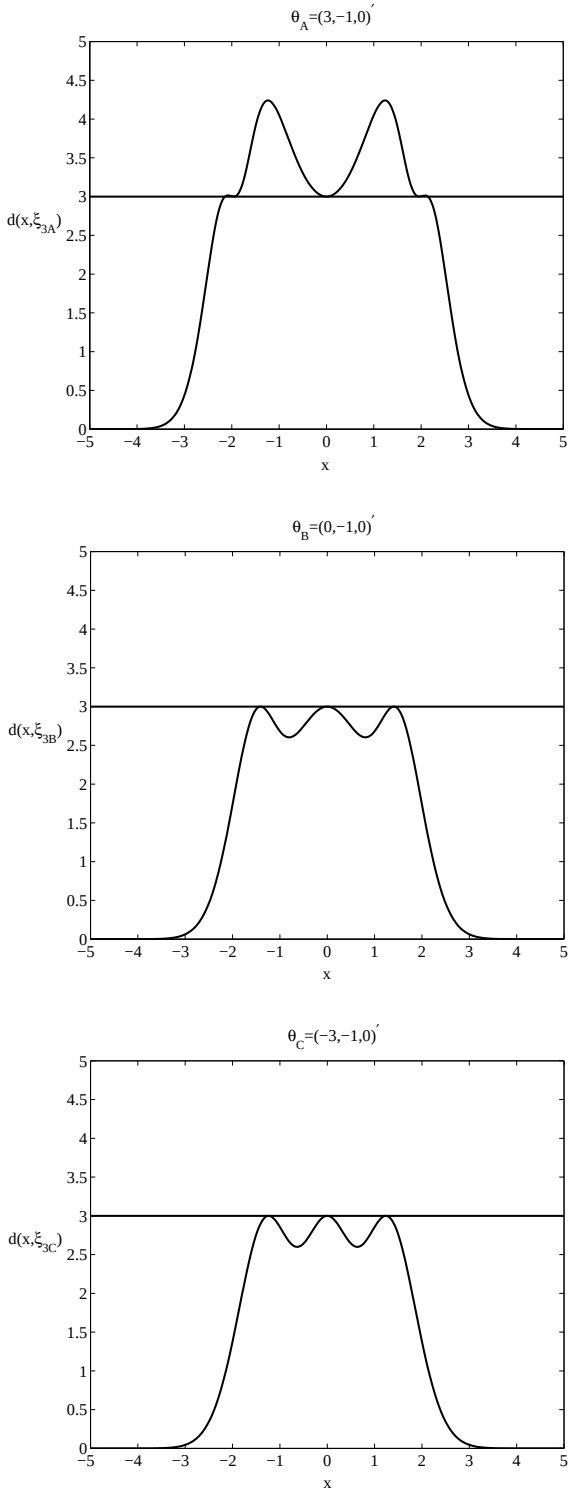


FIGURE 3.2: The standardized predictor variance $d(x, \xi)$ for the designs given in Table 3.1.

$$\frac{\partial f(x_1,x_2,w)}{\partial x_1} =$$
$$(1-2w)\,v^*(x_1)\left[-\left(2q(x_1)\,x_1^2-1\right)\left(x_2^2-x_1^2\right)-2x_1^2\right]+2wv^*(x_2)\,x_2^2\left[q(x_1)\left(x_2^2-x_1^2\right)-2\right]=0$$

$$\frac{\partial f(x_1,x_2,w)}{\partial x_2} =$$
$$(1-2w)\,v^*(x_1)\,x_1^2\left[-q(x_2)\left(x_2^2-x_1^2\right)+2\right]-2wv^*(x_2)\left[\left(2q(x_2)\,x_2^2-1\right)\left(x_2^2-x_1^2\right)-2x_2^2\right]=0$$

$$\frac{\partial f}{\partial w} = v^*(x_1)\,x_1^2\left[(1-2w)(1-6w)\right]+v^*(x_2)\,x_2^2\left[4w(1-3w)\right]=0$$

where

$$q(x)=\frac{1-\exp\{\alpha-x^2\}}{1+\exp\{\alpha-x^2\}}.$$

While these equations do not offer as much of a simplification as in the case with 3 design points, they may serve as an alternative in finding the optimal design. Here, solving the equations (numerically) resulted in the following design

$$\xi_4=\left\{\begin{array}{cccc}-2.061 & -1.324 & 1.324 & 2.061\\ 0.297 & 0.203 & 0.203 & 0.297\end{array}\right\}.$$

The standardized predictor variance $d(x,\xi_{4A})$ is plotted in Figure 3.3. An examination of the plot shows that $d(x,\xi_{4A})\leq p=3$ and that the maxima are attained at the design points.

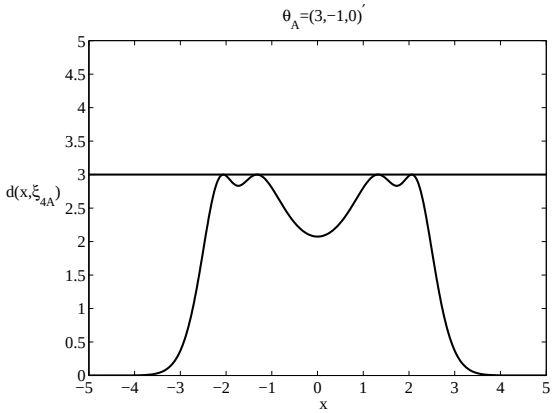


FIGURE 3.3: The standardized predictor variance $d(x,\xi_4)$ for the design ξ_4.

Step 4. Optimality of the designs ξ_{4A}, ξ_{3B} and ξ_{3C} is here verified numerically. It is confirmed that the maxima of the standardized predictor variance

$$d\left(x,\xi\right)=tr\left[\mathbf{m}\left(\boldsymbol{\theta},x\right)\mathbf{M}^{-1}\left(\xi,\boldsymbol{\theta}\right)\right]$$

are attained at the suggested design points in all three cases, i.e. the designs meet the conditions of the General Equivalence Theorem. A common property for the three designs is that all of them are placed symmetrically around $\mu=0$. It can also be noted that the design points are placed at $\pi\left(x\right)=\pi\left(-x\right)=0.226$ for the 3-point designs and at $\pi\left(x_1\right)=1-\pi\left(x_2\right)=0.223$ for the 4-point design.

3.2 c-optimality

For certain applications the objective is to estimate just one of the model parameters. For instance, when seeking the point of optimum response, it is only the parameter μ that is of interest and the other two parameters may be regarded as nuisance parameters. In such cases a c-optimal design is appropriate. For c-optimal designs, the canonical design space offers useful information about the number of design points as well as the placing of them. This section is based on the paper by Fackle Fornius and Nyquist (2008). First, the characteristics of the canonical space is examined. Subsequently, c-optimal designs for estimation of each of the model parameters are treated in detail.

In order to estimate the parameter μ with high precision define $\mathbf{c}=\left(0,0,1\right)'$ and minimize the asymptotic variance of $\mathbf{c}'\widehat{\boldsymbol{\theta}}$, i.e. minimize

$$\mathbf{c}'\mathbf{M}\left(\boldsymbol{\theta},\xi\right)^{-1}\mathbf{c}.$$

Correspondingly, $\mathbf{c}=\left(1,0,0\right)'$ and $\mathbf{c}=\left(0,1,0\right)'$ apply for estimation of the other two parameters, α and β, respectively. Recall the standardized information matrix for the quadratic logistic model given a design ξ and parameter vector $\boldsymbol{\theta}=\left(\alpha,\beta,\mu\right)'$

$$\mathbf{M}\left(\boldsymbol{\theta},\xi\right)=\sum_{i=1}^{n}w_iv\left(x_i\right)\left(1,\left(x_i-\mu\right)^2,-2\beta\left(x_i-\mu\right)\right)'\left(1,\left(x_i-\mu\right)^2,-2\beta\left(x_i-\mu\right)\right).$$

It can be noted that $\mathbf{M}(\boldsymbol{\theta},\xi)$ for this model is proportional to that for a linear regression model

$$y = \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_3 + \varepsilon, \tag{3.3}$$

where γ_1, γ_2, and γ_3 are the model parameters,

$$\begin{aligned} z_1 &= \sqrt{v(x)}\,(\partial\eta(x)/\partial\alpha) = \sqrt{\pi(x)(1-\pi(x))} \\ z_2 &= \sqrt{v(x)}\,(\partial\eta(x)/\partial\beta) = \sqrt{\pi(x)(1-\pi(x))}\,(x-\mu)^2 \\ z_3 &= \sqrt{v(x)}\,(\partial\eta(x)/\partial\mu) = -2\sqrt{\pi(x)(1-\pi(x))}\beta(x-\mu) \end{aligned}$$

and $\varepsilon \sim N(0,\sigma^2)$.

In the original formulation, the design problem consists of selecting the design points $x_1, ..., x_n$ from the design space χ. However, as noted above, the standardized information matrix for the model under consideration is proportional to that for the linear model (3.3). The problem of selecting design points $x_1, ..., x_n$ from the design space χ is therefore equivalent to selecting design points $\mathbf{z}_1, ..., \mathbf{z}_n$, $\mathbf{z} = (z_1, z_2, z_3)'$ from the design space Z defined by

$$Z = \left\{ \mathbf{z} \in \mathbb{R}^{\mathbf{3}} : \mathbf{z} = \sqrt{v(x)} \left(\frac{\partial\eta(x)}{\partial\boldsymbol{\theta}} \right),\ x \in \chi \right\}.$$

Z is called the canonical design space and the design problem is now converted to the selection of canonical design points in Z. In the original design problem, there is a dependence of the optimal design on the true parameters $\boldsymbol{\theta}$. In the canonical version, however, the transformation from χ to Z is parameter dependent yielding a design space that varies with $\boldsymbol{\theta}$. In effect, it has resulted in a transformation of the GLM design problem to one that is analogous to the design problem for a linear model.

The canonical design space Z defines a curve in $\mathbb{R}^{\mathbf{3}}$, see Figures 3.4 to 3.6. These graphs also display the reflection of Z through the origin, $-Z$. Note that $z_1,\ z_2 > 0$ for $z \in Z$. The sets Z and $-Z$ as well as the convex hull of $Z \cup -Z$ are used in the geometric approach presented later in this section. Each point in Z corresponds to a specific value of $\eta(x)$ which in turn corresponds to a specific value of x. The function $\pi(x)$ is symmetric around the point of optimum response, i.e. $\pi(x+\mu) = \pi(-x+\mu)$. The

curve Z shares this property. It can further be noted that the shape of the curve is independent of μ, so defining $t = \beta (x - \mu)^2$ facilitates a more detailed examination.

Suppose first that $\beta < 0$, so that $t \leq 0$ and $\pi (x)$ has a maximum attained at $x = \mu$, or equivalently at $t = 0$. Three cases are shown in Figures 3.4 to 3.6: $\boldsymbol{\theta}_A = (3, -1, 0)'$, $\boldsymbol{\theta}_B = (0, -1, 0)'$and $\boldsymbol{\theta}_C = (-3, -1, 0)'$. Consider first Figure 3.4 which displays model A. It is seen that the curve Z makes a closed loop, approaching the origin as $t \to -\infty$, z_1 is at most 0.5 which appears when $t = -\alpha$ and the curve has a symmetry point at $t = 0$. For the next case, model B, displayed in Figure 3.5, the symmetry point coincides with the maximum of z_1 at $t = 0$, which is now given by $z_1 = \exp (\alpha/2) / (1 + \exp (\alpha))$. Model C is viewed in Figure 3.6. The curve now has a similar shape as in the previous case, but it is shrunk, and the maximum of z_1 still appears at $t = 0$. In the case $\beta > 0$, $t \geq 0$ and $\pi (x)$ has a minimum at $x = \mu$, or $t = 0$. The case with $\alpha > 0$ is equivalent to the case with $\alpha < 0$ when $\beta < 0$ and similarly, the case with $\alpha < 0$ is equivalent to the case with $\alpha > 0$ for $\beta < 0$.

Elfving (1952) provides a method to derive c-optimal design points and weights from the geometry of Z. The procedure is as follows. Take the set Z, its reflection through the origin, $-Z$, and then determine the convex hull generated by Z and $-Z$. Let Z^* denote the boundary of this convex hull. Draw the vector $\mathbf{c}$ to find out where it perforates the boundary of the convex hull. Stretching this vector may be needed if it does not reach Z^* in the first place. The point of intersection, $\mathbf{c}^*$, either lies in Z or the part of the convex hull that connects Z and $-Z$. If the point is in Z the design point is obtained directly. Otherwise it is a convex combination of some extreme points of either Z or $-Z$. If the point is in $-Z$, $-\mathbf{z}$ will be taken as the design point. Elfving also showed that once the design points are determined to be $\mathbf{z}_i$, $i = 1, ..., n$, the optimal design weights are obtained as

$$\textstyle\sum_i w_i \mathbf{z}_i = \mathbf{c}^*, \quad w_i > 0, \quad \sum_i w_i = 1.$$

Here, the convex hull that encloses the curves Z and $-Z$ comprises a three-dimensional body. The surface of this body is partly flat and partly curved. Two-dimensional plots of the different views generated by the three pairs of axes are given in Figure 3.7. In the plots the convex hull is completed

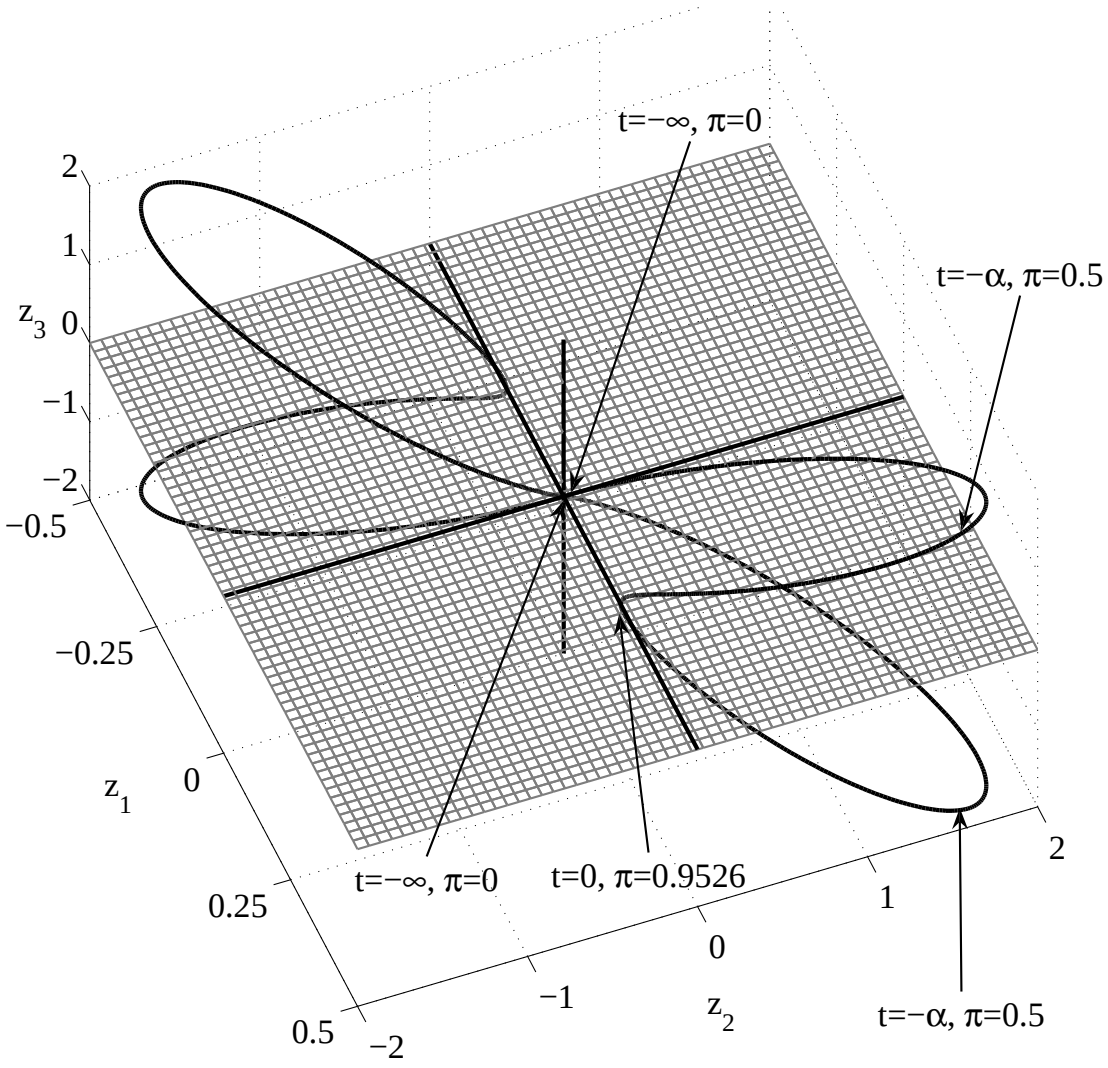


FIGURE 3.4: The canonical design space Z and -Z, $\boldsymbol{\theta}_A = (3, -1, 0)$.

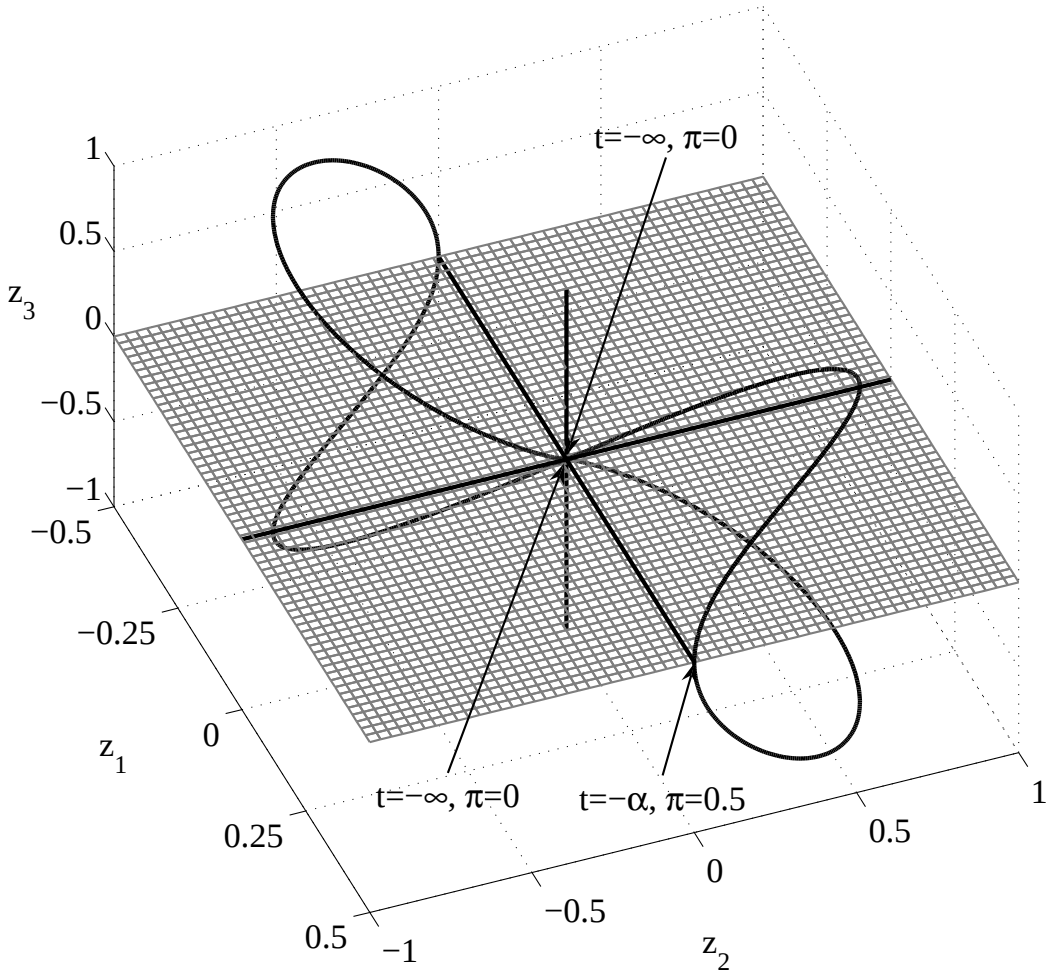


FIGURE 3.5: The canonical design space Z and -Z, $\boldsymbol{\theta}_B = (0, -1, 0)$.

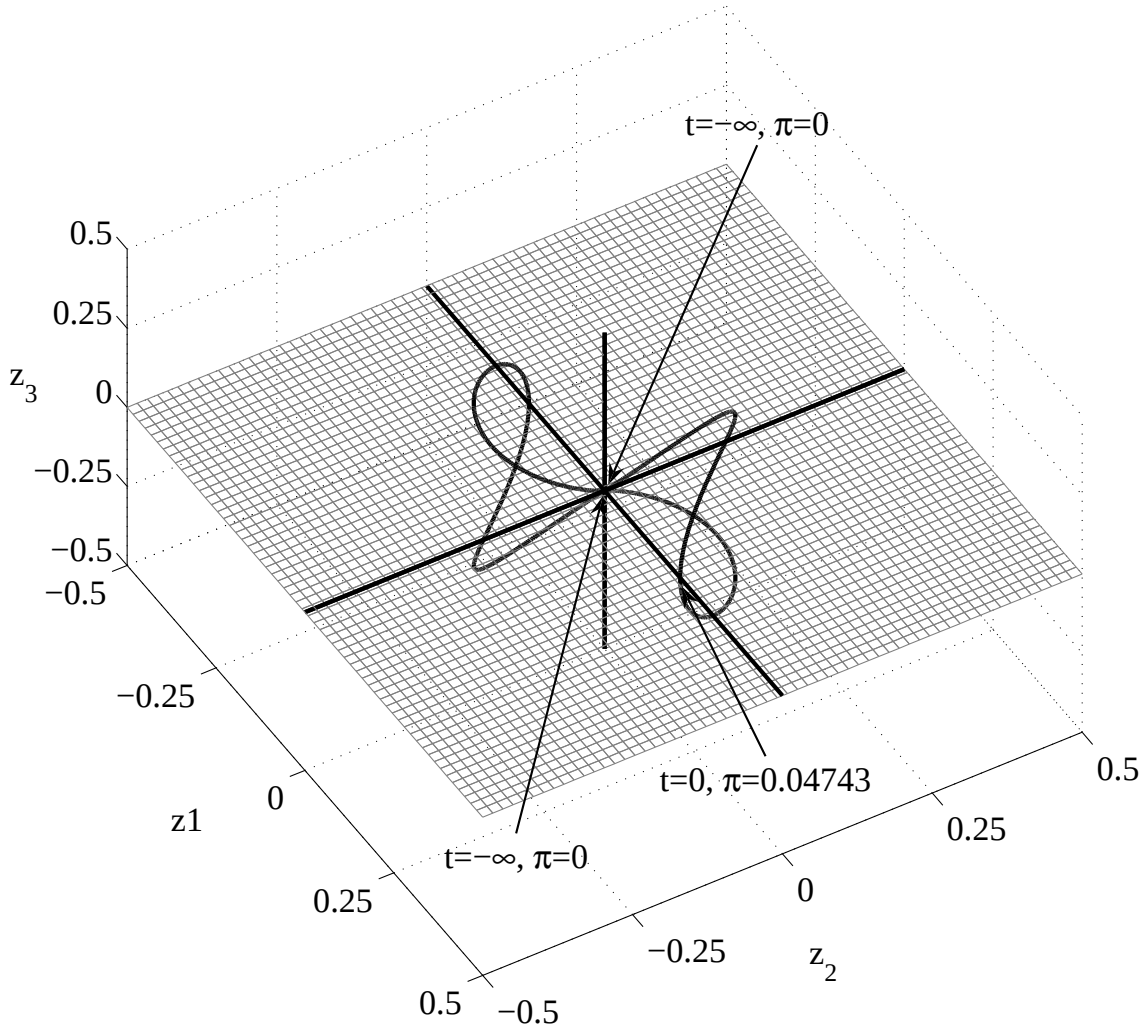


FIGURE 3.6: The canonical design space Z and -Z, $\boldsymbol{\theta}_C = (-3, -1, 0)$.

with dotted lines. For example, in the upper panel displaying the $z_1 - z_2$ viewpoint, the boundary of the convex hull is made up of two flat parts (the dotted lines) and two curved parts (where it coincides with the sets Z and $-Z$). The two flat parts are the contours of two planes, henceforth referred to as the top and bottom plane.

The graphs in Figure 3.7 are useful to identify the location of the optimal design points. The points of contact between the dotted lines and the solid lines are potentially critical points. The particular cases of estimating one of the model parameters are described below. Especially, estimation of μ, the point of maximum or minimum probability $\pi(x)$, often plays a crucial role in practical applications.

3.2.1 $\mathbf{c} = (1, 0, 0)'$

The c-optimal design for estimation of α, the parameter that controls the height of the response curve at the optimum point, proves to be a one-point design. This is clear from an examination of the canonical design space. Examples for the parameter sets $\boldsymbol{\theta}_A = (3, -1, 0)'$, $\boldsymbol{\theta}_B = (0, -1, 0)'$ and $\boldsymbol{\theta}_C = (-3, -1, 0)'$ are shown in Figure 3.9. The point of intersection between $\mathbf{c}$ and Z^* is in Z and appears at $t = 0$. The value of $\mathbf{c}^*$ at the intersection is found to be $\mathbf{c}^* = (\exp(\alpha/2)/(1 + \exp(\alpha)), 0, 0)'$. For model B this yields $\mathbf{c}^* = \mathbf{z} = (0.5, 0, 0)$ and the design that allocates all observations to $x = 0$. Similarly, for models A and C $\mathbf{c}^* = \mathbf{z} = (0.2125, 0, 0)'$, yielding one point designs at $x = 0$. In general, c-optimal estimation of α implies a one point design with observations taken at the point at which $t = 0$, that is at $x = \mu$,

$$\xi_\alpha^* = \left\{ \begin{array}{c} \mu \\ 1 \end{array} \right\}.$$

Taking observations at the optimum point to study the height of the response curve at this point goes well with intuition.

3.2.2 $\mathbf{c} = (0, 1, 0)'$

When estimating β, the parameter related to the width of the response curve, the optimal design is made up of either 3 or 4 design points. The height of the response curve determines whether it is 3 or 4 points. The canonical design space helps to explain when the number of design points is

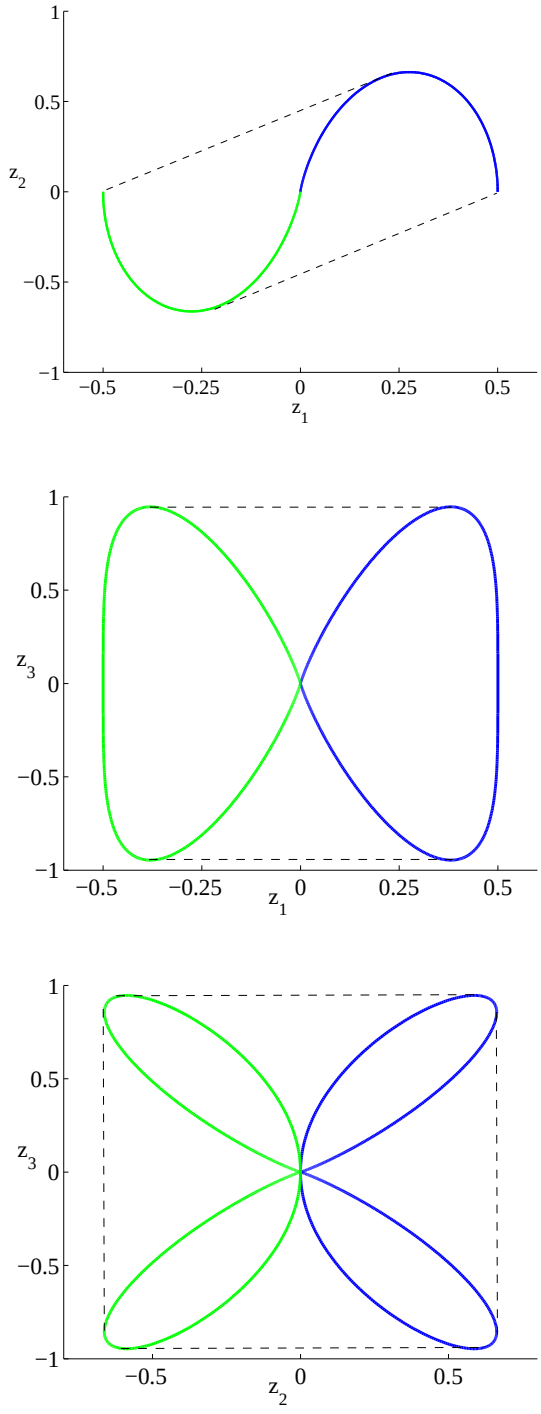


FIGURE 3.7: The sets Z and -Z with solid lines, the convex hull is completed with dotted lines for $\boldsymbol{\theta}_B = (0, -1, 0)$.

altered. Here the vector $\mathbf{c} = (0, 1, 0)'$ penetrates Z^* at the top plane. This plane has either 3 or 4 points of contact with Z and $-Z$, which thus yields 3 or 4 design points. Seen from the $z_1 - z_2$ viewpoint the curve Z starts at the origin and makes a loop before it reaches the z_1-axes, see Figure 3.10. The length of this loop is crucial for the number of contact points with the top plane. If the top plane hits $-Z$ at the symmetry point where $t = 0$ this gives us one design point. If the top plane does not reach this point there will be two points of contact with $-Z$. In both cases there are two contact points between Z and the top plane. The dividing line between 3 and 4 point designs occurs when the loop is long enough, such that $z_1(0) < 0.2762$ (approximately) and there are two values of z_2 for $z_1(0)$. This occurs when $\alpha > 2.4$ (approximately) for $\beta < 0$ and $\alpha < -2.4$ (approximately) for $\beta > 0$. The derivations of the design points for the case $\beta < 0$ are described in detail below. Similar calculations apply to the case $\beta > 0$.

i) $\alpha > 2.4$**, 4 design points**

In the left panel of Figure 3.10 the canonical design space for $\boldsymbol{\theta}_A = (3, -1, 0)'$ is depicted. The 4 design points $\mathbf{z}_1, ..., \mathbf{z}_4$ are marked in the graph. These points are symmetric around the point of optimum response μ. The design points are taken at, say $t = \pm t_{1,\beta}$ and $t = \pm t_{2,\beta}$ or equivalently at $x_1 = \mu \pm \sqrt{|t_{1,\beta}/\beta|}$ and $x_2 = \mu \pm \sqrt{|t_{2,\beta}/\beta|}$. The slope of the plane along the z_1-direction is given by $z_2(t_{\max})/z_1(t_{\max})$ where $t_{\max}$ is the value of t that maximizes z_1. Since $z_2(t_{\max}) = (t_{\max}/\beta)\, z_1(t_{\max})$ and $t_{\max} = -\alpha$ the slope is $-\alpha/\beta$. The derivative of the curve Z^* at the tangent point with the plane is $(dz_2/dt)\,/\,(dz_1/dt)$, evaluated at $t_{1,\beta}$ (or $t_{2,\beta}$). Equating this expression to the slope yields the equation

$$t_\beta = -\alpha - 2\left(\frac{1 + e^{\alpha + t_\beta}}{1 - e^{\alpha + t_\beta}}\right), \tag{3.4}$$

from which both $t_{1,\beta}$ and $t_{2,\beta}$ can be solved. The resulting c-optimal design for estimating β is then

$$\xi^*_\beta = \left\{ \begin{array}{cccc} \mu - \sqrt{|t_{2,\beta}/\beta|} & \mu - \sqrt{|t_{1,\beta}/\beta|} & \mu + \sqrt{|t_{1,\beta}/\beta|} & \mu + \sqrt{|t_{2,\beta}/\beta|} \\ 0.25 & 0.25 & 0.25 & 0.25 \end{array} \right\}.$$

For the parameters $\boldsymbol{\theta}_A = (3, -1, 0)'$ Equation (3.4) is

$$t_\beta = -3 - 2\left(\frac{1 + e^{3 + t_\beta}}{1 - e^{3 + t_\beta}}\right),$$

the solutions of which, found numerically, are $t_{1,\beta} = -5.3994$ and $t_{2,\beta} = -0.6006$. A c-optimal design for estimating β is therefore in this case the design

$$\xi^*_\beta = \left\{ \begin{array}{cccc} -2.3239 & -0.7750 & 0.7750 & 2.3239 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{array} \right\}.$$

ii) $\alpha < 2.4$**, 3 design points**

The middle and left panels of Figure 3.10 show $\boldsymbol{\theta}_B = (0, -1, 0)'$ and $\boldsymbol{\theta}_C = (-3, -1, 0)'$ and the associated 3 point designs. The top plane now connects with $-Z$ at the symmetry point $-z_1(0)$ where $t = 0$ and $x = \mu$. It has also two symmetric tangent points with Z at, say $t = \pm t_\beta$, or equivalently at $x = \mu \pm \sqrt{|t_\beta/\beta|}$. A determination of t_β from Z^* can be found by noting that the slope of the plane along the z_1-direction is $z_2(t_\beta)/\left(z_1(0) + z_1(t_\beta)\right)$ and the derivative of the curve Z^* at the tangent points is $(dz_2/dt)/(dz_1/dt)$ evaluated at t_β. By equating these two expressions, the following expression is obtained

$$t_\beta = -2\left(\frac{1 + e^{\alpha + t_\beta}}{1 - e^{\alpha + t_\beta}}\right)\left(\frac{z_1(t_\beta)}{z_1(0)} + 1\right)$$

from which t_β can be solved. The resulting c-optimal design for estimating β is then

$$\xi^*_\beta = \left\{ \begin{array}{ccc} \mu - \sqrt{|t_\beta/\beta|} & \mu & \mu + \sqrt{|t_\beta/\beta|} \\ w & 1 - 2w & w \end{array} \right\}.$$

The weight w is determined from

$$w \begin{pmatrix} z_1(t_\beta) \\ z_2(t_\beta) \\ z_3(t_\beta) \end{pmatrix} + (1 - 2w) \begin{pmatrix} -z_1(0) \\ 0 \\ 0 \end{pmatrix} + w \begin{pmatrix} z_1(t_\beta) \\ z_2(t_\beta) \\ -z_3(t_\beta) \end{pmatrix} = k \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix},$$

for some constant k. Hence, it is deduced

$$w = \frac{1}{2}\frac{z_1(0)}{z_1(0) + z_1(t_\beta)}.$$

For the case $\boldsymbol{\theta}_B = (0, -1, 0)'$ the equation for t_β is

$$t_\beta = -2\frac{2e^{t_\beta/2} + e^{t_\beta} + 1}{e^{t_\beta} - 1}.$$

The solution, found numerically, is $t_\beta = 3.087$ yielding the design points ± 1.757 and 0. The design weight w is finally obtained as

$$w = \frac{1}{2}\frac{e^0/(1+e^0)}{e^0/(1+e^0) + e^{3.087}/(1+e^{3.087})} \approx 0.355.$$

The resulting c-optimal design is therefore

$$\xi^*_\beta = \left\{ \begin{array}{ccc} -1.757 & 0 & 1.757 \\ 0.355 & 0.290 & 0.355 \end{array} \right\}.$$

3.2.3 $\mathbf{c} = (0, 0, 1)'$

The geometry of Z^* in Figure 3.11 reveals that a two-point design is c-optimal for estimation of μ. Since $\mathbf{c}$ perforates Z^* on the part joining two extreme points of Z and $-Z$, both of these make up the design points. The points are obtained from maximizing $z_3(t)$, which is a solution to

$$t_\mu = \frac{e^{\alpha+t_\mu}+1}{e^{\alpha+t_\mu}-1}.$$

It should be noted that this equation may have two solutions and it is important to select the solution with the same sign as β since $(x-\mu)^2$ is non-negative. The extreme points at Z and $-Z$ have the coordinates $\mathbf{z}_1 = (z_1(t_\mu), z_2(t_\mu), z_3(t_\mu))'$ and $\mathbf{z}_2 = (-z_1(t_\mu), -z_2(t_\mu), z_3(t_\mu))'$, respectively. Furthermore, it is apparent from the graph that the optimal design weights are $w_1 = w_2 = 0.5$ which can be verified by

$$\textstyle\sum_i w_i \mathbf{z}_i = 0.5\mathbf{z}_1 + 0.5\mathbf{z}_2 = 0.5 \begin{pmatrix} z_1(t_\mu) \\ z_2(t_\mu) \\ z_3(t_\mu) \end{pmatrix} + 0.5 \begin{pmatrix} -z_1(t_\mu) \\ -z_2(t_\mu) \\ z_3(t_\mu) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ z_3(t_\mu) \end{pmatrix} = \mathbf{c}^*.$$

The c-optimal design for estimating μ is therefore

$$\xi^*_\mu = \left\{ \begin{array}{cc} \mu - \sqrt{t_\mu/\beta} & \mu + \sqrt{t_\mu/\beta} \\ 0.5 & 0.5 \end{array} \right\}.$$

Note here that if t_μ has the same sign as β the ratio t_μ/β is positive.

For the case A with $\boldsymbol{\theta}_A = (3, -1, 0)'$

$$t_\mu = \frac{e^{3+t_\mu}+1}{e^{3+t_\mu}-1}$$

with the solution $t_\mu = -3.5748$ and, translating the canonical design points back to the corresponding x-values yields the optimal design

$$\xi_A^* = \left\{ \begin{array}{cc} -1.8907 & 1.8907 \\ 0.5 & 0.5 \end{array} \right\}.$$

For the cases B and C with $\boldsymbol{\theta}_B = (0, -1, 0)'$ and $\boldsymbol{\theta}_C = (-3, -1, 0)'$ similar calculations yield $t_\mu = -1.5434$ and $t_\mu = -1.0360$, respectively. The corresponding designs are

$$\xi_B^* = \left\{ \begin{array}{cc} -1.2423 & 1.2423 \\ 0.5 & 0.5 \end{array} \right\}$$

and

$$\xi_C^* = \left\{ \begin{array}{cc} -1.0178 & 1.0178 \\ 0.5 & 0.5 \end{array} \right\}.$$

3.2.4 Conclusions

Analyzing the geometry of the canonical design space leads to the following conclusions. The number of design points in a c-optimal design for estimating the model parameters varies depending on which parameter to be estimated. One design point is required when estimating the height of the response curve at the maximum or minimum. For estimating the parameter β, which is related to the width of the response curve, either three or four points are required. For a response curve with a maximum, a c-optimal design has four points if the maximum is large enough (probability for response is greater than 0.917 at the maximum), while it has three points otherwise. Analogous results are valid when the response curve has a minimum. For estimating the location of the maximum, two points are required. Furthermore, formulae for finding the design points along with the corresponding design weights have been derived. While the results here focus on the logistic regression model, it would not be difficult to use the same methodology for extending the results to other binary response models.

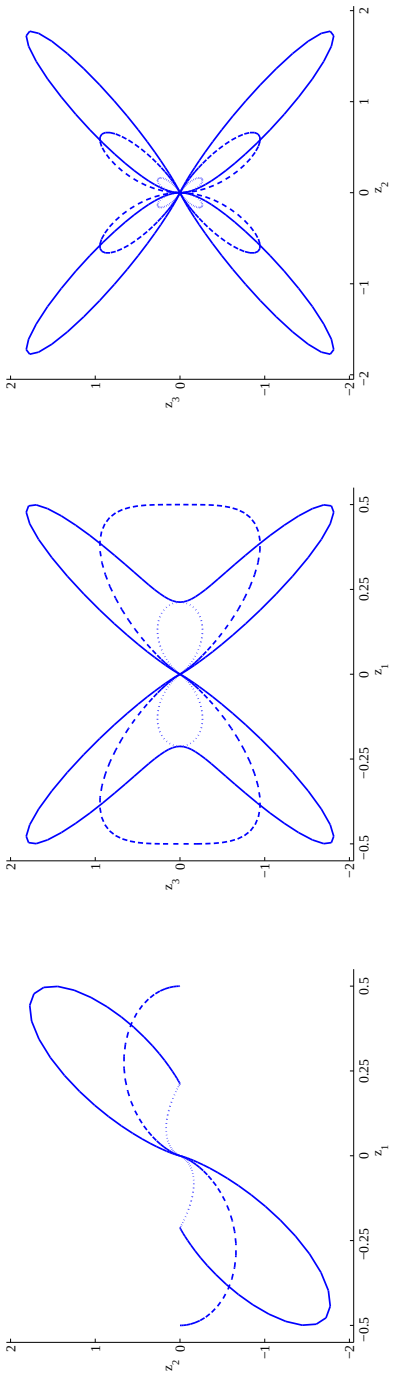
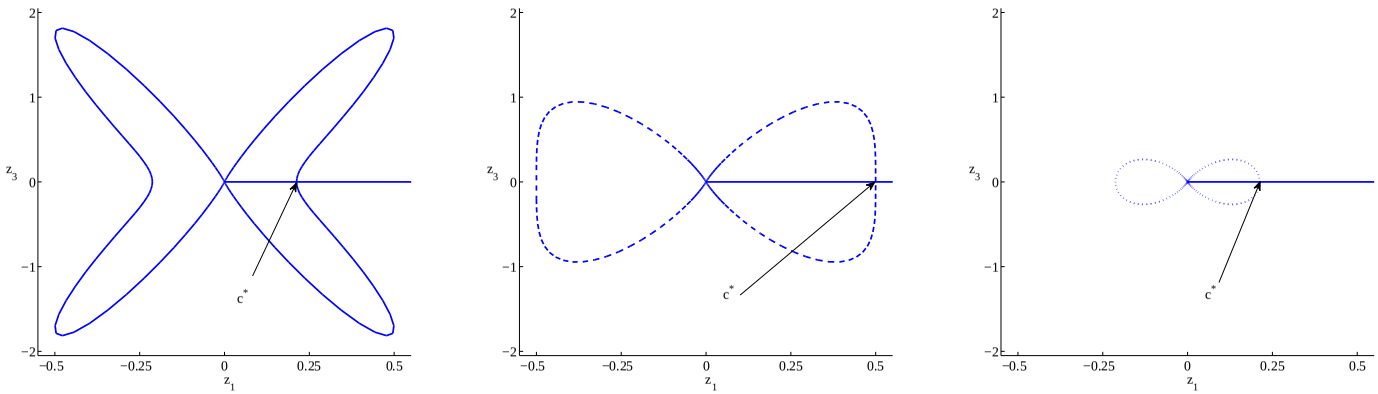
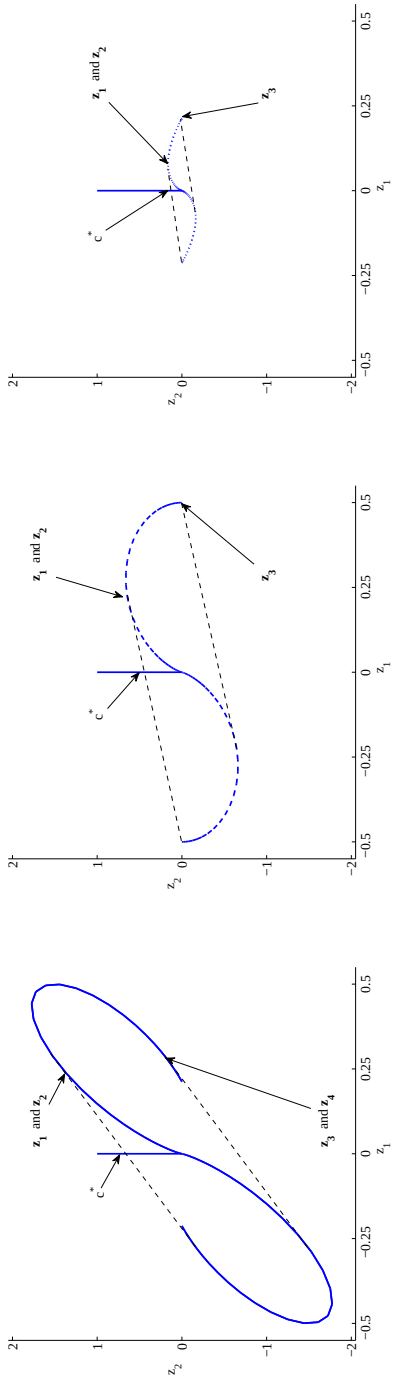


FIGURE 3.8: The sets Z and -Z for $\boldsymbol{\theta}_{\mathbf{A}} = (3, -1, 0)$ (solid line), $\boldsymbol{\theta}_{\mathbf{B}} = (0, -1, 0)$ (dashed line) and $\boldsymbol{\theta}_{\mathbf{C}} = (-3, -1, 0)$ (dotted line)

FIGURE 3.9: The canonical design space, the vector $c = (1, 0, 0)$ and its intersection with Z^*



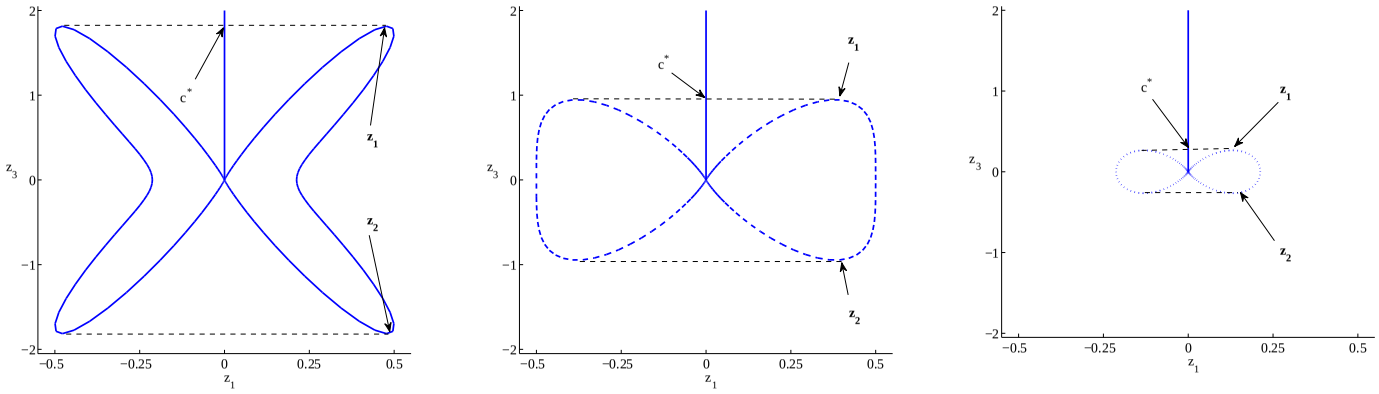
(A) $\boldsymbol{\theta}_{\mathbf{A}} = (3, -1, 0)$ (B) $\boldsymbol{\theta}_{\mathbf{B}} = (0, -1, 0)$ (C) $\boldsymbol{\theta}_{\mathbf{C}} = (-3, -1, 0)$



(A) $\boldsymbol{\theta}_{\mathbf{A}} = (3, -1, 0)$ (B) $\boldsymbol{\theta}_{\mathbf{B}} = (0, -1, 0)$ (C) $\boldsymbol{\theta}_{\mathbf{C}} = (-3, -1, 0)$

FIGURE 3.10: The canonical design space, the vector $c = (0, 1, 0)$ and its intersection with Z^*

FIGURE 3.11: The canonical design space, the vector $c = (0, 0, 1)$ and its intersection with Z^*



(A) $\boldsymbol{\theta}_{\mathbf{A}} = (3, -1, 0)$ (B) $\boldsymbol{\theta}_{\mathbf{B}} = (0, -1, 0)$ (C) $\boldsymbol{\theta}_{\mathbf{C}} = (-3, -1, 0)$

Chapter 4

Parameter Estimation in Small Samples

D-optimality is based on asymptotic properties of the maximum likelihood estimator. In practice, sample sizes are often small due to time and money constraints raising the question how these designs will work in a small-sample setting. Furthermore, the optimal designs depend on the true parameters which are unknown, sometimes making it challenging to actually obtain an optimal design. A simulation study is performed with the purpose to address these issues and the results are reported in this chapter. The properties of the maximum likelihood (ML) estimator are examined for various sample sizes. A couple of non-optimal designs are compared to the optimal design in terms of mean squared error. However, situations may occur where no ML estimate exists for one or more parameters, primarily in small samples. The probabilities to encounter non-existing ML estimates are compared for the different designs. Another estimator, based on a modification of the score function introduced by Firth (1993), is evaluated as a solution to this problem for standard logistic regression by Heinze and Schemper (2002). This estimator, denoted by FL, is also tested in the simulations. This chapter is a development of the paper by Fackle Fornius (2007).

4.1 The ML Estimator

For the quadratic logistic regression model the ML estimates are obtained as the solutions to the score equations $U_j = \partial l/\partial\theta_j = 0,\ j = 1, 2, 3,$

$$\begin{aligned}
\frac{\partial l}{\partial \alpha} &= \sum_{i=1}^{N} (y_i - \pi_i) = 0 \\
\frac{\partial l}{\partial \beta} &= \sum_{i=1}^{N} (y_i - \pi_i)(x_i - \mu_i)^2 = 0 \\
\frac{\partial l}{\partial \mu} &= \sum_{i=1}^{N} -2\beta (y_i - \pi_i)(x_i - \mu) = 0.
\end{aligned}$$

The solutions to the score equations can be obtained iteratively by the method of Fisher scoring, which was introduced in Chapter 2. Let $\widetilde{\mathbf{X}}_N$ denote a $N \times p$ matrix with rows

$$\widetilde{\mathbf{x}}_i' = \left(\frac{\partial \eta_i}{\partial \boldsymbol{\theta}}\right),$$

then the Fisher information matrix for this model can be expressed as

$$\mathbf{I}(\xi, \boldsymbol{\theta}) = \widetilde{\mathbf{X}}_N' \mathbf{V}_N \widetilde{\mathbf{X}}_N.$$

Given a previous estimate, $\boldsymbol{\theta}^{(m-1)}$, the estimate $\boldsymbol{\theta}^{(m)}$ is given by

$$\boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^{(m-1)} + \left[\mathbf{I}^{(m-1)}\right]^{-1} \mathbf{U}^{(m-1)}. \tag{4.1}$$

where $\mathbf{I}^{(m-1)}$ is the Fisher information matrix and $\mathbf{U}^{(m-1)}$ is the vector of scores (U_j), both evaluated at $\boldsymbol{\theta}^{(m-1)}$. Alternatively, equation (4.1) can be written as

$$\boldsymbol{\theta}^{(m)} = \left(\widetilde{\mathbf{X}}_N' \mathbf{V}_N \widetilde{\mathbf{X}}_N\right)^{-1} \widetilde{\mathbf{X}}_N' \mathbf{V}_N \mathbf{z}. \tag{4.2}$$

Note that $\widetilde{\mathbf{X}}_N$ and the GLM weight matrix $\mathbf{V}_N$ are evaluated at $\boldsymbol{\theta}^{(m-1)}$. The vector $\mathbf{z}$ is of dimension $N \times 1$ with elements

$$z_i = \widetilde{\mathbf{x}}_i' \boldsymbol{\theta}^{(m-1)} + \frac{\left(y_i - \pi_i^{(m-1)}\right)}{\left[\pi_i^{(m-1)}\left[1 - \pi_i^{(m-1)}\right]\right]}.$$

The ML estimates $\widehat{\boldsymbol{\theta}}$ can now be obtained by iteratively updating equation (4.2) until the relative difference between successive estimates $\boldsymbol{\theta}^{(m)}$ and $\boldsymbol{\theta}^{(m-1)}$ falls below a convergence criterion (e.g. $1e-5$).

The asymptotic sampling distribution of $\sqrt{N}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)$ is normal with mean $\mathbf{0}$ and covariance matrix

$$V\left(\widehat{\boldsymbol{\theta}}\right)=\left(\widetilde{\mathbf{X}}_n'\mathbf{V}_n\mathbf{W}_n\widetilde{\mathbf{X}}_n\right)^{-1}.$$

That is, the ML estimator is a consistent and asymptotically efficient estimator of $\boldsymbol{\theta}$. Note that different designs will lead to different asymptotic sampling distributions of the ML estimator. Variance estimates can be obtained by the diagonal elements in

$$\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)=\left(\widetilde{\mathbf{X}}_n'\widehat{\mathbf{V}}_n\mathbf{W}_n\widetilde{\mathbf{X}}_n\right)^{-1}$$

where $\widehat{\mathbf{V}}_n$ denotes the GLM weight matrix $\mathbf{V}_n$ evaluated at $\widehat{\boldsymbol{\theta}}$.

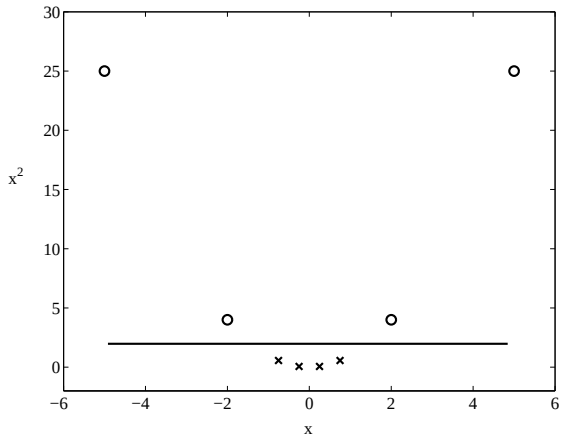
Whether ML estimates exist depends on the pattern of the data points in the sample. A data set can be categorized as belonging to one of three types of data configurations; complete separation, quasi-complete separation or overlap, as described in Albert and Anderson (1984). It is only when data belong to the third configuration that the ML estimate is finite and unique. The responses are binary and the data can thus be divided into two response groups, one including the points where $Y_i=1$ and one including the points where $Y_i=0$. If there is a vector that correctly allocates all observations to their respective response group complete separation is present, that is, if there is a vector $\boldsymbol{\alpha}$ so that $\boldsymbol{\alpha}'\mathbf{x}_i>0$ for all $Y_i=1$ and $\boldsymbol{\alpha}'\mathbf{x}_i<0$ for all $Y_i=0$. Quasi-complete separation occurs when there is a vector $\boldsymbol{\alpha}$ such that $\boldsymbol{\alpha}'\mathbf{x}_i\geq 0$ for all $Y_i=1$ and $\boldsymbol{\alpha}'\mathbf{x}_i\leq 0$ for all $Y_i=0$. If the data configuration is neither complete separation nor quasi-complete separation the data points are overlapped and the ML estimate exists and is unique.

Example. Consider an 8-point design defined as

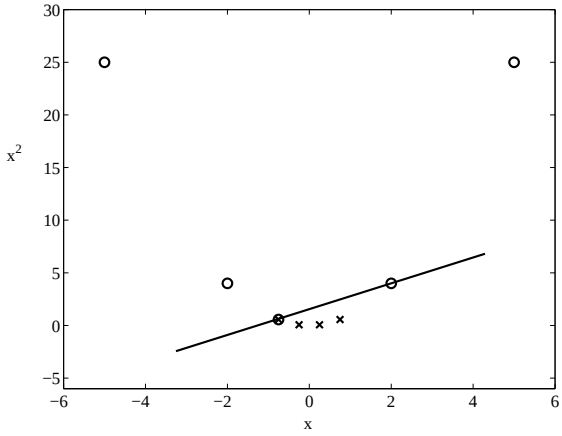
$$\xi_8=\left\{\begin{array}{cccccccc} -5 & -2 & -0.75 & -0.25 & 0.25 & 0.75 & 2 & 5 \\ 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 \end{array}\right\}.$$

Examples on the three types of data patterns are shown in Figure 4.1 for this 8-point design. A ring symbolizes that zeroes are observed at the design

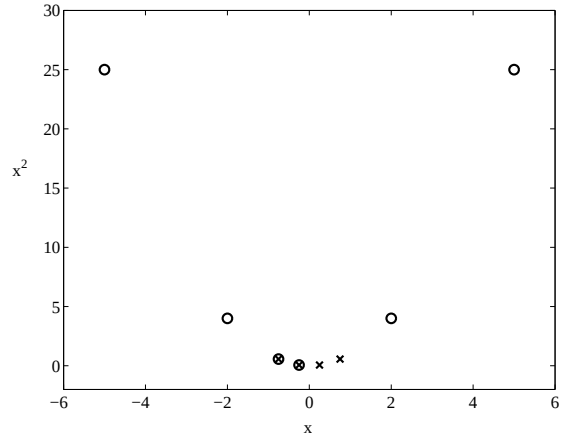
point, a cross symbolizes that ones are observed. In panel A there are only zeroes at the points ± 2 and ± 5 and only ones at the rest of the design points. As can be seen there is a line that completely separates the points ± 2 and ± 5 from the other points. As a contrast, in panel C, no such line can be found. Panel C shows an example of overlap, which is desired in order to obtain existing ML estimates.



(A) Complete separation



(B) Quasi-complete separation



(C) Overlap

FIGURE 4.1: Examples on data configurations for the 8-point design, if there are only zeros observed at a point a ring is displayed, if there are only ones observed a cross is displayed and if both response types are present a cross in a ring is displayed.

4.2 The FL Estimator

Firth (1993) introduced a modification to the score function as a means to reduce the small-sample bias of maximum likelihood estimates. The resulting estimator is proposed as an alternative to the ML estimator. Firth advocated using this alternative estimator directly instead of using corrective methods (i.e. to correct the ML estimate after it is computed). For instance, in the event of non-existing estimates corrective methods are not feasible.

The modification of the score function stems from imposing a "penalty" term to the standard likelihood function. The so called penalized likelihood function is defined as

$$L^{*}\left(\boldsymbol{\theta};\mathbf{y}\right)=L\left(\boldsymbol{\theta};\mathbf{y}\right)\left|\mathbf{I}\left(\xi,\boldsymbol{\theta}\right)\right|^{1/2},$$

with corresponding log-likelihood

$$l^{*}=l+0.5\ln\left|\mathbf{I}\left(\xi,\boldsymbol{\theta}\right)\right|.$$

The effect of the additional term is asymptotically insignificant. The modified score function is given by

$$\mathbf{U}^{*}=\mathbf{U}+0.5tr\left\{\mathbf{I}\left(\xi,\boldsymbol{\theta}\right)^{-1}\mathbf{D}\right\},$$

where the matrix $\mathbf{D}$ has elements

$$D_{ij}=\frac{\partial\mathbf{I}\left(\xi,\boldsymbol{\theta}\right)_{ij}}{\partial\theta_{j}}.$$

For the quadratic logistic model it yields the modified score function

$$\mathbf{U}^{*}=\sum_{i=1}^{N}\left[\left(y_{i}-\pi_{i}\right)\left(\frac{\partial\eta_{i}}{\partial\boldsymbol{\theta}_{j}}\right)+h_{i}\left(0.5-\pi_{i}\right)\right]$$

That is, the score equations are now given by

$$\begin{aligned}
\frac{\partial l}{\partial\alpha} &= \sum_{i=1}^{N}\left(y_{i}-\pi_{i}+h_{i}\left(0.5-\pi_{i}\right)\right)=0\\
\frac{\partial l}{\partial\beta} &= \sum_{i=1}^{N}\left(y_{i}-\pi_{i}+h_{i}\left(0.5-\pi_{i}\right)\right)\left(x_{i}-\mu_{i}\right)^{2}=0\\
\frac{\partial l}{\partial\mu} &= \sum_{i=1}^{N}-2\beta\left(y_{i}-\pi_{i}+h_{i}\left(0.5-\pi_{i}\right)\right)\left(x_{i}-\mu\right)=0.
\end{aligned}$$

The h_i terms are the diagonal elements of the "hat" matrix $\mathbf{H}$

$$h_i = diag\left(\mathbf{H}\right) = diag\left(\mathbf{V}_N^{1/2}\left(\widetilde{\mathbf{X}}_N'\mathbf{V}_N\widetilde{\mathbf{X}}_N\right)^{-1}\widetilde{\mathbf{X}}_N'\mathbf{V}_N^{1/2}\right).$$

The addition of $h_i\left(0.5-\pi_i\right)$ to the score equations in effect causes the FL estimates to be shrunk towards zero, compared to the standard ML estimates. Most importantly, it solves the problem of separation, i.e. the estimates are guaranteed to exist for binary response models, see Firth (1993). FL estimates, $\widehat{\boldsymbol{\theta}}^*$, can now be found iteratively as described in the preceding section, taking

$$\boldsymbol{\theta}^{*(m)} = \boldsymbol{\theta}^{*(m-1)} + \left[\mathbf{I}^{(m-1)}\right]^{-1}\mathbf{U}^{*(m-1)}$$

as the starting point instead of (4.1). Variance estimates are accordingly found as

$$\widehat{V}\left(\widehat{\boldsymbol{\theta}}^*\right) = \left(\widetilde{\mathbf{X}}_n'\widehat{\mathbf{V}}_n\mathbf{W}_n\widetilde{\mathbf{X}}_n\right)^{-1}$$

where the GLM weight matrix $\widehat{\mathbf{V}}_n$ here is evaluated at $\widehat{\boldsymbol{\theta}}^*$.

4.3 Simulation Setup

The aims of the simulation study are to compare D-optimal designs to a couple of non-optimal designs and to evaluate the performance of the ML estimator, both in terms of probability of existence and mean squared error. In addition, the performance of the ML estimator is compared to the performance of the FL estimator.

Four sets of parameters are considered in the simulations, these are given in Table 4.1 and displayed in Figure 4.2. The sets are chosen to represent different variations of the shape of the response curve. The response curve is considered to be high when the maximum value of π is close to 1 and low when the maximum value is close to 0. Given the scale on x the response curves have different widths. Two of them are called wide and two are called narrow. The meaning of these labels should be understood in a relative sense.
For each of the four sets of true parameters three designs are considered. One is the D-optimal design which depends on the parameters and thus

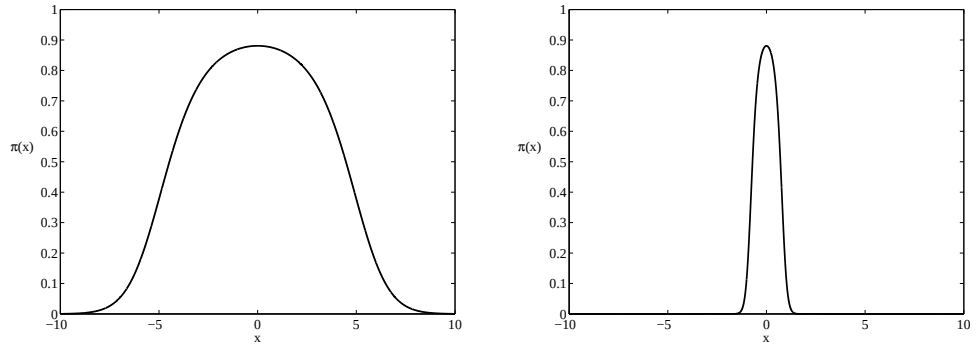
TABLE 4.1: Four parameter sets labeled according to their associated characteristics of $\pi(x)$.

TYPE OF RESPONSE CURVE	PARAMETER SET
"HIGH-WIDE"	$\boldsymbol{\theta} = (2, -0.1, 0)'$
"HIGH-NARROW"	$\boldsymbol{\theta} = (2, -4, 0)'$
"LOW-WIDE"	$\boldsymbol{\theta} = (-2, -0.1, 0)'$
"LOW-NARROW"	$\boldsymbol{\theta} = (-2, -4, 0)'$

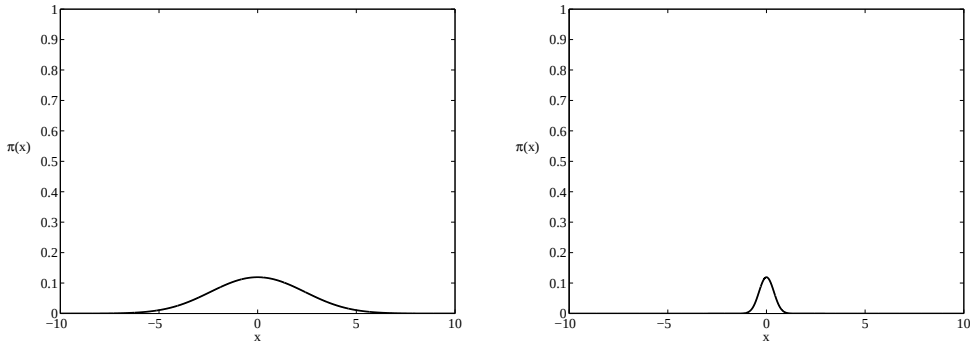
is different in all four cases. In practise, when the true parameters are unknown, a strategy might be to include more than the optimal number of design points to increase the chance that some points are good. Two non-optimal designs composed of 7 (ξ_7) and 8 (ξ_8) design points are compared to the D-optimal designs. The 8-point design is symmetric around the optimum point whereas the 7-point design is not. All designs are given in Table 4.2.

For every combination of true parameter values and design four sample sizes are considered, two small samples $N = 10$ and $N = 20$ and two larger, $N = 50$ and $N = 100$. The exact sampling distributions of the ML and FL estimators are obtained for the small sample sizes by generating all possible samples, the parameter estimates for each sample are then weighted with the probability of obtaining the current sample. For the larger sample sizes simulations are performed instead because the number of possible samples grows very large. The simulation results are based on 10000 runs. All calculations are performed in Matlab.

The proportions of the sample to be allocated to the design points are given by the design weights. The number of observations to be taken at each design point are thus given by $r_i = w_i N$. However, adjustments has to be made to r_i because the number of observations need to be integer values. The resulting designs will then be approximations to the designs given in Table 4.2. In Table 4.3 the number of observations per design point are shown.



(A) $\boldsymbol{\theta} = (2, -0.1, 0)'$ (B) $\boldsymbol{\theta} = (2, -4, 0)'$



(C) $\boldsymbol{\theta} = (-2, -0.1, 0)'$ (D) $\boldsymbol{\theta} = (-2, -4, 0)'$

FIGURE 4.2: The response curves corresponding to the parameter sets given in Table 4.1.

TABLE 4.3: Number of observations taken at each design point.

N	3-POINT DESIGN	4-POINT DESIGN	7-POINT DESIGN	8-POINT DESIGN
10	3/4/3	3/2/2/3	1/2/1/2/1/2/1	2/1/1/1/1/1/1/2
20	7/6/7	6/4/4/6	3/3/3/2/3/3/3	2/3/2/3/2/3/2/3
50	17/16/17	16/9/9/16	7/7/7/8/7/7/7	7/6/6/6/6/6/6/7
100	33/34/33	31/19/19/31	14/14/15/14/15/14/14	12/13/12/13/12/13/12/13

TABLE 4.2: The designs used in the simulations.

PARAMETER SET	DESIGN	
"HIGH-WIDE" $\boldsymbol{\theta} = (2, -0.1, 0)'$	D-OPTIMAL	$\xi^* = \left\{ \begin{array}{cccc} -5.7185 & -2.7017 & 2.7017 & 5.7185 \\ 0.3138 & 0.1862 & 0.1862 & 0.3138 \end{array} \right\}$
"HIGH-NARROW" $\boldsymbol{\theta} = (2, -4, 0)'$	D-OPTIMAL	$\xi^* = \left\{ \begin{array}{cccc} -0.9042 & -0.4272 & 0.4272 & 0.9042 \\ 0.3138 & 0.1862 & 0.1862 & 0.3138 \end{array} \right\}$
"LOW-WIDE" $\boldsymbol{\theta} = (-2, -0.1, 0)'$	D-OPTIMAL	$\xi^* = \left\{ \begin{array}{ccc} -3.9819 & 0 & 3.9819 \\ 1/3 & 1/3 & 1/3 \end{array} \right\}$
"LOW-NARROW" $\boldsymbol{\theta} = (-2, -4, 0)'$	D-OPTIMAL	$\xi^* = \left\{ \begin{array}{ccc} -0.6296 & 0 & 0.6296 \\ 1/3 & 1/3 & 1/3 \end{array} \right\}$
ALL	7-POINT	$\xi_7 = \left\{ \begin{array}{ccccccc} -4 & -1 & 0.5 & 1 & 1.5 & 3 & 6 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \end{array} \right\}$
ALL	8-POINT	$\xi_8 = \left\{ \begin{array}{cccccccc} -5 & -2 & -0.75 & -0.25 & 0.25 & 0.75 & 2 & 5 \\ 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 \end{array} \right\}$

Some issues emerged during the simulations which called for a decision. It should be remembered that these decisions affect the simulation results. If the sequence of ML estimates does not converge, indicating non-existence, it is rather apparent that such an estimate is failed. Even if the estimation algorithm converge, the estimate may not be useful for practical purposes. If a quadratic response curve with a maximum is assumed and the parameter estimates indicate a minimum ($\widehat{\beta} > 0$), or the estimate of the point of maximum response ($\widehat{\mu}$) is degenerated, not much information is gained. For these reasons such parameter estimates are considered unusable and are excluded if any of the following events occur.

- The estimation algorithm does not converge within the iteration limit (set to 100 iterations) which is a sign that no finite parameter estimates exist. Note that it can only come into question for the ML estimates.
- The estimate $\widehat{\beta} > 0$, then the estimated response curve has a minimum instead of a maximum at $\widehat{\mu}$.
- The estimate $\widehat{\mu}$ is degenerated (the cutoff was set to $|\widehat{\mu}| > 10$). In

such cases the estimated point of maximum response is even outside the interval of non-zero response of the true response curve, where the probability of response is effectively zero.

4.4 Simulation Results

4.4.1 Non-existence of the ML Estimates

The problem of separation resulting in non-existence for the ML estimator occurs for the quadratic logistic model considered here. To what extent depends on the sample size, the number of design points and the true parameters. When the sample size is $N = 10$ divided among three points the number of possible samples is equal to

$$(r_1 + 1) \cdot (r_2 + 1) \cdot (r_3 + 1) = 4 \cdot 5 \cdot 4 = 80.$$

The only way for the data to be overlapped and thereby for the ML estimate to exist is to have both response types ($Y = 1$ and $Y = 0$) observed at all three points. At a design point where three observations are taken there are two variants where both response types are represented; 1 or 2 ones out of 3. When four observations are taken at a design point there are three such cases; 1, 2 or 3 ones out of 4. In total there are $2 \cdot 3 \cdot 2 = 12$ distinct samples where the ML estimate exists. Let Z_i be the number of ones observed at design point x_i, then $Z_i \sim bin(r_i, \pi_i)$ where $\pi_i = e^{\eta_i}/(1 + e^{\eta_i})$. The probability that the ML estimate exists, $P(ML)$ can now be computed as

$$P(ML) = [P(Z_1 = 1) + P(Z_1 = 2)] \cdot [P(Z_2 = 1) + P(Z_2 = 2) + P(Z_2 = 3)] \cdot [P(Z_3 = 1) + P(Z_3 = 2)].$$

This probability is shown for all combinations of design and true parameters for $N = 10$ in section a) of Table 4.4. The probability to obtain a ML estimate is essentially zero for the "low-narrow" model irrespective of what design is used. It is not possible to estimate the parameters with only 10 observations in this case. For the "low-wide" model this probability is at best around 10 percent when one of the two non-optimal designs are used. The reason why the non-optimal designs perform better in this respect is that the D-optimal design consists of only 3 points which makes it more difficult to obtain overlapped data compared to the designs with more points.

This is particularly the case for the wide models because for these models $P(Y = 1) > 0$ for a wider range of x values than for the narrow models. As a consequence there are more points to choose from for which there are a possibility of observing both zeros and ones and thus avoiding complete or quasi-complete separation in the data. This is also reflected in the pattern for the "high-wide" model where the probability of obtaining a ML estimate is approximately 50 % for the non-optimal designs compared to 35 % for the D-optimal design. For the "high-narrow" model on the other hand the D-optimal design outperforms the other two having 35 % chance of obtaining parameter estimates as against 11 or 1 %. The low models are more problematic compared to the high models because $P(Y = 1)$ is low and therefore there will be many points where only zeros are observed.

In section b) of Table 4.4 the probabilities that the ML estimate exists when $N = 20$ are presented. The probability to obtain parameter estimates has increased but is still extremely small for the "low-narrow" model. For the "low-wide" model these probabilities have increased about three times for the two non-optimal designs which still are preferable. The D-optimal design has the highest probability for the "high-narrow" model and the three designs are almost equivalent for the "high-wide" model.

For $N = 50$ the probability of obtaining parameter estimates can be estimated by the percentage shares of the simulation runs where the ML estimate existed. These are shown in column a) of Table 4.5. There are hardly ever problems with non-existence for the "high-wide" model with almost 100 % existence for all three designs. The same thing applies to the "high-narrow" model if the D-optimal design is used. The situation is not quite as good for the "low-wide" model where the 81 % associated with the 8-point design is the highest proportion. Nevertheless, it is a lot better than for the "low-narrow" case where existence in 12 % of the times (for the D-optimal design) is the maximum.

For $N = 100$ the percentage shares of the simulations where the ML estimate existed are given in column b) of Table 4.5. There are no longer any problems with non-existence for the "high-wide" model for any of the designs. If the D-optimal or 8-point design is chosen for the "high-narrow" model the problem is also avoided. By choosing one of the non-optimal designs for the "low-wide" model almost 100 % existence of the ML estimate can be expected. When the D-optimal design was used for the "low-narrow"

model existence occurred in only 35 % of the times even though the sample size is quite large.

TABLE 4.4: The number of possible samples, the number of samples where the ML estimate exists and the probability that the ML estimate exists for N=10 and N=20.

TYPE OF RESPONSE CURVE	DESIGN	# SAMPLES	# SAMPLES WHERE THE ML ESTIMATE EXISTS	P(ML)
		a) N=10		
	D-OPT	144	68	0.35
"HIGH-WIDE"	7-P	432	544	0.49
	8-P	576	488	0.49
	D-OPT	144	68	0.35
"HIGH-NARROW"	7-P	432	544	0.014
	8-P	576	488	0.11
	D-OPT	80	12	$2.5 \cdot 10^{-3}$
"LOW-WIDE"	7-P	432	544	0.11
	8-P	576	488	0.096
	D-OPT	80	12	$2.5 \cdot 10^{-3}$
"LOW-NARROW"	7-P	432	544	$2.4 \cdot 10^{-5}$
	8-P	576	488	$2.6 \cdot 10^{-3}$
		b) N=20		
	D-OPT	1225	927	0.76
"HIGH-WIDE"	7-P	12288	11944	0.82
	8-P	20736	20384	0.84
	D-OPT	1225	927	0.74
"HIGH-NARROW"	7-P	12288	11944	0.046
	8-P	20736	20384	0.42
	D-OPT	448	180	0.016
"LOW-WIDE"	7-P	12228	11944	0.27
	8-P	20736	20384	0.36
	D-OPT	448	180	0.016
"LOW-NARROW"	7-P	12228	11944	$4.4 \cdot 10^{-5}$
	8-P	20736	20384	0.015

TABLE 4.5: The percentage share of the simulations where the ML estimate existed for N=50 and N=100.

TYPE OF RESPONSE CURVE	DESIGN	% OF THE SIMULATION SAMPLES WHERE THE ML ESTIMATE EXISTED	
		a) N=50	b) N=100
"HIGH-WIDE"	D-opt	98	100
	7-p	99	100
	8-p	99	100
"HIGH-NARROW"	D-opt	98	100
	7-p	34	69
	8-p	84	98
"LOW-WIDE"	D-opt	12	36
	7-p	73	97
	8-p	81	99
"LOW-NARROW"	D-opt	12	35
	7-p	≈ 0	≈ 0
	8-p	8	22

4.4.2 Estimation Results

Tables 4.6 to 4.11 report the results for sample sizes $N = 10$ and $N = 20$. These are based on complete enumeration of all possible samples, thus the results are exact. The mean squared error, mean and variance of the ML and FL estimators are given in the tables. Furthermore, the mean and variance of the variance estimators are presented in the tables. The theoretical approximate variance derived using asymptotic results (abbreviated as AV) is also included as a reference point. Tables 4.12 to 4.17 report analogous results based on the 10000 simulation runs for sample sizes $N = 50$ and $N = 100$. The occurrences of degenerated estimates (according to the definition in the previous section) is also documented. The probability of obtaining a degenerated estimate is given for sample sizes $N = 10$ and $N = 20$ and the relative frequency of such estimates in the simulations is given for sample sizes $N = 50$ and $N = 100$. For the ML estimator the probability that the estimate does not exist is reported separately. All the reported results are based only on those estimates that are not degenerated.

Thus, the results should be interpreted as *given* that the estimate is not degenerated and exists (in case of the ML estimator). When referring to the estimators from now on this is assumed implicitly. Plots of the sampling distributions of the estimators can be found in the Appendix. Below follows some summarizing comments to the results.

$N = 10$, Tables 4.6 to 4.8

The D-optimal design is most often associated with the lowest mean squared error, as expected. The exceptions are estimation of the parameter α for all models but the "high-narrow" model and estimation of μ for he "high-narrow" model. However, these differences are quite small. Especially when compared to the gains that can be made by using the D-optimal design when estimating any of the other parameters. It can be noted that the bias is really small for the parameter μ in all cases when any of the two symmetric designs (D-optimal or 8-point) is used. The variances of the estimators are almost always quite different from the theoretical approximate variance, that is the asymptotic results are not applicable for such small samples. Neither do the variance estimators work any good.

The FL estimator performs much better than the ML estimator with reference to the probability that an estimate is not degenerated for the high models and the symmetric designs (D-optimal and 8-point). It is also better for the low models, though it is far from good, the risk of obtaining degenerated estimates is still very high. For the 7-point design the FL estimator is substantially better than the ML estimator in terms of the number of "good" estimates for the "high-wide" model only. None of the two estimators can be favored based on comparing their mean squared errors.

It should be noted that the probability to obtain an estimate is extremely small in several cases (particularly for the ML estimator and both the low models). In these cases there exist only a limited number of estimates, which is also reflected in the plots of the sampling distributions (see the Appendix).

$N = 20$, Tables 4.9 to 4.11

For the most part, the mean squared errors of the estimators are lowest when the D-optimal design is used. The exceptions occur for estimation of

α, as was also the case for $N = 10$. The two non-optimal designs alternately succeeds better in these events. Overall, the differences in mean squared errors are fairly large in favor of the D-optimal design. The 7-point design is generally the worst choice. The shape of the distributions are also more symmetric for the the D-optimal and 8-point designs. There are still disagreements between the variance of the estimators in comparison with the asymptotic variance as well as with the mean of the variance estimators.

The FL estimator is superior to the ML estimator in terms of the probability of usable estimates, especially when the D-optimal design is used**.** On the other hand, for the two most problematic cases ("low-narrow" model and the non-optimal designs) the FL estimator also fails. Generally, the problem with non-existence and/or degenerated estimates remains, even though it is somewhat reduced. Again, none of the estimators is consistently better than the other in terms of mean squared error.

$N = 50$**, Tables 4.12 to 4.14**

The magnitude of the mean squared error varies between the designs, still being generally lowest for the D-optimal design. Once again, the majority of the exceptions appear for the parameter α. On the whole, the D-optimal design is better than the 8-point design which is usually better than the 7-point design. When the proportion of "good" estimates is close to 1, as for the D-optimal design and the high models, the variances of the estimators are close to asymptotic variance. The variance estimator also works well for these examples. In the corresponding plots the distributions of the estimators are coming closer to resemble a normal distribution.

The problem of degenerated estimates has nearly completely vanished for all the "high-wide" models and the "high-narrow" model together with the D-optimal design. Though, the percentage of failed estimates is still remarkably large for several of the low models.

$N = 100$**, Tables 4.15 to 4.17**

Comparing the designs, the superiority of the D-optimal design prevails. Estimation of the parameter α remains to be an exception to this rule for some cases. The reduction in mean squared error if the D-optimal design is

chosen is appreciable for several cases. The same can be said about the 8-point design compared to the 7-point design. As for the sample size $N = 50$, the variances of the estimators come close to the asymptotic variance when the percentage of useful estimates amounts to 100 %. In addition, the variance estimators succeed well in the same cases. Some of the plots in the appendix are almost symmetric and well-shaped, while the opposite is true for others (particularly for the "low-narrow" model).

For the largest sample size, there are still severe problems with non-existence for the ML estimator for some combinations, e.g. for the low models and the D-optimal design. In such cases the FL estimator offers quite a large improvement in the number of usable estimates, sometimes at the expense of a higher mean squared error. For the "high-wide" model, the mean squared errors are comparable for the two estimators (a little lower for the FL estimator). For the rest, which of the two estimators is associated with the lowest mean squared error shifts by model and design.

4.5 Discussion

Non-existence of the ML estimator proved to be a substantial problem for the quadratic logistic model and the sets of true parameters examined here, especially for small samples. How severe the problem is depends on the true parameters and the design. The non-optimal designs considered here were sometimes better than the D-optimal design in this respect, due to the increased number of design points. The models where the response curve $\pi(x)$ is low were more problematic, in particular the "low-narrow" model where existence practically never occurred for the smallest sample size ($N = 10$) and only in 35 % of the times for the largest sample size ($N = 100$). The practical consequence is that large samples, demanding big time and money efforts, need to be taken and yet there may be a large risk of not obtaining estimates.

The D-optimal design was generally preferred over the non-optimal designs in terms of mean squared error. Some exceptions occurred, mainly for estimation of the parameter α. Furthermore, the differences between the designs were relatively small in these cases. Unless estimation of α is the main objective, there is potential for considerable improvements by choosing the

D-optimal design. It was particularly the case for estimation of the parameter μ, corresponding to estimation of the point of maximum response. Among the non-optimal designs the better choice was the symmetric 8-point design, specifically with regard to estimation of μ. The results were not as clear-cut when it came to the problems with non-existence for the ML estimator. The choice of design did have a large impact on the severity of these problems. The complications of degenerated estimates for the FL estimator were the least serious when the D-optimal design was used.

There turned out to be large discrepancies between the asymptotic sampling distributions of the estimators and the small-sample distributions (given that estimates exist and are not degenerated). It was only for large N and when the proportion of useful estimates was close to 100 %, that the simulation sample variance was close to the theoretical approximate variance. This is problematic since the optimal designs are derived based on the assumption that the theoretical approximate variance is true. The accuracy of the variance estimator was also affected by the proportion of degenerated estimates.

The FL estimator is proposed as a solution to the non-existence problems. It performed better in this respect, but instead it suffered from other problems resulting in poor estimates. The extent of these problems was related to the proportion of non-existence for the ML estimator, the worst examples are also associated with a high proportion of degenerated FL estimates. However, in several cases the proportion of failed estimates was greatly reduced, especially for the D-optimal design.

All in all, the parameter dependence issue set aside, choosing the D-optimal design would be preferable. Sequential experimentation is available as a possible solution to the problem of parameter dependence. Sequential designs is the topic of the next two chapters.

TABLE 4.6: Maximum likelihood estimation (ML) and estimation based on Firths' modified likelihood (FL).
Sample size: N=10
Design: D-optimal
a) The probability that the ML estimate does not exist
b) The probability that the estimate is degenerated

Model	"HIGH-WIDE" $\boldsymbol{\theta} = (2, -0.1, 0)'$		"HIGH-NARROW" $\boldsymbol{\theta} = (2, -4, 0)'$		"LOW-WIDE" $\boldsymbol{\theta} = (-2, -0.1, 0)'$		"LOW-NARROW" $\boldsymbol{\theta} = (-2, -4, 0)'$	
Estimator	ML	FL	ML	FL	ML	FL	ML	FL
a	0.65		0.65		1.00		1.00	
b	0.03	0.04	0.02	0.03	0.00	0.61	0.00	0.61
MSE	**0.90** **0.00** **1.33**	**1.14** **0.00** **1.08**	**0.77** **2.84** **0.14**	**1.09** **2.99** **0.07**	**4.49** **0.00** **0.01**	**2.00** **0.00** **1.68**	**4.49** **4.66** **0.00**	**2.00** **2.42** **0.04**
$E\left(\widehat{\boldsymbol{\theta}}\right)$	1.33 −0.07 0.00	1.93 −0.09 0.00	1.49 −2.84 0.01	1.96 −3.71 0.01	0.10 −0.05 0.00	−0.64 −0.07 0.00	0.10 −1.98 0.00	−0.64 −2.93 0.00
$V\left(\widehat{\boldsymbol{\theta}}\right)$	0.45 0.00 1.33	1.14 0.00 1.08	0.51 1.50 0.14	1.09 2.90 0.07	0.10 0.00 0.01	0.14 0.00 1.68	0.10 0.58 0.00	0.14 1.29 0.04
$AV\left(\widehat{\boldsymbol{\theta}}\right)$	2.50 0.00 0.65		2.06 6.01 0.02		2.38 0.03 10.01		2.38 55.57 0.25	
$E\left[\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)\right]$	2.10 0.00 5.06	2.98 0.00 8.02	3.21 6.33 27.70	3.45 7.69 9.03	1.03 0.01 5.75	1.16 0.01 28.62	1.03 11.32 0.14	1.16 16.70 0.71
$V\left[\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)\right]$	0.10 0.00 $1.1 \cdot 10^2$	1.65 0.00 $4.9 \cdot 10^3$	$1.4 \cdot 10^2$ 1.60 $> 10^4$	$2.8 \cdot 10^2$ 9.92 $> 10^4$	0.01 0.00 2.10	0.01 0.00 $4.5 \cdot 10^3$	0.01 0.36 0.00	0.07 1.12 2.85

TABLE 4.7: Maximum likelihood estimation (ML) and estimation based on Firths' modified likelihood (FL).
Sample size: N=10
Design: 7-point
a) The probability that the ML estimate does not exist
b) The probability that the estimate is unusable

Model	"HIGH-WIDE" $\boldsymbol{\theta} = (2, -0.1, 0)'$		"HIGH-NARROW" $\boldsymbol{\theta} = (2, -4, 0)'$		"LOW-WIDE" $\boldsymbol{\theta} = (-2, -0.1, 0)'$		"LOW-NARROW" $\boldsymbol{\theta} = (-2, -4, 0)'$	
Estimator	ML	FL	ML	FL	ML	FL	ML	FL
a	0.51		0.99		0.89		1.00	
b	0.11	0.06	0	0.69	0.01	0.85	0	1.00
MSE	**1.26** **0.01** **9.19**	**0.65** **0.00** **2.30**	**3.61** **12.20** **0.32**	**6.61** **15.65** **3.61**	**4.12** **0.17** **2.00**	**2.26** **0.01** **7.95**	**3.98** **12.44** **0.27**	**1.74** **15.81** **4.23**
$E\left(\widehat{\boldsymbol{\theta}}\right)$	2.36 −0.13 −0.79	2.03 −0.10 0.04	0.19 −0.51 −0.44	−0.51 −0.05 −1.08	−0.08 −0.39 0.57	−0.55 −0.03 −0.00	−0.01 −0.47 −0.45	−0.69 −0.02 −1.19
$V\left(\widehat{\boldsymbol{\theta}}\right)$	1.14 0.01 8.56	0.65 0.00 2.30	0.33 0.02 0.12	0.31 0.01 2.46	0.46 0.08 1.67	0.16 0.00 7.95	0.01 0.01 0.07	0.01 0.00 2.81
$AV\left(\widehat{\boldsymbol{\theta}}\right)$	1.22 0.01 2.35		10.11 13.85 0.04		2.07 0.07 14.21		94.22 $2.6 \cdot 10^2$ 1.57	
$E\left[\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)\right]$	29.57 0.03 $1.9 \cdot 10^2$	$2.9 \cdot 10^2$ 0.01 $> 10^4$	1.71 0.65 1.55	0.78 0.01 0.29	2.02 0.41 7.02	0.79 0.01 $> 10^4$	1.49 0.63 1.71	0.66 0.01 0.34
$V\left[\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)\right]$	$> 10^4$ 0.01 $> 10^4$	$> 10^4$ 0.00 $> 10^4$	0.65 0.04 0.22	0.78 0.01 0.29	$1.1 \cdot 10^2$ 0.15 $3.2 \cdot 10^3$	64.65 0.00 $> 10^4$	0.02 0.01 0.05	0.03 0.00 1.31

TABLE 4.8: Maximum likelihood estimation (ML) and estimation based on Firths' modified likelihood (FL).
Sample size: N=10
Design: 8-point
a) The probability that the ML estimate does not exist
b) The probability that the estimate is degenerated

Model	"HIGH-WIDE" $\boldsymbol{\theta} = (2, -0.1, 0)'$		"HIGH-NARROW" $\boldsymbol{\theta} = (2, -4, 0)'$		"LOW-WIDE" $\boldsymbol{\theta} = (-2, -0.1, 0)'$		"LOW-NARROW" $\boldsymbol{\theta} = (-2, -4, 0)'$	
Estimator	ML	FL	ML	FL	ML	FL	ML	FL
a	0.51		0.89		0.90		1.00	
b	0.08	0.05	0	0.01	0.01	0.53	0	0.80
MSE	**0.62** **0.01** **4.03**	**0.65** **0.00** **2.81**	**0.96** **6.06** **0.13**	**4.51** **15.42** **0.06**	**7.04** **0.73** **6.15**	**1.23** **0.00** **4.78**	**8.02** **4.92** **0.19**	**0.79** **15.81** **0.16**
$E\left(\widehat{\boldsymbol{\theta}}\right)$	1.79 −0.12 0.00	1.96 −0.09 0.00	1.09 −1.57 0.00	−0.05 −0.07 0.00	0.56 −0.67 0.00	−0.96 −0.03 0.00	0.82 −1.80 0.00	−1.13 −0.02 0.00
$V\left(\widehat{\boldsymbol{\theta}}\right)$	0.57 0.01 4.03	0.66 0.00 2.81	0.12 0.14 0.13	0.32 0.00 0.06	0.51 0.41 6.15	0.15 0.00 4.78	0.04 0.09 0.19	0.04 0.00 0.16
$AV\left(\widehat{\boldsymbol{\theta}}\right)$	1.63 0.00 1.01		5.05 23.97 0.05		2.02 0.04 13.72		7.90 $1.7 \cdot 10^2$ 0.59	
$E\left[\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)\right]$	6.36 0.02 33.84	21.17 0.01 $> 10^4$	1.98 2.73 0.29	0.82 0.01 0.03	5.68 0.92 76.38	0.98 0.01 $5.4 \cdot 10^2$	2.03 3.72 0.28	1.05 0.01 0.02
$V\left[\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)\right]$	$1.3 \cdot 10^3$ 0.01 $> 10^4$	$> 10^4$ 0.00 $> 10^4$	0.03 1.81 0.01	0.01 0.00 0.00	$> 10^4$ 2.18 $> 10^4$	18.25 0.00 $> 10^4$	0.04 0.74 0.01	0.00 0.00 0.00

TABLE 4.9: Maximum likelihood estimation (ML) and estimation based on Firths' modified likelihood (FL).
Sample size: N=20
Design: D-optimal
a) The probability that the ML estimate does not exist
b) The probability that the estimate is degenerated

Model	"HIGH-WIDE" $\boldsymbol{\theta} = (2, -0.1, 0)'$		"HIGH-NARROW" $\boldsymbol{\theta} = (2, -4, 0)'$		"LOW-WIDE" $\boldsymbol{\theta} = (-2, -0.1, 0)'$		"LOW-NARROW" $\boldsymbol{\theta} = (-2, -4, 0)'$	
Estimator	ML	FL	ML	FL	ML	FL	ML	FL
a	0.24		0.26		0.98		0.98	
b	0.01	0.01	0.00	0.00	0.00	0.15	0.00	0.15
MSE	**0.73** **0.00** **0.71**	**0.97** **0.00** **0.59**	**0.73** **2.20** **0.03**	**1.10** **2.88** **0.02**	**0.81** **0.01** **0.10**	**0.81** **0.00** **0.51**	**0.81** **9.10** **0.00**	**0.82** **6.15** **0.01**
$E\left(\widehat{\boldsymbol{\theta}}\right)$	2.05 −0.10 −0.01	2.05 −0.10 0.02	1.94 −4.03 0.00	2.03 −4.01 0.00	−1.27 −0.03 0.00	−1.02 −0.09 0.00	−1.27 −1.25 0.00	−1.60 −2.38 0.00
$V\left(\widehat{\boldsymbol{\theta}}\right)$	0.72 0.00 0.71	0.97 0.00 0.59	0.73 2.20 0.03	1.10 2.88 0.02	0.27 0.00 0.10	0.16 0.00 0.95	0.27 1.56 0.00	0.66 3.54 0.01
$AV\left(\widehat{\boldsymbol{\theta}}\right)$	1.02 0.00 0.38		1.25 3.02 0.01		1.59 0.02 4.29		1.59 27.42 0.11	
$E\left[\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)\right]$	1.23 0.00 2.19	1.32 0.00 1.96	15.15 3.92 24.12	2.37 4.13 1.69	1.05 0.01 48.15	1.01 0.01 13.25	1.38 10.36 1.71	1.57 16.85 $> 10^4$
$V\left[\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)\right]$	0.89 0.00 $1.3 \cdot 10^3$	0.74 0.00 $4.8 \cdot 10^2$	$> 10^4$ 2.08 $> 10^4$	$> 10^4$ 5.67 $> 10^4$	0.05 0.00 $1.0 \cdot 10^3$	24.47 0.00 $> 10^4$	$2.3 \cdot 10^3$ 2.09 $5.3 \cdot 10^3$	15.83 29.92 $> 10^4$

TABLE 4.10: Maximum likelihood estimation (ML) and estimation based on Firths' modified likelihood (FL).
Sample size: N=20
Design: 7-point
a) The probability that the ML estimate does not exist
b) The probability that the estimate is degenerated

Model	"HIGH-WIDE" $\boldsymbol{\theta} = (2, -0.1, 0)'$		"HIGH-NARROW" $\boldsymbol{\theta} = (2, -4, 0)'$		"LOW-WIDE" $\boldsymbol{\theta} = (-2, -0.1, 0)'$		"LOW-NARROW" $\boldsymbol{\theta} = (-2, -4, 0)'$	
Estimator	ML	FL	ML	FL	ML	FL	ML	FL
a	0.18		0.95		0.73		1.00	
b	0.07	0.03	0	0.10	0.02	0.58	0	0.99
MSE	**0.78** **0.01** **2.95**	**0.68** **0.00** **1.88**	**1.28** **5.27** **0.05**	**8.51** **15.32** **0.33**	**1.96** **0.30** **3.32**	**0.73** **0.01** **2.93**	**1.79** **11.71** **0.40**	**0.23** **15.83** **0.57**
$E\left(\widehat{\boldsymbol{\theta}}\right)$	2.29 −0.12 −0.34	2.11 −0.10 −0.16	1.14 −1.77 −0.03	−0.84 −0.09 −0.09	−0.73 −0.44 0.23	−1.**3**1 −0.04 0.45	−0.71 −0.59 −0.29	−1.65 −0.02 0.34
$V\left(\widehat{\boldsymbol{\theta}}\right)$	0.70 0.00 2.83	0.67 0.00 1.85	0.53 0.30 0.05	0.44 0.01 0.32	0.35 0.19 3.26	0.26 0.00 2.72	0.14 0.10 0.32	0.10 0.00 0.55
$AV\left(\widehat{\boldsymbol{\theta}}\right)$	0.65 0.00 0.89		4.06 7.50 0.03		1.13 0.03 6.58		64.22 $2.2 \cdot 10^2$ 1.28	
$E\left[\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)\right]$	5.12 0.00 42.65	$6.2 \cdot 10^2$ 0.00 $> 10^4$	1.88 1.92 0.14	0.54 0.01 0.03	1.10 0.48 15.25	0.59 0.01 2.90	1.13 0.67 1.07	0.66 0.00 0.21
$V\left[\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)\right]$	$> 10^4$ 0.00 $> 10^4$	$> 10^4$ 0.00 $> 10^4$	0.24 0.29 0.07	0.05 0.00 0.01	1.21 0.65 $4.2 \cdot 10^4$	0.13 0.00 $5.6 \cdot 10^2$	0.04 0.41 0.19	0.02 0.00 0.06

TABLE 4.11: Maximum likelihood estimation (ML) and estimation based on Firths' modified likelihood (FL).
Sample size: N=20
Design: 8-point
a) The probability that the ML estimate does not exist
b) The probability that the estimate is degenerated

Model	"HIGH-WIDE" $\boldsymbol{\theta} = (2, -0.1, 0)'$		"HIGH-NARROW" $\boldsymbol{\theta} = (2, -4, 0)'$		"LOW-WIDE" $\boldsymbol{\theta} = (-2, -0.1, 0)'$		"LOW-NARROW" $\boldsymbol{\theta} = (-2, -4, 0)'$	
Estimator	ML	FL	ML	FL	ML	FL	ML	FL
a	0.16		0.58		0.64		0.99	
b	0.07	0.02	0	0.00	0.02	0.56	0	0.90
MSE	**0.66** **0.01** **2.53**	**0.64** **0.00** **1.75**	**0.55** **2.93** **0.05**	**3.89** **14.89** **0.33**	**2.03** **0.72** **2.94**	**0.77** **0.01** **13.41**	**1.75** **6.17** **0.13**	**0.28** **15.82** **0.62**
$E\left(\widehat{\boldsymbol{\theta}}\right)$	2.26 −0.13 −0.04	2.11 −0.10 −0.08	1.58 −2.91 −0.02	0.11 −0.14 0.32	−0.71 −0.64 −0.02	−1.22 −0.03 −1.13	−0.74 −1.58 0.07	−1.50 −0.02 −0.40
$V\left(\widehat{\boldsymbol{\theta}}\right)$	0.58 0.01 2.53	0.64 0.00 1.75	0.38 1.75 0.05	0.32 0.01 0.23	0.37 0.43 2.94	0.16 0.00 12.12	0.15 0.30 0.13	0.03 0.00 0.46
$AV\left(\widehat{\boldsymbol{\theta}}\right)$	0.65 0.00 0.80		2.03 9.82 0.02		0.83 0.03 7.88		3.25 69.49 0.25	
$E\left[\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)\right]$	2.57 0.01 25.55	$> 10^4$ 0.00 $> 10^4$	1.36 6.01 0.07	0.37 0.01 0.04	1.34 1.11 30.28	0.37 0.01 8.77	0.95 4.24 0.30	0.52 0.00 0.13
$V\left[\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)\right]$	$4.7 \cdot 10^4$ 0.00 $> 10^4$	$> 10^4$ 0.00 $> 10^4$	0.17 16.64 0.00	0.04 0.00 0.00	20.49 3.78 $> 10^4$	$1.7 \cdot 10^3$ 0.00 $> 10^4$	0.00 4.77 0.02	0.00 0.00 0.06

TABLE 4.12: Simulation results of maximum likelihood estimation (ML) and estimation based on Firths' modified likelihood (FL).
Sample size: N=50
Design: D-optimal
a) Percent of the simulations where the ML estimate did not exist
b) Percent of the simulations where the estimate degenerated

Model	"HIGH-WIDE" $\boldsymbol{\theta} = (2, -0.1, 0)'$		"HIGH-NARROW" $\boldsymbol{\theta} = (2, -4, 0)'$		"LOW-WIDE" $\boldsymbol{\theta} = (-2, -0.1, 0)'$		"LOW-NARROW" $\boldsymbol{\theta} = (-2, -4, 0)'$	
Estimator	ML	FL	ML	FL	ML	FL	ML	FL
a	0.02		0.02		0.88		0.88	
b	0	0.00	0	0	0.02	0.10	0.02	0.10
mse	**0.68** **0.00** **0.15**	**0.60** **0.00** **0.14**	**0.69** **1.75** **0.00**	**0.60** **1.39** **0.00**	**0.35** **0.00** **1.02**	**0.43** **0.00** **1.42**	**0.48** **6.12** **0.30**	**0.44** **3.60** **0.05**
$mean\left(\widehat{\boldsymbol{\theta}}\right)$	2.16 −0.11 0.01	2.01 −0.10 0.01	2.18 −4.36 −0.00	2.03 −4.03 −0.00	−1.85 −0.05 0.01	−1.79 −0.08 0.01	−1.84 −1.97 −0.00	−1.78 −3.23 −0.00
$var\left(\widehat{\boldsymbol{\theta}}\right)$	0.66 0.00 0.15	0.60 0.00 0.14	0.66 1.62 0.00	0.59 1.38 0.00	0.32 0.00 1.02	0.39 0.00 1.42	0.46 2.00 0.30	0.40 3.01 0.05
$AV\left(\widehat{\boldsymbol{\theta}}\right)$	0.55 0.00 0.12		0.55 1.26 0.00		0.59 0.01 1.77		0.59 10.92 0.04	
$mean\left[\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)\right]$	0.67 0.00 0.16	0.64 0.00 0.18	0.68 1.61 0.00	0.64 1.48 0.00	0.63 0.00 $1.3 \cdot 10^2$	0.62 0.01 95.98	$1.1 \cdot 10^2$ 7.10 $2.9 \cdot 10^2$	1.17 9.38 3.27
$var\left[\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)\right]$	0.10 0.00 0.06	0.15 0.00 0.08	0.11 0.48 0.00	0.15 0.52 0.00	0.12 0.00 $> 10^4$	0.20 0.00 $> 10^4$	$> 10^4$ 3.71 $> 10^4$	88.36 11.70 $3.0 \cdot 10^2$

TABLE 4.13: Simulation results of maximum likelihood estimation (ML) and estimation based on Firths' modified likelihood (FL).
Sample size: N=50
Design: 7-point
a) Percent of the simulations where the ML estimate did not exist
b) Percent of the simulations where the estimate degenerated

Model	"HIGH-WIDE" $\boldsymbol{\theta} = (2, -0.1, 0)'$		"HIGH-NARROW" $\boldsymbol{\theta} = (2, -4, 0)'$		"LOW-WIDE" $\boldsymbol{\theta} = (-2, -0.1, 0)'$		"LOW-NARROW" $\boldsymbol{\theta} = (-2, -4, 0)'$	
Estimator	ML	FL	ML	FL	ML	FL	ML	FL
a	0.01		0.66		0.27		1.00	
b	0.01	0.00	0	0.00	0.03	0.22	0	0.95
mse	**0.45** **0.00** **0.99**	**0.31** **0.00** **0.74**	**0.87** **1.27** **0.01**	**6.69** **13.22** **0.35**	**0.64** **0.28** **2.71**	**0.35** **0.01** **4.46**	**0.01** **12.97** **0.43**	**0.21** **15.84** **1.53**
$mean\left(\widehat{\boldsymbol{\theta}}\right)$	2.23 −0.11 −0.20	2.06 −0.10 −0.11	1.88 −3.43 −0.01	−0.49 −0.40 −0.46	−1.43 −0.40 0.10	−1.76 −0.06 0.06	−1.90 −0.40 −0.66	−2.43 −0.02 −0.57
$var\left(\widehat{\boldsymbol{\theta}}\right)$	0.40 0.00 0.95	0.31 0.00 0.73	0.86 0.95 0.01	0.50 0.30 0.14	0.31 0.19 2.70	0.29 0.00 4.45	0 0 0	0.02 0.00 1.21
$AV\left(\widehat{\boldsymbol{\theta}}\right)$	0.25 0.00 0.38		1.57 2.50 0.01		0.44 0.01 2.80		19.37 60.62 0.42	
$mean\left[\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)\right]$	0.65 0.00 3.98	0.36 0.00 2.32	1.56 2.20 0.01	0.33 0.09 0.02	0.66 0.38 15.60	0.33 0.01 3.53	0.98 0.29 1.34	0.62 0.00 90.90
$var\left[\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)\right]$	65.04 0.00 $2.2 \cdot 10^3$	0.77 0.00 $6.1 \cdot 10^3$	0.28 0.52 0.00	0.02 0.05 0.00	1.38 0.55 $> 10^4$	0.01 0.00 $1.5 \cdot 10^2$	0 0 0	1.39 0.00 $> 10^4$

TABLE 4.14: Simulation results of maximum likelihood estimation (ML) and estimation based on Firths' modified likelihood (FL).
Sample size: N=50
Design: 8-point
a) Percent of the simulations where the ML estimate did not exist
b) Percent of the simulations where the estimate degenerated

Model	"HIGH-WIDE" $\boldsymbol{\theta} = (2, -0.1, 0)'$		"HIGH-NARROW" $\boldsymbol{\theta} = (2, -4, 0)'$		"LOW-WIDE" $\boldsymbol{\theta} = (-2, -0.1, 0)'$		"LOW-NARROW" $\boldsymbol{\theta} = (-2, -4, 0)'$	
Estimator	ML	FL	ML	FL	ML	FL	ML	FL
a	0.01		0.16		0.19		0.92	
b	0.00	0.00	0	0	0.02	0.17	0	0.61
mse	**0.37** **0.00** **0.53**	**0.31** **0.00** **0.39**	**0.44** **2.85** **0.01**	**0.97** **7.63** **0.05**	**0.62** **0.86** **2.55**	**0.32** **0.01** **4.11**	**0.54** **4.96** **0.10**	**0.17** **15.78** **0.36**
$mean\left(\widehat{\boldsymbol{\theta}}\right)$	2.16 −0.11 0.02	2.04 −0.10 0.01	2.03 −4.11 0.00	1.40 −2.00 0.00	−1.42 −0.58 −0.00	−1.76 −0.06 0.02	−1.47 −2.01 −0.00	−2.33 −0.03 0.01
$var\left(\widehat{\boldsymbol{\theta}}\right)$	0.35 0.00 0.53	0.31 0.00 0.39	0.44 2.83 0.01	0.60 3.64 0.05	0.29 0.63 2.55	0.26 0.00 4.11	0.26 1.00 0.10	0.07 0.00 0.36
$AV\left(\widehat{\boldsymbol{\theta}}\right)$	0.27 0.00 0.28		0.84 4.00 0.01		0.34 0.01 3.04		1.32 27.90 0.10	
$mean\left[\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)\right]$	0.37 0.00 1.39	0.31 0.00 0.70	0.88 4.31 0.02	0.50 1.17 0.05	0.66 0.85 21.80	0.49 0.00 $1.5 \cdot 10^2$	0.68 5.60 0.22	0.42 0.00 30.28
$var\left[\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)\right]$	2.15 0.00 $1.5 \cdot 10^3$	0.05 0.00 31.37	0.15 4.98 0.00	0.33 7.05 0.00	3.77 4.92 $> 10^4$	0.49 0.00 $> 10^4$	0.01 13.46 0.03	0.01 0.00 $3.2 \cdot 10^2$

TABLE 4.15: Simulation results of maximum likelihood estimation (ML) and estimation based on Firths' modified likelihood (FL).
Sample size: N=100
Design: D-optimal
a) Percent of the simulations where the ML estimate did not exist
b) Percent of the simulations where the estimate was unusable

Model	"HIGH-WIDE" $\boldsymbol{\theta} = (2, -0.1, 0)'$		"HIGH-NARROW" $\boldsymbol{\theta} = (2, -4, 0)'$		"LOW-WIDE" $\boldsymbol{\theta} = (-2, -0.1, 0)'$		"LOW-NARROW" $\boldsymbol{\theta} = (-2, -4, 0)'$	
Estimator	ML	FL	ML	FL	ML	FL	ML	FL
a	0		0.00		0.64		0.65	
b	0	0	0	0	0.04	0.05	0.03	0.05
mse	**0.33** **0.00** **0.07**	**0.28** **0.00** **0.07**	**0.33** **0.78** **0.00**	**0.28** **0.64** **0.00**	**0.19** **0.00** **0.80**	**0.21** **0.00** **1.27**	**0.19** **2.87** **0.08**	**0.22** **2.84** **0.11**
$mean\left(\widehat{\boldsymbol{\theta}}\right)$	2.10 −0.10 0.00	2.01 −0.10 0.00	2.10 −4.18 0.00	2.01 −4.01 0.00	−1.95 −0.07 0.01	−1.88 −0.09 0.02	−1.95 −2.94 −0.00	−1.87 −3.69 0.00
$var\left(\widehat{\boldsymbol{\theta}}\right)$	0.32 0.00 0.07	0.28 0.00 0.07	0.32 0.74 0.00	0.28 0.64 0.00	0.19 0.00 0.80	0.20 0.00 1.27	0.19 1.75 0.08	0.20 2.75 0.11
$AV\left(\widehat{\boldsymbol{\theta}}\right)$	0.26 0.00 0.06		0.26 0.61 0.00		0.28 0.00 0.91		0.28 5.46 0.02	
$mean\left[\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)\right]$	0.29 0.00 0.07	0.28 0.00 0.07	0.29 0.69 0.00	0.28 0.66 0.00	0.28 0.00 5.65	0.31 0.00 10.12	1.77 4.57 14.63	6.12 5.77 15.74
$var\left[\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)\right]$	0.01 0.00 0.00	0.01 0.00 0.00	0.01 0.03 0.00	0.01 0.02 0.00	0.01 0.00 $3.4 \cdot 10^2$	0.14 0.00 $4.4 \cdot 10^3$	$1.7 \cdot 10^3$ 0.88 $> 10^4$	$> 10^4$ 3.58 $> 10^4$

TABLE 4.16: Simulation results of maximum likelihood estimation (ML) and estimation based on Firths' modified likelihood (FL).
Sample size: N=100
Design: 7-point
a) Percent of the simulations where the ML estimate did not exist
b) Percent of the simulations where the estimate degenerated

Model	"HIGH-WIDE" $\boldsymbol{\theta} = (2, -0.1, 0)'$		"HIGH-NARROW" $\boldsymbol{\theta} = (2, -4, 0)'$		"LOW-WIDE" $\boldsymbol{\theta} = (-2, -0.1, 0)'$		"LOW-NARROW" $\boldsymbol{\theta} = (-2, -4, 0)'$	
Estimator	ML	FL	ML	FL	ML	FL	ML	FL
a	0		0.31		0.03		1.00	
b	0.00	0.00	0	0	0.01	0.05	0	0.82
mse	**0.17** **0.00** **0.31**	**0.14** **0.00** **0.28**	**0.75** **0.96** **0.00**	**1.03** **1.80** **0.03**	**0.28** **0.26** **1.63**	**0.23** **0.01** **2.76**	**0.53** **9.92** **0.18**	**1.22** **15.81** **1.61**
$mean\left(\widehat{\boldsymbol{\theta}}\right)$	2.10 −0.11 −0.10	2.02 −0.10 −0.05	2.10 −4.01 0.00	1.91 −3.80 −0.02	−1.74 −0.32 0.10	−1.87 −0.08 −0.05	−2.02 −0.90 −0.32	−3.08 −0.02 −0.63
$var\left(\widehat{\boldsymbol{\theta}}\right)$	0.16 0.00 0.30	0.14 0.00 0.28	0.74 0.96 0.00	1.02 1.76 0.03	0.22 0.21 1.62	0.21 0.01 2.76	0.53 0.31 0.08	0.05 0.00 1.22
$AV\left(\widehat{\boldsymbol{\theta}}\right)$	0.13 0.00 0.19		0.76 1.28 0.01		0.22 0.01 1.39		10.31 33.37 0.22	
$mean\left[\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)\right]$	0.15 0.00 0.39	0.14 0.00 0.38	0.86 1.40 0.01	0.83 1.41 0.02	0.32 0.27 9.14	0.31 0.01 77.03	0.92 0.80 0.72	0.55 0.00 $1.0 \cdot 10^2$
$var\left[\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)\right]$	0.03 0.00 14.73	0.02 0.00 11.20	0.13 0.26 0.00	0.21 0.74 0.00	0.34 1.74 $> 10^4$	0.53 0.00 $> 10^4$	0.00 0.30 0.58	0.18 0.00 $> 10^4$

TABLE 4.17: Simulation results of maximum likelihood estimation (ML) and estimation based on Firths' modified likelihood (FL).
Sample size: N=100
Design: 8-point
a) Percent of the simulations where the ML estimate did not exist
b) Percent of the simulations where the estimate degenerated

Model	"HIGH-WIDE" $\boldsymbol{\theta} = (2, -0.1, 0)'$		"HIGH-NARROW" $\boldsymbol{\theta} = (2, -4, 0)'$		"LOW-WIDE" $\boldsymbol{\theta} = (-2, -0.1, 0)'$		"LOW-NARROW" $\boldsymbol{\theta} = (-2, -4, 0)'$	
Estimator	ML	FL	ML	FL	ML	FL	ML	FL
a	0		0.02		0.01		0.78	
b	0	0	0	0	0.01	0.04	0	0.23
mse	**0.16** **0.00** **0.21**	**0.13** **0.00** **0.18**	**0.47** **2.15** **0.01**	**0.54** **3.17** **0.01**	**0.24** **0.36** **1.46**	**0.19** **0.01** 2.49	**0.32** **3.43** **0.07**	**0.86** **15.73** **0.80**
$mean\left(\widehat{\boldsymbol{\theta}}\right)$	2.08 −0.10 −0.01	2.01 −0.10 −0.01	2.14 −4.28 −0.00	1.92 −3.68 −0.00	−1.74 −0.35 0.00	−1.87 −0.08 0.01	−1.84 −2.65 0.01	−2.85 −0.04 −0.10
$var\left(\widehat{\boldsymbol{\theta}}\right)$	0.15 0.00 0.21	0.14 0.00 0.18	0.45 2.08 0.00	0.53 3.08 0.01	0.18 0.29 1.46	0.17 0.01 2.49	0.30 1.61 0.07	0.15 0.09 0.79
$AV\left(\widehat{\boldsymbol{\theta}}\right)$	0.13 0.00 0.15		0.40 1.92 0.00		0.17 0.01 1.56		0.63 13.41 0.05	
$mean\left[\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)\right]$	0.14 0.00 0.22	0.14 0.00 0.21	0.49 2.23 0.00	0.42 1.75 0.01	0.26 0.24 20.89	0.25 0.01 $3.8 \cdot 10^2$	0.49 6.28 0.12	0.34 0.02 75.36
$var\left[\widehat{V}\left(\widehat{\boldsymbol{\theta}}\right)\right]$	0.00 0.00 3.42	0.00 0.00 0.20	0.08 0.98 0.00	0.15 2.17 0.00	0.11 1.02 $> 10^4$	0.38 0.00 $> 10^4$	0.02 11.73 0.02	0.01 0.17 $> 10^4$

Chapter 5

Sequential Designs

One of the main interests for RSM is to determine the optimum operating conditions, i.e. to locate the point of maximum or minimum response. Examples on response variables to be optimized are different quality aspects of a product, such as yield or strength. The fact that the optimal design depends on the unknown true model parameters for the quadratic logistic model is troublesome. Of course, there would be no point in performing experiments if the true parameters were known. One approach to solve this problem is to use sequential designs. Starting with a preliminary estimate or guess of the parameters a locally optimal design can be constructed. This design is then used to update the parameter estimates, which in turn leads to another locally optimal design, new parameter estimates and so on. Sequential designs for binary data with the purpose to estimate a percentile of the response curve are treated in Wu (1985).

The theory of stochastic approximation provides an alternative method for estimation of the optimum point of a response function. Stochastic approximation started with the work of Robbins and Monro (1951) and Kiefer and Wolfowitz (1952) and is a nonparametric sequential approach. The design points are determined successively according to a recursive scheme such that the resulting sequence will converge to the point of optimum response. A stochastic approximation method is evaluated for binary data in the case of estimating a percentile of the response curve in Wu (1985). Wu (1986) goes through the connections and differences between the stochastic approximation method and a parametric approach based on maximum likelihood estimation.

Here, two sequential design approaches are considered. The first sequential design is an adaptive nonparametric stochastic approximation approach based on the recursion of Kiefer and Wolfowitz. In accordance with a recursive scheme successive observations are made on the response variable in a way that the sequence of design points $x_1, x_2, ...$ converges to μ. This approach is adaptive in the sense that it makes use of the information gained so far by adjusting the step from x_r to x_{r+1}. The second approach is parametric; optimal designs are derived sequentially based on the assumption that the most recent parameter estimates are true. Locally c-optimal designs consisting of two equally weighted designs points are computed at each stage, using two versions of the standardized information matrix. The first is the regular version that reflects the information in the candidate design and the second one is a weighted version that also takes into account the information in the already observed points.

5.1 A Nonparametric Sequential Design

Let $f(x)$ denote a response function that is unknown. Robbins and Monro (1951) give a stochastic approximation method for finding the solution $x = \theta$ to the equation $f(x) = d$, where d is a constant. It is assumed that $f(x) = d$ has a unique solution θ and that for every x observations can be made on a random variable $Y(x)$ such that $E[Y(x)] = f(x)$. Starting at an arbitrary x_1 consecutive observations are made on $Y(x)$ at $x_2, x_3, ...$ in such a way that x_r converges to θ as $r \to \infty$. At x_r the next design point, x_{r+1}, is chosen according to the following scheme

$$x_{r+1} = x_r + a_r (y_r - d), \tag{5.1}$$

where $\{a_r\}$ is a fixed infinite sequence of decreasing positive constants satisfying $\sum a_r^2 < \infty$. The choice of the constants $\{a_r\}$ is important for the performance of the sequence $\{x_r\}$. In the sense of achieving minimal asymptotic variance it is optimal to set a_r equal to $-[rf'(\theta)]^{-1}$, see e.g. Chung (1954) and Sacks (1958). However, $f'(\theta)$ is generally unknown and needs to be estimated. One possible estimator is the least squares estimator of the slope in the regression Y on x. An adaptive version of the Robbins-Monro procedure, where the estimate of $f'(\theta)$ is updated at each step by using the

least squares estimator, is given by

$$\begin{aligned} x_{r+1} &= x_r - \left(n\widehat{\beta}_r\right)^{-1}(y_r - d), \\ \text{where} \quad \widehat{\beta}_r &= \frac{\sum y_i (x_i - \overline{x}_r)}{\sum (x_i - \overline{x}_r)^2}, \quad \overline{x}_r = \frac{\sum x_r}{n}. \end{aligned} \tag{5.2}$$

This procedure is described by Anbar (1978) and was reviewed and evaluated for binary data in Wu (1985). The adaptive procedure (5.2) was proven to be asymptotically equivalent to the nonadaptive Robbins-Monro procedure (5.1) with $a_r = -[rf'(\theta)]^{-1}$ by Anbar (1978) and Lai and Robbins (1979).

The Robbins-Monro procedure was further developed by Kiefer and Wolfowitz (1952) and extended to the case of estimating the maximum of a response function. The unknown response function $f(x)$ has its maximum at the point θ and $f(x)$ is assumed to be strictly increasing (decreasing) for $x < \theta$ $(x > \theta)$. Furthermore it is assumed that observations can be taken on the random variable $Y(x)$ at any level x and that $E[Y(x)] = f(x)$. The principle is the same as for the Robbins-Monro procedure, that is successive observations are made on $Y(x)$ according to a specified scheme. The difference is that observations are made in pairs at each step. Starting at an arbitrary x_1, the following $x_2, x_3, ...$ are obtained by making observations at $x_r \pm c_r$, i.e. both $Y(x_r - c_r)$ and $Y(x_r + c_r)$ are observed at each step. The sequence $\{x_r\}$ is defined as

$$x_{r+1} = x_r + a_r \frac{y(x_r + c_r) - y(x_r - c_r)}{2c_r} = x_r + a_r z_r \tag{5.3}$$

and converges to θ as $r \to \infty$. $\{a_r\}$ and $\{c_r\}$ are preassigned infinite sequences of positive numbers such that

$$\begin{aligned} c_r &\to 0, \\ \sum a_r &= \infty, \\ \sum a_r c_r &< \infty, \\ \sum a_r^2 c_r^2 &< \infty. \end{aligned}$$

For example $a_r = r^{-1}$ and $c_r = r^{-1/3}$ satisfy these conditions. The random variable $Z_r = [Y(x_r + c_r) - Y(x_r - c_r)]/(2c_r)$ can be viewed as an

approximation to the derivative of the response function at x_r. This reduces the problem to find the maximum of $f(x)$ to that of finding the solution to the equation $f'(x) = 0$. The recursion (5.3) can then be thought of as a special case of the Robbins-Monro method for finding the solution $x = \mu$ to $f'(x) = 0$ by making successive observations on Z.

In this thesis the procedure of Kiefer and Wolfowitz is combined with the adaptive Robbins-Monro procedure (5.2) described above for estimating the maximum of the response function $\pi(x)$, abbreviated as KW hereafter. At each step m observations are taken at $x_r \pm c_r$ so that the numerical derivative becomes $Z_r = [Y(x_r + c_r) - Y(x_r + c_r)] / (2mc_r)$ where $Y(x_r + c_r) \sim bin(m, \pi(x_r + c_r))$ and $Y(x_r - c_r) \sim bin(m, \pi(x_r - c_r))$. A graphical illustration is given in Figure 5.1. Starting at an arbitrary x_1, the value on x_r is then updated via the adaptive Robbins-Monro procedure

$$x_{r+1} = x_r - \left(n\widehat{\beta}_r\right)^{-1} z_r$$

where $\widehat{\beta}_r$ is the least squares estimator of the slope in the regression Z on x. The estimate of the optimum point at the $r : th$ step will then be given by

$$\widehat{\theta}_r = x_r.$$

Kiefer and Wolfowitz (1952) put some conditions on the response function that prevents it from being too steep or to flatten out towards zero. If the curve is too steep it may cause unduly large changes in x and any observations taken where the response curve is zero will be uninformative, making it impossible to know in which direction to take the next step. Kiefer and Wolfowitz comment however, that it will be sufficient if the conditions are fulfilled in an interval $[D_1,\ D_2]$. No observations will be taken outside the limits of this interval. The flatness of the logistic response curve may pose some problems unless there is knowledge about an appropriate interval before the experiment is started. In some situations it may be the case that such information is not available. The success of this procedure is also dependent upon good choices of the starting values x_1, β_1 and c_1. If the starting point is too far off, there will be essentially zero probability of obtaining a response so that no information can be gained. The starting value for c_r should be small enough to avoid unduly large changes in x and large enough to enable the sequence to move away from a bad starting

point. The relation between the number of observations taken at each step and the number of steps is another aspect that probably will be important.

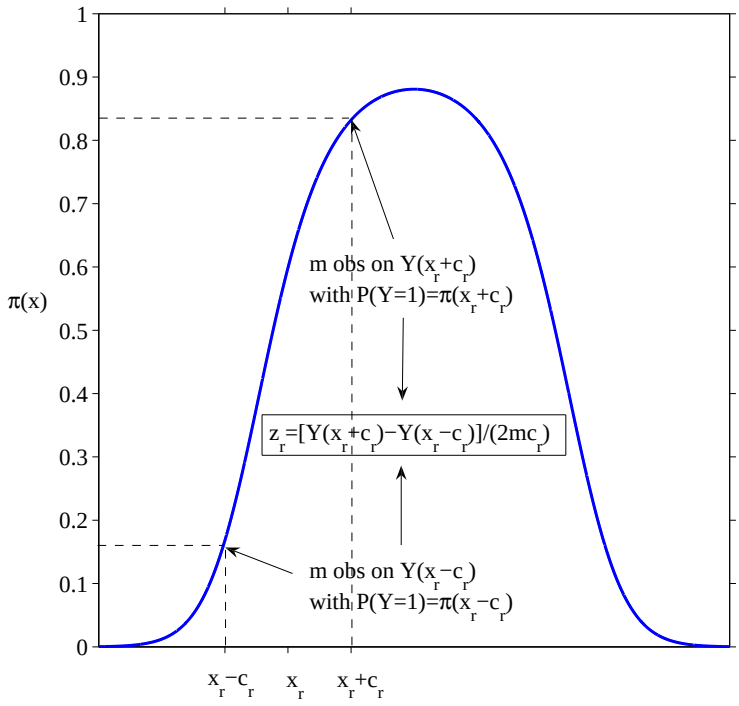


FIGURE 5.1: A graphical illustration of the KW approach.

5.2 c-optimal Sequential Designs

The point of maximum response of the quadratic response curve is equal to the parameter μ. Estimating μ with minimum variance would be desirable justifying the use of a c-optimal design with $\mathbf{c} = (0, 0, 1)'$. As an attempt to handle the problem of parameter dependence, the c-optimal design is updated sequentially. It was established in Chapter 3 that a c-optimal design for estimation of the point $x = \mu$ consists of two points with equal weight. In essence, two points are taken at a certain distance from each other, a distance that does not necessarily decrease, as opposed to the KW approach. This parametric sequential approach, called COPT, can be

described by the following steps:

1. Choose an initial design $\xi_{(initial)}$
2. Estimate the parameters $\widehat{\boldsymbol{\theta}}^{(initial)}$
3. Minimize $\widehat{\mathbf{c}}^T\mathbf{M}^{-1}\left(\xi, \widehat{\boldsymbol{\theta}}^{(initial)}\right)\widehat{\mathbf{c}}$ to find the locally c-optimal design given $\widehat{\boldsymbol{\theta}}^{(initial)} \rightarrow \xi_1^* = \left\{ \begin{array}{cc} x_1 & x_2 \\ 0.5 & 0.5 \end{array} \right\}$
4. Take m observations at the design points in ξ_1^* and estimate the parameters: $\widehat{\boldsymbol{\theta}}^{(1)}$
5. Minimize $\widehat{\mathbf{c}}^T\mathbf{M}^{-1}\left(\xi, \widehat{\boldsymbol{\theta}}^{(1)}\right)\widehat{\mathbf{c}}$ to find the locally c-optimal design given $\widehat{\boldsymbol{\theta}}^{(1)} \rightarrow \xi_2^* = \left\{ \begin{array}{cc} x_3 & x_4 \\ 0.5 & 0.5 \end{array} \right\}$

etc.

The parameter estimates $\widehat{\boldsymbol{\theta}}$ are the usual maximum likelihood estimates. The estimate of the optimum point after r steps is then given by $\widehat{\mu}_r$.

If the initial design contains N^* observations and $2m$ observations are taken at each step there are $N^* + 2mr = N_{tot}$ observations in total after r steps. The observed standardized information matrix can then be expressed as

$$\mathbf{M}_{obs}\left(\xi, \widehat{\boldsymbol{\theta}}\right) = \sum_{i=1}^{N_{tot}} \frac{1}{N_{tot}} \widehat{\pi}_i \left(1 - \widehat{\pi}_i\right) \begin{pmatrix} 1 \\ (x_i - \widehat{\mu})^2 \\ -2\widehat{\beta}\left(x_i - \widehat{\mu}\right) \end{pmatrix}' \begin{pmatrix} 1 \\ (x_i - \widehat{\mu})^2 \\ -2\widehat{\beta}\left(x_i - \widehat{\mu}\right) \end{pmatrix}$$

where $\widehat{\pi}_i = \left(e^{\mathbf{x}_i'\widehat{\boldsymbol{\theta}}}\right) / \left(1 + e^{\mathbf{x}_i'\widehat{\boldsymbol{\theta}}}\right)$. Another version of the previous sequential design, called COPT2, makes use of the observed information matrix in the following way. $\mathbf{M}\left(\xi, \widehat{\boldsymbol{\theta}}\right)$ is replaced by a weighted information matrix computed as

$$\mathbf{M}_W\left(\xi, \widehat{\boldsymbol{\theta}}\right) = \frac{N_{tot}}{N_{tot} + 2m}\mathbf{M}_{obs}\left(\xi_{obs}, \widehat{\boldsymbol{\theta}}\right) + \frac{2m}{N_{tot} + 2m}\mathbf{M}\left(\xi, \widehat{\boldsymbol{\theta}}\right)$$

and $\widehat{\mathbf{c}}^T \mathbf{M}_W^{-1} \left(\xi, \widehat{\boldsymbol{\theta}} \right) \widehat{\mathbf{c}}$ is minimized instead. $\mathbf{M}_{obs}$ is the information matrix based on all observations up to this point and ξ_{obs} is the design consisting of these observations. This is potentially an improvement because it also takes into account the information from the preceding steps. Besides the changed information matrix everything is the same as for the COPT procedure.

Chapter 6

A Simulation Study of Sequential Designs

6.1 Simulation Setup

In this chapter the efficiencies of the two proposed sequential designs are evaluated in a simulation study. The results of the simulations are also reported in Fackle Fornius (2008). There are many questions to bring clarity about including: Is any of the approaches superior? Is there an optimal choice of c for the KW approach? Is the weighted version of the information matrix preferred over the unweighted? What is best, taking few steps with many observations at each step, or taking many steps with few observations? Are the performances robust to misspecifications of the model?

Two parameter sets are considered here: $\boldsymbol{\theta}_A = (2, -0.1, 0)'$ and $\boldsymbol{\theta}_B = (2, -4, 0)'$. The sequential design approaches described in the preceding chapter are analyzed in the simulations. Different sample sizes ranging from $N = 200$ to $N = 1000$ as well as different batch sizes are explored. To study the effects of misspecifications in the model an alternative to the linear predictor is also tested.

All the three approaches; KW, COPT and COPT2 are started from the same initial design before they take separate paths. Two different initial designs are used, see Table 6.1. Design 1 is constructed to be better than Design 2 for estimation of the point of maximum response μ. It is better in the sense that it is symmetric around $\mu = 0$. c-optimal designs are

similarly characterized by design points placed symmetrically around μ, but limited to two points. However, it is not possible to obtain maximum likelihood estimates of the model parameters with a c-optimal design because two points are not sufficient to estimate three parameters. Design 2 is constructed based on the same principle but pushed aside, intended to represent a beforehand assumption that μ is close to 2. The initial designs consist of a fairly large amount of observations ($N^* = 150$) to avoid the problem of non-existing maximum likelihood estimates, which is also a motive for choosing many design points.

TABLE 6.1: Two initial designs.

Design 1	$\left\{\begin{array}{c} -5 \\ 5/150 \end{array}\right.$	$\begin{array}{c} -3 \\ 10/150 \end{array}$	$\begin{array}{c} -1 \\ 20/150 \end{array}$	$\begin{array}{c} -0.5 \\ 40/150 \end{array}$	$\begin{array}{c} 0.5 \\ 40/150 \end{array}$	$\begin{array}{c} 1 \\ 20/150 \end{array}$	$\begin{array}{c} 3 \\ 10/150 \end{array}$	$\left.\begin{array}{c} 5 \\ 5/150 \end{array}\right\}$
Design 2	$\left\{\begin{array}{c} -3 \\ 5/150 \end{array}\right.$	$\begin{array}{c} -1 \\ 10/150 \end{array}$	$\begin{array}{c} 1 \\ 20/150 \end{array}$	$\begin{array}{c} 1.5 \\ 40/150 \end{array}$	$\begin{array}{c} 2.5 \\ 40/150 \end{array}$	$\begin{array}{c} 3 \\ 20/150 \end{array}$	$\begin{array}{c} 5 \\ 10/150 \end{array}$	$\left.\begin{array}{c} 7 \\ 5/150 \end{array}\right\}$

The c-optimal procedures are executed as follows. At the start, observations are made on the response variable Y at the design points specified by the initial design. Initial maximum likelihood estimates $\widehat{\boldsymbol{\theta}}^{(initial)}$ are then calculated. Both the COPT and COPT2 approaches are now ready to start from Step 3 (as described in Section 5.2) with the minimization of $\widehat{\mathbf{c}}^T \mathbf{M}^{-1}\left(\xi, \widehat{\boldsymbol{\theta}}^{(initial)}\right)\widehat{\mathbf{c}}$. The formulae derived in Chapter 3 can be used to carry out this task. m response values are now generated at each of the two c-optimal design points, resulting in a new c-optimal design and so on. $2m$ observations are generated per batch until the total number of observations amounts to N.

The KW approach needs at least two observations on

$$Z = \left[Y\left(x + c\right) - Y\left(x - c\right)\right] / \left(2mc\right)$$

before a least squares estimate of the slope parameter in the Z on x regression can be obtained. As discussed before the KW approach is also sensitive to the choices of x_1 and β_1. Some effort is therefore put in selecting good

starting values. A preliminary estimate of the point of maximum response is obtained from the initial design. The first design point in KW is set equal to this estimate:

$$x_1 = \widehat{\mu}_{initial}.$$

This point plus the two points given by $0.7\pi\,(\widehat{\mu}_{initial})$, as depicted in Figure 6.1, comprise a start design for the KW approach. There is no elaborate reason for choosing 0.7 specifically, other than that it seemed to work good in comparison with several alternatives. Observations are made at these three design points that will result in three observations on Z which then can be used to obtain the starting value β_1. For $r = 2, 3, ...$ the design points are obtained according to (5.3) from Section 5.1. Several values on c_r are tested, namely $c_r = cr^{-1/3}$ with $c = 1, 3, 5, 7, 9, 11, 13, 15$. Three batch sizes are considered: $m = 5$, $m = 10$ and $m = 20$.

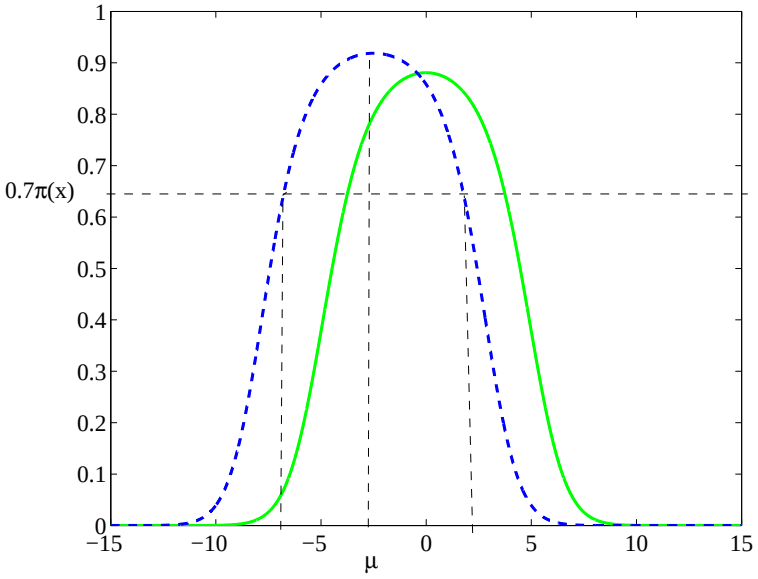


FIGURE 6.1: A representation of the choice of first design points for the KW approach. The true response curve is given by the solid line and the estimated curve is given by the broken line.

To examine the effects of a misspecified model another model with a different predictor η is also evaluated. Model C defined according to

$$\begin{aligned} \eta_i &= -abs\,(x)\,, \\ \pi_i &= \frac{e^{\eta_i}}{1+e^{\eta_i}}, \end{aligned}$$

is displayed in Figure 6.2 together with models A and B. The optimum point has the same location as for the logistic models but the response curve is lower at the maximum and it has a different shape. The responses are generated according to this model while the COPT and COPT2 approaches still (incorrectly) assume that the three parameter logistic model is true. The KW approach is nonparametric apart from the initial stage where the (inaccurate) logistic assumption remains unchanged.

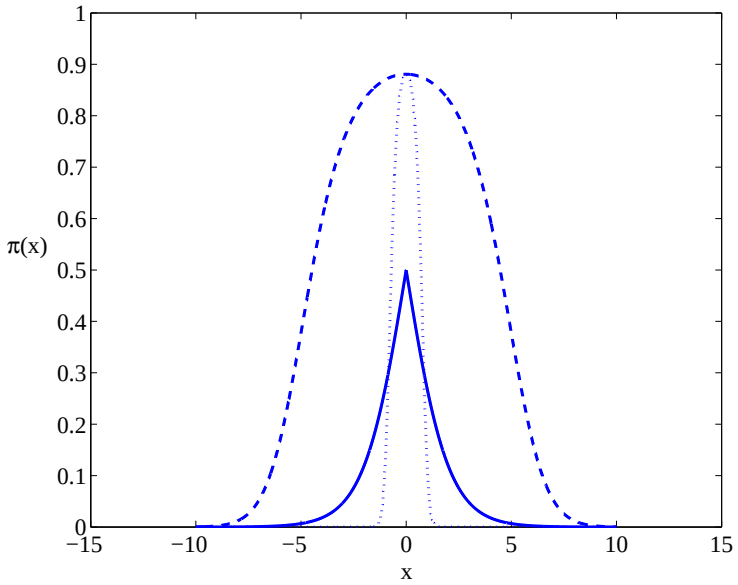


FIGURE 6.2: Models A (broken line), B (dotted line) and C (solid line).

6.2 Results

The simulation results are presented in Tables 6.2 to 6.6. Each table shows the mean squared error (mse) of $\widehat{\mu}$ based on 500 simulated samples. The mse of $\widehat{\mu}$ before any of the sequential procedures is started, which is based

on the initial maximum likelihood estimates of the parameters, is presented in the second column. In the following columns the mse of $\widehat{\mu}$ is given after in turn $N = 200$, $N = 600$ and $N = 1000$ observations in the sequential designs, in addition to the $N^* = 150$ observations that was made at the initial stage. If $\widehat{\mu}_{initial}$ is a "bad" point that is far from μ it may cause difficulties. For the parametric approaches there may be numerical problems related to the parameter estimation. These problems appear in either that the maximum number of iterations is reached or that the computations result in NaN. NaN stands for "not a number" and turns up in Matlab when the computations do not produce a numerical result. Being at a position where the probability to obtain a response is nearly zero, it is impossible for the KW approach to know in which direction to take next step; for such cases this approach never leaves the starting point. It can also happen that the KW sequence ends up at such a position even though the starting point is "good". $\widehat{\mu}$ is considered deteriorated when $|\widehat{\mu}| > 10$. The number of deteriorated starting points is shown in brackets in the second column. The simulation samples that failed, either because of numerical problems or because $\widehat{\mu}$ is deteriorated are discarded and the number of such samples is shown in brackets. In each table there are four sections; $a - d$ where a, b and c contain the results for $m = 20$, $m = 10$ and $m = 5$, respectively. As a reference, section d contains the mse of $\widehat{\mu}$ based on taking all $N^* + N$ observations at the initial design and making one maximum likelihood estimate of μ. The $N^* + N$ observations are divided among the initial design points according to the same distribution as for the case with N^* observations. This nonsequential alternative is tested to make certain that one cannot do just as good without bothering about any sequential issues.

Model A, design 1. The mse for the two parametric approaches; COPT and COPT2, are remarkably similar for all N. The choice between different values of m does not seem to have any impact on the mse. For the KW approach it is evident that the constant c is important, the best choice of c varies from 7 to 11. If c is too small or too large there is a greater risk that the sequence degenerates to a point that is too far off from where it cannot recover. This happens more often when m is small. The best KW approach has about twice the mse of the COPT approaches for $m = 20$ and $m = 10$ and up to three times for $m = 5$. Both the sequential procedures beats the alternative of making all observations at one time.

Model A, design 2. The number of "bad" starting points has increased compared to the previous case because initial design 2 is inferior to design 1. The COPT approaches manage to sort out some of these problematic cases whereas the number stays the same or even increases for the KW approach. The mse for COPT and COPT2 are still similar. Although the mse is greater compared to when design 1 was used at the initial stage it is about the same size as before for all N. For the KW method the best choice of c seems to be somewhere between 9 and 13. The mse for the KW approach does not differ that much compared to when initial design 1 was used for the well-behaved cases. However it should be noted that the problematic cases have increased in number. The COPT approaches are preferred over the KW approach for all N. The sequential procedures are still performing better than the non-sequential alternative for all N with mse in view, while there exist no failing cases when all observations are made at the same time. All three alternatives for m give about the same results.

Model B, design 1. The mse is lower even at the initial stage for this model and it decreases even more as N increases for the COPT approaches. The KW scheme remains at $\widehat{\mu}_{initial}$ almost every time which is also reflected in the fact that the mse is not changing. The COPT routines are associated with the lowest mse followed by the non-sequential alternative. The results are also the same regardless of the size of m.

Model B, design 2. Another problem that can occur is that of nonexisting maximum likelihood estimates. If the data pattern is such that the responses are separated from the nonresponses it is not possible to obtain any maximum likelihood estimates, see Albert and Anderson (1984). It is only when data are overlapped such that it is not possible to separate the two different response types, that parameter estimates exist. For design 2 there are mainly two data points (± 1) where both responses and non-responses can be observed, at all other design points the probability for a response is nearly zero. Almost all samples will therefore be separated, only about 2 percent of the samples are overlapped which is the reason for not presenting these results. This shows how important it is to take the problems with non-existing estimates seriously.

Model C, design 1. The mse of $\widehat{\mu}$ at the initial design is lower when compared to model A but higher in comparison with model B. The COPT and COPT2

procedures continue to perform equally well. As more observations are made the mse decreases although the reduction is not as large as for model A, so the mse is a little higher than for the other two models for all N. It would be surprising if the parametric approaches were completely unaffected by the misspecification of the model. The nonparametric approach on the other hand, that would be expected to handle the erroneous model the best, works no good. For $m = 20$ the mse even increases with N for some constants c and for some c it is unchanged. For $m = 10$ or $m = 5$ the mse is at best slightly reduced. The response curve for model C is lower and has a different shape which results in that z becomes equal to zero more often. As a consequence it is more common that the sequence gets stuck at some point. The descending order of performance turned out to be as follows; parametric sequential > parametric nonsequential >nonparametric sequential.

Model C, design 2. The fact that design 2 is poor is reflected in a higher mse at the initial stage. COPT and COPT2 recover to almost the same level as when design 1 was used in the following steps. The results for the nonparametric procedure show the same tendency that the mse sometimes gets even higher as N grows or just decreases a little. The same order of performance that puts the parametric sequential design before the parametric nonsequential followed by the nonparametric method prevails.

6.3 Conclusions

The parametric approach based on constructing a c-optimal design at each step was superior in all cases that were examined here. It even excelled the nonparametric approach in the case with an incorrect model. The results suggest that the problems of parameter dependence can be worked out by using any of the two parametric approaches. The effect of having the wrong beforehand idea of the location of the optimum point almost disappeared after taking 200 observations sequentially. The sequential parametric approaches always outperformed the nonsequential strategy of maximum likelihood estimation of μ based on making all observations at once. The potential improvement from using the observed information matrix in the construction of designs failed to appear. In fact the performances of COPT and COPT2 were almost identical.

Choosing a good constant c is of crucial importance for the success of the KW approach. In this setting a good choice turned out to be choosing c somewhere around 10. The main advantage with the nonparametric approach is its simplicity, that is, not having to make any distributional assumptions. Opposite to what might be expected it was more sensitive to the misspecified model than the parametric approaches. There were several cases where the KW approach was outperformed by the nonsequential alternative of making all observations at the initial stage. It was dependent upon good initial estimates and thereby on a good initial design in the sense of avoiding a large number of degenerated estimates. There were also many cases for which the KW approach did not leave the starting point at all. Possibly, an interval with appropriate limits that prevents the sequence to degenerate could be a remedy.

Non-existing maximum likelihood estimates may cause problems as was apparent by the lack of results for model B in combination with design 2. There were no such problems for model A in combination with design 2 and the only difference between these two response curves is the scale on the control variable. This implies that considering the measurement scale of the control variable might be worthwhile.

The different combinations of the number of observations taken at each step and the number of steps did not have that large impact on the results for neither of the approaches, though the problems with degenerated cases tended to be more frequent for $m = 5$.

The empirical results imply that there are gains to be made by using a sequential c-optimal approach for finding the point of optimum response. More extensive evaluations and evaluations of theoretical nature are required.

TABLE 6.2: Simulation results of three sequential designs for estimating the optimum point in model A with initial design 1, mse based on 500 samples .

Design	initial design	$N = 200$	$N = 600$	$N = 1000$
		a) m=20		
$COPT$	0.41 (1)	0.020 (1)	0.0075 (0)	0.0044 (0)
$COPT2$	0.41 (1)	0.023 (1)	0.0071 (0)	0.0043 (0)
$KW, c = 1$	0.41 (1)	1.25 (1)	0.87 (1)	0.84 (1)
$KW, c = 3$	0.41 (1)	0.27 (1)	0.11 (1)	0.088 (1)
$KW, c = 5$	0.41 (1)	0.075 (1)	0.043 (1)	0.037 (1)
$KW, c = 7$	0.41 (1)	0.040 (1)	0.020 (1)	0.017 (1)
$KW, c = 9$	0.41 (1)	0.086 (1)	0.015 (1)	0.010 (1)
$KW, c = 11$	0.41 (1)	0.24 (1)	0.023 (1)	0.011 (1)
$KW, c = 13$	0.41 (1)	0.35 (1)	0.072 (1)	0.021 (1)
$KW, c = 15$	0.41 (1)	0.40 (1)	0.17 (1)	0.057 (1)
		b) m=10		
$COPT$	0.55 (1)	0.024 (0)	0.0084 (0)	0.0048 (0)
$COPT2$	0.55 (1)	0.022 (0)	0.0071 (0)	0.0044 (0)
$KW, c = 1$	0.55 (1)	2.40 (10)	1.73 (14)	1.91 (14)
$KW, c = 3$	0.55 (1)	0.21 (4)	0.16 (4)	0.14 (4)
$KW, c = 5$	0.55 (1)	0.10 (5)	0.060 (5)	0.056 (5)
$KW, c = 7$	0.55 (1)	0.047 (2)	0.029 (2)	0.025 (2)
$KW, c = 9$	0.55 (1)	0.034 (2)	0.019 (2)	0.013 (2)
$KW, c = 11$	0.55 (1)	0.072 (2)	0.015 (2)	0.011 (2)
$KW, c = 13$	0.55 (1)	0.20 (1)	0.020 (1)	0.011 (1)
$KW, c = 15$	0.55 (1)	0.33 (1)	0.040 (1)	0.014 (1)
		c) m=5		
$COPT$	0.51 (1)	0.021 (0)	0.0075 (0)	0.0045 (0)
$COPT2$	0.51 (1)	0.023 (0)	0.0071 (0)	0.0047 (0)
$KW, c = 1$	0.51 (1)	4.66 (35)	3.67 (40)	3.87 (46)
$KW, c = 3$	0.51 (1)	0.47 (3)	0.27 (3)	0.24 (3)
$KW, c = 5$	0.51 (1)	0.15 (8)	0.11 (8)	0.085 (8)
$KW, c = 7$	0.51 (1)	0.10 (4)	0.052 (4)	0.050 (4)
$KW, c = 9$	0.51 (1)	0.048 (3)	0.030 (3)	0.025 (3)
$KW, c = 11$	0.51 (1)	0.099 (3)	0.022 (3)	0.016 (3)
$KW, c = 13$	0.51 (1)	0.16 (1)	0.043 (1)	0.041 (1)
$KW, c = 15$	0.51 (1)	0.10 (2)	0.017 (2)	0.012 (2)
		d) all $N^* + N$ observations at the initial design		
	0.45 (1)	0.14 (0)	0.055 (0)	0.037 (0)

TABLE 6.3: Simulation results of three sequential designs for estimating the optimum point in model A with initial design 2, mse based on 500 samples.

Design	initial design	$N = 200$	$N = 600$	$N = 1000$
		a) m=20		
$COPT$	2.67 (14)	0.021 (14)	0.016 (10)	0.0043 (9)
$COPT2$	2.67 (14)	0.022 (13)	0.0081 (9)	0.0046 (8)
$KW, c = 1$	2.67 (14)	1.92 (32)	1.18 (32)	1.07 (32)
$KW, c = 3$	2.67 (14)	0.55 (29)	0.29 (29)	0.27 (29)
$KW, c = 5$	2.67 (14)	0.25 (23)	0.21 (22)	0.20 (22)
$KW, c = 7$	2.67 (14)	0.27 (20)	0.18 (20)	0.18 (20)
$KW, c = 9$	2.67 (14)	0.12 (20)	0.022 (20)	0.013 (20)
$KW, c = 11$	2.67 (14)	0.68 (16)	0.19 (16)	0.17 (16)
$KW, c = 13$	2.67 (14)	0.98 (18)	0.25 (18)	0.18 (18)
$KW, c = 15$	2.67 (14)	1.38 (16)	0.42 (16)	0.23
		b) m=10		
$COPT$	2.86 (23)	0.022 (10)	0.0080 (4)	0.0042 (4)
$COPT2$	2.86 (23)	0.023 (11)	0.0068 (2)	0.0044 (2)
$KW, c = 1$	2.86 (23)	4.02 (48)	3.14 (53)	2.59 (57)
$KW, c = 3$	2.86 (23)	0.25 (42)	0.19 (42)	0.17 (42)
$KW, c = 5$	2.86 (23)	0.099 (32)	0.27 (31)	0.27 (31)
$KW, c = 7$	2.86 (23)	0.067 (24)	0.038 (23)	0.040 (23)
$KW, c = 9$	2.86 (23)	0.11 (24)	0.030 (24)	0.020 (24)
$KW, c = 11$	2.86 (23)	0.11 (21)	0.022 (21)	0.018 (21)
$KW, c = 13$	2.86 (23)	0.25 (23)	0.035 (23)	0.017 (23)
$KW, c = 15$	2.86 (23)	0.55 (23)	0.052 (23)	0.021 (23)
		c) m=5		
$COPT$	2.26 (21)	0.10 (8)	0.0080 (7)	0.0044 (7)
$COPT2$	2.26 (21)	0.021 (9)	0.0069 (9)	0.0041 (9)
$KW, c = 1$	2.26 (21)	4.99 (66)	5.02 (75)	5.08 (82)
$KW, c = 3$	2.26 (21)	0.60 (33)	0.31 (34)	0.30 (34)
$KW, c = 5$	2.26 (21)	0.14 (32)	0.10 (32)	0.099 (32)
$KW, c = 7$	2.26 (21)	0.22 (27)	0.25 (27)	0.24 (27)
$KW, c = 9$	2.26 (21)	0.10 (28)	0.080 (28)	0.076 (28)
$KW, c = 11$	2.26 (21)	0.053 (26)	0.033 (26)	0.025 (26)
$KW, c = 13$	2.26 (21)	0.17 (26)	0.019 (26)	0.016 (26)
$KW, c = 15$	2.26 (21)	0.16 (26)	0.024 (26)	0.020 (26)
		d) all $N^* + N$ observations at the initial design		
	2.67 (12)	0.87 (1)	0.21 (0)	0.11 (0)

TABLE 6.4: Simulation results of three sequential designs for estimating the optimum point in model B with initial design 1, mse based on 500 samples .

Design	initial design	$N = 200$	$N = 600$	$N = 1000$
		a) m=20		
$COPT$	0.0019 (1)	0.00041 (0)	0.00015 (1)	0.00010 (1)
$COPT2$	0.0019 (1)	0.00040 (0)	0.00015 (1)	$9.8 \cdot 10^{-5}$ (1)
$KW, c = 1$	0.0019 (1)	0.020 (1)	0.24 (1)	0.30 (1)
$KW, c = 3$	0.0019 (1)	0.0019 (1)	0.0019 (1)	0.0018 (1)
$KW, c = 5$	0.0019 (1)	0.0019 (1)	0.0019 (1)	0.0019 (1)
$KW, c = 7$	0.0019 (1)	0.0019 (1)	0.0019 (1)	0.0019 (1)
$KW, c = 9$	0.0019 (1)	0.0019 (1)	0.0019 (1)	0.0019 (1)
$KW, c = 11$	0.0019 (1)	0.0019 (1)	0.0019 (1)	0.0019 (1)
$KW, c = 13$	0.0019 (1)	0.0019 (1)	0.0019 (1)	0.0019 (1)
$KW, c = 15$	0.0019 (1)	0.0019 (1)	0.0019 (1)	0.0019 (1)
		b) m=10		
$COPT$	0.0020 (1)	0.00046 (1)	0.00018 (1)	$9.3 \cdot 10^{-5}$ (1)
$COPT2$	0.0020 (1)	0.00048 (1)	0.00016 (1)	$9.8 \cdot 10^{-5}$ (1)
$KW, c = 1$	0.0020 (1)	0.39 (3)	0.71 (3)	0.76 (3)
$KW, c = 3$	0.0020 (1)	0.0020 (1)	0.0021 (1)	0.0014 (1)
$KW, c = 5$	0.0020 (1)	0.0020 (1)	0.0020 (1)	0.0020 (1)
$KW, c = 7$	0.0020 (1)	0.0020 (1)	0.0020 (1)	0.0020 (1)
$KW, c = 9$	0.0020 (1)	0.0020 (1)	0.0020 (1)	0.0020 (1)
$KW, c = 11$	0.0020 (1)	0.0020 (1)	0.0020 (1)	0.0020 (1)
$KW, c = 13$	0.0020 (1)	0.0020 (1)	0.0020 (1)	0.0020 (1)
$KW, c = 15$	0.0020 (1)	0.0020 (1)	0.0020 (1)	0.0020 (1)
		c) m=5		
$COPT$	0.0019 (3)	0.00041 (3)	0.00015 (3)	0.00011 (3)
$COPT2$	0.0019 (3)	0.00048 (3)	0.00017 (3)	$9.9 \cdot 10^{-5}$ (3)
$KW, c = 1$	0.0019 (3)	0.13 (6)	0.20 (7)	0.21 (7)
$KW, c = 3$	0.0019 (3)	0.0019 (3)	0.0027 (3)	0.0048 (3)
$KW, c = 5$	0.0019 (3)	0.0019 (3)	0.0019 (3)	0.0018 (3)
$KW, c = 7$	0.0019 (3)	0.0019 (3)	0.0019 (3)	0.0019 (3)
$KW, c = 9$	0.0019 (3)	0.0019 (3)	0.0019 (3)	0.0019 (3)
$KW, c = 11$	0.0019 (3)	0.0019 (3)	0.0019 (3)	0.0019 (3)
$KW, c = 13$	0.0019 (3)	0.0019 (3)	0.0019 (3)	0.0019 (3)
$KW, c = 15$	0.0019 (3)	0.0019 (3)	0.0019 (3)	0.0019 (3)
		d) all $N^* + N$ observations at the initial design		
	0.0020 (3)	0.00080 (0)	0.00043 (0)	0.00024 (0)

TABLE 6.5: Simulation results of three sequential designs for estimating the optimum point in model C with initial design 1, mse based on 500 samples.

Design	initial design	$N = 200$	$N = 600$	$N = 1000$
		a) m=20		
$COPT$	0.13 (0)	0.035 (0)	0.014 (0)	0.0087 (0)
$COPT2$	0.13 (0)	0.036 (0)	0.015 (0)	0.0096 (0)
$KW, c = 1$	0.13 (0)	0.98 (4)	1.34 (5)	1.31 (6)
$KW, c = 3$	0.13 (0)	0.19 (0)	0.21 (0)	0.19 (0)
$KW, c = 5$	0.13 (0)	0.21 (0)	0.24 (0)	0.22 (0)
$KW, c = 7$	0.13 (0)	0.14 (0)	0.17 (0)	0.16 (0)
$KW, c = 9$	0.13 (0)	0.13 (0)	0.12 (0)	0.10 (0)
$KW, c = 11$	0.13 (0)	0.13 (0)	0.13 (0)	0.12 (0)
$KW, c = 13$	0.13 (0)	0.13 (0)	0.13 (0)	0.13 (0)
$KW, c = 15$	0.13 (0)	0.13 (0)	0.13 (0)	0.13 (0)
		b) m=10		
$COPT$	0.12 (0)	0.030 (0)	0.013 (0)	0.0084 (0)
$COPT2$	0.12 (0)	0.032 (0)	0.013 (0)	0.0090 (0)
$KW, c = 1$	0.12 (0)	1.14 (6)	1.35 (8)	1.49 (9)
$KW, c = 3$	0.12 (0)	0.21 (2)	0.23 (3)	0.30 (3)
$KW, c = 5$	0.12 (0)	0.10 (0)	0.34 (2)	0.46 (2)
$KW, c = 7$	0.12 (0)	0.10 (0)	0.060 (1)	0.052 (1)
$KW, c = 9$	0.12 (0)	0.12 (0)	0.097 (0)	0.082 (0)
$KW, c = 11$	0.12 (0)	0.12 (0)	0.10 (0)	0.091 (0)
$KW, c = 13$	0.12 (0)	0.12 (0)	0.11 (0)	0.11 (0)
$KW, c = 15$	0.12 (0)	0.12 (0)	0.12 (0)	0.11 (0)
		c) m=5		
$COPT$	0.13 (0)	0.033 (0)	0.012 (0)	0.0079 (0)
$COPT2$	0.13 (0)	0.034 (0)	0.014 (0)	0.0082 (0)
$KW, c = 1$	0.13 (0)	2.70 (13)	2.95 (18)	2.86 (19)
$KW, c = 3$	0.13 (0)	0.63 (3)	1.01 (4)	1.09 (4)
$KW, c = 5$	0.13 (0)	0.25 (0)	0.54 (3)	0.61 (3)
$KW, c = 7$	0.13 (0)	0.11 (0)	0.15 (0)	0.14 (1)
$KW, c = 9$	0.13 (0)	0.12 (0)	0.10 (0)	0.084 (0)
$KW, c = 11$	0.13 (0)	0.13 (0)	0.10 (0)	0.098 (0)
$KW, c = 13$	0.13 (0)	0.13 (0)	0.12 (0)	0.11 (0)
$KW, c = 15$	0.13 (0)	0.13 (0)	0.12 (0)	0.11 (0)
		d) all $N^* + N$ observations at the initial design		
	0.13 (0)	0.067 (0)	0.031 (0)	0.022 (0)

TABLE 6.6: Simulation results of three sequential designs for estimating the optimum point in model C with initial design 2, mse based on 500 samples.

Design	initial design	$N = 200$	$N = 600$	$N = 1000$
		a) m=20		
$COPT$	0.59 (2)	0.047 (0)	0.015 (0)	0.010 (0)
$COPT2$	0.59 (2)	0.042 (0)	0.015 (0)	0.0097 (0)
$KW, c = 1$	0.59 (2)	2.23 (9)	2.09 (9)	2.04 (9)
$KW, c = 3$	0.59 (2)	0.63 (6)	0.61 (6)	0.69 (6)
$KW, c = 5$	0.59 (2)	0.51 (5)	0.46 (5)	0.41 (5)
$KW, c = 7$	0.59 (2)	0.66 (2)	0.56 (2)	0.58 (2)
$KW, c = 9$	0.59 (2)	0.49 (3)	0.51 (3)	0.48 (3)
$KW, c = 11$	0.59 (2)	0.54 (2)	0.53 (2)	0.46 (2)
$KW, c = 13$	0.59 (2)	0.57 (2)	0.54 (2)	0.52 (2)
$KW, c = 15$	0.59 (2)	0.59 (2)	0.55 (2)	0.49 (2)
		b) m=10		
$COPT$	0.63 (0)	0.044 (0)	0.015 (0)	0.0091 (0)
$COPT2$	0.63 (0)	0.046 (0)	0.016 (0)	0.010 (0)
$KW, c = 1$	0.63 (0)	2.63 (11)	2.82 (14)	2.82 (14)
$KW, c = 3$	0.63 (0)	0.76 (4)	0.90 (4)	1.09 (4)
$KW, c = 5$	0.63 (0)	0.54 (3)	0.55 (4)	0.55 (4)
$KW, c = 7$	0.63 (0)	0.53 (0)	0.44 (0)	0.44 (1)
$KW, c = 9$	0.63 (0)	0.41 (1)	0.32 (1)	0.27 (1)
$KW, c = 11$	0.63 (0)	0.38 (2)	0.28 (3)	0.24 (3)
$KW, c = 13$	0.63 (0)	0.53 (1)	0.46 (1)	0.42 (1)
$KW, c = 15$	0.63 (0)	0.54 (1)	0.49 (1)	0.46 (1)
		c) m=5		
$COPT$	0.52 (2)	0.045 (0)	0.016 (0)	0.0095 (0)
$COPT2$	0.52 (2)	0.035 (0)	0.014 (0)	0.0088 (0)
$KW, c = 1$	0.52 (2)	5.00 (16)	5.19 (19)	5.32 (20)
$KW, c = 3$	0.52 (2)	1.54 (7)	1.98 (10)	2.05 (10)
$KW, c = 5$	0.52 (2)	0.71 (2)	0.89 (2)	0.92 (4)
$KW, c = 7$	0.52 (2)	0.36 (2)	0.45 (2)	0.50 (2)
$KW, c = 9$	0.52 (2)	0.30 (4)	0.24 (4)	0.47 (5)
$KW, c = 11$	0.52 (2)	0.56 (2)	0.50 (2)	0.48 (2)
$KW, c = 13$	0.52 (2)	0.50 (2)	0.44 (2)	0.40 (2)
$KW, c = 15$	0.52 (2)	0.62 (2)	0.60 (2)	0.66 (2)
		d) all $N^* + N$ observations at the initial design		
	0.65 (1)	0.20 (0)	0.070 (0)	0.040 (0)

Chapter 7

Concluding Remarks

This concluding chapter summarizes the results and gives some suggestions for further research. This thesis has treated optimal design of experiments when the purpose is to estimate the model parameters in the quadratic logistic model. Special attention was given to estimation of the point of optimum response. Optimizing the quadratic response curve corresponds to searching for the optimum operating conditions in the RSM context. RSM is often concerned with processes involving several variables. An immediate extension is to consider more than one control variable to increase the number of applications.

Locally optimal designs were derived in Chapter 3. The procedure to obtain locally D-optimal designs was illustrated step by step. Some formulae for obtaining the designs, given the assumption of either 3 or 4 design points, were provided. The solutions to these formulae can be obtained numerically, which is rather straightforward if one has access to some mathematical software (e.g. Matlab or Mathcad). An advantage is that the formulae can be implemented directly. In contrast to using for example algorithms, which involve issues like finding a good initial design, deciding when convergence is achieved etc.

It was shown that when deriving D-optimal designs it suffices to consider the standardized response π^* based on $\boldsymbol{\theta}^*$. Thus, the locally D-optimal design depends only on the parameter α since $\boldsymbol{\theta}^*$ only includes α. A 3-point design with equal design weights was shown to be D-optimal for the two low response curves. The optimal design for estimation of the high

response curve required four points and with the design weights now being different. A common feature for the designs was that the design points were placed symmetrically around the point of maximum response. It should be remembered that the possibility that the locally D-optimal design consists of more than four points for some set of true parameters cannot be ruled out. Though, no such example has yet been encountered. It is open for further research to establish the number of design points of the locally D-optimal designs.

In Chapter 3 locally c-optimal designs were derived via the canonical design space. The purpose of estimating one single model parameter was considered. Exploring the canonical design space graphically led to a deeper understanding of the design problem. For instance, a c-optimal design for estimation of the parameter μ turned out to consist of two equally weighted design points placed symmetrically around μ. Estimation of α requires one design point and the c-optimal design for estimation of β varies between three and four points (depending on the probability of response at the maximum). Formulae to find both the design points and the associated design weights were presented. The canonical design space can also be applied for D-optimality. The D-optimal design would then be found as the contact points between the canonical design space and the smallest covering ellipsoid. An analysis of the canonical space could bring clarity about the arrangement of the optimal design points and weights. Presumably, it could also lead to formulae for obtaining the D-optimal design points. Furthermore, it is likely that there is a turning point in terms of maximum probability of response where the number of design points is altered. Possibly, the canonical space can give information about this turning point.

There are two main problems with the locally optimal designs, the first is the parameter dependence and the second the fact that the derivation of the designs involves the asymptotic information matrix. A simulation study was performed with the purpose to address these issues. Two parameter estimators were used in the simulations, the standard maximum likelihood estimator and an alternative estimator that instead maximizes a penalized likelihood. A difficulty with the ML estimator is that it is not guaranteed to exist for all samples. The alternative FL estimator was included because it has been suggested as a solution to this problem.

The non-existence of the ML estimator came about as a severe problem for all the investigated cases. The probability that the ML estimator exists was dependent on the choice of design and the true parameters, besides the sample size. From this viewpoint the non-optimal designs sometimes benefited from the many design points. Bayesian designs typically include more design points, therefore it would be interesting to evaluate such designs for this model. The risk of coming across non-existence has to be taken into account if the ML estimator is to be used. Consider facing the situation of having already performed an experiment and no estimates are available. Good strategies for choosing additional points need to be developed for such situations. Designs that are optimal with respect to maximizing the probability to obtain estimates could also be of interest.

The FL estimator do provide estimates in all cases. However, using the FL estimator was associated with other complications which made the estimates useless for practical purposes. These complications often coincided with a high proportion of non-existence for the ML estimator. For some of the most severe cases the FL estimator was, if not as bad as, not a large improvement compared to the ML estimator. On the other hand, there were also examples where the FL estimator offered a great improvement. To sum up, the FL estimator did not offer a completely satisfactory alternative. Estimation of the quadratic response curve was particularly difficult for the low models. In general, it can not be expected to be known at the start of the experiment whether the curve is high or low.

In terms of mean squared error the optimal designs were superior compared to the non-optimal designs for the most cases. There were generally rather large disagreements between the small sample variances of the parameter estimators and their asymptotic analogue. This may cause problems since the asymptotic information matrix is the basis for the construction of the optimal designs. Although the differences diminished when N was increased, they were still appreciable for several cases when $N = 100$ and the response curve was low. Much work remains concerning the implications of the departures from the asymptotic results in small samples. It is of interest to investigate if the locally optimal designs derived on the basis of the true sampling distribution would deviate a lot from those based on asymptotics. Another potentially important aspect which has not been investigated here is the use of approximations to the continuous designs. The

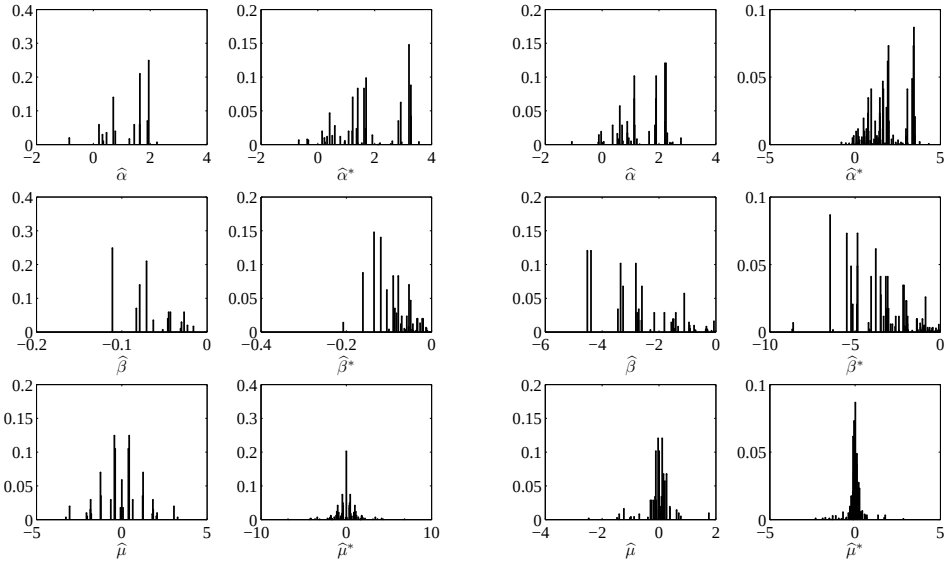
efficiencies of the approximate exact designs ought to be compared to the optimal continuous designs.

In Chapter 5 two sequential approaches were proposed as a solution to the parameter dependence issue. One is a parametric approach based on c-optimality and the other one is nonparametric. The advantage of the nonparametric approach is to avoid having to make any distributional assumptions whereas a parametric approach makes more use of the information. The two sequential strategies were evaluated in a simulation study (Chapter 6) for the case when the aim is to find the point of maximum response $x = \mu$. In the parametric approach, c-optimal designs are derived sequentially based on the assumption that the most recent parameter estimates are true. Two versions of the standardized information matrix were evaluated. The nonparametric alternative is an adaptive stochastic approximation approach. Different combinations of initial design, number of batches and batch sizes were considered. Furthermore, another alternative to the linear predictor was included.

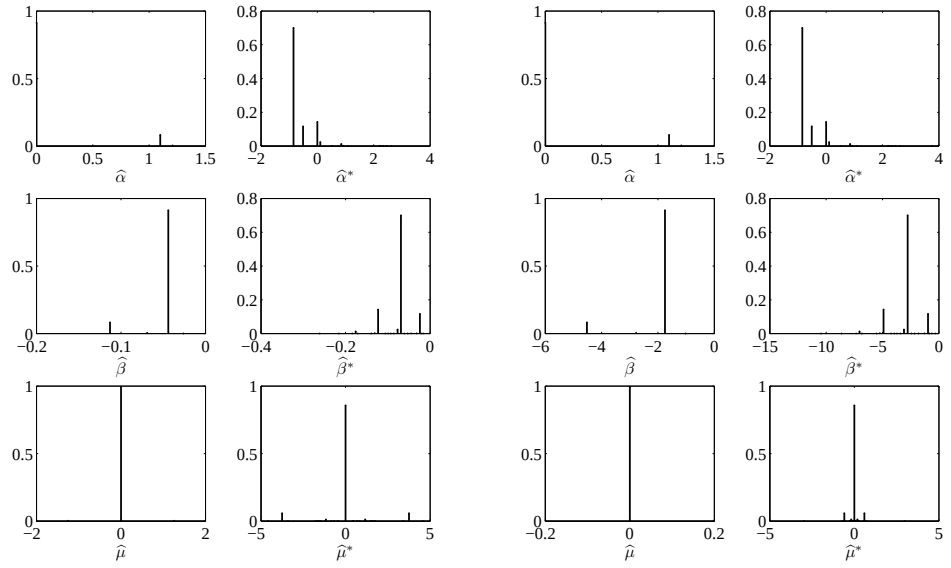
The results of the simulations are really promising for the c-optimal sequential designs that consistently outperformed the other alternatives. The choice between the two versions of the information matrix did not matter as they were almost identical. The success of the stochastic approximation approach was limited to some cases and it was sometimes inferior to the nonsequential approach. In addition it was more sensitive to the initial design and to the misspecified model. Furthermore it required the specification of a constant that proved to have a large impact on the results. There were not any large differences pertaining to the division between the number of batches and the batch size for any of the approaches. A practical consideration that should be mentioned in this context is that sequential designs with a large number of steps are limited to applications were the response time is short. It is called for more extensive evaluations of the sequential designs, regarding for example smaller sample sizes. For example, the choices of initial design and number of batches are likely to be more influential for smaller sample sizes. Theoretical evaluations are also desired but complicated by the dependency between the sequential steps.

The parameter dependence issue is probably perceived as problematic from a practitioners point of view. Undeniably, it is a drawback if one needs to know the location of the point of optimum response prior to the experiment in order to estimate the same point. The lessons learned here imply that sequential designs can circumvent these problems. To conclude, it is the belief of the author that the theory of optimal design can be made to good use for binary response experiments.

Appendix

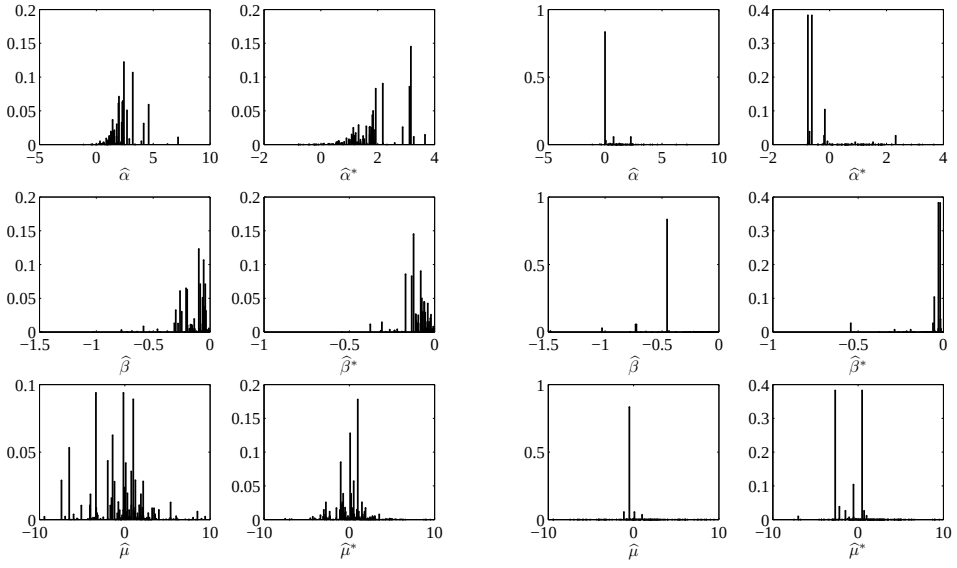


(A) "HIGH-WIDE" (B) "HIGH-NARROW"

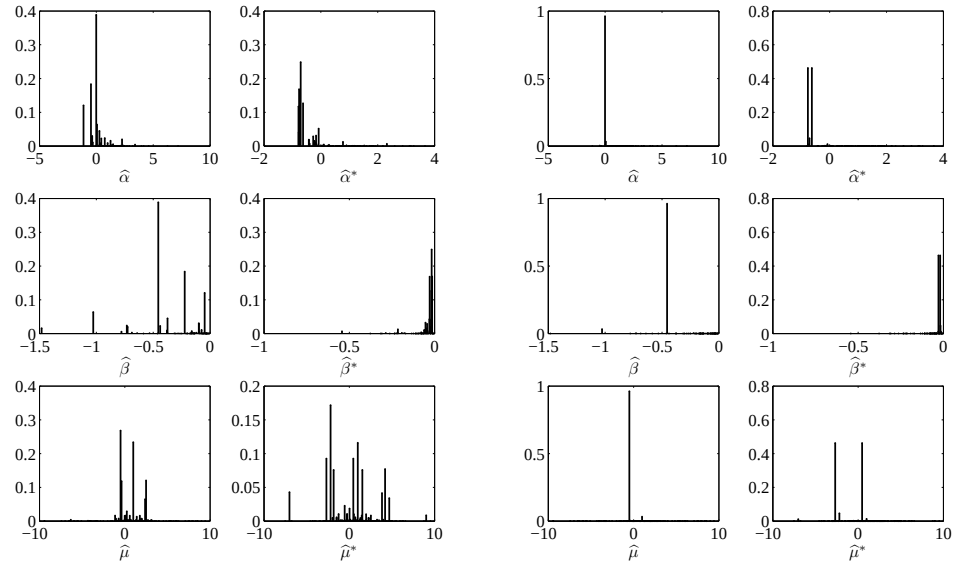


(C) "LOW-WIDE" (D) "LOW-NARROW"

FIGURE 7.1: Distribution of the ML and FL estimators.
Sample size: N=10
Design: D-optimal

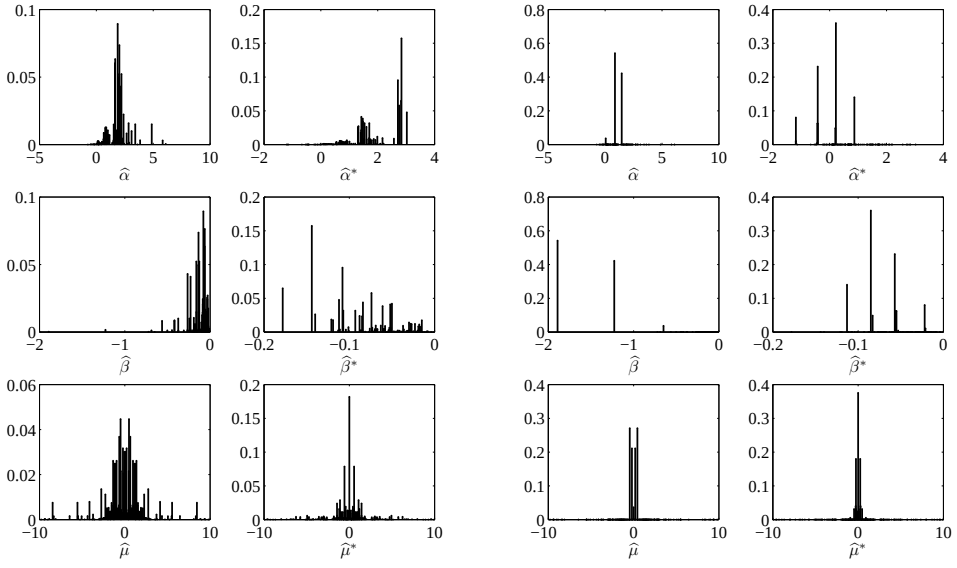


(A) "HIGH-WIDE" (B) "HIGH-NARROW"

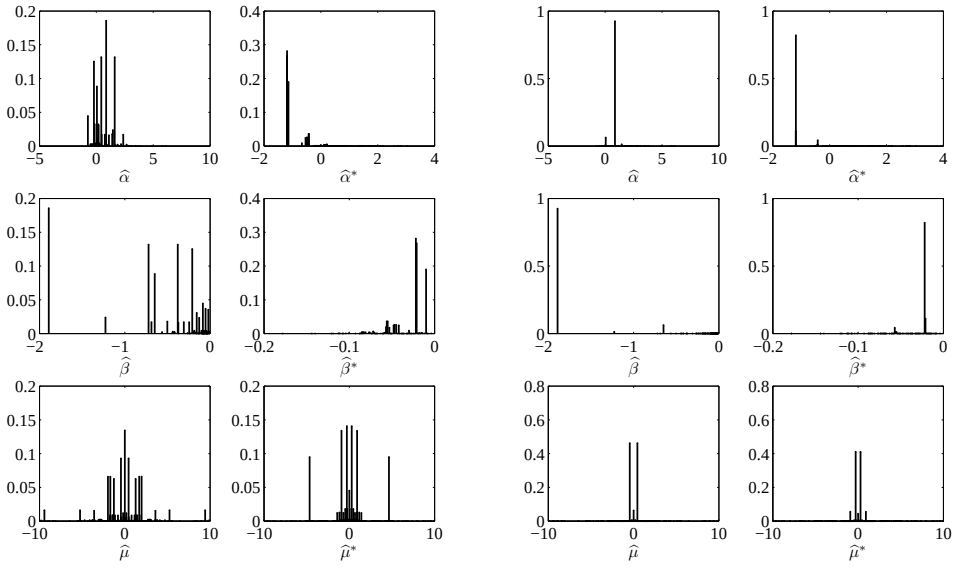


(C) "LOW-WIDE" (D) "LOW-NARROW"

FIGURE 7.2: Distribution of the ML and FL estimators.
Sample size: N=10
Design: 7-point

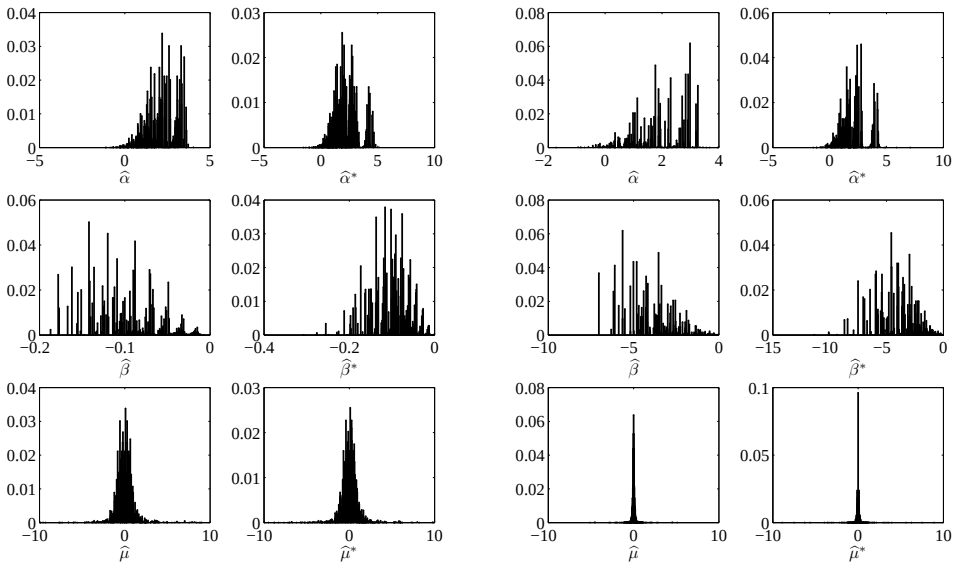


(A) "HIGH-WIDE" (B) "HIGH-NARROW"

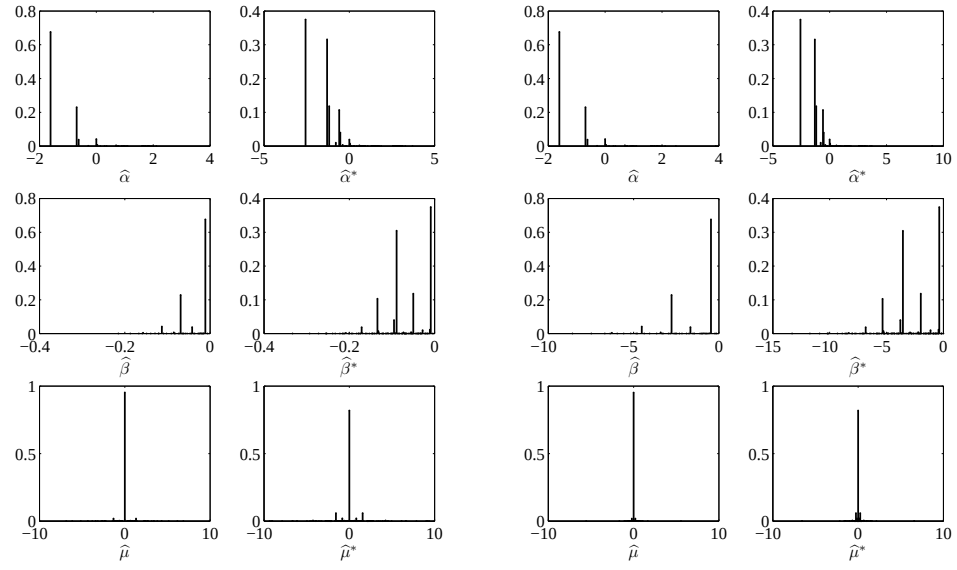


(C) "LOW-WIDE" (D) "LOW-NARROW"

FIGURE 7.3: Distribution of the ML and FL estimators.
Sample size: N=10
Design: 8-point

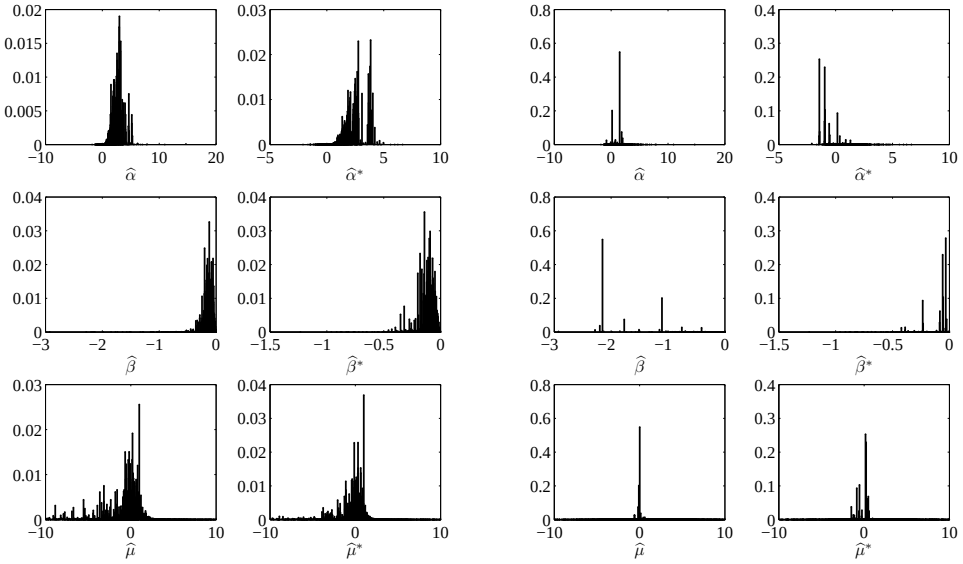


(A) "HIGH-WIDE" (B) "HIGH-NARROW"

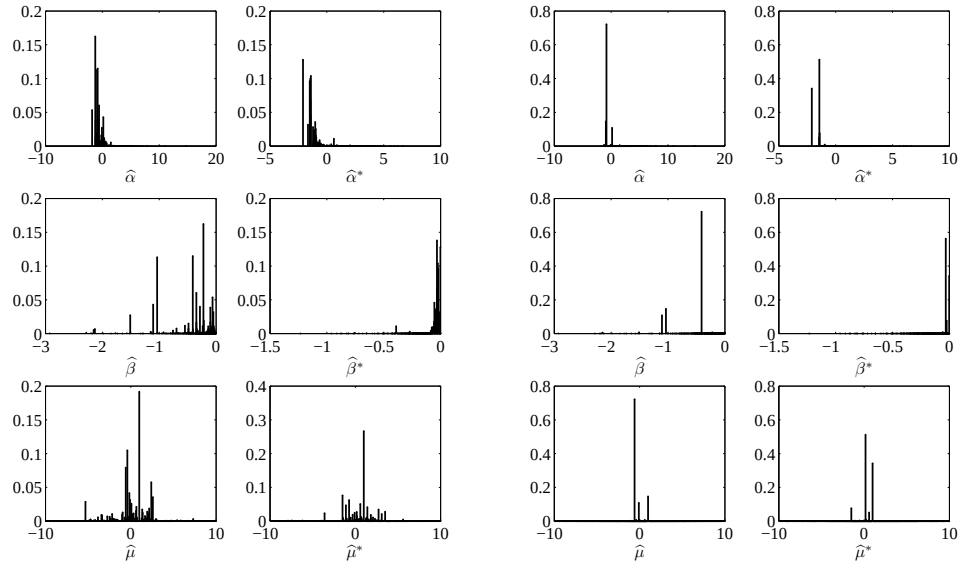


(C) "LOW-WIDE" (D) "LOW-NARROW"

FIGURE 7.4: Distribution of the ML and FL estimators.
Sample size: N=20
Design: D-optimal

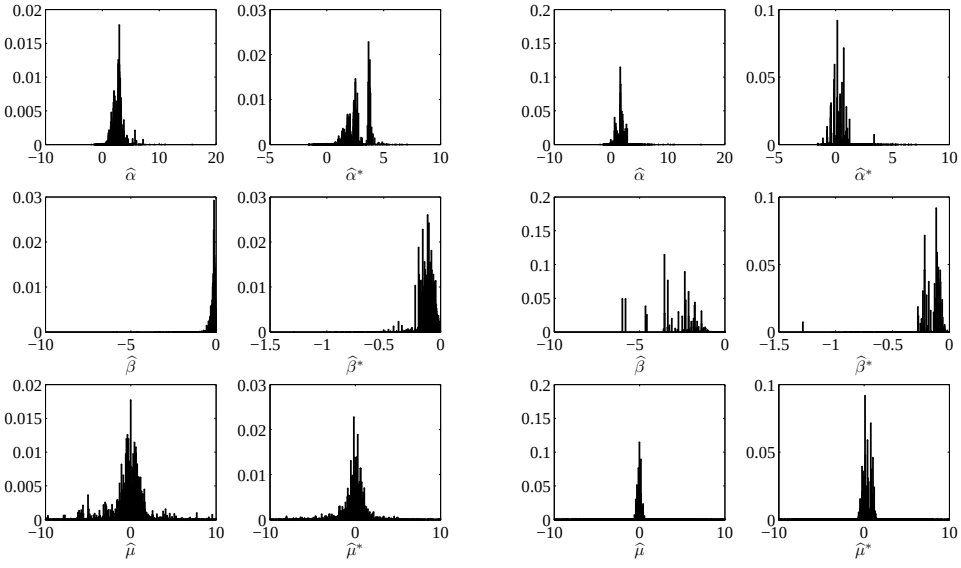


(A) "HIGH-WIDE" (B) "HIGH-NARROW"

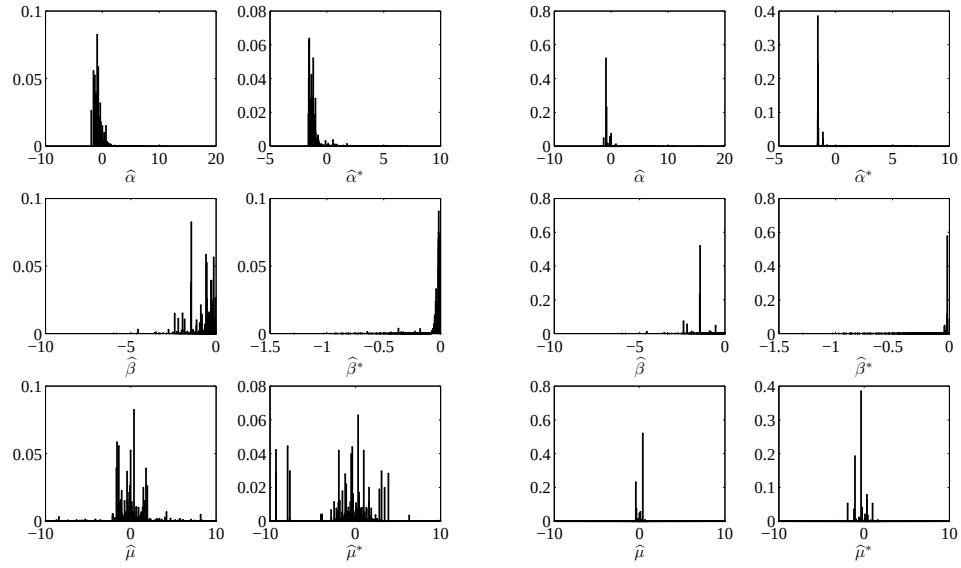


(C) "LOW-WIDE" (D) "LOW-NARROW"

FIGURE 7.5: Distribution of the ML and FL estimators.
Sample size: N=20
Design: 7-point

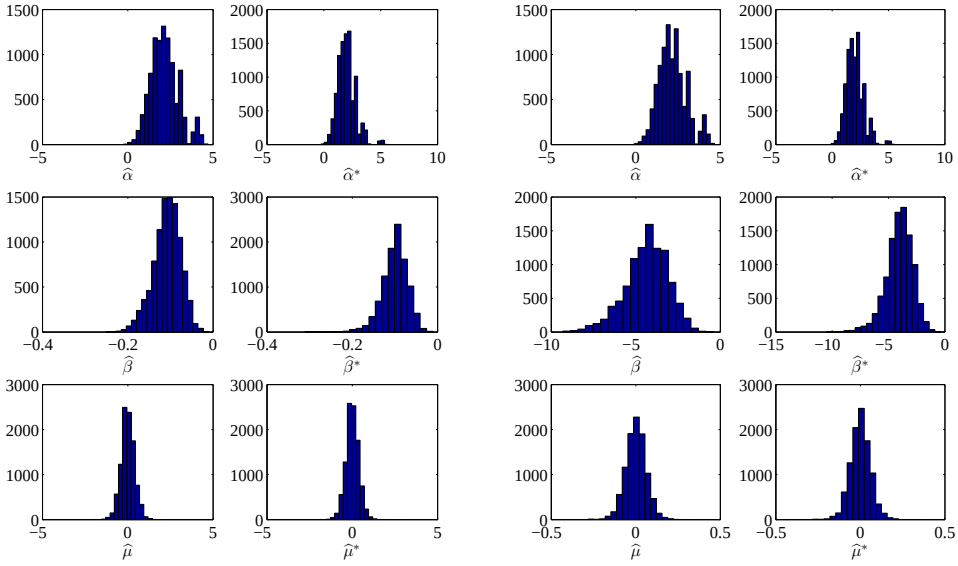


(A) "HIGH-WIDE" (B) "HIGH-NARROW"

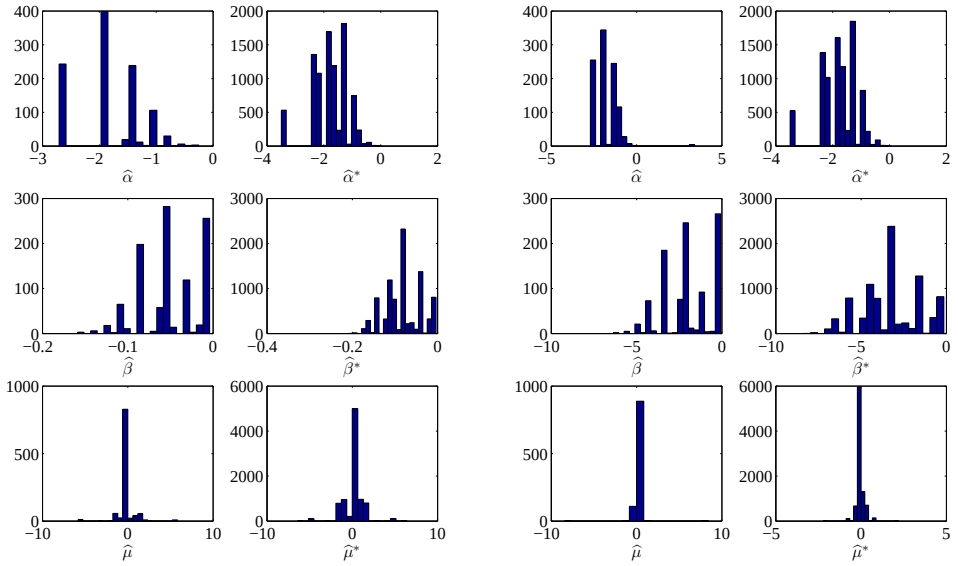


(C) "LOW-WIDE" (D) "LOW-NARROW"

FIGURE 7.6: Distribution of the ML and FL estimators.
Sample size: N=20
Design: 8-point

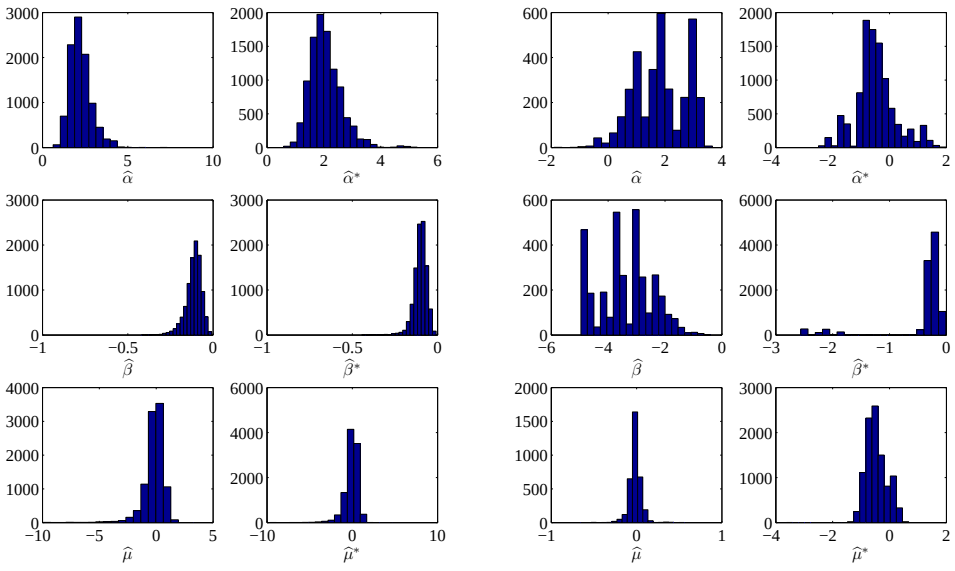


(A) "HIGH-WIDE" (B) "HIGH-NARROW"



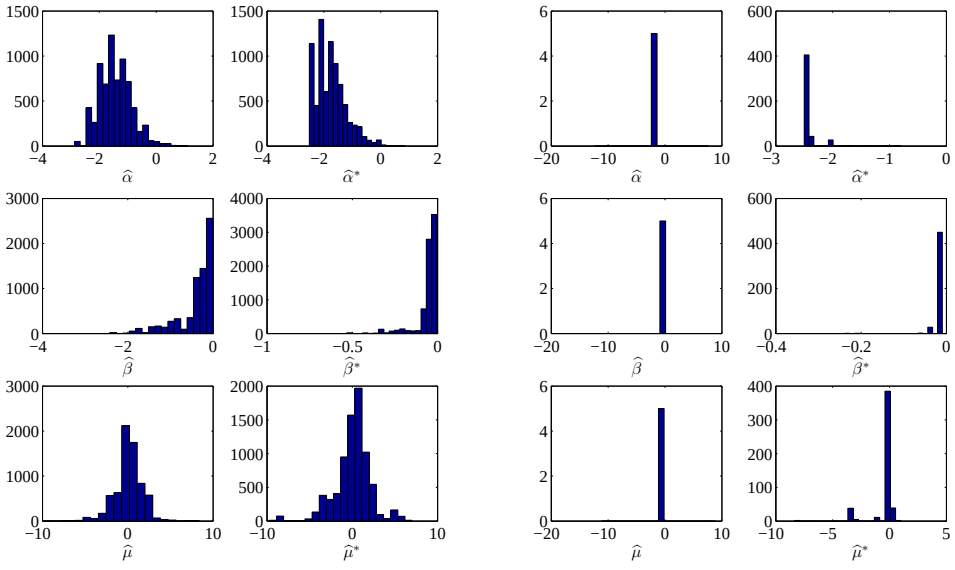
(C) "LOW-WIDE" (D) "LOW-NARROW"

FIGURE 7.7: Distribution of the ML and FL estimators based on 10000 simulation runs.
Sample size: N=50
Design: D-optimal



(A) "HIGH-WIDE"

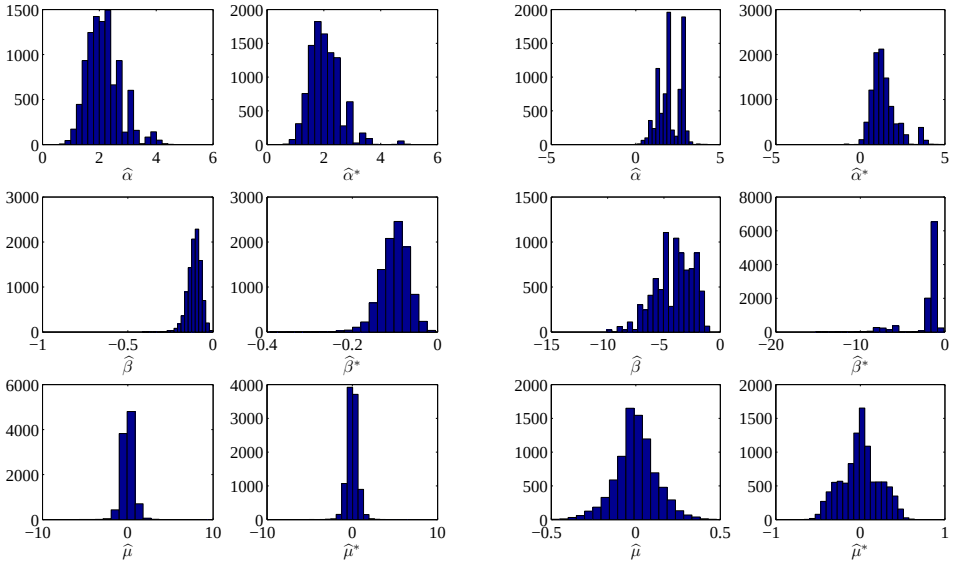
(B) "HIGH-NARROW"



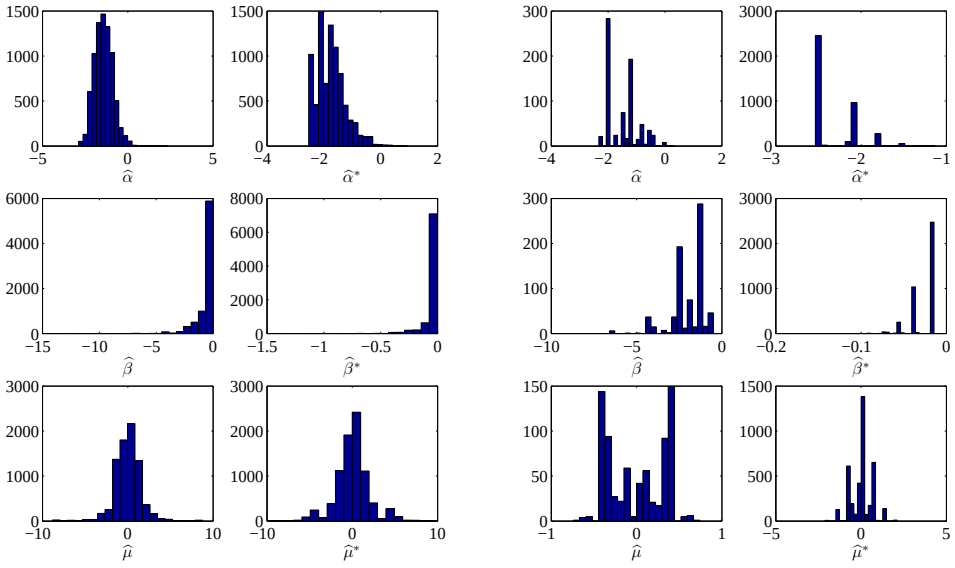
(C) "LOW-WIDE"

(D) "LOW-NARROW"

FIGURE 7.8: Distribution of the ML and FL estimators based on 10000 simulation runs.
Sample size: N=50
Design: 7-point

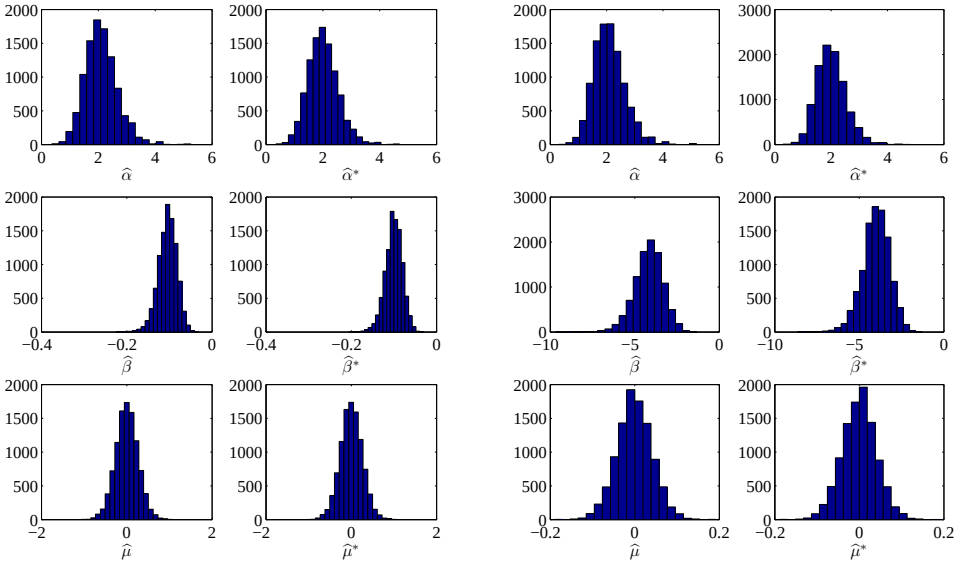


(A) "HIGH-WIDE" (B) "HIGH-NARROW"

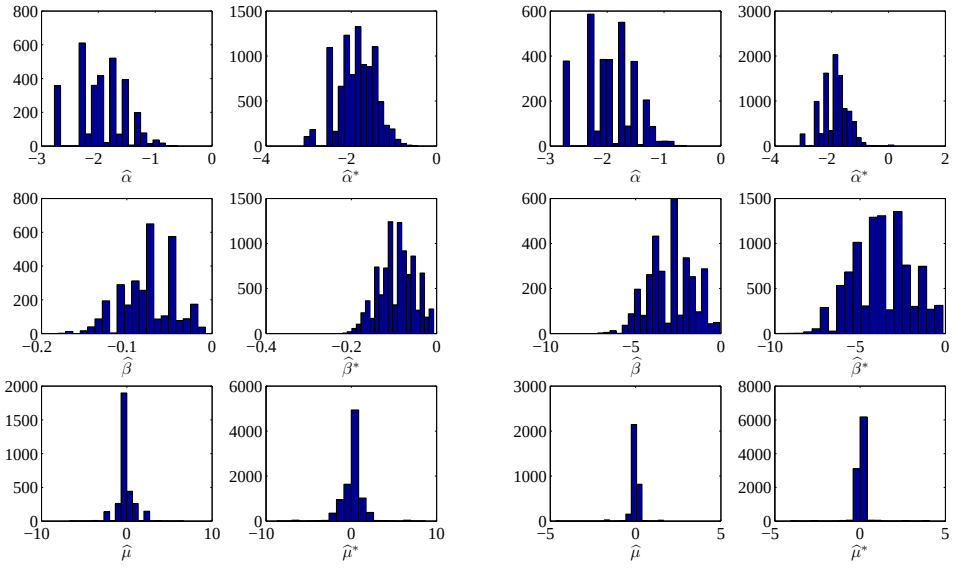


(C) "LOW-WIDE" (D) "LOW-NARROW"

FIGURE 7.9: Distribution of the ML and FL estimators based on 10000 simulation runs.
Sample size: N=50
Design: 8-point

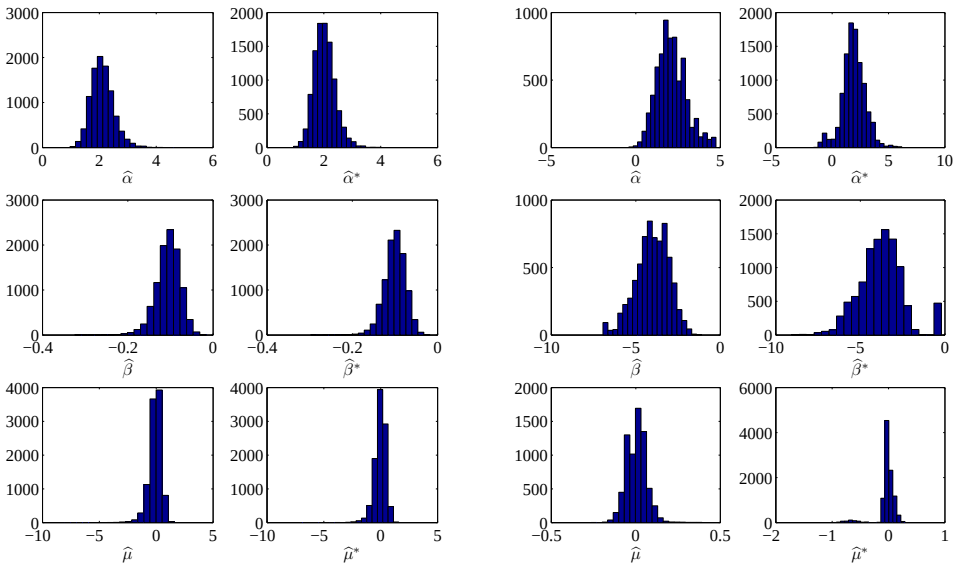


(A) "HIGH-WIDE" (B) "HIGH-NARROW"

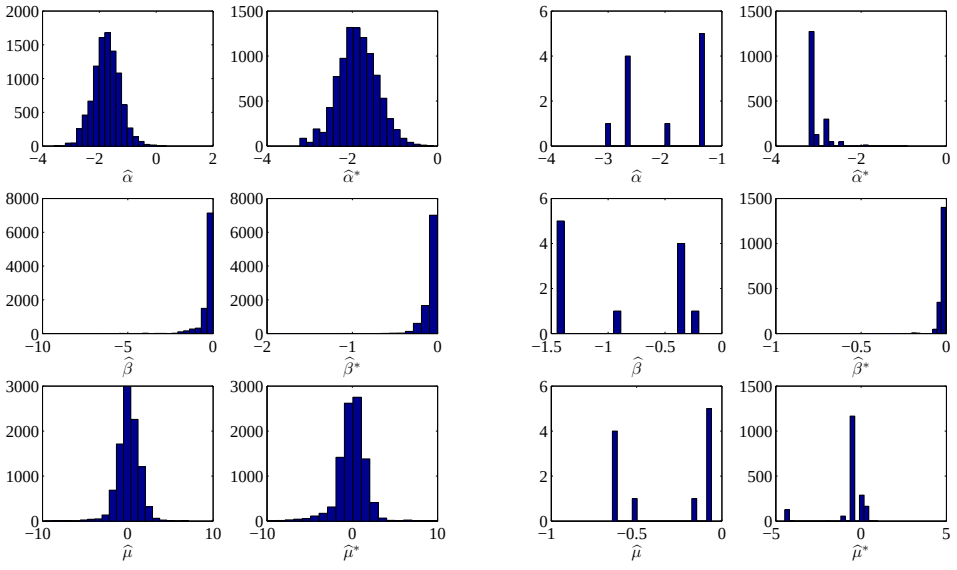


(C) "LOW-WIDE" (D) "LOW-NARROW"

FIGURE 7.10: Distribution of the ML and FL estimators based on 10000 simulation runs.
Sample size: N=100
Design: D-optimal

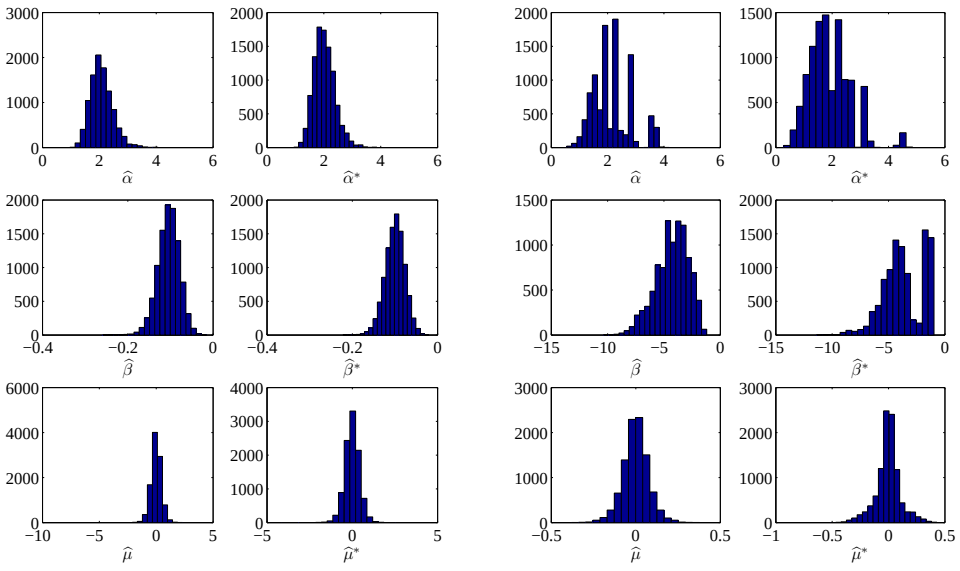


(A) "HIGH-WIDE" (B) "HIGH-NARROW"

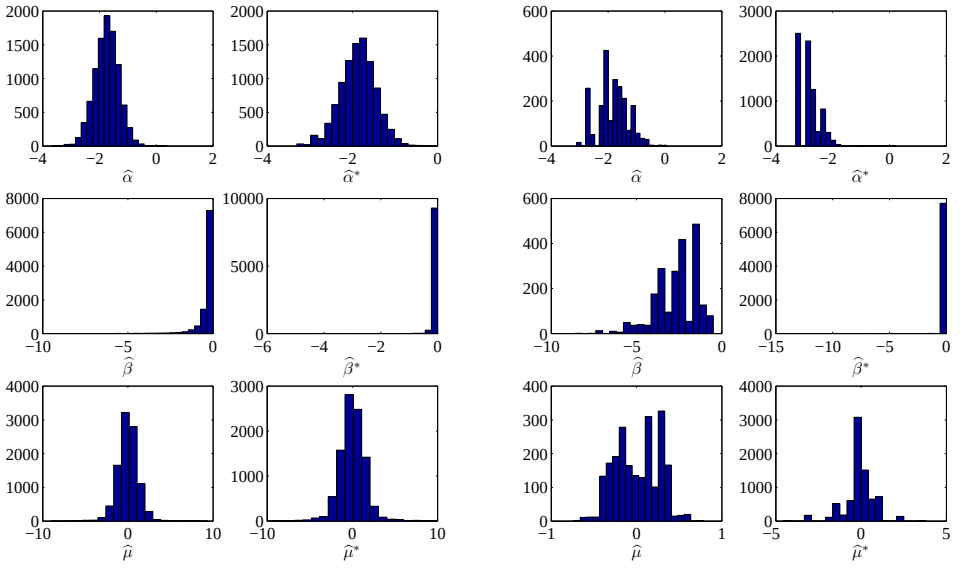


(C) "LOW-WIDE" (D) "LOW-NARROW"

FIGURE 7.11: Distribution of the ML and FL estimators based on 10000 simulation runs.
Sample size: N=100
Design: 7-point



(A) "HIGH-WIDE" (B) "HIGH-NARROW"



(C) "LOW-WIDE" (D) "LOW-NARROW"

FIGURE 7.12: Distribution of the ML and FL estimators based on 10000 simulation runs.
Sample size: N=100
Design: 8-point

Bibliography

Anbar, D. (1978). A Stochastic Newton-Raphson Method. *Journal of Statistical Planning and Inference* 2, 153-163.

Albert, A. and Anderson, J.A. (1984). On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika* 71, 1-10.

Atkinson, A.C. and Donev, A.N. (1992). *Optimum Experimental Designs*. Oxford University Press, New York.

Biedermann, S., Dette, H., and Zhu, W. (2005). Geometric Construction of Optimal Designs for Dose-Response Models with Two Parameters. Research Report, Fakultät für Mathematik, Ruhr-Universität Bochum.

Box, G.E.P. and Draper, N.R. (1987). *Empirical Model-Building and Response Surfaces*. Wiley, New York.

Box, G.E.P. and Wilson, K.B. (1951). On the Experimental Attainment of Optimum conditions. *Journal of the Royal Statistical Society. Series B* 13, 1-45.

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

Dobson, A.J. (2002). *An Introduction to Generalized Linear Models*. Chapman and Hall, London.

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

Fackle Fornius, E. (2007). D-optimal Designs for Quadratic Logistic Regression Models. *International Journal of Statistical Sciences.* 6, 269-303.

Fackle Fornius, E. (2008). Sequential Designs for Binary Data with the purpose to Maximize the Probability of Response. *Communications in Statistics - Simulation and Computation*. 37, 6.

Fackle Fornius, E. and Nyquist, H. (2008). Using the Canonical Design Space to obtain c-optimal Designs for the Quadratic Logistic Model. To be published in *Communications in Statistics - Theory and Methods*.

Fedorov, V. (1972). *Theory of Optimal Experiments*. Academic Press, New York.

Fedorov, V. and Hackl, P. (1997). *Model-oriented Design of Experiment*. Springer, New York.

Firth D. (1993). Bias Reduction of Maximum Likelihood Estimates,*Biometrika* 20, 272-319.

Fisher R.A. (1935). *The Design of Experiments*. Oliver & Boyd Ltd., Edinburgh.

Ford, I., Kitsos, C.P. and Titterington, D.M. (1989). Recent Advances in Nonlinear Experimental Design. *Technometrics* 31,49-60.

Ford, I. and Torsney, B and Wu, J. (1992). The Use of a Canonical Form in the Construction of Locally Optimal Designs for Non-Linear Problems. *Journal of the Royal Statistical Society. Series B* 54, 569-583.

Heinze, G. and Schemper, M. (2002). A Solution to the Problem of Separation in Logistic Regression. *Statistics in Medicine* 21, 2409-2419.

Khuri, A. (2001). An Overview of the Use of Generalized Linear Models in Response Surface Methodology. *Nonlinear Analysis* 47, 2023-2034.

Kiefer, J. (1959). Optimum Experimental Design. *Journal of the Royal Statistical Society. Series B* 21, 272-319.

Kiefer, J. (1961). Optimum Designs in Regression Problems, II. *Annals of Mathematical Statistics* 32, 298-325.

Kiefer, J. and Wolfowitz, J. (1952). Stochastic Estimation of the Maximum of a Regression Function. *The Annals of Mathematical Statistics* 23, 462-466.

Kiefer, J. and Wolfowitz, J. (1959). Optimum Designs in Regression Problems. *Annals of Mathematical Statistics* 30, 271-294.

Kitsos, C.P., Titterington, D.M. and Torsney, B. (1988). An Optimal Design Problem in Rythmometry. *Biometrics* 44, 657-671.

Lai, T. L., and Robbins, H. (1979). Adaptive Design and Stochastic Approximation. *Annals of Statistics* 7, 1196-1221.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models,* 2nd edition. Chapman and Hall, London.

Myers, W.R., Myers, R.H. and Carter, W.H. Jr. (1994). Some Alphabetic Optimal Designs for the Logistic Regression Model. *Journal of Statistical Planning and Inference* 42, 57-77.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A* 135, 370-384.

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

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

Robbins, H. and Monro, S. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics* 22, 400-407.

Sacks, J. (1958). Asymptotic Distribution of Stochastic Approximation Procedures. *The Annals of Mathematical Statistics* 29, 373-405.

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

Sitter, R. R. and Torsney, B. (1995). Optimal Designs for Binary Response Experiments with two Design Variables. *Statistica Sinica.* 5, 405-419.

Sitter, R. R. and Wu, C.F.J. (1999). Two-Stage Design for Quantal Response Studies. *Biometrics.* 55, 396-402.

Wu, C. F. J. (1985). Efficient Sequential Designs With Binary Data. *Journal of the American Statistical Association* 80, 974-984.

Wu, C. F. J. (1986). Maximum Likelihood Recursion and Stochastic Approximation in Sequential Designs. *Adaptive Statistical Procedures and Related Topics* 8, 298-313.