SPRINGER BRIEFS IN STATISTICS

Christos P. Kitsos

Optimal Experimental Design for Non-Linear Models

Theory and Applications

Springer

SpringerBriefs in Statistics

For further volumes:
http://www.springer.com/series/8921

Christos P. Kitsos

Optimal Experimental Design for Non-Linear Models

Theory and Applications

Springer

Christos P. Kitsos
Department of Informatics
Technological Educational Institute
of Athens
Athens
Greece

ISSN 2191-544X ISSN 2191-5458 (electronic)
ISBN 978-3-642-45286-4 ISBN 978-3-642-45287-1 (eBook)
DOI 10.1007/978-3-642-45287-1
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013956330

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Acknowledgments

I would like to thank Assistant Prof. V. Zarikas (Lamia), and part-time Assistant Profs. T. Toulias, Y. Hatzikian, and K. Kolovos (Athens), for useful discussions during the preparation of this monograph. I would like to thank the reviewers for their time working with the text. They made very sharp and useful comments, which are very much appreciated. The encouragement of Springer, BioMathematics/Statistics, is very much appreciated.

Contents

Chapter 1
Introduction

Abstract This chapter defines the Nonlinear Experimental Design problem and traces its evolution. It is emphasized that the (continuous) Design Theory started with a Non-linear problem.

In a brief look back at the origin of the methods discussed in this monograph, it seems that the first real-life problems discussed were nonlinear rather than linear. The Least Square Method originated from a nonlinear problem, thanks to the Mozart of Mathematics: Carl Friedrich Gauss (1777–1855). Experimental Design theory, so called by its founder Sir Ronald Aylmer Fisher (1890–1962), also originated from a nonlinear problem. I shall briefly trace these problems.

One of the most famous data sets is the one collected in 1793 for defining a new measurement unit: the meter, see the contribution of Stigler (1981) to the history of Statistics. What is also very important is that for this particular data set:

- In recent statistical terms, a nonlinear function had to be finally estimated, between the modulus of arc length (s), latitude (d) and meridian quadrant (L).
- A linear approximation was used in 1755 for this problem and a second order expansion was applied in 1832.

The meter was eventually defined as equal to one 10,000,000th part of the meridian quadrant, i.e. the distance from the North Pole to the Equator along a parallel of longitude passing through Paris, and the linear approximation was in the form:

$$\eta(L) = \frac{s}{d} = \theta_1 + \theta_2 \cdot \sin^2 L$$

With θ_1 being the "length of a degree at the equator" and θ_2 "the excess of a degree at the pole over one at the equator". Then the ellipticity (e) was estimated through θ_1 and θ_2 by the following nonlinear relationship:

$$\frac{1}{e} = 3 \cdot \frac{\theta_1}{\theta_2} + \frac{3}{2}$$

C. P. Kitsos, *Optimal Experimental Design for Non-Linear Models*, SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-45287-1_1,

It is clear that Gauss, see Plackett (1972) for details, treated the data in his own remarkable mathematical way, considering a nonlinear function.

The first statistical treatment of a nonlinear function came from the pioneer of modern Statistics, R. A. Fisher. He started work at Rothamsted Experimental Station in 1919, and around 1922 he came across what is known as the dilution series problem. A brief description is as follows. For a small volume u taken from the volume V of a liquid containing n tiny organisms (such as bacteria), the probability p that u contains no organisms can be evaluated as

$$p = \left(1 - \frac{u}{V}\right)^n \cong \exp(-nu/V) = \exp(-\theta u).$$

The parameter θ, the density per unit volume, has to be estimated (see details in Sect. 7.2). The question is how one should perform the experiment to get the best possible estimate. The probability p in the above relation is expressed as a nonlinear function. Fisher solved this nonlinear problem in 1922, using a concept of his own: his information.

Since Fisher's pioneering work in experimental design, Statistics has become involved in all experimental sciences: Chemistry, Biology, Pharmacology, Industry, Psychology, Toxicology and so on. Of course statisticians do not provide methods for designing experiments in isolation. However, in cooperation with the experimenter, for whom the objective of an experiment is clear, the statistician provides the most informative pattern of the experiment so that the required objective can be achieved.

The objectives of the experimenter can be:

1. To obtain an estimate for a response y in some particular region using the input variable vector $u = (u_1, u_2, \ldots, u_k)$. This is the response surface problem introduced by Box and Draper (1959).
2. To determine the best mathematical model describing most precisely the investigated phenomenon. This is the discrimination problem between rival models and was reviewed by Hill (1976).
3. In some sense, all or a subset of the parameters are to be estimated as well as possible. This is the optimal experimental design problem originated by Smith (1918).

The above-mentioned objectives are common to linear and nonlinear experimental designs (LED and NLED), i.e. when the assumed suitable (and correct) model, describing the underlying phenomenon, is linear or nonlinear with respect to its parameters.

There is no such volume of review work in NLED, although work on experimental designs started with a nonlinear problem, thanks to Gauss and Fisher. Some work has been performed by Kitsos (1986), Ford et al. (1989) reviewed the problem, Kitsos (1989) worked with the sequential procedures and Abdelbasit and

Plackett (1981, 1983) reviewed the nonlinear case for certain types of problems, but they gave no attention to regression type experiments. Moreover Wu (1985a, b) worked on binary response problems. A recent review was published by Kitsos (2010b, 2012), while Chaudhuri and Mykland (1995) provided useful results on the sequential nonlinear design, and Hu (1998) faced the same problem from the Bayesian point of view.

However, the nonlinear experimental design problem finds applications in many fields: as a regression problem (continuous case) in chemical kinetics (chemistry, biology), as a binary model (discrete case) in testing explosives (the so-called *up and down method*), in biological assays (for the percentile points see Tsutakawa (1980) and Kitsos (1999) among others), fatigue experiments, educational studies and life testing. The calibration problem is eventually a nonlinear problem, see Kitsos and Kolovos (2010), while for a geophysical application see Zarikas et al. (2010). The calibration problem needs particular investigation (the underlying model is linear, but a nonlinear function is possible and needs to be estimated) but it is not discussed in this monograph. It is interesting to realize that there are nonlinear models, the Michaelis–Menten model being one, which obtain numerically a solution that is not necessarily unique, see Biebler et al. (2008).

The target of this monograph is to provide a solid statistical background and the appropriate applications on the theory of NLED; to provide general optimality criteria, in a theoretical framework; to provide methods and discuss problems associated with nonlinear problems, appropriate for the application and to link the NLED theory with the practical problems of various applications. So "a bridge" is offered between a particular field of interest and NLED. It is considered as a "second level" book, and therefore some knowledge from Statistics, Mathematics (and Linear Algebra for the theoretical background of Chap. 8) would be helpful for the reader.

The emphasis will be on the target (3) described above and its related difficulties. I tried to keep the design of this monograph "balanced": thus equal weight is put on to the theoretical insight and to the applications. Keep in mind that the theory was developed to face the applications rather, than just to construct a mathematically elegant development.

As there is a strong link with the Linear Experimental Design, I shall also refer to it when it is needed. A large portion of the monograph is devoted to discussion, so as to create a rapport with the reader; doing so helps to clarify the ideas introduced.

This monograph includes 23 Examples, 18 Tables with results from working simulation studies, 7 Figures trying to explain the developed mathematical insight, and 3 Simulation Studies fully discussed and explained in detail, in order to link Theory and Applications, in reference to Optimal Experimental Design for Non-Linear Models.

References

Abdelbasit KM, Plackett RL (1981) Experimental design for categorized data. Int Stat Rev 49:111–126

Abdelbasit KM, Plackett RL (1983) Experimental design for binary data. JASA 78:90–98

Biebler KE, Schreiber K, Bode R (2008) Solutions to the Michaelis–Menten kinetics are not necessarily unique. In: International conference of computing methods in sciences and engineering (ICCMSE)

Box GEP, Draper NR (1959) A basis for the selection of a response surface design. JASA 54:622–654

Chaudhuri P, Mykland PA (1995) On efficient design on nonlinear experiments. Stat Sinica 5:421–440

Ford I, Kitsos CP, Titterington DM (1989) Recent advances in nonlinear experimental design. Technometrics 31:49–60

Hill PDH (1976). Optimal experimental design for model discrimination. Ph.D. thesis, University of Glasgow

Hu I (1998) On sequential designs in nonlinear problems. Biometrika 85:496–503

Kitsos CP (1986) Design and inference in nonlinear problems. Ph.D. thesis, University of Glasgow, UK

Kitsos CP (1989) Fully sequential procedures in nonlinear design problems. Comp Stat Data Anal 8:13–19

Kitsos CP (1999) Optimal designs for estimating the percentiles of the risk in multistage models in carcinogenesis. Biometrical J 41:33–43

Kitsos CP (2010) Adopting hypergeometric functions for sequential statistical methods. Bull Greek Math Soc 57:251–264

Kitsos CP (2012) On the optimal continuous experimental design problem. Discussiones Mathematicae Probab Stat 31(1–2):59–70

Kitsos CP, Kolovos KG (2010) An optimal calibration design for pH meters. Instrum Sci Technol 38(6):436–447

Plackett RL (1972) The discovery of the method of least squares. Biometrika 59:239–251

Smith K (1918) On the standard deviation of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. Biometrika 12:1–85

Stigler SM (1981) Gauss and the invention of least squares. Ann Stat 9:465–474

Tsutakawa RK (1980) Design of dose levels for estimating a percentage point of a logistic quantal response curve. Appl Stat 29:25–33

Wu CFJ (1985a) Efficient sequential designs with binary data. JASA 80:974–984

Wu CFJ (1985b) Asymptotic inference from sequential design in a nonlinear situation. Biometrika 72:553–558

Zarikas V, Gikas V, Kitsos CP (2010) Evaluation of the optimal design cosinor model for enhancing the potential of robotic theodolite kinematic observation. Measurement 43(10):1416–1424

Chapter 2
Motivation

Abstract The appropriate notation is introduced linking the Mathematical problem with the Experimentation. The linearization of the underlying model and Fisher's information is discussed and examples are discussed.

2.1 Introduction

Experimenters working in Laboratories, without being necessarily Mathematicians, know that performing any experiment, in principle the following are needed:

- The experimental unit, appropriate for each particular field.
- The range of experimental conditions, known as design space.
- The measurements or responses (y) obtained at certain values of the explanatory or input variables (u).

Usually, in Chemical problems, the experimenter designs in blocks, and the experimental unit is the apparatus which provides the measurements. In Psychology or Medicine the experimental units are the individual (usually particular animals) under investigation, see Example 1. Thus the experiment is performed through single binary observations. In Medicine or Toxicology the binary response models are more often adopted, linking the "discrete" data, see Examples 4, 7 and 10 below. In Chemistry Kinetics, see Sect. 4.7 below, the "continuous" nonlinear case is usually applied; see also Examples 8 and 13.

The Book of Science has been written with the Mathematical alphabet—do not be afraid of this globalization! It can be easily proved even for the experimentalist. Thus, in this Chapter the notation is introduced and the necessary assumptions and definitions for the mathematical formulation of the nonlinear experimental design problem. The appropriate discussion helps, especially the non-mathematicians, to understand the problem at hand.

C. P. Kitsos, *Optimal Experimental Design for Non-Linear Models*,
SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-45287-1_2,

2.2 Notation

The Euclidian space $U \subseteq R^k$ in which the predictor variables (covariates or explanatory variables or independent variables, or input variables) $u = (u_1, u_2, \ldots, u_k)$ take their values is known as *experimental region* or *design space*. An example from chemical kinetics is "time" and from medicine "dose". The *parameter space* $\Theta \subseteq R^P$ is the set where the involved parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_P)^T$ take their values. Let Ξ be the family of measures ξ such that

$$\xi(u) \geq 0 \ , \ u \in U \ \text{ and } \ \int \xi(du) = 1. \tag{2.2.1}$$

Such a measure ξ is referred as a *design measure,* while the pair (U, ξ) will be called *the design*. The *support* of the design (U, ξ), Supp(ξ), is the set of points u for which the design measure is positive $\xi(u) > 0$. This is only a theoretical consideration. It might be a design point optimal, but with zero design measure, $\xi(u) = 0$, at this point u, and therefore with no practical use, as the experimenter is not taking observations at this point. Practically speaking the design measure acts as the proportion of observations devoted to the optimal design points. That is why sometimes the following way to present a design measure is adopted

$$\xi = \left\{ \begin{matrix} u_1 & u_2 & \ldots & u_k \\ \\ p_1 & p_2 & \ldots & p_k \end{matrix} \right\}, \sum p_i = 1, \, p_i = \frac{n_i}{n}$$

What are the optimal design points? One can certainly perform his/her experiment at different values of the input variables. The optimal values are those values which will provide a "useful predefined result"—example: minimize the variance of the parameters.

Let us denote by Mat(v, m) the set of $v \times m$ size matrices and let $\theta \in$ Mat(p, 1) be the vector of parameters, $u_i \in$ Mat(1, k) the predictor variable, i = 1, 2, …, n, with n the sample size.

For the response y it is assumed that either $y \in \Psi \subseteq R$ or $y \in \{0, 1, \ldots, \lambda\}$, $\lambda \geq 1$. When the response y is supposed to take any value in Ψ, it is also supposed that a regression model (in general nonlinear) exists consisting in the deterministic portion $f(u, \theta)$ and the stochastic portion, e, known as error, linked through the (continuous) regression formulation

$$y_i = f(u_i, \theta) + e_i. \tag{2.2.2}$$

Let us assume that $f(u, \theta) = \theta^T g(u)$ with g being a (vector) continuous function of u, θ^T is the transpose vector of θ. Then the nonlinear problem is reduced to the so called linear problem.

When $\Psi = \{0, 1, 2, \ldots, \lambda\}$ the multiresponse problem is obtained. The most common case is $\lambda = l$, i.e. a binary response. In this case the outcome takes only

two values, $Y_i = 1$ or 0, and it is linked with the covariates and the parameters through a (discrete) probability model "T" with

$$P(Y_i = 1) = T(u_i, \theta), P(Y_i = 0) = 1 - T(u_i, \theta), \tag{2.2.3}$$

where u_i is the value of u going with observation Y_i, i = 1, 2, …,n

Example 1 Typical situation in Bioassays is to consider the models:

$$\begin{aligned} &\textit{Logit model}: T(u, \theta) = \{1 + \exp(-\theta_1(u - \theta_2))\}^{-1}, \text{ or} \\ &\textit{Probit model}: T(u, \theta) = (\sqrt{2}\pi\theta_2)^{-1} \int_{-\infty}^{u} \exp(-(u - \theta_1)^2/(2\theta_2^2))du. \end{aligned} \tag{2.2.4}$$

In both cases the parameter vector is $\theta = (\theta_1, \theta_2)$, with θ_1 the location parameter and θ_2 the scale parameter.

2.3 Assumptions

Every theoretical approach is based on a number of assumptions. It is proposed to keep the number of these assumptions limited. The following main assumptions will be considered throughout this book. Let us refer to them as Assumption 1 or 2 etc. when they are recalled. When limiting results for the sequence of estimators θ_n are considered the parameter space Θ is assumed to be compact. This is needed, as in compact set (closed and bounded set) any sequence from this set converges within this set, due to Bolzano-Weierstrass Theorem. For the errors the main assumption which is imposed is:

Assumption 1 The errors e_i are independent and identically distributed (iid) with $E(e_i) = 0$ and $V(e_i) = \sigma^2 > 0$, i = 1, 2, …, n.

Under Assumption 1 for model (2.2.2) is: $\eta = \eta(u, \theta) = E(y) = f(u, \theta)$.
And for the model (2.2.3) is: $\eta = \eta(u, \theta) = E(y) = T(u, \theta)$.

Assumption 2 The iid (independent identically distributed) errors are normally distributed with mean 0 and variance $\sigma^2 > 0$.

As far as the function f concerned is needed to be smooth in the neighborhood of θ_t, the true value of the parameter θ. That is why it is assumed:

Assumption 3 The function $f(u, \theta)$ is continuous in Θ at θ_t and the second order derivatives of f with respect to θ exist at and near θ_t.

For the introduced function T, concerning the binary response problems, recall relation (2.2.3), it is assumed that:

Assumption 4 T(u) is a monotonic differentiable function.

Moreover the function T(u) can be considered as a cumulative distribution function of the random variable Z defined through the random variable Y as follows

$$Y = \begin{cases} 1 & \text{if} \quad Z \leq u \\ 0 & \text{if} \quad Z > u. \end{cases}$$

Then: $P[Y = 1] = P[Z \leq u] = T(u)$, say, and $P[Y = 0] = P[Z > u] = 1 - T(u)$.

Thus: $E(Y) = 1 \times T(u) + 0 \times (1 - T(u)) = T(u) = \eta$ and $Var(Y) = T(u) \times (1 - T(u))$.

So the expected value and the variance of the new random variable Y are evaluated through the function T(u).

Assumption 5 The assumed model used to plan the design is correct.

This Assumption is essential in practice, as the experimentalist, very often, tries to elucidate which model is appropriate, with no statistical background for such a target.

The theoretical framework is based on these assumptions, which can be easily followed by the experimenter. Therefore it is recommended that the experimentalist should see if are fulfilled.

2.4 On the Existence of Estimators

When the experimental data is collected, the question arises whether it is possible to obtain the appropriate estimates of the involved parameters. Therefore the existence of the estimators requires a particular theoretical investigation. For the (continuous) model (2.2.2) the following quantity is introduced

$$S_n(\theta) = \sum (y_i - f(u_i, \theta))^2 = \|y - f(u, \theta)\|_2 \tag{2.4.1}$$

where $\|\cdot\|_2$ is the 1_2-norm. An estimate $\hat{\theta}$ is the *least squares estimate* (LSE) if

$$S_n(\hat{\theta}) = \min\{S_n(\theta)\,;\, \theta \in \Theta\} \tag{2.4.2}$$

Under Assumption 2 it is known that this LSE coincides with the maximum likelihood estimator (MLE). It can be proved that there exists such a LSE, under certain conditions, see Wu (1981), who relaxed the conditions imposed in the pioneering paper of Jennrich (1969).

For the binary response problem Silvapulle (1981) provided that conditions under which the likelihood function L,

$$L \propto \Pi\{T(u_i, \theta)\}^{y_i}\{1 - T(u_i, \theta)\}^{1-y_i} \tag{2.4.3}$$

can provide maximum likelihood estimators (Appendix A2). Roughly speaking this occurs when the intersection of the sets of values taken by the explanatory variables corresponding to successes and to failures is not the null set (see footnotes of Tables A4.1, A4.2, A4.4, in Appendix 4). This happens to be a necessary and sufficient condition for the logit and probit models. Practically, a proportion can not be estimated when only a data set of "successes" or only "failures" is available. Moreover the estimators are not appropriate if the number of "successes" ("failures") is very small comparing to the total number of observations.

Now, having ensured that the likelihood equation can provide MLE and denoting by l the log-likelihood the following matrix is defined

$$S(\widehat{\theta}, \xi_n, y) = -\left(\frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j} \Big| \theta = \widehat{\theta}\right) \tag{2.4.4}$$

where ξ_n is the design measure on n observations, called the *sample information matrix*.

Example 2 Maximum likelihood estimates for the logit model can be obtained through the, well known from the linear case, *normal equations*

$$\sum T_i = \sum y_i \sum u_i T_i = \sum y_i u_i,$$

with $T_i = T(u_i; \theta)$ as in (2.2.4).

One of the most important measures in Statistics is (the parametric) Fisher's information measure. There is also the entropy type of Fisher's information measure, but this is beyond the target of this book, see Kitsos (2011a, b). The Fisher's information matrix is introduced in the next section.

2.5 Fisher's Information Matrix

There are two different types of Fisher's information measures: the parametric one, which will be presented next, and the entropy type which has been extended (Kitsos and Tavoularis 2009), and is not discussed in this monograph. Both are strongly related with the involved uncertainty in the physical phenomenon under investigation. Sometimes, due to the cost of the experiment, the experimentalist needs to run a small number of trials, and eventually estimate, adequately, the involved parameters—one can imagine how useful the theoretical limiting results are in such cases! But Fisher's information is always there, helpful to the analysis, even with a small sample size, see Sect. 6.4.

In non-linear problems the variance depends on the unknown parameter the experimenter wants to estimate, i.e. $\sigma^2 = \sigma^2(u, \theta)$. That is σ^2 depends on the design point and the parameter vector. In the linear case it is assumed independent of the parameter vector θ. In practice it may or may not be possible to assume that

is "known", provided a guess, or knowledge from a previous experiment. In principle, this is the main difference between the linear and the non-linear case.

Let $\nabla\eta$ denote the vector of partial derivatives (recall Assumption 1)

$$\nabla\eta = \left(\frac{\partial\eta}{\partial\theta_1}, \frac{\partial\eta}{\partial\theta_2}, \ldots, \frac{\partial\eta}{\partial\theta_p}\right)^T. \tag{2.5.1}$$

Then for the exponential family of models Fisher's information matrix is defined to be

$$I(\theta, u) = \sigma^{-2}(\nabla\eta)(\nabla\eta)^T. \tag{2.5.2}$$

Moreover in many of the nonlinear problems the covariate u and the parameter θ appear together linearly in the form $\theta^T u$. Thus the following proposition holds:

Proposition 2.1 *If* $\eta = \eta(\theta^T u)$ *it can be proved that* $\nabla\eta = [w(\theta^T u)]^{1/2} u$ *with* $w(z) = [\partial\eta/\partial z]^2$, $z = \theta^T u$ *and therefore:*

$$I(\theta, u) = \sigma^{-2} w(\theta^T u) u u^T.$$

This result describes that fact that if the nonlinear model is *"intrinsic linear"* as far as the parameter concerns, $\eta = \eta(\theta^T u)$, then Fisher's information matrix can be evaluated proportional to the matrix produced only by the input vector u, i.e. equal to $\sigma^{-2} w(\theta^T u) u u^T$.

The concept of the *average-per-observation information matrix* will play an important role in the developed scenario concerning the nonlinear experiment design problem. It is defined for the ξ_n, the n-point design measure, to be equal:

For the discrete case:

$$M(\theta, \xi_n) = n^{-1} \sum I(\theta, u_i),$$

For the continuous case:

$$M(\theta, \xi) = \int_U I(\theta, u)\xi(du), \ \xi \in \Xi. \tag{2.5.3}$$

The theoretical insight of Caratheodory's Theorem (Appendix 1), so essential for the linear experiment design, can also be used for the average information matrix in nonlinear problems, see Titterington (1980a). Now, suppose that the matrix $M = M(\theta, \xi)$ is partitioned in the form:

$$M = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} \tag{2.5.4}$$

with $M_{11} \in$ Mat (s, s), $M_{12} \in$ Mat(s, p − s), $M_{22} \in$ Mat(p − s, p − s), $1 \leq s < p$. Then the following matrix is defined:

$$M_S = M_S(\theta, \xi) = M_{11} - M_{12}M_{22}^{-}M_{12}^{T} \tag{2.5.5}$$

with M_{22}^{-} being the generalized inverse of M_{22}. The information matrix I, as in (2.5.2), is considered partitioned in the same fashion. This partition is helpful when interest lies in estimating the leading $s < p$ parameters in the vector θ, as it will be explained in Sect. 3.6.

2.6 Linearization of the Model

The idea of the (design) matrix X, including all he input variables, is essential when one is working with linear models, see Seber and Wild (1989) among others. In nonlinear models the researcher can not define a matrix X in the same fashion. This can be done only approximately through the partial derivatives of θ, with θ taking its "true" value, θ_t. Then the following $n \times p$ matrix is defined

$$X = (x_{ij}) = \frac{\partial f(u_i, \theta)}{\partial \theta_j} \mid_{\theta=\theta_t}. \tag{2.6.1}$$

Then the matrix $X = X(\theta)$ is formed as a function of θ. The function $f(u, \theta)$, recall (2.2.2), can be "linearized" through a Taylor series expansion, in the neighborhood of θ_t as:

$$f(u, \theta) = f(u, \theta_t) + \sum (\theta_j - \theta_{tj})(\partial f(u, \theta)/\partial \theta_j)|\theta = \theta_t \tag{2.6.2}$$

Following the pattern of linear regression models in the nonlinear regression case, an approximation to the covariance matrix, of the estimates of the parameters, can be defined as:

$$C \cong [X^T(\theta_t)X(\theta_t)]^{-1}\sigma^2. \tag{2.6.3}$$

Moreover for all nonlinear problems a useful approximation to the covariance matrix is $C^{-1} \cong nM(\theta_t, \xi)$. Eventually for the average per-observation information matrix M for the design measure ξ holds:

$$M(\theta, \xi) = \frac{1}{n}(\nabla\eta)^T\nabla\eta = \frac{\sigma^2}{n}\mathbf{C}^{-1}(\widehat{\theta}, \xi) \tag{2.6.5}$$

Relation (2.6.5) describes that (asymptotically) the covariance matrix is the inverse of the average per observation information matrix.

2.6.1 Examples

To clarify the above theoretical background, the following examples are given, helpful for the inquiring reader.

Example 3 Recall Examples 1 and 2, as far as the logit model concerns. Given that the linear approximation of the form

$$[1 + \exp(-\theta_1(u - \theta_2))]^{-1} \cong 1/2 + 1/6\ \theta_1(u - \theta_2)$$

holds under the assumption $|\theta_1(u - \theta_2)| \leq 3$, then the normal equations, for the logit model, approximately are:

$$n/2 + (\theta_1/6)\sum(u_i - \theta_2) = \sum y_i$$
$$(1/2)\sum u_i + (\theta_1/6)\sum u_i(u_i - \theta_2) = \sum u_i y_i$$

Solving the simultaneous equations the estimates of the parameters (θ_1, θ_2) are obtained.

In the following examples the idea of Fisher's information matrix is clarified. These examples will be reconsidered in the sequel of this development.

Example 4 Consider that model, in which $P(y = 1) = T(\theta^T u)$. That is the model T is "*intrinsic linear*": the model can be non-linear, but the parameters and the input variable are related linearly.

Let: $\theta_1 + \theta_2 u_1 = z = \theta^T u$ and $T'(z) > 0$, $\theta = (\theta_1, \theta_2)$, $u = (1, u_1)$.
Then the log-likelihood ℓ, will be $\ell = \log\{\ T(z)^Y[1 - T(z)]^{1-Y}\} + \text{const.}$
Therefore Fisher's information matrix equals to

$$I(\theta, \xi) = E\{(\nabla\ell)(\nabla\ell)^T\} = \alpha(\theta)\ uu^T, \text{with } \alpha(\theta) = T'^2[T(1 - T)].$$

Application: T might be either the logit or probit model.

Example 5 For the nonlinear regression model $\eta = \theta_1 - \exp(-\theta_2 u), u \in [-1, 1]$

Fisher's information matrix is evaluated equal to

$$I(\theta, \xi) = (\nabla\eta)(\nabla\eta)^T = \begin{pmatrix} 1 & u\exp(-\beta)(-\theta_2 u) \\ u\exp(-\theta_2 u) & u^2\exp(-\theta_2 u) \end{pmatrix}$$

Note that $I(\theta, \xi)$ is a function only of $\theta_2, I(\theta, \xi) = I(\theta_2, \xi)$, therefore the covariance matrix, recall relation (2.6.5), depends on the parameters, but only on the component θ_2 of the parameter vector (θ_1, θ_2). Notice that in linear models the covariance matrix does not depend on the parameters.

Example 6 The nonlinear regression model used to describe growth phenomena is

$$\eta = \theta_1 \exp(\theta_2 u), u \in [a, b]$$

The x_{ij} vectors $i = 1, 2, \ldots, n$, $j = 1, 2$ can be formed according to (2.6.1) as

$$x_{i1} = \partial\eta/\partial\theta_1 = \exp(\theta_2 u_i), \quad x_{i2} = \partial\eta/\partial\theta_2 = \theta_1 u_i \exp(\theta_2 u_i)$$

The quantity $S_n(\theta) = \sum (y_i - \theta_1 \exp(\theta_2 u_i))^2$ is formed. In order to find the Least Square Estimator $\hat{\theta}$, which minimizes $S_n(\theta)$, the partial derivatives of $S_n(\theta)$ are needed, which provide the "normal equations"

$$\sum (y_i - \theta_1 \exp(\theta_2 u_i)) \exp(\theta_2 u_i) = 0$$
$$\sum (y_i - \theta_1 \exp(\theta_2 u_i)) \theta_1 u_i \exp(\theta_2 u_i) = 0.$$

Moreover, in this case, from the *Hessian matrix* and its' expected value it is eventually evaluated (recall (2.6.5)) that

$$\sigma^{-2} nM(\theta, \xi_n) = \begin{pmatrix} \sum \exp(2\theta_2 u_i) & \theta_1 \sum u_i \exp(2\theta_2 u_i) \\ \theta_1 \sum u_i \exp(2\theta_2 u_i) & \theta^2 \sum u^2 \exp(2\theta_2 u_i) \end{pmatrix}.$$

Also, from (2.4.4) is evaluated that

$$\sigma^{-2} S(\widehat{\theta}, \xi_n, y) = \begin{pmatrix} \sum \exp(2\widehat{\theta}_2 u_i) & \hat{\theta}_1 \sum u_i \exp(2\widehat{\theta}_2 u_i) \\ \widehat{\theta}_1 \sum u_i \exp(2\theta_2 u_i) & A(y, u, \widehat{\theta}) \end{pmatrix} \tag{2.7.1}$$

where $-A(y, u, \widehat{\theta}) = \widehat{\theta}_1 \sum u_i^2 \exp(\widehat{\theta}_2 u_i) \{y_i - 2\widehat{\theta}_1 \exp(\widehat{\theta}_2 u_i)\}$.

Example 7 Dose-Finding in Phase I/II Clinical Trials

The goals of a clinical trial, Thall and Russel (1998), are:

1. To perform an experiment that satisfies specific safety and efficacy requirements,
2. To consider the trial early as it is likely that no dose is both safe and efficacious,
3. To estimate, with some risk, the rates of the events at the pre-decided level of dose.

The dose response curve, for the logit model (recall Example 1), is the cumulative odds model, to describe and to approach, the unknown dose-response function. So the underlying model, describing the experiment, is nonlinear. A binary indicator describes the levels of severe toxicity. The appropriate Clinical Trial imposes a target to estimate that optimal level of the dose, among different candidate dose levels, that satisfies both the efficacy and toxicity criteria.

For examples of non-linear models in Cancer Bioassays see Kitsos et al. (2009).

Example 8 Oxidation of benzene.

This model is, among the chemical reaction models, one with the most necessary needed information: Includes 4 input variables and 4 parameters:

$$\text{Model: } \eta = \frac{\theta_1 \exp(-\theta_3 u_3)\theta_2 \exp(-\theta_4 u_3) u_1 u_2}{\theta_1 \exp(-\theta_3 u_3) u_1 + u_4 \theta_2 \exp(-\theta_4 u_3) u_2}$$

The input variables and the design space U are defined below, while the four involved parameters belong to the parameter space $\Theta \subseteq R^4$,

$$u = (u_1, u_2, u_3, u_4) \in U = \Delta_1 \times \Delta_2 \times \Delta_3 \times \{5.75\} \subseteq R^4, \text{ with}$$
$$\Delta_1 = \left[10^{-3}, 16 \times 10^{-3}\right], \Delta_2 = \left[10^{-3}, 4 \times 10^{-3}\right], \Delta_3 = [623, 673].$$

Moreover the introduced notation is:

η	The initial reaction rate
u_1	The concentration of oxygen
u_2	The concentration of benzene
$u_3 = 1/T - 0.0015428$	T is the absolute temperature of the reaction
u_4	The observed stoichiometric number
θ_1, θ_2, θ_3, θ_4	Model parameters arising from Arrhenius' law.

Usually the real situation is not such complicated, see Sect. 4.7, i.e. indeed the parameters involved are 2 or 3 and usually no particular knowledge concerning one of them is needed. For the n-th order chemical kinetic models see Manerswammy et al. (2009), while for a compilation of the Chemical Kinetics models and the corresponding design theory see Kitsos and Kolovos (2013).

The above examples provide evidence that the nonlinear optimal experimental design theory covers a broad range of applications as it will be also clarified in the sequence.

References

Jennrich RJ (1969) Asymptotic properties of non-linear least-squares estimators. Ann Math Stat 40:633–643

Kitsos CP (2011a) Invariant canonical form for the multiple logistic regression. Math Eng Sci Aerosp (MESA) 2(3):267–275

Kitsos CP (2011b) On Fisher's information measures. e-Proceedings of digital technologies 2011, Zilina, Slovakia, 10–11 Nov 2011, pp 178–183

Kitsos CP, Kolovos KG (2013) A compilation of the D-optimal designs in chemical kinetics. Chemical Engin Comm 200:1–20

Kitsos CP, Tavoularis NK (2009) Logarithmic Sobolev inequalities for information measures. IEEE Trans Inform Theory 55(6):2554–2561

Kitsos CP, Toulias TL, Lolas G (2009) Modeling cancer bioassays. Biometrie und Medizinische Informatik Greifswalder Seminarberichte, Heft, vol 15, pp 75–90

Manerswammy A, Munson-McGee S, Steiner R, Johnson CI (2009) D-optimal design of DSC experiments for nth order kinetics. J Therm Anal Calorium 97:895–902

Seber CAF, Wild CJ (1989) Nonlinear regression. Wiley, London

Silvapulle MJ (1981) On the existence of the maximum likelihood estimators for the Binomial response models. J.R.S.S.B. 43:310–313

Thall PF, Russel K (1998) A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. Biometrics 54:251–264

Titterington DM (1980) Aspects of optimal design in dynamic systems. Technometrics 22:287–299

Wu CFJ (1981) Asymptotic theory of nonlinear least squares estimates. Ann Stat 9:501–513

Chapter 3
Locally Optimal Designs

Abstract The locally optical design is discussed and the alphabetic optimal designs are introduced, and the Geometrical insight is discussed. The canonical form for a binary response problem is introduced, and its "affine" character is discussed.

3.1 Introduction

The aim of the optimal experiment is to estimate appropriately the parameters included in the model. It might be considered all p parameters, or a set of $s < p$ linear combinations of the p parameters. In the sequel, when only s of the parameters are to be estimated. It will be assumed that they are the first s components of the vector $\theta = (\theta_1, \theta_2, \ldots, \theta_P)$ and the notation $\theta_{(S)} = (\theta_1, \ldots, \theta_S)$ shall be used. The average information matrix, related approximately to the covariance matrix, as in (2.5.3), is a natural starting point for the establishment of an optimality criterion. That is, some real-valued function of $M(\theta, \xi)$ can serve as a criterion, to decide why one design is optimal comparing with others. Now, one might ask: why the design is called *locally optimal*?

It is the θ-dependence which leads to the term "locally optimal": the optimal design depends on the true value of θ and therefore it might be optimal "locally", in the sense that it is optimal "around", in the *neighborhood*, of the parameter value.

This θ-dependence is the main point of difference between linear experimental design, originated in Smith (1918) and the nonlinear case, originated in Fisher (1922).

Thus in this chapter the appropriate criteria describing why the use of one design is the best one, in comparison with other candidate designs, are provided. In addition, the geometrical interpretation of these design criteria is briefly discussed.

C. P. Kitsos, *Optimal Experimental Design for Non-Linear Models*,
SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-45287-1_3,

3.2 Formulation

Suppose that one wishes to estimate a set of linear combinations of the parameter vector $\theta = (\theta_1, \ldots, \theta_P)$. This might lead to an estimation of the vector θ itself, some linear combinations of the p components of θ or to $s \leq p$ components. Let $Q \in Mat(s,p)$ $1 \leq s \leq p$ be the matrix of the known coefficients defining and that the quantities of interest are the linear forms, presented in compact form, $Q\theta$. If $rank(Q) = p$, when $s = p$, the matrix Q is nonsingular. If $s < p$ then let us suppose that $rank(Q) = s$. On the basis of the experiment the average (per observation) information matrix $M = M(\theta, \xi)$ is obtained. Notice that the inverse of the matrix M might not exist (when det(M) = 0). In the sequel θ is regarded as taking its true value. Then the following operator J_Q applied to M can be defined through the above matrix Q:

$$J_Q(M) = QM^{-}(\theta, \xi)Q^T \tag{3.2.1}$$

with M^{-} a generalized inverse of M and $Q^T \in Mat(p,s)$

Given the above notation a real valued function is needed, ω say, applied to J_Q to be used as an optimality criterion. The function ω has been chosen to be a convex decreasing one on the set of nonnegative definite matrices, $NMat(s,s)$ say, i.e. $\omega(A) \leq \omega(B)$ if the matrix $A - B \in NMat(s,s)$ and $A, B \in NMat(s,s)$.

Definition 3.2.1 The design measure ξ^* is called ω-optimal if and only if (iff):

$$\omega\{J_Q[M(\theta, \xi^*)]\} = \min\{\omega\{QM^{-}(\theta, \xi)Q^T\}, \xi \in \Xi\}. \tag{3.2.2}$$

The following special cases of the concave function ω and the matrix Q are examined, which lead to the traditional definitions, adopted from the linear case. The special cases for ω are:

$$\omega(.) = \begin{cases} \omega_1(.) = \log\left[\det\left(QM^{-}Q^T\right)\right] \\ \omega_2(.) = tr\left(QM^{-}Q^T\right) \\ \omega_3(.) = \max\ eigenvalue\ of\left(QM^{-}Q^T\right) \\ \omega_4(.) = \sup\left[tr\left(I(\theta, u)QM^{-}Q^T\right)\right] \end{cases} \tag{3.2.3}$$

The most often considered special cases for Q are

$$Q = \begin{cases} A \in Mat(p,p), rank(A) = p \\ I \in Mat(p,p) - the\_identity\_matrix \\ A \in Mat(s,p), rank(A) = s \\ [I_S : 0], I_S \in Mat(s,s), 0 \in Mat(s, p-s) \\ c \in Mat(p, I) \end{cases} \tag{3.2.4}$$

The above cases are the most often appeared under the theoretical framework presented here. For practical applications see the following Sects. 3.3 and 3.4.

Usually experimenters from Biology or Chemistry are referring to the logdet criterion, with no explanation on how it has been decided; see for geometrical explanation Sect. 3.7.

3.3 Special Cases

The following definitions are given for $\omega(.)$ as in (3.2.3) and Q = A as in (3.2.4):

Definition 3.3.1 The design measure ξ^* is called (iff: if and only if)

$D_A(\theta)$-optimal iff $\omega_1(.)$ is considered,
$A_A(\theta)$-optimal iff $\omega_2(.)$ is considered,
$E_A(\theta)$-optimal iff $\omega_3(.)$ is considered,
$G_A(\theta)$-optimal iff $\omega_4(.)$ is considered. □

When $Q = I \in Mat(p, p)$, the operator $J_I(.)$ provides the generalized inverse of the matrix M. In this case the so called Φ criterion is obtained, see Ford et al. (1989) for a review. Under this notation, the traditionally described as Φ criterion, Kiefer and Wolfowitz (1960), is $\phi = \omega \circ J_I$, where $\circ$ denotes the composition of two functions, see also Silvey (1980, Chap. 3). Thus the corresponding to ω_i, i = 1, 2, 3, 4 of (3.2.5) the classical case is obtained

$$\phi_i = \omega_i \circ J_I, \quad \text{i} = 1, 2, 3, 4. \tag{3.3.1}$$

The "alphabetic" traditional definitions of these criteria are the following.

Definition 3.3.2 An optimal design measure ξ^* is called

$D(\theta)$-optimal iff $\varphi_1(.)$ is considered,
$A(\theta)$-optimal iff $\varphi_2(.)$ is considered
$E(\theta)$-optimal iff $\varphi_3(.)$ is considered,
$G(\theta)$-optimal iff $\varphi_4(.)$ is considered.

The reference to the parameter is only to emphasize that these criteria depend on the parameter, while in the linear case the terms D, G, A, etc.-optimality are used.

For the optimality criteria $D(\theta)$ and $G(\theta)$. White (1973) extended Kiefer and Wolfowitz's (1960) *equivalence theorem* as it follows:

Theorem 3.3.1 *For the optimal design measure ξ^* the following are equivalent.*

1. *ξ^* is $D(\theta)$-optimal.*
2. *ξ^* is $G(\theta)$-optimal.*
3. *$Sup[d(u, \xi, \theta)] = p = \dim \theta$, where $\dim\theta$ is the dimension of $\theta \in \Theta \subseteq R^P$ and $d(u, \xi, \theta) = tr\left\{I(\theta, u)M(\theta, \xi)^{-1}\right\}$ (with M assumed to be nonsingular).* □

When $A \in Mat(p, p)$ nonsingular, there is no difference between the D-optimality of a linear transformation A, $DA(\theta)$ and $D(\theta)$-optimality. This is true due to the fact that

$$\det\left[AM^{-1}(\theta, \xi)A^T\right] = \det\left[M^{-1}(\theta, \xi)\right] \cdot [\det(A)]^2 \tag{3.3.2}$$

This is not the case with other optimality criteria and this result is an essential difference between the D-optimality and the other criteria.

$D(\theta)$-optimality minimizes the volume of the approximate confidence ellipsoids for θ, centered at the estimate of θ, $\hat{\theta}$ say. Moreover, the information matrix $M(\theta, \xi^*)$ corresponding to the optimal design ξ^* is unique—when θ takes its true value—since ω in this case is a strictly concave function. The duality theory for the linear case, first tackled by Silvey (1972), and established by Sibson (1972) can be applied in the nonlinear case when $\theta = \theta_t$, for the other criteria as well. Thus:

$G(\theta)$ Optimality minimizes the maximum approximate variance of the estimated future response, the interpretation for both $D(\theta)$ and $G(\theta)$ optimality has been made, under Assumption 2, that of normal errors.

$A(\theta)$ Optimality minimizes the sum of approximate variances of the parameter estimates, as in the linear case, see Titterington (1980a).

$E(\theta)$ Optimality seeks to minimize the variance of the worst-estimated linear combination $c^T\theta$, with $c^Tc = 1$, see also Dette and Haines (1994).

Silvey (1980) reviews these criteria for the linear case, while Pukelsheim (1993) offers a compact theoretical review of the criteria, mainly for the linear case.

A particular case is that of $[I_S : 0]$, $I_S \in Mat(s, s)$ the unit matrix and $0 \in Mat(s, p - s)$ the zero matrix. This case is discussed in Sect. 3.4.

Now let us consider the case $Q = c \in Mat(p, 1)$, i.e. Q is a vector. Assuming that $M^{-1}(\theta, \xi)$ exists and recalling Definition 3.2.1, it is

$$\omega\{J_C[M(\theta, \xi^*)]\} = \min\left\{c^TM^{-1}(\theta, \xi)c, \xi \in \Xi\right\} \tag{3.3.3}$$

with ω the identity function. This criterion minimizes the approximate variance of a linear combination of $\hat{\theta}$ and it is known as $c(\theta)$-optimality. Special interest is given to the c-optimality, due to the excellent Geometrical interpretation of the design measure. For completely discussed applications see Example 10 below and Chap. 8.

3.4 Applications

In the linear case the above criteria are independent of θ and thus the reference to them is D, G, A, E-optimality. The nonlinear case can be treated as the linear, when it is supposed that θ is known.

The $D(\theta)$-optimality criterion has been the most commonly used in practice, ever since the pioneering work of Box and Lucas (1959), who obtained locally optimal designs when $n = p$ for a number of nonlinear models. For the exact locally optimal design when $n = p$ maximization of $\det(X^T X)$ (see (2.6.1) for the definition of the matrix X) is the equivalent to maximization of det (X) because, $\Delta = \det(X^T X) = [\det(X)]^2$. Notice that $\det(X)$ exists, as it is supposed that the created simplex is due to the assumption n = p.

Atkinson and Hunter (1968) suggested that n should be $n = rp$, the sample size is a replication of the p-term design. Hence one should perform the experiment of Box and Lucas for the p-point p-term model and replicate the experiment r times. Box (1968) gave a generalization for $n = rp + k,\ r > 1,\ 0 \le k \le p - 1$.

Most of the work dealing with the chemical kinetic models is summarized in Sect. 4.7. Given the "true" θ the optimal design points for the covariates involved has been listed. $A(\theta)$-optimality has been suggested by Titterington (1980a) for dynamic systems. Little attention has been paid to $E(\theta)$-optimality in applications. □

Example 9 Consider the logit and probit models [recall Example 1, see also McCullagh and Nedler (1989)] under $D(\theta)$-optimality. This is essential because of their use in applications. For the quantal response model of the form $T = T(\theta^T u)$ (recall Example 4), the $D(\theta)$-optimal design is concentrated at two points, namely:

$$u_1 = (u_0 - \theta_1)/\theta_2, \quad u_2 = (-u_0 - \theta_1)/\theta_2 \tag{3.4.1}$$

with $\xi_1 = \xi_2 = 0.5$, which means allocate half observation at each optimal design point, i.e. a D-optimal design. Then the value of the determinant of the average per observation information matrix is $D_o = \det M(\theta, u_0) = \left\{u_0 \cdot a(u_0)^2/\theta_2\right\}$ with $a(u_0) = a(u_0, \theta)$ as in Example 4.

Let us assume that the function $D_o = D_o(u_0)$ has a unique maximum at $\hat{u}_0$. Then the two D-optimal points turn to be of the form

$$\frac{(\pm\hat{u}_0 - \theta_1)}{\theta_2} \in U$$

Moreover for the *logit* case it is: $\hat{u}_0 = 1.54$, For the *probit* case it is: $\hat{u}_0 = 1.14$.

If the design space U is symmetric about $-\theta_1/\theta_2$ and $(\pm\hat{u}_0 - \theta_1)/\theta_2 \in U = [\kappa, \lambda]$ then the $D(\theta)$-optimal design is based on the end points of the interval, i.e. $u_1 = \kappa,\ u_2 = \lambda$.

Notice that both logit and probit models are based on a two point symmetric D-optimal design. The situation is similar for the

Double Exponential $\left(\frac{1}{2}\exp(-|u|)\right)$ at the points $\hat{u}_0 = \pm 0.768$ and for

Double Reciprocal $\left(\frac{1}{2}(1 + |u|)^{-2}\right)$ at the points $\hat{u}_0 = \pm 0.390$.

No symmetry takes place for the

$$\textit{Complementary Log–Log } (\exp(u - e^2)) \text{ model:}$$

The D-optimal design allocates half observations at $\hat{u}_0 = -1.338$ and $\hat{u}_0 = 0.980$.

For the binary response optimal design problem, under logistic regression see Mathew and Sinha (2001). □

Example 10 The c(θ) optimality criterion can be used when the percentile point of the logistic curve is asked to be estimated. The 100p percentile point, L_p of the response curve T(u) is defined as the solution of the equation: $\mathrm{T}(L_p) = \mathrm{p}$, when T(u) is the logistic. Then it is

$$\mathrm{T}(L_p; \theta) = \left\{1 + \exp(-(\theta_0 + \theta_1 L_p))\right\}^{-1} = \mathrm{p} \tag{3.4.2}$$

Therefore

$$L_p = -\theta_1^{-1}\left[\theta_0 + \ln(p^{-1} - 1)\right] = \tilde{L}_p(\theta_0, \theta_1) \tag{3.4.3}$$

Thus L_p has been expressed as a non linear function of θ_0, θ_1. In bioassays the median is the most common percentile of interest. It is easy to see that for $\mathrm{p} = 0.5$ relation (3.4.3) is reduced to

$$L_{0.5} = -(\theta_0/\theta_1) \tag{3.4.4}$$

for the logit case. Clearly, designing as well as possible to get the best estimate of L_p, has a practical use. The vector $\nabla\tilde{L}_p$is evaluated as

$$\nabla\tilde{L}_p = -\theta_1^{-1}(1, L_p)^T \tag{3.4.5}$$

Let n_i be the number of observations at u_i for $\mathrm{i} = 1, 2, \ldots, \mathrm{k}$. Then, for the MLE $\widehat{\theta} = \left(\widehat{\theta}_0, \widehat{\theta}_1\right)$, it is known, Kitsos (1986) that for $\mathrm{k} \geq 2$, and $\sum n_i = n$ large

$$\mathrm{Var}(\widehat{\theta}_0, \widehat{\theta}_1) \cong \left\{\sum \mathrm{T}(u_i)(1-\mathrm{T}(u_i))u_i\, u^T n_i\right\}^{-1} \tag{3.4.6}$$

in which $u_1 = (1, u_{i1})^T$. From (3.4.5) and (3.4.6) one obtains that

$$\mathrm{Var}\ (\widehat{L}_p) \cong (\nabla\tilde{L}_p)^T \mathrm{Var}\ (\widehat{\theta}_0, \widehat{\theta}_1)\nabla\tilde{L}_p \tag{3.4.7}$$

Therefore minimization of $\mathrm{Var}(\hat{L}_p)$ is approximately equivalent to minimization of

$$c^T M^{-1}(\xi)\ \mathrm{c} \tag{3.4.8}$$

with $\mathrm{c} = (1, L_p)^T$ and $\mathrm{M}^{-1}(\xi)$ given by the right had side of (3.4.6). This is equivalently to find the c(θ)-optimal design, for the desired $\theta = (\theta_0, \theta_1)$. □

From the above Example 10 it is clear that the c-optimal design, depends on the vector c (the "ray" say, that hits the induced design space, see also Appendix II). The question is to find for a design, acting as a basis for a vector c, and therefore all the other designs will be considered, as a transformation of the "basic" one. This design will be referred as the "canonical" form and it will be discussed below.

3.5 Canonical Form of a Design (c(θ)-Optimality)

It is known, Federov (1972, p 81), that D-optimal designs are invariant to any non-degenerate linear transformation of the parameters.

It is in fact convenient, to have a design criterion, which will remain invariant under certain transformations of the design space. Then it is possible to have a design acting as a basis, a "*canonical form*" of the design, which, when it is transformed, it would produce another one, a "daughter designs". So a family of "invariant" designs is introduced. The transformation defined below is based on the above discussion and it is of the form:

$$\mathrm{h} : U \subseteq R^k \to Z \subseteq R^k : \mathrm{u} \to \mathrm{h(u)} = \mathrm{z} = \mathrm{Bu} \tag{3.5.1}$$

with B $\in$ Mat (p, p), nonsingular.

Consider that nonlinear models in which the parameters appear in the linear form with the covariate u, such as $\theta^T u = \theta_0 + \theta_1 u_1$. Take as a design criterion that of c(θ)-optimality i.e. the general concave criterion function is of the form

$$\Phi(M_u) = c^T M_u^{-1} c \tag{3.5.2}$$

with $M_u = M_u(\theta, \xi)$ the average information matrix in U-space equals to

$$M_u = \sum \alpha(\theta_0 + \theta_1 u_{1i}) \mathrm{c}\, c^T \tag{3.5.3}$$

and α(.) as in Example 4, and $c_i = (1, u_{1i})^T$, c = (1, u_1)^T. Let us then define the matrix B as

$$\mathrm{B} = \begin{pmatrix} 1 & 0 \\ \theta_0 & \theta_1 \end{pmatrix} \tag{3.5.4}$$

and

$$\mathrm{z} = \mathrm{Bu} = \begin{pmatrix} 1 & 0 \\ \theta_0 & \theta_1 \end{pmatrix} \begin{pmatrix} 1 \\ u_1 \end{pmatrix} = \begin{pmatrix} 1 \\ \theta_0 + \theta_1 u_1 \end{pmatrix} = \begin{pmatrix} 1 \\ z_1 \end{pmatrix} \tag{3.5.5}$$

Then the criterion (3.5.2) can be written as:

$$\begin{aligned} c^T M_u^{-1} c &= c^T \left[\sum_i \alpha(\cdot) c_i c_i^T \right]^{-1} c \\ &= c^T B^T (B^T)^{-1} \left[\sum_i \alpha(\cdot) c_i c_i^T \right]^{-1} B^{-1} B c \\ &= (Bc)^T \left(\sum_i \alpha(\cdot) [Bc_i][Bc_i]^T \right)^{-1} (Bc) = c_z^T M_z^{-1} c_z \end{aligned} \tag{3.5.6}$$

The ray c and the average per observation information matrix on the Z space equal to $c_z = (1, \theta_0 + \theta_1 u_1)^T = (1, z_1)^T$, $M_z = \sum_i \alpha(.) z_i z_i^T, z_i = (1, z_{1i})^T$ respectively.

Thus the equivalence of c(θ) optimality in U space and Z space has been proved. This is of practical use as a design can be constructed on a 'suitable' design space with θ_0, θ_1 fixed and then transformed back to the design of interest. □

Now let us improve, extend and apply this line of thought. One of the most appropriate statistical methods to evaluate the Relative Risk (RR) in practice, for example in the dose–response experiments involving one or more variables, is the logistic regression. The construction of local *D*- or *c*-optimal designs of the binary response models have been discussed by Sitter and Torsney (1995a, b), while Ford et al. (1992) introduced the one variable logistic model canonical form. This idea is extended to *p*-variable logistic model.

Indeed: for the Logit transformation, the logistic regression model is obtained; see also Breslow and Day (1980). In such a case extending Example 1, the model is

$$y = \log \frac{p(x)}{1 - p(x)} = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p + \sigma e, e \sim N(0, 1) \tag{3.5.7}$$

In order to obtain a group of affine transformations forming the canonical form of the model (3.5.7), the following Proposition and Theorem, see Kitsos (2011a), for details and proofs, are applied.

Proposition 3.5.1 *The set of the (affine) transformations G as*

$$G = \left\{ g = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & & & & \vdots \\ 0 & \cdots & & 1 & 0 \\ \beta_1 & \beta_2 & \cdots & \beta_p & \sigma \end{pmatrix} = \begin{pmatrix} \mathbb{I}_p & \mathbb{O}^t \\ \theta & \sigma \end{pmatrix}, \quad g \in \mathbb{R}^{(p+1)^2} \right\}, \tag{3.5.8}$$

with $\mathbb{I}_p = \text{diag}(1, 1, \ldots, 1) \in R^{p \times p}$, $\mathbb{O} = (0, \ldots, 0) \in R^{1 \times p}$, $\theta = (\beta_1, \beta_2, \ldots, \beta_p) \in R^{1 \times p}$ and $\sigma \in R^+$, *R the set of real numbers, forms a group, under matrix multiplication.*

Theorem 3.5.1 *Consider any positive definite matrix M with* $\det M \neq 0$ *and vector c with the appropriate dimension, so that,* $N = c^{\mathrm{T}} M^{-1} c$ *is valid. Then, N remains invariant if c and M transformed under g.*

This theoretical result practically means that the experimenter can work as follows: per-form the experiment with the "easiest" scale and position parameters—those with some prior knowledge from previous experience. Then, transfer the obtained results with an element from the group of the affine transformations and still you have an optimal design. Returning to the initial design space, through the inverse transformation, still you are moving within a class of optimal designs. Thus it is recommended a two stage design:

1. Devote a portion of your experiments with the initial guesses
2. Perform the remaining experiments within the class of the design generated by the group of transformations G.

When the model obtained from the part (2) above is fitted, still the canonical form is appreciated: to evaluate how far you moved from it, how the model is renewed. □

3.6 Designs for Subsets of Parameters

The asymptotic generalized variance of the estimators of the s-component vector $\theta_{(S)} = (\theta_1, \ldots, \theta_S)$ is defined to by the quantity

$$\mathrm{V}(\theta, \xi) = n^{-1} \det[M_s] \tag{3.6.1}$$

where M_s is as in (2.5.10). Note that M_s has to be nonsingular so that the vector $\theta_{(s)}$ to be estimable. With the operator notation introduced above, it is easy to see that when $\mathrm{A} = [I_s : 0]$, $J_A(M) = M_s$. Thereafter the notation ϕ_s as $\Phi_s = \omega \, o \, J_{[I_s:0]}$ is used.

For the cases ϕ_i, i = 1,2,3,4 the following definition are given

Definition 3.6.1 The optimal design measure ξ^* is called

$D_s(\theta)$-optimal if ϕ_{1s} is considered,
$A_s(\theta)$-optimal if ϕ_{2s} is considered
$E_s(\theta)$-optimal if ϕ_{3s} is considered
$G_s(\theta)$-optimal if ϕ_{4s} is considered. □

White (1973) stated an equivalence theorem for $D_s(\theta)$ and $G_s(\theta)$ optimality, similar to that for the linear case, while Begg and Kalish (1984) applied the $D_s(\theta)$ optimality for the logistic problem.

3.7 Geometrical Aspects

For the linear model (2.3.3) the geometry was built up not on the design space U, but on its image through g, $U_0 = g(U)$ say, known as *induced design space*. Furthermore, for this transformation it is proved that the information matrix M(ξ) is preserved expressed in terms of the family of design measures

$$\Xi_0 = \left\{\xi_0 = \xi g^{-1}, \xi \in \Xi\right\} \tag{3.7.1}$$

The geometrical approach to the linear optimal design of experiments has been extensively covered by Titterington (1975, 1980a). The geometrical insight into the linear problem is based on the pioneering equivalence theorems of Kiefer and Wolfwitz (1960), for G- and D-optimality, while Karlin and Studden (1966), worked for D_s optimality.

Under the above discussion the following "duality' principles play an important role in the realm of linear experimental design, and are summarized in the following Theorem.

Theorem 3.7.1 *Let U_0 be a compact set which spans R^P (which spans the leading s-dimensional co-ordinate subspace).* Then:

1. *The D-optimal design problem for U_0 is the dual of the minimal ellipsoid problem for U_0* (Sibson 1972).
2. *The D_s optimal design problem is the dual of the thinnest cylinder problem* (Silvey and Titterington 1973).

In both cases the two problems share a common extreme value. □

Based on Theorem 3.7.1 the minimal ellipsoid problem is that of finding an ellipsoid of minimal content centered at the origin, containing U_0. The kernel matrix of this particular ellipsoid is the inverse of the information matrix, $M^{-1}(\xi)$, equivalently the covariance matrix—which in linear case does not depend on the involved parameter. The only possible effective support points of D-optimal designs are points in U which, through g, correspond to the points where the minimal ellipsoid "hits" U_0.

The thinnest cylinder problem is that of finding a cylinder of minimum cross-sectional content—with cylinder axis required to pass through the origin—which spans R^s and which contains U_0. The only possible effective support points, are points in U whose images in U_0 are points where the thinnest covering cylinder hits U_0.

These ideas from the linear case are discussed to clarify where this ellipsoid is, and how it is influenced by the design measure and the information matrix, see also Fig. 3.1. Now one can proceed to the nonlinear case, adopting the linear development.

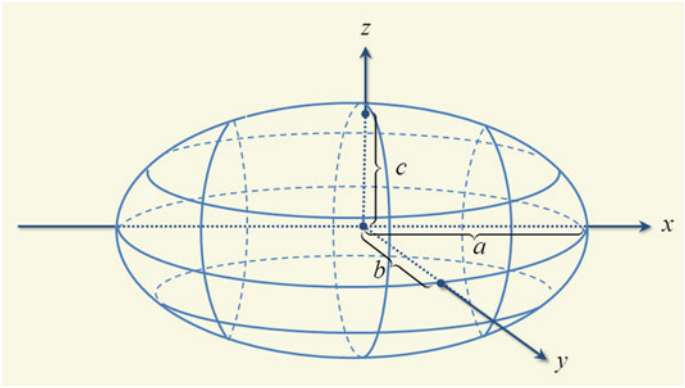

Fig. 3.1 The confidence ellipsoid in 3 dimensions

The fact that the information matrix depends on the unknown parameter θ, and since f is not linear as it is g, encourage us to approach the problem slightly differently.

When θ takes its true value all the geometric aspects covered by Titterington (1980a, b) can be applied to the nonlinear case. When somebody ends up the experiment with an estimate of θ, this estimate forms the geometry in a "local" sense, as the estimate might be any θ in the neighborhood of θ_t. Thus D(θ) optimality corresponds to the minimal local ellipsoid. This line of thought is discussed extensively:

Let $\lambda_i,\ i = 1, 2, \ldots, p$ be the eigenvalues of $\mathrm{M}(\xi, \vartheta)$ and therefore $1/\lambda_i,\ i = 1, 2, \ldots, p$ are the eigenvalues of $\mathrm{M}^{-1}(\xi, \vartheta)$. The eigenvalues $1/\lambda_i,\ i = 1, 2, \ldots, p$ are proportional to the squares of the lengths of the axes of the confidence ellipsoid (see also Sect. 6.3). For example, in principle, the three- dimensions ellipsoid has the form

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1,$$

The volume of this ellipsoid is $V = \frac{4}{3}\pi abc = \frac{4}{3}\pi\sqrt{\lambda_1\lambda_2\lambda_3}$, were the eigenvalues are equal to: $1/\lambda_1 = 1/a, 1/\lambda_2 = 1/b, 1/\lambda_3 = 1/c$ and define the principal directions of the ellipsoid, see Fig. 3.1 below.

Therefore, due to the eigenvalue properties, the alphabetic optimality criteria can be geometrically interpreted as:

- $D(\vartheta)$-optimality: $\min \prod_{i=1}^{p} \frac{1}{\lambda_i}$.

  Minimizes the generalized variance of the parameter estimates ($\mathrm{asdet}\,\mathrm{M}^{-1} = \prod \frac{1}{\lambda_i}$). Equivalently, the volume of the ellipsoid is minimum.

- $A(\vartheta)$-optimality: $\min \sum_{i=1}^{p} \frac{1}{\lambda_i}$.
  Minimizes the sum of the variance of the parameter estimates (as $\mathrm{trM}^{-1} = \sum \frac{1}{\lambda_i}$).
- $G(\vartheta)$-optimality is equivalent to $D(\vartheta)$.
- $E(\vartheta)$-optimality: $\min\left\{\max \frac{1}{\lambda_i}\right\}$. Minimizes the variance of the least well-estimated contrast, $c^{\mathrm{T}}c = 1$ for the given vector c.

Recall that for given real numbers (scalars) $a_i,\ i = 1, 2, \ldots, n$, with $\sum_{i=1}^{n} a_i = const.$ then $\prod_{i=1}^{n} a_i^{v_i}$ with $v_i,\ i = 1, 2, \ldots, n$ rational numbers, receives its maximum value, only when $a_i = a_j,\ i \neq j$. In such a case with $a_i = \lambda_i$, the above mentioned eigenvalues, $\mathrm{D}(\vartheta)$-optimality is obtained.

The ellipsoid is reduced to a sphere when $\lambda_i = \lambda,\ i = 1, 2, \ldots, n$. So, $\mathrm{D}(\vartheta)$-optimality has no relation with $\mathrm{A}(\vartheta)$-optimality, intuitively. In distribution theory, this is known as elliptically contoured and spherically contoured distributions, with covariance matrix Σ or $\sigma^2 \mathbb{I}_n$, see Fang et al. (1990). The above analysis was based on the confidence regions, while Muller and Kitsos (2004) worked for optimality criteria based on tolerance regions. The geometry of c-optimality is covered by Elfving (1952).

This excellent theorem remains invariant to time, and has been used extensively to the (eventually non-linear problem of) calibration problem, Kitsos and Kolovos (2010). □

Elfving (1952) stated the geometrical characterization of c-optimality as follows:

Consider the model

$$\eta = E(y|u) = f^T(u)\Theta, \quad u \in U.$$

As far as c-optimality is concerned, the optimal design weights of the observations can be obtained through the following geometrical argument:

Draw the convex hull Γ of the space $U_o = f(U)$ and its reflection in the origin, $-U_O$.

Draw the vector c and let T_1 be the point of intersection of c with Γ. Then T_1 lies on the line which joins $A \in U_o$ and $B' \in ({}' - U_O)$ and which forms part of the surface Γ.

Note that B' is the reflection of B, see Fig. 3.2 below.

Then the c-optimal design is located at A and B with weights f, 1-f respectivelywhere $f/(1-f) = \mathrm{T_1B'/AT_1}$ and optimum variance: $[\mathrm{OT/OT_1}]^2$.

With the vector c being the (OT) in Fig. 3.2 and the point $\mathrm{T_1}$ is the intersection of the vector c with the created convex hull, from the induced design space and its reflexion.

The Elfving's theorem, was related to quadratic loss, see Studden (1971), while the geometry of the E-optimality was extensively discussed by Dette and Studden (1993).

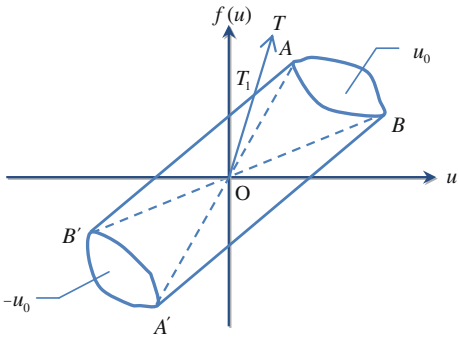

Fig. 3.2 Elfving's theorem

The dependence on θ influences the geometry in the nonlinear case, as the geometric ideas are based on the approximation (asymptotically) of the covariance matrix by $[nM(\theta, \xi)]^{-1}$.

A particular special case of the partially nonlinear models is discussed below.

3.8 Partially Nonlinear Models

The question is, if all the involved parameters, in a nonlinear model are equally important when an optimal experimental design approach is adopted. The answer is that, roughly speaking only the non-linearly involved parameters are essential. Indeed:

Hill (1980) defined a model to be partially nonlinear for the k parameters, k < p, iff

$$\nabla f(u, \theta) = B(\theta) h(u, \theta_k) \tag{3.8.1}$$

where $B(\theta)$ is a matrix not depending on u but just on $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$, θ_k is the vector of the k parameters which appear in a nonlinear way and h is a vector of functions depending on θ_k. When (3.8.1) holds the $D(\theta)$-optimal design depends only on θ_k.

Example 11 Consider the model describing decay (or growth) phenomena with $f(u, \theta) = \theta_1 \cdot \exp(\theta_2 u)$. Then one can evaluate

$$\nabla f(u, \theta) = \begin{pmatrix} 1 & 0 \\ 0 & \theta_1 \end{pmatrix} \begin{pmatrix} \exp(\theta_2 u) \\ u \exp(\theta_2 u) \end{pmatrix}$$

Therefore, according to the above discussion, the $D(\theta)$-optimal design will depend only on θ_2. This means that initial knowledge only for one parameter is needed to define the optimal design points for this particular model. □

Khuri (1984) provided, for the model (3.8.1), a sufficient condition of the following form for $D_s(\theta)$-optimality: Let θ_s be the "linear terms", θ_{p-s} be the "nonlinear terms" and consider $B(\theta)$ partitioned as $B(\theta) = \left[B_1^T(\theta), B_2^T(\theta)\right]$, $B_1 \in Mat(s,p)$, $B_2 \in Mat(p-s,p)$. Suppose the corresponding $M(\theta,\xi)$ is as in (2.5.3). A sufficient condition for a locally $D_s(\theta)$ optimal design (for θ_s) to be dependent only on θ_{p-s} is that matrix $B_2(\theta)$ should be expressible in the form $B_2(\theta)P(\theta)\left[I_{p-s} : K\right]$, where $P(\theta) \in Mat(p-s,p-s)$ and is nonsingular, I_{p-s} is the identity matrix and $K \in Mat(p-s,s)$ does not involve θ.

Example 12 Consider the Michaelis–Menten model with

$$f(u,\theta) = \theta_1 u/(\theta_2 + u).$$

For this model one can evaluate

$$\nabla f(u,\theta) = \left[u(\theta_2+u)^{-1}, \theta_1 u(\theta_2+u)^{-2}\right]^T = \begin{pmatrix} 1 & 0 \\ 0 & \theta_1 \end{pmatrix} \cdot \begin{pmatrix} u(\theta_2+u)^{-1} \\ u(\theta_2+u)^{-2} \end{pmatrix}$$

Thus (3.8.1) holds and therefore the D-optimal design, for the vector $\theta = (\theta_1, \theta_2)$, depends only on θ_2. □

Now a more general case is considered, a model with both a linear part and a nonlinear part. This model is referred as the proper partially nonlinear, and let us suppose to be as follows:

$$f(u,\theta) = 1(u,\beta_1) + m(u,\beta_2) \tag{3.8.2}$$

where $\beta_1 = (\theta_1, \theta_2, \ldots, \theta_s)$, $\beta_2 = \left(\theta_{s+1}, \ldots, \theta_p\right)$ i.e. $\theta = (\beta_1, \beta_2)$ and $1(u,\beta_1) = \theta_0 + \theta_1 u + \cdots + \theta_s u_s$ is a linear function of β_1 and $m(u,\beta_2)$ is any nonlinear function of β_2 The following proposition holds.

Proposition 3.8.1 *For the model* (3.8.4) *the D*(θ)*-optimal design depends on* β_2. *Moreover the* $D_s(\theta)$*-optimal design that estimates* β_1, *also depends on* β_2.

Proof $M = M(\theta,\xi) = M(\beta_2,\xi)$ thus trivially, ∇f does not depend on β_1. Therefore the $D(\theta)$-optimal design depends on β_2. Moreover $M_{22} = M_{22}(\beta_2,\xi)$ and thus the ratio $\det(M)/\det(M_{22})$ is a function of β_2 only. Therefore the $D_s(\theta)$-optimal design depends on β_2. □

Example 13

1. Let $f(u,\theta) = \theta_1 - \exp(-\theta_2 u,) \quad \theta_2 > 0.$
   The D-optimal design depends only on θ_2
2. Let $f(u,\theta) = \theta_1 + \theta_2 u + \sin(\theta_3 u) + \cos(\theta_4 u).$

It is easy to see that M_{22} is a function of (θ_3, θ_4). Therefore $D(\theta)$ and $D_2(\theta)$-optimality depends on $\beta_2 = (\theta_3, \theta_4)$ only. □

3.9 Discussion

Experimental design in the linear case started as an optimum allocation of the observations at the optimum design points, which are eventually roots of a hypergeometric function, Kitsos (2010).

In his pioneering work Fedorov (1972) summarized and extended all the linear work. The main target through this theoretical framework is to obtain the appropriate algorithms to get the optimum design measure for estimating θ. Moreover Fedorov (1972) provided the first algorithm, but it was only in Wu and Wynn (1978) that a general dichotomous convergence theorem was obtained, concerning the convergence of the sequence of design measures.

The theoretical framework in the linear case is completed by the duality theory which first came to light in Lagrangian theory, see Silvey (1972), Sibson (1972), Silvey and Titterington (1973), Pukelsheim and Titterington (1983). The nonlinear theory suffers from the dependence on the parameters, which should be estimated!

Recall (2.6.3), the average information matrix. Thus any function of $M(\theta, \xi)$ has to be based on the knowledge of the parameters that need to be estimated. This θ-dependence also occurs when, the underlying model is a linear regression models and interest lies in a nonlinear function (also known as a nonlinear aspect) of its parameters. An example is the curvature of the second degree linear model, Ford and Silvey (1980), while Fornius (2008) worked on optimality concerning the quadratic logistic.

References

Atkinson AC, Hunter WG (1968) The design of experiments for parameter estimation. Technometrics 10:271–289

Begg CB, Kalish LA (1984) Treatment allocation for nonlinear models in clinical trials: the logistic model. Biometrics 40:409–420

Box MJ (1968) The occurrence of replications in optimal designs of experiments to estimate parameters in non-linear models. J Roy Stat Soc: B 30:290–302

Box GEP, Lucas HL (1959) Design of experiments in nonlinear situation. Biometrika 49:77–90

Breslow NE, Day NE (1980) Statistical methods in cancer research. IARC, No 32. Lyon

Dette H, Haines ML (1994) E-optimal designs for linear and nonlinear models with two parameters. Biometrika 81:739–754

Dette H, Studden WJ (1993) Geometry of E-optimality. Ann Stat 21:416–433

Elfving G (1952) Optimum allocation in linear regression theory. Ann Math Stat 23:255–262

Fang K-T, Kotz S, Ng KW (1990) Symmetric multivariate and related distributions. Chapman and Hall, London

Fedorov VV (1972) Theory of optimal experiments. Academic Press, London

Fisher RA (1922) On the mathematical foundation of theoretical statistics. Phil Trans Roy Soc Lond Ser A 22:309–368

Ford I, Silvey SD (1980) A sequentially constructed design for estimating a nonlinear parameter function. Biometrlka 67:381–388

Ford I, Kitsos CP, Titterington DM (1989) Recent advances in nonlinear experimental design. Technometrics 31:49–60

Ford I, Torsney B, Wu CFJ (1992) The use of canonical form in the construction of locally optimal designs for nonlinear problems. J Roy Stat Soc: B 54:569–583

Fornius FE (2008) Optimal design of experiments for the quadratic logistic model. Ph.D. Thesis, Stockholm University, Stockholm
Hill PDH (1980) D-optimal designs for partially nonlinear regression models. Technometrics 22:275–276
Karlin S, Studden WJ (1966) Optimal experimental designs. Ann Math Stat 37:783–815
Khuri AI (1984) A note on D-optimal designs for partially nonlinear regression models. Technometrics 26:59–61
Kiefer J, Wolfowitz J (1960) The equivalence of two extreme problems. Canadian J Math 12:363–366
Kitsos CP (1986) Design and inference in nonlinear problems. Ph.D. Thesis, University of Glasgow, UK
Kitsos CP (2010) Adopting hypergeometric functions for sequential statistical methods. Bull Greek Math Soc 57:251–264
Kitsos CP (2011a) Invariant canonical form for the multiple logistic regression. Math Eng Sci Aerosp (MESA) 2(3):267–275
Kitsos CP (2011b) On Fisher's information measures. In: e-Proceedings of digital technologies 2011, 10–11 Nov 2011, Zilina, Slovakia, pp 178–183
Kitsos CP, Kolovos KG (2010) An optimal calibration design for pH meters. Instrum Sci Technol 38(6):436–447
Mathew T, Sinha BK (2001) Optimal designs for binary data under logistic regression. J Stat Plan Infer 93:295–307
McCullagh P, Nedler JA (1989) Generalized linear models, 2nd edn. Chapman and Hall, London
Muller CH, Kitsos CP (2004) Optimal design criteria based on tolerance regions. In: Di Bucchianno A, Lauter H, Wynn H (eds) MODA7-advances in model-oriented design and analysis. Physica, Heidelberg, pp 107–115
Pukeisheim F, Titterington DM (1983) General differential and Lagrangian theory for optimal experiment design. Ann Stat 11:1060–1068
Pukelsheim F (1993) Optimal design of experiments. Wiley, New York
Sibson R (1972) Discussion of paper by HP Wynn. J Roy Stat Soc: B 34:181–183
Silvey SD (1972) Discussion of paper by HP Wynn. J Roy Stat Soc: B 34:174–175
Silvey SD (1980) Optimal design. Chapman and Hall, London
Silvey SD, Tittertington DM (1973) A geometric approach to optimal design theory. Biometrika 60:21–32
Sitter RR, Torsney B (1995a) D-optimal deigns for generalized linear models. In: Kitsos CP, Muller WG (eds) MODA4-advances in model-oriented data analysis. Springer, Berlin, pp 87–102
Sitter R, Torsney B (1995b) Optimal design for binary response experiments with two design variables. Stat Sinica 5:405–419
Smith K (1918) On the standard deviation of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. Biometrika 12:1–85
Studden WJ (1971) Elfving's theorem and optimal designs for quadratic loss. Ann Math Stat 42:1613–1621
Titterington DM (1975) Optimal design: some geometrical aspects of D-optimality. Biometrika 62:313–320
Titterington DM (1980a) Aspects of optimal design in dynamic systems. Technometrics 22:287–299
Titterington DM (1980b) Geometric approaches to design of experiment. Math Oper Stat 11:151–163
White LV (1973) An extension to the general equivalence theorem for nonlinear models. Biometrika 60:345–348
Wu CFJ, Wynn HP (1978) The convergence of general step-length algorithms for regular optimum design criteria. Ann Stat 6:1273–1285

Chapter 4
Static Design

Abstract This chapter defines the Static design and various approaches are discussed. A number of applications from Chemical Kinetics are provided.

4.1 Introduction

In the linear case, with a typical example the p-term polynomial regression, the D-optimal design has a tendency to use as optimal design points, the "end" points of the design space, among others, when $p \geq 2$. Under some considerations, (Fedorov 1972. Th. 2.2.3) the design points for D-optimality can be defined as roots of a particular polynomial (Legendre, Jacobi, Laguerre, Hermite). Particular examples, under the regression framework, have been demonstrated in Atkinson and Donev (1992). The design at the p-term polynomial regression allocates measure 1/p at these points. This result helps us to consider that there is a hypergeometric function that passes through the optimal design points, Kitsos (2011).

The situation is different in nonlinear problems. The design points can only be defined under the "true" value of the unknown parameter θ. Therefore a "guess" about θ has to be supplied. The aim is then to gain knowledge about θ with an efficient estimate, $\hat{\theta}$, say, so that the covariance matrix $\hat{C}ov = Cov(\hat{\theta})$ will be approximated by $n^{-1}M^{-1}(\theta_t, \xi)$ with θ_t the true value, see Franceshini and Macchietto (2008) for an applied point of view. This θ-dependence requires the development of two alternative procedures for the construction of experimental designs in practice:

- *Procedure 1* Choose the optimal design points. Conduct the experiment once at these points. This procedure shall be referred as *static design*.
- *Procedure 2* Choose the possible optimal initial design points, based on a guess of the parameter. Conduct the experiment at these points and estimate the parameters. Re-assess the optimal design points, using the estimates of the parameters obtained at the pervious stage. Conduct the experiment at these new

C. P. Kitsos, *Optimal Experimental Design for Non-Linear Models*,
SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-45287-1_4,

points and get new estimates. Continue the procedure until a predefined stopping rule is satisfied. This procedure shall be referred as *sequential design*.

In this chapter the static designs will be tackled. Sequential designs will be discussed in Chap. 5.

4.2 Locally Optimal Approach

In practice the true θ is unknown. Thus a guess for θ might be submitted either from previous experimental work or from theoretical considerations. The local optimal design criteria discussed in Chap. 3, provide a line of thought on how this guess, instead of the true θ, is applied: to evaluate the estimator $\hat{\theta}$ on the basis of this experiment. Therefore static designs can be obtained by using a guess for θ, in the initial design. In Sect. 4.7 the optimal design points for a number of nonlinear regression models from chemical kinetics, are summarized, see Kitsos (1986, 1995). The locally optimum settings of the covariates are also given. Based on these guesses for the parameter, the appropriate optimal design points have been evaluated in a number of papers in the literature, as there is not a global framework for the entire non-linear model, as in linear case.

4.3 Lauter's Approach

An attempt to avoid θ-dependence has been discussed through S-optimality, "averaging" over all possible values of the parameter, Lauter (1974).

Definition 4.2.1 The design measure ξ^* is called S-optimal iff

$$\mathrm{S}(\xi*) = \max \mathrm{S}(\xi), \xi \in \Xi \quad \text{where } \mathrm{S}(\xi) = \int \ln|M(\theta, \xi)| v(d\theta) \tag{4.3.1}$$

and v is a given measure defined on some σ-algebra of R^{p}. □

For the introduced S-optimality she proved the following equivalence theorem:

Theorem 4.3.1 *Let* $\mathrm{d}_1(\theta, \xi, \mathrm{u}) = (\nabla \mathrm{f})^T \mathrm{M}^{-1}(\theta, \xi)(\nabla \mathrm{f}), \Xi_0 = \{\xi : \int d_1(\theta, \xi, u) v(d\theta) \langle \infty\}$

Then for $\Xi_0 \neq \varnothing$ *and* $|S(\xi)| \langle \alpha$ *for every* $\xi \in \Xi_{\mathrm{o}}$ *the following three conditions are equivalent*:

1. $\xi \in \Xi_0$ *is S-optimal*.
2. $\xi *$ minimizes $\max\{\int d_1(\theta, \xi, u) v(d\theta), \ \mathrm{u} \in \mathrm{U}\} = \Gamma$
3. $\Gamma = \mathrm{p} \int v(d\theta)$. □

To avoid θ-dependence, a prior distribution can be assumed for θ and then proceed using an average information matrix, independent of θ, of the form $\mathrm{M}(\xi) = \mathrm{E}_{\theta}[\mathrm{M}(\theta, \xi)]$. Moreover, for any weighting function w(.) on the parameter space Θ, which may or may not be a formal prior density. Indeed one can use the fact that

$$\mathrm{M}(\xi) = \int \mathrm{M}(\theta, \xi)\mathrm{w}(\mathrm{d}\theta) \tag{4.3.2}$$

and construct a new criterion, with φ as in (3.3.1)

$$\Phi_{\mathrm{w}}(\xi) = \int \varphi[\mathrm{M}(\theta, \xi)]\mathrm{w}(\mathrm{d}\theta) \tag{4.3.3}$$

Another approach based on discrimination was proposed by Stone and Morris (1985).

4.4 Stone and Morris Approach

Stone and Morris (1985) proposed two alternative criteria for the static design problem. One of these criteria is based on the log-likelihood and the other one on the sum of squares. Both require extra knowledge of two values θ', θ'' of the parameter of interest θ. They also include the possibility of a nuisance parameter δ. The first criterion function, which must be maximized, is

$$\mathrm{C_L} = \mathrm{E}(\mathrm{LR}|\,\theta', \delta) - \mathrm{E}\big(\mathrm{LR}|\,\theta'', \delta\big), \quad \text{with LR} = 1\mathrm{og}[\mathrm{p}(\mathrm{y}|\,\theta', \delta') \,/\, \mathrm{p}(\mathrm{y}|\theta'', \delta'')] \tag{4.4.1}$$

were LR is the logarithm of the likelihood ratio for θ' and θ'', with δ evaluated at δ' and δ'', the conditional maximum likelihood estimates for δ. The "design for discrimination" character of this criterion is obvious. Moreover the assumption of a common δ in (4.4.1), which must be pre-specified, must reduce the practical utility of this criterion. Their second criterion also requires pre-specification of two θ's, θ', θ''. This criterion is

$$\mathrm{C_S} = \inf\left\{\sum [n_i(\theta', \delta') - n_i(\theta'', \delta'')]^2 \delta', \delta'' \in \right\} \tag{4.4.2}$$

where $\mathrm{n_i}$ (θ, δ) denotes the expectation of the i-th observation and A is a pre-specified set. For this criterion, if there is no nuisance parameter, the result may be a singular design from which θ will be inestimable. Note that the specification of the set A may present practical difficulties as well.

The fact that both $\mathrm{C_L}$ and $\mathrm{C_S}$ are based on two specified values for θ, which the experimenter eventually aim to estimate, makes these discrimination criteria rather weak, as far a inference is concerned. Both the above criteria accepted a criticism by Ford et al. (1989).

4.5 Maxi–Min Criterion

Another alternative method of avoiding the θ-dependence problem is the maxi-min design approach. That is the solution of

$$\max \min \left[h_{\Phi}(M[\theta, \xi])\right], \xi \in \Xi, \theta \in \Theta \tag{4.5.1}$$

is required, where by $h_{\Phi}(.)$ is a function h of the criterion Φ. The maxi-min design from (4.5.1) will provide that design, whose minimum value of $h_{\Phi}(.)$, is greater than that of any other. Even if the locally optimum criterion Φ is invariant under transformations of the parameter space, it is not necessary that the maximin criterion can be invariant. The locally optimal values of Φ, the criterion function, may vary considerably with θ, indicating that some of the θ values may dominate the construction of the good design. A function h_{Φ}, which is of great use, in this respect, is the efficiency measure defined as

$$h_{\Phi}(M[\theta, \xi]) := \text{Eff}(\theta, \xi) = \Phi[M(\theta, \xi)] \Delta \Phi[M(\theta, \xi^*)] \tag{4.5.2}$$

with $\Delta = / \text{ or } -$ and $\xi* = \xi * (\theta)$ the locally optimum design for θ. □

Example 14 The maxi-min efficiency criterion for the calibration problem.

Consider the simple regression model with

$$n = E(y|u) = \theta_0 + \theta_1 u_1 \, u_1 \in U = [-1 \, , \, 1]$$

where U is the design space. The target is to estimate the value of $u_1 = u_0$ given $n = \text{H}$ constant value i.e.

$$u_0 = (H - \theta_0)/\theta_1$$

For the one-stage design the use of the criterion function Φ, (recall Sect. 3.3) either as D-optimality for (θ_0, θ_1) or c-optimality for estimating u_0. The D-optimal design is of interest because it will investigate the effectiveness of the D-optimal design as measured by the c-optimality criterion. For the calibration design problem the maxi-min efficiency design turns out to be the D-optimal design, Kitsos (1986). □

4.6 Constant Information Designs

A constant information design is the one where the information of the associated with a nonlinear design, $M(\theta, \xi)$, is at least approximately independent of θ.

Fisher (1922) came across this property in the dilution series experiment, which will be discussed extensively, see Sect. 7.2. Abdelbasit and Plackett (1981, 1983) discuss and extend Fisher's work, and they propose constant information designs because "the asymptotic dispersion matrix of the estimators is then the same, whatever the values of the parameters". There is no doubt that this is an interesting property, although is mainly applied in Quality Control problems. One can keep

his conservations, that this might be considered as the only goal for a design of experiment, see Ford et al. (1989). Typical alternative approach is the Sequential Design, were the variance varies.

A design with a constant information structure does not remain invariant under nonlinear transformations. Moreover under the maxi-min criteria, a number of designs which are at least as good as equal information designs can be obtained.

Most of these methods have not been applied to any real application. Their statistical background is not that widely applied, as the appropriate software has not been developed yet. The appropriate alternative to the static designs is the sequential way of designing, which is extensively discussed in the next chapter.

4.7 Applications from Chemical Kinetics

In this section some of the nonlinear models from the Chemical Kinetics are presented. The underlying mathematical model, the appropriate initial conditions for those parameters that influence the design points (recall Sect. 3.8) and the optimal design points for the input variables, were the experimenter needs to know in prior, in order to conduct the experiment are provided.

Example 15 The Mitscherlish equation of diminishing returns:

$$\eta = \theta_1 + \theta_2 \exp(\theta_3 u), U = [\kappa, \lambda] \subseteq R, \Theta \subseteq R^- x R^+$$

where

η	is the expected amount of growth
θ_1	Hypothetical growth from an infinite amount of fertilizer
$\theta_1 + \theta_2$	Rate at which additional increment of fertilizer decreases
u	Amount of added fertilizer

Initial conditions for: only for the nonlinear parameters i.e. $\theta_0 = (\theta_1, \theta_2, \theta_{03})$

D-Optimal design points are: $\kappa, -\frac{1}{\theta_{03}} + \frac{\kappa e^{\theta_{03}} - \lambda e^{\theta_{03}\lambda}}{\exp(\theta_{03}\kappa) - \exp(\theta_{03}\lambda)}, \lambda.$ □

Example 16 The growth (or decay) law:

$$\eta = \theta_1 \exp(\theta_2 u), with\ (\theta_2 \rangle 0)\ or\ (\theta_2 \langle 0), U = [\kappa, \lambda] \subseteq R, \Theta \subseteq R^+ x R$$

where

η	is the amount of substance growth (or decay)

Initial conditions for: only for the nonlinear parameter $\theta_0 = (\theta_1, \theta_{02})$

D-Optimal design points are:

(1) For $\theta_{02} > 0$ are: $\lambda - \frac{1}{\theta_{02}}, \lambda$

(2) For $\theta_{02} < 0$ are: $\kappa, \kappa - \frac{1}{\theta_{02}}$. □

Example 17 Irreversible reaction B → C:

$$\eta = \exp\{-\theta_1 t_1 \exp[-\theta_2(1/T - 1/T_O)]\},$$
$$u = (t_1, T) \in U \subseteq R^+ x \Delta,\ \Delta = [380, 450].\ \Theta \subseteq R^2$$

where

η [B]
θ_1 the rate at specific temperature T_0
θ_2 E/R being the proportion of the activation energy

Initial conditions for: $T_0 = 400$, and $\theta_0 = (1.0, 16{,}000)$

D-Optimal design points are: Notice that there are two input variables: t and T, i.e. (t, T), and the optimal design points are
(0.419, 420), (8.209, 380). □

Example 18 Chemical reaction of the form A → B → C:

$$\eta = [\theta_1/(\theta_1 - \theta_2)]\{\exp(-\theta_2 t) - \exp(-\theta_1 t)\},\ u = t \in U = R^+.\ \Theta \subseteq R^2$$

where

η the amount of B present after time t, expressed as a function of the total material present when initially (t = 0) only material A is present (n = [B])
θ_1 Rate constant A → B,
θ_2 Rate constant B → C

Initial conditions for: $\theta_0 = (0.7, 0.2)$. Then the D-Optimal design points are:

(1) For a D-optimal design for estimating both parameters design conduct the experiment at 6.858 and at 1.229.
(2) For a D1-optimal design for estimating the parameter θ_1 conduct the experiment at 1.17 and at 7.74.
(3) For a D1-optimal design for estimating the parameter θ_2 conduct the experiment at 3.52 and at 5.64. □

Example 19 Chemical reaction of the form R → P_1 + P

$$\eta = (\theta_1 \theta_3 P_1)/(1 + \theta_1 P_1 + \theta_2 P_2),\ u = (P_1, P_2) \in \Delta x \Delta.,\ \Delta = [0, 3],$$
$$\textit{and } \Theta \subseteq R^3$$

(The catalytic dehydration of Hexyl Alcohol Reaction).
where

η Speed of chemical reaction
P_1 Partial pressure of product P1,
P_2 Partial pressure of product P
θ_1 The absorption equilibrium constant for the reactant R
θ_2 The absorption equilibrium for the product P1
θ_3 The effective reaction rate constant

Initial conditions for the parameters: $\theta_0 = (2.9, 12.2, 0.69)$

The D-Optimal design points are the following:

(1) To estimate the vector θ, then perform the design at the pairs of points with values: (0.2, 0.0), (3.0, 0.0), (3.0, 1.0)
(2) To estimate two components of the vector θ, let us say (θ_1, θ_2) when $\theta_3 = 0.69$ then perform the design at the pairs of points with values: (0.345, 0.0), (3.0, 0.795). □

Example 20 BET (Brunauer-Ermet-Teller) equation:

$$\eta = (\theta_2\theta_1 P_1)/[(1 - P_1)(1 + (\theta_2 - 1)P_1)], u = P_1 \in \Delta \subseteq R. \Delta = [0.05, 030], \Theta \subseteq R^2$$

where

η is the volume of gas absorbed on the solid P1
P_1 P/P0 relative pressure
θ_1 the monolayer capacity,
θ_2 constant characteristic of the gas–solid

When initial conditions are provided, only for the parameter $\theta_2 = 2$ then:

The D-Optimal design points are:
at 0.13 and at 0.30. □

There are certainly more models from chemical kinetics, see Kitsos and Kolovos (2013), Munson-McGee et al. (2011) among others, but these examples clarify how the experimentalist can gain time and money when the optimal design theory is adopted. Kitsos and Kolovos (2013) provided an extensive analysis on the chemical structure of a compilation of models, as well as the appropriate D-optimal design for the chemical kinetic model under consideration. It is clear there that most of the chemical kinetic models are with one or two variables and very rare experimenter faces the problem of more than four variables.

In the following chapter the sequential approach of design is introduced.

References

Abdelbasit KM, Plackett RL (1981) Experimental design for categorized data. Int Stat Rev 49:111–126
Abdelbasit KM, Plackett RL (1983) Experimental design for binary data. JASA 78:90–98
Atkinson AC, Donev AN (1992) Optimum experimental designs. Clarendon Press, Oxford
Fedorov VV (1972) Theory of optimal experiments. Academic Press, Waltham
Fisher RA (1922) On the mathematical foundation of theoretical statistics. Phil Trans Roy Soc London Ser A 22:309–368
Ford I, Kitsos CP, Titterington DM (1989) Recent advances in nonlinear experimental design. Technometrics 31:49–60
Franceshini G, Macchietto S (2008) Mpdel-based design of experiments for parameter precision: State of the art. Chem Eng Sci 63:4846–4872

Lauter E (1974) A method of designing experiments for nonlinear models. Math Operationsforsch U Stat 5:697–708

Kitsos CP (2010) Adopting hypergeometric functions for sequential statistical methods. Bull Greek Math Soc 57:251–264

Kitsos CP, Kolovos KG (2013) A compilation of the D-optimal designs in chemical kinetics. Chem Engin Comm 200:1–20

Kitsos CP (1986) Design and inference in nonlinear problems. Ph.D. thesis, University of Glasgow, UK

Kitsos CP (1995) On the support points of D-optimal nonlinear experimental designs for chemical kinetics. In: Kitsos CP, Muller WG (eds) MODA4—Advances in model-oriented data analysis, pp 71–76

Munson-McGee SH, Mannarswammy A, Andersen PK (2011) D-optimal design for sorption kinetic experiments: Cylinders. J Food Eng 104:201–207

Stone M, Morris JA (1985) Explanatory studies in non-sequential non-local non-linear design. J Stat Plann Infer 12:1–9

Chapter 5
Sequential Designs

Abstract The Sequential Principle of design is adopted as the point estimation method and the theoretical insight is discussed, in a compact way. The Stochastic Approximation iterative scheme is discussed as a particular case.

5.1 Introduction

The objective is to construct a design that eventually estimates, the unknown parameter vector θ, adequately. Adopting the sequential procedure an initial design has to be chosen using prior knowledge of θ to get an estimate of the parameters. This estimate is useful as an initial guess to redesign, re-estimate and so on. Important questions are:

- How is the initial design chosen?
- What measures of optimality are used?
- How is the design revised or continued?
- How inference will be assessed?

This Chapter attempts to answer these questions. One can proceed by either designing in batches of observations or adding a single observation at a time into the design: this is known as fully sequential design. This definition can be revised by adding not a single observation per stage, but a number of observations equal with the involved number of parameters, see Kitsos (1989). In such a case it can be proved that if the initial design is D-optimal, then the limiting design is also a D-optimal one.

C. P. Kitsos, *Optimal Experimental Design for Non-Linear Models*, SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-45287-1_5,

5.2 Background

Let us assume that the initial design has been constructed and an estimate $\hat{\theta}$ has been obtained. When a new design point is added (in terms of a batch of observations or a single observation) a new Fisher information matrix is obtained and a new estimator is evaluated—either a LSE or a MLE. Thus a sequence of least squares estimates $\hat{\theta}_{\mathrm{n}}$ is obtained. The establishment of a strong consistency for the sequence of estimators (that might be vectors), provided that the parameter space is compact, i.e. $\hat{\theta}_{\mathrm{n}} \xrightarrow{as} \theta$, as $\mathrm{n} \to \infty$ is an essential request.

The sequence of average information matrices obtained in this way is also a strongly consistent sequence i.e. as $\mathrm{n} \to \infty$ relations (5.2.1) and (5.2.2) hold as below

$$M\left(\hat{\theta}_n, \xi\right) \xrightarrow{as} M(\theta, \xi) \tag{5.2.1}$$

$$\sqrt{n}(\hat{\theta}_n - \theta \xrightarrow{L} \mathrm{N}\left(0, \sigma^2 \mathrm{M}^{-1}(\theta, \xi)\right) \tag{5.2.2}$$

where L means convergence in distribution, to a multivariate normal distribution as in (5.2.2), see for details Wu (1981). For the various definitions for the statistical convergence of a sequence see Schervish (1995) among others.

It is suggested that the initial design should be built up at the optimum points of the corresponding locally optimal design, on the basis of an initial guess for θ. The question is "how the next design point is chosen?" The answer is: choose as the next design point the one that minimizes the estimator's generalized variance. That is a D(θ)-optimality criterion is used for choosing the next design point. This defines the following algorithm:

1. Define initial values θ_0 for the parameter vector θ and perform the experiment.
2. From the initial design obtain an estimate, $\hat{\theta}_1$ of θ. Then
3. Choose as the next design point $\mathrm{u}_{\mathrm{n+1}}$, $\mathrm{n} = 1, 2\ldots$ the one which minimizes

$$d(\hat{\theta}_n \cdot \xi_n \cdot u_{n+1}) = \mathrm{n}^{-1}\left[\nabla f\left(\hat{\theta}_n, u_{n+1}\right)\right]^{\mathrm{T}} \mathrm{M}^{-1}\left(\hat{\theta}_n, \xi_n\right)\left[\nabla f\left(\hat{\theta}_n, u_{n+1}\right)\right]$$

4. Perform the experiment at $\mathrm{u}_{\mathrm{n+1}}$ and get $\hat{\theta}_{n+1}$
5. Repeat Steps 3 and 4 the required number of times.

Under very strong assumptions it can be proved that, as $\mathrm{n} \to \infty$.

$$\det \mathrm{M}(\theta, \xi_{\mathrm{n}}) \xrightarrow{a \cdot s} \det \mathrm{M}(\theta, \xi^*) \tag{5.2.3}$$

where ξ^* is the optimal measure. This limiting result holds for the particular case of D(θ)-optimality, when the initial design is D-optimal the limiting is also, providing the stochastic approximation approach has been adopted, see Kitsos (1989) for details, and Sect. 5.5. The study of the involved sequences, in a sequential design, is a difficult task, as there are involved at the same time the sequence of parameters, design points and design measures; see Chaudhuri and Mykland (1995).

5.3 Extensions

Establishment of the convergence of the sequence $\mathrm{M}(\hat{\theta}_n, \xi_n)$ to $\mathrm{M}(\theta, \xi^*)$ under some criterion function exhibits many technical difficulties. One main virtue of the linear theory is the dichotomous convergence theorem of Wu and Wynn (1978) for any criterion function and for the sequence of information matrices $\mathrm{M}(\xi_n)$—notice that there is no parameter dependence in linear case.

The idea of the directional derivative is essential in the optimal design theory. It is introduced briefly below:

Let g be a function from R^n to $(-\infty, \infty)$ and let $x = (x_1, x_2, \ldots, x_n)$. Then the (Gateaux) directional derivative is defined at x in the direction of y to be:

$$G(x, y) = \lim_{q \to 0^+} q^{-1}\{g(x - qy) - g(x)\}$$

If g is differentiable then it can be proved that: $G(x, y) = \sum y_i(\partial g(x)/\partial x_i)$

The Frechet direction derivative is defined as: $F(x, y) = G(x, y - x)$

Moreover, for a concave function g and x a point where g is finite then $G(x, y)$ exists for all y; this is valid whether or not g is differentiable at x.

Notice that the entries x, y might be matrices, as it happens in the optimality theory. In the design contest, with the criterion ϕ the directional derivative Φ is corresponded to the above mentioned notation, whereas $g = \phi$, $G = \phi$ and x, y can be matrices. Notice that the entries x, y can be matrices, such as the average per observation information matrix M. □

Based on the above discussion, Titterington (1980a), for any criterion function φ and its corresponding directional derivative Φ, it is suggested to choose as the next design point u_{n+1}, that one which minimizes the quantity

$$d(\hat{\theta}_n \cdot \xi_n \cdot u_{n+1}) = \Phi[\mathrm{M}(\hat{\theta}_n, \xi_n),\ \mathrm{I}(\hat{\theta}_n, u_{n+1})] \tag{5.3.1}$$

There are two difficult features concerning the sequential design.

- Firstly, as any design point comes into the design on the basis of a previous estimate of the parameter, the design points are not statistically independent. Thus the "information" matrix, as defined earlier, is not Fisher's information matrix, in the sense, that it does not necessarily provide an approximate covariance matrix.

- Secondly, at the s-th stage, say, the estimator θ_S has to be evaluated. This can be done through the Newton–Raphson method, which might diverge if a poor initial guess was considered.

These two problems are related to inference about θ and will be faced in practice in Chap. 7. Thus a number of simulations are performed, explaining the existent theoretical problems. Moreover Ford et al. (1985) discussed the procedures for obtaining valid inferences when the sequential nature of design is adopted. Simply the results suggested are: work sequentially for the design and perform the calculations for the average per observation information matrix as if it is not a sequential design.

As far as the second point concerns, the Newton–Raphson iteration scheme is the numerical method which supplies the estimate at stage s, say, through the iteration

$$\hat{\theta}_{S,k+1} = \hat{\theta}_{S,k} - \mathrm{S}^{-1}\mathrm{q_s} \quad \mathrm{k} = 1, 2\ldots \tag{5.3.2}$$

where $\hat{\theta}_{S,k}$ is an estimate of the k-th iteration at the s-stage and S is the appropriately-evaluated Hessian of the log-likelihood, which has to present a non-zero determinant, and $\mathrm{q_S}$ the vector of first partial derivatives, recall Example 6. For a more thorough discussion on the Newton–Raphson method in nonlinear problem see any book on Numerical Analysis. A statistical version of the Newton–Raphson method, known as Stochastic Approximation method (or the Robbins–Monro scheme) will be discussed in Sect. 5.5.

The sequential idea of designing has also been faced from a Bayesian point of view, Hu (1998). The criterion for parameter estimation is to choose that value θ_m of θ, for which the posterior density is a maximum. Moreover, in principle for large sample situations the posterior distribution for $\theta, \pi(\theta|y_n, u_n)$, given a prior distribution should be approximately normal with

$$\pi(\theta|y_n, u_n) \cong N\left(\theta_m, \left(B + S(\theta_m, \xi_n, y_n)^{-1}\right)\right) \tag{5.3.3}$$

where the matrix B reflects the prior information, see Ford et al. (1989), for a compact review. Relation (5.3.2) is essential in practice: the method converges when the initial value is in the neighborhood of the limiting value. Otherwise it diverges, and this creates problems. The solution is usually to apply a truncation of the design space, and the iteration to try to "bring the estimators within the truncated space", see also Chap. 7. Moreover it is recommended that with 2–3 iterations adopting the Bisection method, a "suitable initial guess", can be obtained see Simulation studies I and II in Chap. 7.

5.4 Binary Response Problems

Experiments with dichotomous outcomes can be faced in a variety of practical situations. In these cases the "response" and the "non-response" outcome can be presented in different ways. The existence of the parameters is essential, and must be taken into consideration Sillvapule's Theorem, see Appendix 2. Some practical situations concerning the binary response problem are as follows:

- *In testing explosives*: Usually a weight is dropped on the explosive mixture from a certain height. The dichotomous variable takes value "explode" or "not explode".
- *In entomological assays*: A critical dose level can be associated with the insect of interest. The response is "killed" or "not killed".
- *In fatigue experiments*: The strength of a certain material is tested. This response is "strong" or "not strong".
- *In educational studies*: The tutor might have questions of the form: "right" or "wrong".
- *In life testing*: Experiment on the life cycle of a photographic film or safety equipment such as fire extinguishers.
- *In Toxicological studies*: Experiments are performed as far as Cancer studies concern on the binary response: "tumor", or "not tumor".

In this family of problems the main interest is usually devoted to the estimation of a percentile L_p of the response curve. Usually this percentile is the median $L_{0.5}$, while in Cancer problems the "low dose extrapolation problem", can be considered as a percentile problem, see Kitsos (1999). The commonly used sequential methods are the following:

– Spearman–Karber's method
– Up-and-Down method (UD)
– Stochastic Approximation (SA).

The Spearman–Karber estimator was used in the early 1940s. Therefore it is not described here. The Up-and-Down sequential scheme, mainly used in industrial problems, can be described by the model:

$$u_{n+1} = \begin{cases} u_n - \delta & if \quad Y_n = 1 \\ u_n + \delta & if \quad Y_n = 0 \end{cases} \tag{5.4.1}$$

It was first applied on testing explosives, but it can be used to dose response problems. The choice of the "step size" δ is a problem. One suggestion is that it should be a rough estimate of σ, while the method has been discussed as an example of a Markov process: from the definition of the method each run depends only on the current situation. Eventually it was proved that the method is a Stochastic Approximation method, see Sect. 5.6.

As far as applications are concerned in binary response problems Wu (1985) suggested a local approximation, to any unknown response, by the logistic, when the quartile L_p with $p \in [0.1, 0.9]$ is to be estimated and he obtained fully efficient estimates using the Stochastic Approximation scheme, which will be discussed in Sect. 5.5.

5.5 Stochastic Approximation

Let us consider the problem of finding the root of the equation $Q(x) = q$, with $Q(.)$ a known function and q a given real number. This problem is based on non random variables, and therefore it is non-stochastic. That is a "non-stochastic approximation" is assumed when solving $Q(x) = q$ as above. But there are real life cases, originated to Pharmacology, see the early work of Guttman and Guttman (1959), were random variables are involved where $Q(x) = q$. Notice that if $Q(,)$ is a cumulative distribution function and $0 < q < 1$, then the solution corresponds to the percentile point and thus a *Stochastic Approximation* is considered.

The Stochastic Approximation (SA) method can be applied to an experiment that is fully sequential, i.e. when knowledge about θ is built by adding one experimental unit at each stage of the experiment. Roughly speaking this is a stochastic version of the Newton–Raphson (NR) iteration introduced by Robbins and Monro (1951) in their pioneer work on SA. The method is an elegant one, as far as the theoretical development is concerned. Avoiding the technicalities, and the strong mathematical background needed for the full development of the method, the practical aspects and essential applications of the method are discussed in Chap. 7.

A brief review of the theoretical insight of the method is presented next, again avoiding most of the technicalities.

The SA method tries to evaluate the root, say θ, of the following equation,

$$E\{Y(u)\} := T(u) = p \quad p \in R. \tag{5.5.1}$$

where θ is unique and the function T, as well as p are provided. Recall Sect. 2.3 for the notation, and E mean Expected value, as usually. Notice the same notation θ for the root of the Eq. (5.5.1), as it is the involved parameter to be estimated. The essential conditions needed are:

(C1) $\inf|T(u)—p| \geq \delta > 0$
(C2) $T(u)$ non-decreasing and $T'(\theta) = b > 0$.

Then it can be proved that a sequence α_n, $n = 1, 2,\ldots$ of real numbers exists with

$$\alpha_n > 0, \ \sum \alpha_n = \infty, \ \sum \alpha_n^2 \langle \infty \tag{5.5.2a}$$

in such a way that the sequence of stimuli (for the n-th observation u_n going with the n-th response y_n) the next observation will be at

$$u_{n+1} = u_n - \alpha_n(y_n - p) \quad n = 2,\ 3, \ldots,\ u_1 \text{ arbitraire} \tag{5.5.2b}$$

and converges to θ in mean square, i.e. as $n \to \infty$

$$\lim E(u_n - \theta)^2 = 0. \tag{5.5.3}$$

Recall that (C2) is actually Assumption 4 mentioned in Sect. 2.3. The physical meaning of the sequence α_n can be thought as the "weight" associated with trial n. A typical α_n might be $\alpha_n = n^{-1}$ or more generally any sequence satisfying the relation

$$c'/n \le \alpha_n \le c''/n. \quad c', c'' \text{ constants}$$

Considering a more general form from the sequence

$$\alpha_n = c/(nb), \quad b = T'(\theta) = T'(u)|_{u=\theta} \tag{5.5.4}$$

and assuming that

$$(A)\ Yi = a + b(u_i - \theta) + e_i, \quad b = T'(\theta) \tag{5.5.5}$$

with e_i under Assumption 1 of Sect. 2.3, then it can be proved that, as $n \to \infty$,

$$\lim E(u_n - \theta)^2 = 0, \quad \lim \mathrm{Var}(u_n - \theta) = \sigma^2 c^2/\left[nb^2(2c - 1)\right], \quad \text{with} \quad c > 0.5 \tag{5.5.6}$$

In Statistics, in principle, when sequences are examined, their "normality in limit" is a very desirable property. Indeed under a number of assumptions the sequence described turns in limit to a Normal distribution. Namely: with any α_n of the form $\alpha_n = n^{-(1-\varepsilon)}$, $\varepsilon < 1/2$, and one more (!) assumption

$$(B)\ E\{Y(u) - T(u)\}^2 = \sigma^2 > 0, \quad \text{for every u.}$$

it can be proved that, as $n \to \infty$

$$n^{(1-\varepsilon)/2}(u_n - \theta) \xrightarrow{L} N(0, \sigma^2/2b). \tag{5.5.7}$$

Moreover for the "simple" sequence, with C being a constant,

$$\alpha_n = C/n, \quad n = 1, 2 \ldots \tag{5.5.8a}$$

with $C > 1/2$ K where $K \le \inf[(T(u) - p)/(u - \theta)]$, then, as $n \to \infty$,

$$\sqrt{n}(u_n - \theta) \xrightarrow{L} N(0, \sigma^2 C^2/(2bC - 1)), \quad bC > 1/2, \tag{5.5.8b}$$

From (5.5.8b) it is obvious that the asymptotic variance is minimized with an optimal choice of C, C_{opt} say, when

$$C_{opt} = b^{-1} = [T'(\theta)]^{-1} \tag{5.5.8c}$$

Recall that p is given but the c.d.f T will usually not be known, which happens in real life problems. So the common problem of nonlinear situations is faced. Without the knowledge of some quantities (here T, θ) the optimal procedure (here C_{opt} and consequently u_{n+1}) can not be obtained. Thus the problem of creating the sequence (5.5.5) contains the intrinsic problem of creating an approximation for C_{opt}.

Approximations need iterations, iterations are sequences, so a "daughter" sequence of u_n is needed, β_n say, which might converge, hopefully, to b. Simultaneously the "parent" sequence, will converge to θ. The main virtue of SA is that SA is a kind of regression of Y on u, and at the same time, C_{opt} is a kind of slope for the unknown T(θ). Thus at stage n + 1 one can have as "C" the slope coefficient of the regression line formed from the data u_i and $Y(u_i)$ i = 1, 2,… namely

$$C \cong \hat{\beta}_n = \sum (u_i - \bar{u}_n) Y_i / \sum (u_i - \bar{u}_n)^2, \quad n \geq 2, \ \bar{u}_n = n^{-1} \sum u_i \tag{5.5.9}$$

Under a number of strict assumptions, described in Kitsos (1986), a number of limiting results have been proved, the main one due to Lai and Robbins (1981) as n → ∞

$$(i)\sqrt{n}(u_n - \theta) \xrightarrow{L} N(0, \sigma^2/b^2) \quad (ii)\lim u_n = \theta \tag{5.5.10}$$

Moreover it holds that

$$\lim \{\sum (u_i - \theta)^2 / \log n\} = \sigma^2/b^2. \tag{5.5.11}$$

The quantity $\Sigma(u_i - \theta)^2$ has been named the cost of the experiment. □

Example 21 Let $l_n(\theta)$ be the log-likelihood of the n observations for a model with c.d.f $p(y_i|u_i, \theta)$. Then for the n + 1 observation the log-likelihood will be l_{n+1} and equals

$$l_{n+1}(\theta) = \sum \log p(y_i|u_i, \theta) = l_n(\theta) + p(y_{n+1}|u_{n+1}, \theta). \tag{5.5.12}$$

Let $\hat{\theta}_n$, $\hat{\theta}_{n+1}$ be the MLE obtained from $l_{n+1}(\theta)$ and $l_n(\theta)$ respectively. Taking the derivatives of the two sides of (5.5.12) one can see that

$$\frac{\partial l_{n+1}(\theta)}{\partial \theta} = \frac{\partial l_n(\theta)}{\partial \theta} + \frac{\partial \log p(y_{n+1}|u_{n+1}, \theta)}{\partial \theta}$$

From the definition of MLE

$$
\begin{aligned}
0 &= \frac{\partial \ln(\theta)}{\partial \theta}\Big|_{\hat{\theta}_{n+1}} + \frac{\partial \log p(y_{n+1}|u_{n+1},\theta)}{\partial \theta}\Big|_{\hat{\theta}_{n+1}} \\
&= \frac{\partial l_{n+1}(\hat{\theta}_n + \hat{\theta}_{n+1} - \hat{\theta}_n)}{\partial \theta} + S_C(y_{n+1}|u_{n+1},\hat{\theta}_{n+1}) \qquad (5.5.13) \\
&\cong 0 + (\hat{\theta}_{n+1} - \hat{\theta}_n)\frac{\partial^2 \ln(\hat{\theta}_n)}{\partial \theta} + S_C(y_{n+1}|u_{n+1},\hat{\theta}_n)
\end{aligned}
$$

based on a first order Taylor expansion about θ_n, where by $S_C(.|.)$ the score function for a single observation is denoted. From (5.5.13) the following appropriate recursion is obtained

$$
\hat{\theta}_{n+1} = \hat{\theta}_n - \frac{1}{\partial^2 \ln(\hat{\theta}_n)/\partial\theta} \quad S_C(y_{n+1}|u_{n+1},\hat{\theta}_n)
$$

Approximating the Hessian of the above relation by Fisher's information

$$
\hat{\theta}_{n+1} = \hat{\theta}_n + I^{-1}(\hat{\theta}_n)\, S_C(y_{n+1}|u_{n+1},\hat{\theta}_n) \quad n = 0,1,2\ldots \qquad (5.5.14)
$$

Iterative scheme (5.5.14) is a typical one: the new estimate is based on the old one, adjusted by the information "in hands already". □

Example 22 Consider the nonlinear regression model $y = \exp(-\theta u) + e$.

Construct a SA scheme converging to the root of the equation $\partial l/\partial\theta = 0$, i.e. to the MLE of the parameter θ. For the error term e it is assumed that follows Assumption 2. The notation of Example 13 is used. Namely:

$$
\begin{aligned}
l_n(\theta) &= \text{const.} - (1/(2\sigma^2))\Sigma(y_i - \exp(-\theta u_i))^2, \partial \ln/\partial\theta \\
&= -\sigma^{-2}\Sigma u_i \exp(-\theta u_i)\,(y_i - \exp(-\theta u_i))
\end{aligned}
$$

$\partial^2 \ln/\partial\theta^2 = -\sigma^{-2}\Sigma[(-u_i^2\exp(-\theta u_i))\,(y_i - \exp(-\theta u_i)) + u_i^2\exp(-2\theta u_i)]$ and therefore Fisher's information is evaluated as

$$
I(\theta) = \sigma^{-2}\Sigma u_i^2 \exp(-2\theta u_i)
$$

Applying the recursion formula (5.5.14) it is

$$
\hat{\theta}_{n+1} = \hat{\theta}_n - I^{-1}(\hat{\theta}_n)[u_{n+1}\exp(-\hat{\theta}_n u_{n+1})\,(y_{n+1} - \exp(-\hat{\theta}_n u_{n+1}))] \qquad (5.5.15)
$$

The information $u_i^2\ \exp(-2\theta u_i)$ is asked to be minimized in each stage, as an optimum design rule. Therefore taking the logarithm of the information and evaluating the root of the derivative, it can be shown that, the optimum design rule occurs when

$$
U_{i+1} = 1/\hat{\theta}_i \quad i = 0,1\ldots \qquad (5.5.16)
$$

Substituting (5.5.16) in (5.5.15)

$$\hat{\theta}_{n+1} = \hat{\theta}_n - [(1/\hat{\theta}_n)\, \mathrm{e}^{-1}(\mathrm{y}_{\mathrm{i}+1} - \mathrm{e}^{-1})]/\sum^{n+1}(1/\hat{\theta}^2_{i-1})\exp(-2\hat{\theta}_n/\hat{\theta}_{i-1})] \quad (5.5.17)$$

as one point recursion. Assuming that in a long run $(\hat{\theta}_n/\hat{\theta}_{i-1} \cong 1)$, the scheme (5.5.17) is approximated by

$$\hat{\theta}_{n+1} = \hat{\theta}_n - \left[(1/\hat{\theta}_n)\mathrm{e}(\mathrm{y}_{\mathrm{n}+1} - \mathrm{e}^{-1})\right]/\left[\sum^{n+1}(1/\hat{\theta}^2_{i-1})\right] \quad (5.5.18)$$

In the long run the summation $\sum^{n+1}(1/\hat{\theta}^2_{i-1})$ will be approximately equal to $(\mathrm{n}+1)\ \hat{\theta}_n^{-2}$. Therefore, relation (5.5.18) can be approximated by

$$\hat{\theta}_{n+1} = \hat{\theta}_n - [\hat{\theta}_n \mathrm{e}/(\mathrm{n}+1)](\mathrm{y}_{\mathrm{n}+1} - \mathrm{e}^{-1}). \quad (5.5.19)$$

Scheme (5.5.19) is a stochastic approximation scheme in which the sequence α_{n} is

$$\alpha_{\mathrm{n}} = \hat{\theta}_n e/(n+1)$$

Moreover, assuming that the sequence of estimates lies in between a lower bound, Θ_1 and an upper bound Θ_{u} then it holds

$$\Sigma\alpha_{\mathrm{n}} \geq \Theta_1 \mathrm{e}\sum 1/(\mathrm{n}+1) \text{ and } \Sigma\alpha_{\mathrm{n}} \leq \Theta_u^2 \mathrm{e}^2 \sum 1/(\mathrm{n}+1)^2$$

Thus for the truncated sequence in which $\hat{\theta}_{n+1}$ is defined by (5.5.19) unless the right-hand side is less than Θ_1 (in which case one has to define $\hat{\theta}_{n+1} = \Theta_1$) or is greater than Θ_{u} (in which case one has to define $\hat{\theta}_{n+1} = \Theta_{\mathrm{u}}$).

Thus condition (5.5.2a) holds due to (5.5.15), therefore the SA scheme converges to the root of the equation $\partial l/\partial\theta = 0$, i.e. to the MLE θ. □

Example 23 The Up and Down method is a Stochastic Approximation Scheme.

Indeed:

Recall iteration scheme (5.5.2b). When the "regression" Eq. (5.5.2b) is of the form $T(L_p) = p, \quad p \in (0,1)$ i.e. the 100p percentile of the response is to be evaluated and (5.5.2b) is reduced to

$$\mathrm{Lp}^{(\mathrm{n}+1)} = \mathrm{Lp}^{(\mathrm{n})} - \mathrm{a}_{\mathrm{n}}(\mathrm{y}_{\mathrm{n}} - \mathrm{p})$$

When $\mathrm{p} = 0.5$ then the median, $m = \mathrm{L}_{0.5}$, is to be estimated through

$$\begin{aligned} \mathrm{m}_{\mathrm{n}+1} &= \mathrm{m}_{\mathrm{n}} - \mathrm{a}_{\mathrm{n}}(\mathrm{y}_{\mathrm{n}} - 1/2) \\ &= \mathrm{m}_{\mathrm{n}} - (\mathrm{a}_{\mathrm{n}}/2)(2\mathrm{y}_{\mathrm{n}} - 1) \\ &= \mathrm{m}_{\mathrm{n}} - \delta(2\mathrm{y}_{\mathrm{n}} - 1) \end{aligned} \quad (5.5.20)$$

Considering as an appropriate sequence $\mathrm{a}_{\mathrm{n}} = 2\delta$, relation (5.5.20) is equivalent to (5.4.1). Thus the SA scheme has been reduced to the UD.

5.6 Application to Cancer Problems

The Low Dose Extrapolation Designs are adopting different probability models, Olin et al. (1995). One of them is Weibull distribution, also applied in Relative Risk theory. Early estimates of a Virtually Safe Dose (VSD) were based on the confidence intervals. The estimation of the confidence interval depends on the assumed model, from the class of Multistage Models, Wosniok et al. (1999).

The essential differences between the candidate models heavily influence the extrapolation towards zero. A linear approximation is usually considered, extrapolating towards to zero, for all the models. Therefore it is very difficult to choose the right model, as all of them, provide similar values in the neighborhood of zero dose level.

Let X be the individual tolerance dose considered as a random variable, then the function $F(x) = P(X \leq x)$, with x being the dose level administered is of practical use, see Example 2. Recall, that the function F(.) can be considered as a cumulative distribution function and is, in principle, not known, but approximated.

The binary response problem "malignant tumor", "no-malignant tumor" can be formulated as follows:

$$Y_i = \begin{cases} 1, & ''tumor'' \quad with \quad probability \quad F(x) \\ 0, & ''no-tumor'' \quad with \quad probability \quad 1-F(x) \end{cases}$$

This scheme is typical to Bioassays. As the function F(x) *is* not known, it is assumed that it can be approximated by different models, which might offer different results. The appropriate F(x) can be selected, applying Kolmogorov–Smirnov test, see Wosniok et al. (1999). In most of the cases a logit model is assumed, and for a Bayesian probit model see Consonni and Marin (2007).

In Life Testing problems the *Weibull Model*, from the class of *Multistage Models*, is widely used. This is of the form:

$$F(x) = 1 - \exp(-\theta x^s) \tag{5.6.1}$$

with s being a shape parameter and if:

$s > 1$ the model is sub-linear,
$s < 1$ is subralinear,
$s = 1$ the model coincides with the so-called *one-hit model*, the exponential distribution.

The one-hit model postulates that cancer is the result of a single cell, while the Weibull model with the extra shape parameter, has a different hazard function and it is widely used in a number of life-testing applications.

Suppose that the target is to estimate the L_p, $p \in (0, 1)$ percentile for the Weibull model. This is equivalent to solve the equation $G(x) = F(x) - p = 0$. As F(x) is the Cumulative distribution function the solution is the desired percentile, see also Sect. 5.5. In practice only observations of G are given, the values

$yi = G(x_i) + e_i$ are available, with e_i being the errors with mean zero and variance σ^2. The Stochastic Approximation Scheme, discussed in Sect. 5.5, is therefore adopted to solve this equation, see for details Kitsos (1999).

Proposition 5.6.1 *For the Weibull Model the sequence of the* $100p$ *percentiles:*

$$L_{p,n+1} = L_{p,n} - \left(n\theta(1-p)\,L_p^{s-1}\right)^{-1}(y_n - p) \quad n = 1, 2, \ldots$$

converge (*in mean square*) *to the real p-th percentile point* L_p *and this design minimizes the variance, i.e. is a D-optimal design.*

This iterative scheme converges, in mean square, to the appropriate percentile, which in the case of Low Dose Extrapolation should be "small", example $p = 0.01$.

References

Chaudhuri P, Mykland PA (1995) On efficient design on nonlinear experiments. Stat Sinica 5:421–440

Consonni G, Marin J-M (2007) Mean-field variational approximate Bayesian inference for latent variable model. Comput Statistics Data Anal 52:790–798

Ford I, Titterington DM, Wu CFJ (1985) Inference and sequential design. Biometrika 72:545–551

Ford I, Kitsos CP, Titterington DM (1989) Recent advances in nonlinear experimental design. Technometrics 31:49–60

Guttman L, Guttman R (1959) An illustration of the use of stochastic approximation. Biometrics 15:551–559

Hu I (1998) On sequential designs in nonlinear problems. Biometrika 85:496–503

Kitsos CP (1986) Design and inference in nonlinear problems. Ph.D. thesis, University of Glasgow, UK

Kitsos CP (1989) Fully sequential procedures in nonlinear design problems. Comp Stat Data Anal 8:13–19

Kitsos CP (1999) Optimal designs for estimating the percentiles of the risk in multistage models in carcinogenesis. Biometrical J 41:33–43

Lai TL, Robbins H (1981) Adaptive design and stochastic approximation. Ann Stat 7:1196–1221

Olin S, Farland W, Park C, Rhomberg L, Scheuplein R, Start T, Wilson J (1995) Low-dose extrapolation of cancer risks. ILSI Press, Washington, D.C

Robbins H, Monro S (1951) A stochastic approximation method. Ann Math Stat 22:400–407

Schervish JM (1995) Theory of statistics. Springer, New York

Titterington DM (1980) Aspects of optimal design in dynamic systems. Technometrics 22:287–299

Wosniok W, Kitsos CP, Watanabe K (1999) Statistical issues in the application of multistage and biologically based models. In: Cogliano VJ, Luebeck EG, Zapponi GA (eds) Perspectives on biologically based risk assessment (Chap. 7). Plenum Publishers, Berlin

Wu CFJ (1981) Asymptotic theory of nonlinear least squares estimates. Ann Stat 9:501–513

Wu CFJ (1985) Efficient sequential designs with binary data. JASA 80:974–984

Wu CFJ, Wynn HP (1978) The convergence of general step-length algorithms for regular optimum design criteria. Ann Stat 6:1273–1285

Chapter 6
Approximate Confidence Regions

Abstract The problem of the approximate confidence regions is discussed, under the light of the introduced measures of nonlinearity. The first order autoregressive model and Michaelis–Menten are discussed as particular cases.

6.1 Introduction

When an experiment has been performed statistical inference needs: point estimation and the construction of the appropriate, possibly approximate, confidence (intervals) regions.

In nonlinear problems the linear theory is applied, when constructing the confidence intervals. Here is a particular case faced: Although the nonlinear case is broader than the linear one, the nonlinear problem is not worked on to reduce it to linear, but the effort is to extend results from the linear case to the nonlinear. The cost one has to pay is the approximation involved and its lack of accuracy. The accuracy depends on "how nonlinear" the function under consideration is. Thus, the idea of a *measure of nonlinearity* was introduced, which is discussed in this chapter. Moreover, when the design is constructed sequentially, the question lies on how to evaluate the confidence interval. This is discussed in Sect. 6.5.

6.2 Background

Recall model (2.2.2), i.e. a classical nonlinear regression model, under Assumptions 2, 3 and 5. As the function f(.), the deterministic part of model (2.2.3), is nonlinear, it is of interest to see "how much nonlinear" it is, by a Taylor expansion.

C. P. Kitsos, *Optimal Experimental Design for Non-Linear Models*,
SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-45287-1_6,

The tangent hyper-plane to the *solution locus* (the surface in the sample space generated by the points $\eta_i = f(\theta, u_i)$, i = 1,2,…,n with e regarded as a variable), see Seber and Wild (1989), at point $f(\hat{\theta})$ is given by

$$q(\theta) = f(\hat{\theta}) + \hat{X}'(\theta - \hat{\theta}) \tag{6.2.1}$$

where $f(\theta) = (f(\theta, u_1), f(\theta, u_2), \ldots, f(\theta, u_n))^T$, $\hat{X} = X(\hat{\theta})$ and X as in (2.6.1). In principle θ_t (the true θ) is needed, instead of $\hat{\theta}$, but $\hat{\theta}$ is used in practice. In his pioneering paper, Beale (1960) proposed a dimensionless empirical measure of nonlinearity Λ^*, defined by

$$\Lambda^* = ps^2(d_2/d_4). \tag{6.2.2}$$

With $s^2 = S_n(\hat{\theta})/(n - p)$, the estimated variance, recall (2.4.2), p the number of parameters involved and

$$d_i = \sum \|n(\theta_k) - q(\theta_k)\|^i \quad i = 2, 4 \tag{6.2.3}$$

where $\{\theta_k\}$ is a set of m points in the neighborhood of θ. □

The theoretical measure of nonlinearity according to Beale, Λ say, is the same as Λ^* but with σ^2 in place of s^2, so, in principle, it is more realistic. The minimum value of the theoretical measure of nonlinearity Λ_o, was named the *intrinsic nonlinearity* of the assumed correct model. Although this measure attracted some criticism, it is the first attempt to evaluate and involve to the approximations the nonlinearity of the model. The introduced measure Λ_o is a sort of curvature of the solution locus.

Bates and Watts (1980), using ideas from differential geometry, proved that Λ_o is one quarter of the mean square intrinsic curvature. Moreover they proved that by replicating the design r times the curvature at any point in any direction is reduced by a factor $1/\sqrt{r}$. Thus, replication obtained its own geometrical interpretation; however some examples refer rather to non-linear regression, than to the experimental design theory.

6.3 Confidence Regions

Attempting to construct (approximate) confidence regions the target is always to minimize their volume/area/length (equivalent to various optimality criteria, mainly to D-optimal design). Hence optimal design might lead to minimum approximate confidence regions. If relation (6.2.1) is true, i.e. the model is linear, then a $100(1 - \alpha)$ % confidence region corresponds to

$$S_n(\theta) - S_n(\hat{\theta}) = (\theta - \hat{\theta})^T (X^T X)(\theta - \hat{\theta}) \leq ps^2 F(\alpha; p, n - p) \tag{6.3.1}$$

with $S_n(\theta)$ as in (2.4.1), s^2 an estimate of σ^2, X as in (2.6.1) and $F(\alpha; p, n-p)$ as usual the $100(1-\alpha)\,\%$ of the F distribution, with p and n − p degrees of freedom. Note that in nonlinear problems the estimator of σ^2 is, as in linear case

$$s^2 = \frac{S_n(\hat{\theta})}{(n-p)} \tag{6.3.2}$$

But it is not an unbiased estimator of σ^2. Recall that the matrix X depends on an estimate of θ, and that (6.3.1) is a p-dimensional ellipsoidal, see Fig. 3.1 for $p = 3$ and $M(\xi, \theta)$ is approximated as in (6.3.1). Thus the approximation is based both on the linearity and the dependence of X on θ, and it defines the confidence ellipsoidal "locally" depending either on the estimate or "initial guess" of the parameter vector.

The measure of nonlinearity discussed in Sect. 6.2 is employed to adjust the nonlinear approximated confidence region, introducing an extra parameter as in the l.h.s of (6.3.3) below

$$S_n(\theta) - S_n(\hat{\theta}) \leq \lambda p s^2 F(\alpha; p, n-p). \tag{6.3.3}$$

The extra parameter is

$$\lambda = \begin{cases} 1, & \textit{linearization without Beale's assumption} \\ 1 + \left(\frac{n}{n-1}\right)\Lambda_o & \textit{if } p = 1 \\ 1 + \left[\frac{n(p+2)}{(n-p)p}\right]\Lambda_o & \textit{if } p \geq 2. \end{cases} \tag{6.3.4}$$

Beale's measure of nonlinearity involves the curvature of the nonlinear model, which is less than $\frac{1}{2}(F(a, p, n-p))^{\frac{1}{2}}$ (Seber and Wild 1989, p. 135) with F being the F distribution, with p and n − p degrees of freedom. So the supremum value of Beale's measure of nonlinearity, for models with two or more parameters, can be reduced to

$$B = 1 + \frac{n}{n-2}\frac{1}{\sqrt{F}} \tag{6.3.5}$$

Therefore, the approximate confidence region (6.3.3) can be, eventually, see Kitsos (2001) for detail

$$(\theta - \widehat{\theta})^T I(\widehat{\theta}, \xi)(\theta - \widehat{\theta}) \leq B p s^2 F(a; p, n-p). \tag{6.3.6}$$

When $B = 1$, a linear approximation is considered. This revised evaluation for the confidence intervals was applied to the Michaelis–Menten model, for collected data.

Thus the measure of nonlinearity was developed to adjust an approximation form due to linearity, when a confidence region is constructed. It is clear now that this measure attempts to consider "how much nonlinear" is the model, at that

neighbourhood of the assumed parameter value. In the nonlinear case confidence regions appear to have a "banana-shape" and they are not ellipsoids.

Hamilton and Watts (1985) argue that elliptical confidence regions (under D(θ)-optimality) that are suitable for large samples, are not appropriate for small samples. Thus they tried to construct a quadratic approximation to the volume of small sample confidence regions. They proposed a design criterion of the form

$$H = a \cdot \log\det\{M(\theta, \xi)\} - \log\{1 + b \cdot trQ(\theta, \xi)\} \tag{6.3.7}$$

where a, b are constants, $M(\theta, \xi)$ is the average information matrix and $Q(\theta, \xi)$ is a matrix describing parameter effects. Due to the second term in (6.3.7) the criterion H is not invariant under transformations of θ. Moreover their criterion requires an estimate of σ^2 which is not always available. The measure of nonlinearity discussed in Sect. 6.2 is employed to adjust the nonlinear approximated confidence region.

In Sect. 6.4, the small sample problem is tackled in constructing confidence intervals, adopting a sequential design procedure.

6.4 Simulation Study

The features of a fully sequential design appear in the autoregressive model, were the parameter θ is a real number.

$$y_{i+1} = \theta y_i + \varepsilon_{i+1}, \quad i = 1, 2, \ldots, n \tag{6.4.1}$$

The value of y_1 is given, the errors ε_{i+1} satisfy the Assumption 2. Then at the n-th stage an estimate of θ, θ_n, is given by

$$\hat{\theta}_{n+1} = \sum y_i y_{i+1} / \sum y_i^2 \tag{6.4.2}$$

where the summation runs from 1 to n. Notice that the first order autoregressive model can be applied in Signal Process, Kitsos and Toulias (2012), who discussed the confidence interval and the tolerance interval approach. Here the confidence interval problem is discussed. The sample information I_n/σ^2 can be evaluated with $I_n = \sum y_i^2$. As Lai and Siegmund (1983) point out the following asymptotic result holds for $|\theta| \prec 1$

$$I_n^{1/2}(\hat{\theta}_n - \theta) \longrightarrow LN(o, \sigma^2) \tag{6.4.3}$$

Ford et al. (1985) discuss this model. In the simulation study performed small sample sizes of n = 10, 5 observations were used, for confidence level $\alpha = 0.05$. As the initial value $y_1 = 0.0$ was considered and the error variance was assumed $\sigma^2 = 1$. Different values of θ were taken from the range $-1,5 \leq \theta \leq 1,5$, that is even beyond the appropriate interval −1, +1) for the parameter θ. Confidence limits were evaluated according to:

Table 6.1 Simulation study on autoregressive model (6.4.1). Nominal level $\alpha = 0.05$. Number of simulations N = 1,000. Sample size n = 10, 5, $y_1 = 0.0$, $\sigma = 1$

N	θ	P	MSE	S	K	$\bar{\theta}$
10	−1.5	0.919	0.03	3.88	21.07	−1.44
	−1.0	0.947	0.09	1.39	5.47	−0.84
	−0.5	0.970	0.09	0.63	3.27	−0.42
	0.0	0.967	0.10	−0.06	2.56	0.00
	0.5	0.965	0.10	−0.58	3.01	0.41
	1.0	0.955	0.09	−1.29	5.45	0.85
	1.5	0.928	0.03	−3.96	22.21	1.45
5	−1.5	0.968	0.27	1.13	5.79	−1.29
	−1.0	0.969	0.27	0.08	10.00	−0.82
	−0.5	0.979	0.25	−0.09	6.05	−0.39
	0.0	0.970	0.25	0.22	4.92	0.02
	0.5	0.973	0.33	−0.58	4.95	0.79
	1.0	0.975	0.61	−0.14	4.98	0.38
	1.5	0.957	0.29	−1.23	6.02	1.29

P is the estimated coverage, S is the Average skewness of estimates, K is the Average kurtosis of estimates and $\bar{\theta}$ is the Average $(\hat{\theta}_i)$, i = 1,2,…,n

$$\hat{\theta} \mp t(n-1; 1-a/2)\sqrt{[(RSS/n-1)I/(n-1)]} \tag{6.4.4a}$$

with RSS being the residual sum of squares,

$$RSS = \sum y_i^2 - \hat{\theta} \sum y_{i+1} y_i. \tag{6.4.4b}$$

To test the normality of the sequence θ_n the skewness and the kurtosis were evaluated. From Table 6.1 it is easy to see that as far as the normality is concerned, through the evaluated skewness and kurtosis, the results are unsatisfactory when $\theta \notin (-1, 1)$.

The mean squared error (MSE) is, of course, larger when the sample size is reduced from n = 10 to n = 5 observations.

6.5 The Michaelis–Menten Model of Pharmacokinetics

A general theory for enzyme kinetics was firstly developed by Michaelis and Menten (1913) in their pioneering work. The metabolism of an agent is described by a reaction rate. The basic toxic kinetic model of metabolism is the Michaelis–Menten (MM) model. This is discussed very briefly below:

When an enzyme E combines reversibly with a substrate S, to form an enzyme-substrate complex ES, which can be dissociate or proceed to the product P, the following scheme is assumed

$$\mathrm{E} + \mathrm{S} \underset{k_2}{\overset{k_1}{\rightleftarrows}} \mathrm{ES} \xrightarrow{k_3} \mathrm{E} + \mathrm{P} \tag{6.5.1}$$

with k_1, k_2, k_3 the associated rate constants. Let us denote $K_M = \frac{k_2+k_3}{k_1}$, known as MM constant and $V_{max} = k_3 C_{TOT}$, C_{TOT} = the total enzyme concentration.

Then a plot of the initial velocity of reaction u, against the concentration of substrate, C_S, will provide the MM rectangular hyperbola of the form

$$\eta = \frac{V_{max} C_S}{K_M + C_S}. \tag{6.5.2}$$

Notice that the defined general form as in (2.2.2) holds with $\theta = (V_{max}, K_M)$, $u = C_S$. It is very crucial to mention here that Biebler et al. (2008) pointed out that, when numerical procedures for the solution of the fundamental differential equation involved are adopted, and different initial triples are given, then one is coming across of different results for the MM model.

From an optimal experimental design point of view for η as in (6.5.2) the target is to evaluate, Fisher's information, recall (2.5.1),

$$\nabla\eta = \left(\frac{\partial\eta}{\partial_{\max}}, \frac{\partial\eta}{\partial K_M}\right)^{\mathrm{T}} = \left(\frac{C_s}{K_M + C_s}, -\frac{V_{max} C_s}{(K_M + C_s)^2}\right)^{\mathrm{T}}. \tag{6.5.3}$$

Therefore, the average per observation information matrix, related to Fisher's information, recall (2.5.3), can be evaluated as:

$$\sigma^{-2}\mathrm{n}\mathbf{M}(\theta, \xi) = \sum_{i=1}^{n} \begin{pmatrix} C_{s,i}^2 \tau_i^2 & -V_{max} C_{s,i}^2 \tau_i^3 \\ -V_{max} C_{s,i}^2 \tau_i^3 & V_{max}^2 C_{s,i}^2 \tau_i^4 \end{pmatrix} \tag{6.5.4}$$

with $\tau_i = 1/(K_M + C_{s,i})$, $\theta = (V_{max}, K_M)$, $i = 1,2,\ldots,n$.

Asymptotically, the covariance matrix, recall (2.6.5) is $\mathbf{C} = \mathbf{C}(\hat{\theta}, \xi) = (\mathrm{n}\mathbf{M}(\hat{\theta}, \xi))^{-1}$.

For the MM, with $C_S \in (0, U]$, the locally D-optimal design at $K_M = K_0$ allocates half observations at points U and $\mathrm{opt}C_S$ with

$$\mathrm{Opt}C_S = \frac{K_0 U}{2K_0 + U},$$

were U is the maximum allowable substrate concentration. The corresponding value of the determinant of the D-optimal design is

$$\mathrm{d} = \frac{V_{max}^2 U^6}{16 K_o^2 (K_0 + U)^6}.$$

This value can be compared with the various values given by Endrenyi and Chan (1981) under different approaches. If $U \gg K_0$ the locally D-optimal design ξ is

$$\xi^* = \begin{Bmatrix} U & K_0 \\ 0.5 & 0.5 \end{Bmatrix},$$

i.e. Allocate half of the observations at the maximum value of the concentration of substrate, U and at the initial guess for the MM constant, K_0. This design can be proved very helpful in applications, while for a Bayesian approach and applications see Matthews and Allcock (2004), and for a number of different design approaches for MM model see Lopez-Fidalgo and Wong (2002).

In Chap. 7 it is studied, through two simulations, how the optimal experimental design line of though is applied.

References

Seber CAF, Wild CJ (1989) Nonlinear regression. Wiley, NJ

Beale EML (1960) Confidence regions in non-linear estimation. J. Roy. Stat. Soc. B 22:41–88

Bates DM, Watts DG (1980) Relative curvature measures of nonlinearity. J. Roy. Stat. Soc. B 42:1–25

Hamilton DC, Watts DC (1985) A quadratic design criterion for precise estimation in nonlinear regression models. Technometrics 27:241–250

Kitsos CP, Toulias TL (2012) Confidence and tolerance regions for the signal process. Recent Patents on Signal Processing, vol 2

Ford I, Titterington DM, Wu CFJ (1985) Inference and sequential design. Biometrika 72:545–551

Michaelis L, Menten ML (1913) Kinetics for intertase action. Biochemische Zeiturg 49:333–369

Biebler KE, Schreiber K, Bode R (2008) Solutions to the Michaelis-Menten kinetics are not necessarily unique. In: International conference on computational methods in sciences and engineering (ICCMSE), 2008

Endrenyi L, Chan FY (1981) Optimal design of experiments for the estimation of precise hyperbolic kinetic and binding parameters. J Theor Biol 90:241–263

Kitsos CP (2001) Design aspects for the Michaelis–Menten Model. Biometrical Letters 38:53–66

Matthews JNS, Allcock GC (2004) Optimal designs for Michaelis-Menten kinetic studies. Stat Med 23:477–491

Lai TL, Robbins H (1979) Consistency and asymptotic efficiency of slope estimates in stochastic approximation schemes. Z. Wanhrscheinlichkeitstheorie wer. Gebite 56:329-360

Lopez-Fidalgo J, Wong WK (2002) Design issues for the Michaelis-Menten model. J Theor Biol 215:1–11

Chapter 7
Simulation Studies

Abstract The dilution assessment and the first order growth model are Extensively discussed, through two simulation studies. Various Strategies to tackle the problems were adopted and compared.

7.1 Introduction

Up to this point, a strong theoretical background of the NLED problem has been developed. This Chapter as well as the next one, are devoted to applications, offering food for thought for Toxicology, Biostatistics, Engineering, Environmentrics, Epidemiological studies and not only. This is "to put theory to work" on particular problems. The difficulties which arise and the results obtained are discussed. Both the binary (Simulation I) and the continuous (Simulation II) cases are tackled for one parameter and two parameters respectively. In this sequel the problems simulated, and extensive interpretation of the results are described. The discussed simulations are:

- Simulation I: The dilution series assessment (Sects. 7.2–7.6).
- Simulation II: The first order growth law (Sects. 7.7–7.8).

7.2 The Dilution Series Assessment

Experimenters and statisticians should be indebted to Rothamsted Experimental Station as it has offered a job to the jobless Fisher! Since then, Fisher developed the theory of experimental design and tackled the first nonlinear design problem in 1922. An example of the non-linear problems is that of the dilution series assessment, which is described below.

C. P. Kitsos, *Optimal Experimental Design for Non-Linear Models*,
SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-45287-1_7,

It is desired to determine the concentration of micro-organisms in a solution. In this case various dilutions are sampled. For every dilution whether or not there is sterility is recorded. Let us use the following notation.

u: a small volume that is taken out of a volume V, of liquid which contains N,
N: the number of tiny organisms, from $U = [U_1, U_u]$, the design space.
θ: The density per unit volume, i.e. $\theta = N/V$.
The probability that the volume contains no organisms is

$$p = (1 - u/V)^N \approx \exp(-Nu/V) = \exp(-\theta u)$$

y: The binary response describing the phenomenon defined as

$$\begin{aligned} y = 1 &: \text{ no organism in u (sterile)}, \\ y = 0 &: \text{ organisms in u (fertile)}. \end{aligned}$$

The probability model describing the experiment is therefore

$$p(y|u,\theta) = \begin{cases} \exp(-\theta u) & y = 1 \\ 1 - \exp(-\theta u) & y = 0 \end{cases} \quad \theta \succ 0 \tag{7.2.1}$$

The aim is to estimate θ as appropriate as possible. Model (7.2.1) might describe also the probability that an insect survives a dose of u units of a certain insecticide, while in carcinogenesis problems it appears as the "one hit model". The following results hold:

Claim 1: Fisher's information $I(u,\theta)$, for a single observation, for model (7.2.1) is

$$I(\theta, u) = u^2/(\exp(\theta u) - 1) \tag{7.2.2}$$

Proof From (7.2.1) the log-likelihood function $\ell = \ell(y|u,e) = \log p(y|u,\theta)$ will be

$$\ell = \ell(y|u,e) = \begin{cases} -\theta u & y = 1 \\ \log(1 - \exp(-\theta u)) & y = 0 \end{cases}.$$

Therefore the second degree partial derivative is

$$\partial^2 \ell/\partial\theta^2 = \begin{cases} 0 & y = 1 \\ -[u^2 \exp(-\theta u)]/[1 - \exp(-\theta u)]^2 & y = 0 \end{cases}.$$

Thus Fisher's information equals to

$$\begin{aligned} I(\theta, u) &= E(-\partial^2 \ell/\partial\theta^2) = 0 p(1|u,\theta) \\ &\quad + \left[u^2 \exp(-\theta u)\right]/[1 - \exp(-\theta u)]^2 p(0|u,\theta) \\ &= u^2 \exp(-\theta u)/(1 - \exp(-\theta u)). \end{aligned}$$

Then (7.2.2) is obtained, q.e.d. □

The target is to maximize Fisher's information, so that the variance to be minimum, recall (2.6.5). Thus it is proved that:

Claim 2: The maximum of $I(\theta, u)$ is obtained when

$$\theta u^* = 1.59 \tag{7.2.3a}$$

Proof Indeed from Claim 1 it is:

$\partial I(\theta, u)/\partial u = 0$ i.e. $2 = 2\exp(-\theta u) + \theta u$.

The equation $h(z) = 2\exp(-z) + z - 2 = 0$, with $z = \theta u$ can be solved numerically by the Newton–Raphson iterative scheme. □

The optimum experimental design point then can be evaluated explicatively due to Claim 2, namely:

Claim 3: The optimum design point, the one which minimizes the variance, i.e. that which corresponds to $D(\theta)$—optimality, depends on θ, due to (7.2.3a) and equals to

$$u^* = \begin{cases} 1.59/\theta & \text{if} \quad 1.59/\theta \in U \\ U_l \quad \text{or} \quad U_u & \text{otherwise} \end{cases}. \tag{7.2.3b}$$

The form of the probability model (recall (2.2.3) with $T(u, \theta)$ as in (7.2.1)) will be binomial with success probability $p = p(1|u,\theta)$and number of successes the number of sterile samples. In terms of probability, the value $u^* = 1.59/\theta$ corresponds to probability level $p = 0.2$. Notice that the optimal design point depends on the parameter the experimenter wants to estimate! It seems reasonable to keep the probability levels between [0.025, 0.0975], truncating [0, 1]. This truncation proceeds with the calculations and it evaluates the MLE, recall Silvapulle's Theorem, Sect. 2.4 and Appendix 2, as a whole batch of successes or failures will provide no estimates. Therefore, throughout the simulations the bounds U_l, U_u can be used. Then the end points of the restricted design space U, can be evaluated from the relations

$$\exp(-\theta_t U_u) = 0.025, \ \exp(-\theta_t U_l) = .975, \tag{7.2.4}$$

where θ_t is the "true" value of θ. As such a value $\theta_t = 3.18$ was chosen. This value corresponds to $u^* = 0.5$ from (7.2.3a). Thus from (7.2.4) U_l and U_u are evaluated as

$$U_l = 0.00796 \qquad U_u = 1.160. \tag{7.2.5a}$$

Note that U_l would represent the optimal design point for $\theta = 199.70$ and U_u for $\theta = 1.370$. Therefore the parameter space $\Theta = [\Theta_1, \Theta_u]$ is also restricted, with

$$\Theta_1 = 1.370 \qquad \Theta_u = 199.70. \tag{7.2.5b}$$

Eventually a truncated parameter space is obtained, with no loss of the generality, corresponding to the truncated design space. This means that when the estimators are not within the above limits, presented in (7.2.5b), the experimenter

tries to "bring the estimators within it". This is a useful technique working with practical problems.

7.3 The Strategy of Simulation I

Adopting the sequential procedure for designing, the next design points are chosen the ones which minimize Fisher's information, see also the Algorithm in Sect. 5.2.

Claim 4: In the first stage $s = 1$, the estimator of θ, θ_1 say, can be explicitly evaluated from the data as

$$\hat{\theta}_1 = -u_1 \ln\left(r^{-1} \sum y_i\right). \tag{7.3.1}$$

Proof

(1) The likelihood function and the first and second order derivatives are

$$\begin{aligned}
L(\theta; u, y) &= \Pi \exp(-\theta u_i)^{y_i} (1 - \exp(-\theta u_i))^{1-y_i} \\
&= \exp(-\theta \sum u_i y_i) \Pi (1 - \exp(-\theta u_i))^{1-y_i} \Rightarrow \\
\ell(\theta) &= \log(L(.)) = -\theta \sum u_i y_i + \sum (1 - y_i) \ln(1 - \exp(-\theta u_i)) \Rightarrow \\
\ell'(\theta) &= -\sum u_i y_i - \sum (1 - y_i)[u_i \exp(-\theta u_i)]/[1 - \exp(-\theta u_i)] (*) \\
\ell''(\theta) &= -\sum (1 - yi)[u_i \exp(-\theta u_i)]/[1 - \exp(-\theta u_i)]^2.
\end{aligned}$$

The MLE is obtained from the equation $\ell'(\theta) = 0$. The Newton–Raphson iteration scheme is applied to obtain the solution of the equation $\ell''(\theta) = 0$, namely $\theta_{\nu+1} = \theta_\nu - \ell'(\theta_\nu)/\ell''(\theta_\nu), \theta_0$ given, $\nu = 1, 2 \ldots$.

(2) Evaluation of $\hat{\theta}_1$.

For the first batch of observations, $u_i = u_1$, $i = 1, 2 \ldots$ with the number of replications to be r, $r = 5, 25, 50$, from (*) it is obtained that

$$\begin{aligned}
&\ell'(\theta) = -u_1 \sum^r y_i + u_1 \sum^r (1 - y_i) \exp(-\theta u_1)/(1 - \exp(-\theta u_1)) = 0 \\
&\Rightarrow \sum^r y_i = [\exp(-\theta u_1)/(1 - \exp(-\theta u_1))] \sum^r (1 - y_i) \\
&\Rightarrow \sum^r y_i - \exp(-\theta u_1) \sum^r y_i = \exp(-\theta u_1) \left(r - \sum^r y_i\right) \\
&\Rightarrow r \exp(-\theta u_1) = \sum^r y_i \Rightarrow (7.3.1).
\end{aligned}$$

□

The bounds of θ can be used as the estimates of θ, in the extreme cases, beyond the defined parameter space. It can be proved that if all y_i's are 1 then $\hat{\theta} = 0$ and when all y_i's are 0 then $\hat{\theta} = \infty$. Therefore one can avoid situations, especially with small batches, when the MLE (see above) could not be evaluated. The MLE exists and was evaluated iteratively when $0 \prec \sum y_i \prec n$: this is equivalent with the implication that there is not a whole batch of zeros or units. This point is emphasized, as it is not always considered. The numerical method of Newton–Raphson (NR) was used to solve, at each stage, the likelihood equation (see Claim 4, part (1) of the proof above) for the evaluation of $\hat{\theta}$.

The Newton–Raphson method converges when the initial value lies in the neighborhood of the solution. Therefore this difficulty has to be overcome, which happened often when small batches were used in early stages of the sequential design. The Bisection method was used with a rather "large initial interval", [0.01, 100.3] so that to obtain a "good" initial value. This initial value "feeds" the Newton–Raphson method.

The question arisen is: why the Bisection method is not adopted, through out all the estimation. The answer is that Bisection method leads to mini-max decision problems, which is a completely different theoretical approach, while Newton–Raphson is acting as a stochastic approximation, see also Sect. 5.5.

As far as the design points are concerned the procedure can be described by, recall also (7.2.3b):

$$u_{s+1} = \begin{cases} 1.59/\theta_s & \text{if} \quad u_{s+1} \in U \\ U_1 & \text{if} \quad \hat{\theta}_s \prec \theta_1 \\ U_2 & \text{if} \quad \hat{\theta}_s \prec \theta_u \end{cases} \qquad s = 1, 2, \ldots, s_{max}. \tag{7.3.2}$$

The maximum number of stages, s_{max}, say, depends on the number of replications r chosen. Simulations were carried out for r = 5, 25, 50, and sample size n = 100 therefore s_{max} = 20, 4, 2, respectively.

The estimates corresponding to (7.3.2) were obtained through

$$\hat{\theta}_s = \begin{cases} -u_o \ln\left(\frac{1}{n}\sum yi\right) & if \quad s = 1 \\ \Theta_u & if \quad all \qquad y_i's = 0, s \succ 1 \\ \Theta_l & if \quad all \qquad y_i's = 1 \\ \hat{\theta} \quad evaluated \quad through \quad \begin{cases} NR & if \quad r = 50 \\ Bi\,sec\,tion \quad and \quad NR\ if & r = 5, 25 \end{cases} \end{cases}. \tag{7.3.3}$$

Therefore the design points and the estimates at each point are well defined through the introduced truncation and the discussed numerical techniques.

To investigate the dependence of the design procedure on the initial design points the values $\theta_1 = 2.196$ and $\theta_2 = 7.15$ were chosen as the unknown ("far" from the "true value" 3.18) starting values. These values of θ lead to corresponding design points u = 0.72 and u = 0.22 in (7.3.2) and probability levels

$p = 0.1$ and $p = 0.5$ respectively. Assuming that the "true" value is $\theta_t = 3.18$, this corresponds to a local optimum design point $u^* = 0.5$, with probability level $p = 0.2$. For the final estimate $\hat{\theta}$, i.e. when $n = 100$ observations were used, an approximate confidence interval can be evaluated as

$$\hat{\theta} \mp 1.96/(1/S(\hat{\theta}, \xi_n, y)) \tag{7.3.4}$$

where $S(\hat{\theta}, \xi_n, y)$is the sample information, recall (2.4.4). That is, although in the sequential design the design points are not entering the design independently of the response (7.3.4) offers reliable results. The confidence interval is constructed by "pretending" that the design points were independent of the response. The experiment was repeated 1,000 times.

7.4 Simulation Procedures

Different procedures were applied to tackle the dilution problem under the strategy described above. These strategies or approaches help the researcher to receive food for thought, about the design he has to conduct in prior. These procedures will be referred as P1, P2 etc. In all cases the sample size was chosen $n = 50$ or 100.

P1. *Static design*: The optimal static design (recall Chap. 4) will be the one which takes all the observations at the locally optimal point for the true θ, as in (7.2.3b). Notice that as only one parameter is involved, all the design criteria D-, G, A-optimality coincide. But usually D-optimality is referred as it is easier to be adopted from the experimenters. Therefore the n observations were taken at $u^* = u^*(\theta_{st})$ where θ_{st} is the "starting value" for θ, and the MLE was calculated. Results in Table A4.1, Appendix 4.

P2. *Sequential design, equal batches*: The batch sequential method of designing was adopted. Equal batches were used to reach the total sample size n. Results in Table A4.2, Appendix 4.

P3. *Sequential design, unequal batches*: The design was started out with a batch of 25 or 50 observations. The MLE was evaluated explicitly at the first stage. Thereafter, i.e. when $s = 2, 3, \ldots, s_{max}$, the number of replications, r' say, was taken to be 5. The values of s_{max} are 15 and 10 correspond to the initial batches 25 and 50 observations. Results in Table A4.3, Appendix 4.

P4. *Fully sequential design (Stochastic Approximation)*: Batches of 5, 25, 50 observations were used to start off the design. One observation was then added, i.e. $r' = 1$, to the design and only one step of the Newton–Raphson iteration was used to produce the estimate of θ. This is the Stochastic Approximation scheme discussed in Sect. 5.5. Results in Table A4.4, Appendix 4.

P5. *Fully sequential design (Full Maximum Likelihood at the end)*: The data are as in P4 but the exact MLE is obtained at the end of the experiment. Results in Table A4.5, Appendix 4.

The notation used in Appendix 4 is:

θst	Starting value for θ.
r	Number of observation per batch, r = 5, 25, 50.
ECP	Estimated Coverage Probability i.e. the proportion of times out of the 1000 simulations the true value of θ was captured in the confidence interval. "Outliers" are the cases with only success (failures).
$\overline{\theta}$	The average value of the estimates, $\hat{\theta}_i$, produced in N = 1000 simulations.
S	Estimated skewness of $\hat{\theta}$.
K	Estimated kurtosis of $\hat{\theta}$.
EMSE	Estimated Mean Square Error of the 1000 evaluated $\hat{\theta}_i$'s through:

$$\text{EMSE} = \text{est.Var}(\hat{\theta}) + [\text{est.Bias}(\hat{\theta})]^2$$

The "true" value of $\theta = 3.18$ and the sample size n was either 50 or 100.

7.5 Discussion I

It is easy to see that the total information for θ, and therefore the variance of θ can be evaluated (asymptotically) explicitly as, see (7.2.2):

$$(\text{nI}(\text{u},\theta))^{-1} = [\text{nu}^2/(\exp(\theta_{\text{t}}\text{u}) - 1)]^{-1} \tag{7.5.1}$$

When the design allocates all the n observation at the point u, procedure P1 is adopted. Table 7.1 provides the value of $\text{n}^{-1}\ \text{I}^{-1}$ (u, θ) for the design points selected to start the design, under different sample size. Therefore, a guide for the evaluated mean squares is provided, when a static design is performed, so that it can be compared with the sequential procedures.

Let us consider and investigate each procedure separately.

- P1

Sample sizes n = 100 or 50 and also the extreme sample size of n = 1000 were applied to study the asymptotic behavior of the one point, one stage design. For n = 1000, the reader verifies that EMSE is not too far from the expected value (see Table 7.1). As the sample size gets smaller, EMSE of course increases. The normality of the vector of estimates behaves quite well, when n = 1000, but it gets worse when n drops to 50.

Thus the sample size is very critical especially when nearing to the end points of the probability levels in (7.2.4). This happens when $\text{u}_1 = .72$ which corresponds to probability level p = 0.1. At the extreme case scenario, of a batch of y_i's that all equal to zero, no MLE is evaluated then. Thus the one-stage design does not have the opportunity to improve the estimate, when n = 50, from this pathological situation.

Table 7.1 Evaluating the asymptotic variance

u/n	0.72	0.50	0.22
1000[a]	0.0171	0.0156	0.0208
100	0.171	0.156	0.208
75[b]	0.231	0.208	0.278
50	0.342	0.312	0.416

[a] only for P1, [b] only for P4, P5

- P2

When $n = 100$ and $r = 50$ (or $n = 50$ and $r = 25$) the so called *two stages design* is created. In the two stages design if, in the first stage, the estimator is a "poor" one, the design does not have the opportunity to recover in the next stage. But, when the initial estimate is "reasonable", it is improved in the second stage. This is less likely when the initial batch is 5 or 25 observations. The design behaves similarly with $r = 5, 25$ when $n = 100$.

- P3

There is not a two stage design in P3. When 5 observations are used in the first batch, P3 coincides with P2. When 25 observations were used as the first batch, the EMSE obtained was slightly better than the equal batch procedure P2, when a point far from the true value was considered. When $r = 50$ the design had "enough time to recover" from a possible bad estimate at first stage.

- P4

As only one observation was added at each stage, the sample size $n = 75$ was also used as an intermediate stage between $n = 50$ and $n = 100$. The aim was to check how far things can be improved, by adding only one observation. The performance of the procedure is largely independent of the initial batch size and the value θst, although there is a little more variability when $n = 50$.

- P5

There is little difference between P4 and P5. Under different sample sizes the EMSE are close to that of P4. The comments for P4 are similar to those of P5.

The procedures mentioned above can be divided into two categories.

- One stage design (Procedure P1)
- Sequential design:
  - Block design (Procedures P2 and P3)
  - Fully sequential design (Procedures P4 and P5).

Truncation helps the experimenter to "bring the design back" to the sample space defined, as it is a waste of time to look outside of the parameter space Θ for estimates.

The Bisection method is a rather slow numerical method as far as convergence rate is concerned, whilst Newton–Raphson has at least second order convergence rate. Thus with Bisection the initial value θ_0 to feed Newton–Raphson was in the neighborhood of the solution of the likelihood equation.

As far as normality of the estimators is concerned, acceptable behavior is observed with $n = 100$, and reasonable with $n = 50$. According to R. A. Fisher: "Nothing that we say shall be true, except in the limit when the sample is indefinitely increased; a limit obviously never obtained in practice". This situation is described in P1 where with $n = 1000$ everything seems to be acceptable, except the sample size. That particular idea of "practice" was behind this simulation study and it seems to us that $n = 100$ is quite "large" and $n = 50$ is "reasonable".

After the discussion of the experience of the Simulation Study I, Simulation Study II follows.

7.6 The First Order Growth Law

Biological processes concerning a measure of growth y, of plants or animals can be expressed through a regression set-up, known as growth law.

Consider the period ranging from lower U_1 up to the upper U_u, with the input variable u denoting time in this application. The expected initial value of y (i.e. when $u = 0$) is denoted by $\theta_1 > 0$. The rate of increase of biological process is denoted by $\theta_2 > 0$. The phenomenon can be described by the nonlinear regression model, recall (2.2.2)

$$Y_i = \theta_1 \exp(\theta_2 u_i) + e_i,\ i = 1,\ 2, \ldots, n \quad u \in U = [U_1,\ U_u]. \tag{7.6.1}$$

The error term e_i is under Assumption 2, when inference is made. Under the criterion of $D(\theta)$-optimality the locally optimal two point design allocates half observation at the optimal design points

$$u_1 = U_u - 1/\theta_2, \quad u_2 = U_u \tag{7.6.2}$$

where $\theta_2 > 0$. Therefore supplying a "guess" θ_{20}, for θ_2 a static design can be produced. For the so called decay model, i.e., when $\theta_2 > 0$, the design points are

$$u_1 = U_1, \quad u_2 = U_1 - 1/\theta_2. \tag{7.6.3}$$

The designs (7.6.2) or (7.6.3)—notice the symmetry—depend only on θ_2, as they are considered partially-nonlinear (recall Example 6). The aim is to use this model under different sequential design procedures and to investigate the estimation problem, the distribution of the sequence of parameter-estimators, as well as the construction of the confidence intervals.

7.7 Strategy and Procedures of Simulation II

The optimum design measure $\xi^* = 1/2$ for the two points (7.6.2), corresponds to the D-optimal design. Keeping the sample size n = 40 constant and the following procedures (the notation is Π_1, Π_2, etc. for them) are discussed.

Π_1. *Static Design*: One stage design. Allocate half observations as in (7.6.2) providing θ_{20}, a guess for the true value of the parameter, θ_{2t}.

Π_2. *Two-stage Design*: Use half of the observations in the first stage, i.e. allocate one fourth of observations at each optimal design point. Obtain the estimate of the parameter $\hat{\theta}$. Use them to redesign as in (7.7.2) using $\hat{\theta}_2$ instead of θ_2. The other three procedures are sequential ones and only the number of stages is changed.

Π_3. *Five-stage Design*: Use 8 observations in each stage.

Π_4. *Ten-stage Design*: Use 4 observations in each stage.

Π_5. *Fully-sequential Design*: Two observations at each stage, i.e. allocate one observation at each "optimal" point at each stage.

One can start the design with different θ_2 values, considering them as "far" from the unknown true value, of the form $\theta_{2f} = \theta_{2t}, \theta_{2t} + 2, \theta_{2t} - 2$. Let us denote by $\theta_{2t} = 1, 2, 3, 4$ the "true" value of θ_2. The value of θ_1 was kept constant, $\theta_1 = 10.0$, as the design does not depend on it. The design points were evaluated, due to (7.6.3), according to

$$u^*_{1S-1} = 1 - 1\Big/\hat{\theta}_{2S}, \qquad u^*_{2S-1} = 1 \tag{7.7.1}$$

with $\hat{\theta}_{2S}$ being the estimate of θ_2 at stage s.

Claim I: For the first stage (s = 1), in sequential procedures, or for the static design the estimate for θ can be evaluated explicitly as

$$\hat{\theta}_1 = \bar{y}_{11} \cdot \exp\left(-\hat{\theta}_2 u_{11}\right), \ \hat{\theta}_2 = [\ln \bar{y}_{11} - \ln \bar{y}_{12}]/u_{12} \tag{7.7.2}$$

Proof Let j = 1, 2 denote the two points where the observations are obtained. Let I denote the number of stages, i = 1, …, I and n_i the number of observations at each point in each stage. Then the sum of squares, SS, equals to:

$$\begin{aligned} SS &= \sum_{j=1}^{2} \left\{ \sum_{i=1}^{I} \left[\sum_{K=1}^{n_i} \left(y_{jik} - \theta_1 e^{\theta_2 u_{jL}} \right)^2 \right] \right\} \\ &= \sum_{j=1}^{2} \left\{ \sum_{i=1}^{I} \left[\sum_{K=1}^{n_i} \left(y_{jiK} - \bar{y}_{ij.} \right)^2 + n_i \left(\bar{y}_{ji.} - \theta_1 e^{\theta_2 u_{ji}} \right)^2 \right] \right\} \\ &= \sum_{j=1}^{2} \sum_{i=1}^{I} \sum_{K=1}^{n_i} \left(y_{jiK} - \bar{y}_{ij.} \right)^2 + \sum_{j=1}^{2} \sum_{i=1}^{I} \left(\bar{y}_{ji.} - \theta_1 e^{\theta_2 u_{ji}} \right)^2 \\ &= SS_1 + SS_2. \end{aligned}$$

Considering n_i constant through the experiment, i.e. equal replications, r say, at each point each time. Thus:

$$SS_2 = r\sum_{j=1}^{2}\sum_{i=1}^{I}\left(\bar{y}_{ji.} - \theta_1 e^{\theta_2 u_{ji}}\right)^2.$$

To obtain the estimates at the first stage i = 1, notice that eventually it is desired to minimize SS_2 of the form

$$\left((y_{11.} - \theta_1 \exp(\theta_2 u_{11}))^2 + (y_{21.} - \theta_1 \exp(\theta_2 u_{21}))^2\right)$$

That is, both the terms must be equal to zero i.e.

$$\widehat{\theta}_1 \exp\left(\widehat{\theta}_2 u_{11}\right) = \bar{y}_{11}$$
$$\widehat{\theta}_1 \exp\left(\widehat{\theta}_2 u_{21}\right) = \bar{y}_{21}$$
$$\bar{y}_{11.}/\bar{y}_{21.} = \exp\left(\widehat{\theta}_2(u_{11} - u_{21})\right)$$

Hence relation (7.7.2) is obtained. □

In other stages, when s > 1, the estimates where obtained through the modified *Newton–Raphson scheme*, for solving a multivariate equation, which very briefly states:

Let f be a function $f : R^k \longrightarrow R^k$ with a root ζ i.e. $f(\zeta) = 0$. When k = 1, to evaluate the root ζ the iteration, known as Newton–Raphson, of the form:

$$x_{i-1} = x_i - f(x_i)/f'(x_i),\ \mathrm{i} = 0,\ 1,\ 2\ldots.$$

converges to the root. When k > 1 the above scheme is generalized to:

$$x_{i-1} = x_i - (Df(x_i))^{-1} f(x_i),\ \mathrm{i} = 0,\ 1,\ 2\ldots(*)$$

where $\mathrm{Df}(\mathrm{x_i})$is the n × n Jacobian matrix with elements $\partial f_l/\partial x_{ij}$, l, j = 1, 2,…, k. In this case the root is a vector. Assuming that $\mathrm{Df}(\mathrm{x_i})$is nonsingular. Moreover the iteration scheme (*) can be modified in the form:

$$x_{i+1} = x_i - \lambda(Df(x_i))^{-1} f(x_i),\ \lambda \in (0,1),\ \mathrm{i} = 0,\ 1,\ 2\ldots.$$

Let us settled $\lambda = 0.5$, for this study, as the modification parameter, i.e. a "half-step" of the Newton–Raphson iteration was used to approach the solution, i.e., to solve the normal Eq. (2.7.5). □

Claim II: To minimize SS_2 the following calculation are needed:

$$F_1 = \frac{\partial SS_2}{\partial \theta_1} = -2\sum\sum n_i\left(\bar{y}_{ji.} - \theta_1 e^{\theta_2 u_{ji}}\right)e^{\theta_2 u_{ji}}$$

$$F_2 = \frac{\partial SS_2}{\partial \theta_2} = -2\sum\sum n_i\left(\bar{y}_{ji.} - \theta_1 e^{\theta_2 u_{ji}}\right)\theta_1 u_{ji} e^{\theta_2 u_{ji}}$$

$$n_i = r$$

$$F_1 = -r\sum\sum \bar{y}_{ji.} e^{\theta_2 u_{ji}} + \theta_1 r\sum\sum\left(e^{\theta_2 u_{ji}}\right)^2$$

$$F_2 = -r\theta_1\sum\sum \bar{y}_{ji.} u_{ji} e^{\theta_2 u_{ji}} + \theta_1^2 r\sum\sum u_{ji}\left(e^{\theta_2 u_{ji}}\right)^2$$

The equations $F_1 = 0$ and $F_2 = 0$ are need to be solved, using the Newton–Raphson method. The elements of the Hessian matrix H(i, j) i = 1, 2, j = 1, 2 are:

$$H(1,1) = r\sum\sum\left(e^{\theta_2 u_{ji}}\right)^2$$

$$H(1,2) = -r\sum\sum \bar{y}_{ji} u_{ji} e^{\theta_2 u_{ji}} + 2\theta_1 r\sum\sum u_{ji}\left(e^{\theta_2 u_{ji}}\right)^2$$

$$H(2,1) = -r\sum\sum \bar{y}_{ji} u_{ji} e^{\theta_2 u_{ji}} + 2\theta_1 r\sum\sum u_{ji}\left(e^{\theta_2 u_{ji}}\right)^2$$

$$H(2,2) = -r\theta_1\sum\sum \bar{y}_{ji} u_{ji}^2 e^{\theta_2 u_{ji}} + 2\theta_1^2 r\sum\sum\left(u_{ji} e^{\theta_2 u_{ji}}\right)^2$$

Now, at each stage the estimate $\hat{\theta}_{2S}$ was substituted into the Hessian (see above) when $\hat{\theta}_{S+1} = \left(\hat{\theta}_{1S+1} \cdot \hat{\theta}_{2S+1}\right)$ was to be evaluated, through Newton–Raphson. The average per observation information matrix $M = M(\theta, \xi)$ (recall Example 6) was evaluated at $\theta = \hat{\theta}(S_{max})$, i.e. the estimate at the last stage. Simultaneous and individual approximate confidence intervals were evaluated; recall (6.3.1), through

$$\left(\theta - \hat{\theta}\right)^{T} M(\theta, \xi)\left(\theta - \hat{\theta}\right) \leq 2s^2 F(2, 38; 0, 95) \tag{7.7.3}$$

$$\hat{\theta}_1 \mp 1,96\sqrt{\left(M_{ii}^{-1}(\theta, \xi)s^2\right)}, \quad i = 1, 2 \tag{7.7.4}$$

respectively. The notation: s^2 a suitable estimate of σ^2, i.e. residual sum of squares divined by 38 df, F, as usual, denotes the F distribution, and $M_{ii}^{-1}(.)$ the diagonal elements of $M^{-1}(.)$. Approximate confidence intervals are obtained through (7.7.3) when the design points are predetermined and not obtained sequentially. The coverage probabilities for both $\hat{\theta}_1$ and $\hat{\theta}_2$ individually and jointly were evaluated, see Appendix 5.

The EMSE's (Estimated Mean Square Error) for θ_1 and θ_2 were evaluated, as well as $\log\det M_0$, with M_0 as the right hand side of (2.7.1) with $\theta = \hat{\theta}(S_{max})$. The results are in Tables A5.3–A5.7, in Appendix 5. Table A5.6 provides the measures of efficiency evaluated for θ_2 individually and for the design, i.e. for θ. As such measures it was used

$$\mathrm{Eff}(\theta_i) = \frac{\mathrm{EMSE}(\text{for } \theta_i \text{ in static design})}{\mathrm{EMSE}(\text{for } \theta_i \text{ in design under study})}, \quad i = 1,\ 2 \qquad (7.7.5a)$$

$$\mathrm{Eff}(\theta) = \frac{\log\det \mathrm{M}_0 \text{ of design under study}}{\log\det \mathrm{M}_0 \text{ for static design}} \qquad ((7.7.5b))$$

The results are presented in Appendix 5 Tables 7.7–7.14.

7.8 Discussion II

The logdetM_0, of course, achieves its maximum value, when $\theta_{2F} = \theta_{2T}$, i.e. when $M(\theta, \xi)$ becomes large at the "true" value of the parameter, that is the "local" ellipsoid is minimum at that point, recall Sect. 3.7. The EMSE are, as expected, smaller when $\theta_{2F} = \theta_{2T}$. There is not much difference when the true value is approached either from smaller or larger values.

As far as the coverage probabilities are concerned, on the average, their values are close to 0.95 and all the methods perform well. Among the sequential procedures ($\Pi 3$, $\Pi 4$, $\Pi 5$), the fully sequential procedure, $\Pi 5$, leads to better EMSE. The normality of the vectors of estimates, obtained from 1000 simulations, seems to behave very well. Table A5.8 provides evidence for this as all the kurtosis values are very close to 3 and the skewness values are very close to zero.

The efficiency of the static design for θ_2, when the starting values are "far from the true value" (recall: the true value it is not known) is rather poor. In one-stage design θ_1 is treated as known and there is no chance—as there is other stage—for the estimator to deviate much from its true value, this supports the adoption of the sequential design procedure. Table A5.8 supports the comment that the efficiency in the procedures is getting on the average better in the order $\Pi 1 < \Pi 2 < \Pi 3 < \Pi 4 < \Pi 5$.

Notice that the sequential nature of the design for nonlinear models:

- May often be irrelevant to the manner of obtaining estimators and constructing confidence intervals based on familiar sampling theory methods.
- There are cases were sequential design procedures can result in "tighter" inferences, i.e. shorted confidence intervals. Among them the fully sequential design might provide the tightest inference.

Thus, although the static design for the true θ might be experimentally economical, the absence of knowledge about θ suggests that a sequential procedure should be adopted.

Chapter 8
Optimal Design in Rythmometry

Abstract The "cosinor" model is introduced and the various optimal design for this model are discussed. The application of this model either to Bio-assays, as Rythmometry, or to Engineering is explained.

8.1 Introduction

In this Chapter a specific illustration of a nonlinear design problem is presented. The so-called *cosinor model* (Nelson et al. 1979) has been proposed, at the beginning, as a model for biological time series. An example of such a time series is that of circadian rhythms in normal and asthmatic patients. Recently the model was adopted for an engineering application, see Zarikas et al. (2010). A form of the cosinor model, depending on clock or calendar time, concerning bio-rhythms, has been applied by Hetzel and Clark (1980). Confidence intervals and related statistical analysis, for fitting the nonlinear regression model, has been developed by Nelson et al. (1979).

The problem is studied from the point of view of experimental design: what are the optimum times during the day that the measurements have to be recorded. How many times per day should the measurement take place and how should these times be weighted optimally (i.e. what the design measure is).

Various optimal design procedures are discussed, such as D- and especially c-optimality, from a geometrical point of view, due the Elfving's theorem; see Chap. 3, Fig. 3.2. The efficiencies are compared with the locally optimum design. In the real life data set, originated from an engineering application, a gain in observation has been recorded, Zarikas et al. (2010). As the design depends on time, u will be replaced by t in the sequel. The unit for t is time in days.

C. P. Kitsos, *Optimal Experimental Design for Non-Linear Models*, SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-45287-1_8,

8.2 Background

Some rhythms (either from Biology or from Engineering) can be described by the following cosine model, known as the *cosinor model*:

$$y(t) = \eta(t, \theta) + \varepsilon \quad \text{with } \eta(t, \theta) = \theta_0 + \theta_1 \cos(\omega t + \theta_2) \tag{8.2.1}$$

where:

- y(t) is the response at time t, i.e. the biological/engineering rhythm under consideration
- θ_0 the mesor: The "mean" value about which oscillation occurs
- θ_1 the amplitude: The half difference between the highest and lowest value during the oscillation in a complete cycle (360° or 24 h)
- θ_2 acrophase : Timing of high point in degrees
- ω angular frequency = degrees/unit time. Consider $\omega = 2\pi$ to corresponds to a complete cycle
- ε the error term under Assumption 2, when inference is made and under Assumption 1, when only the design is discussed, $t \in [0, 1]$

The model is illustrated in Fig. 8.1.

From a practical (clinical or engineering) point of view the ratio θ_1/θ_0 is the parameter of interest. This represents the ratio of the amplitude of the cyclic variation to the overall mean. Let us assume a period of one day, i.e. consider $\omega = 2\pi$. The reference point for phase is 0° or 00.00 h since cos0° = 1. Zero time is taken as 00.00 h on the first day, when the study started. It is easy to see, from Fig. 8.1, that the case $\theta_1/\theta_0 < 1$ is the only one which has a physical meaning in the problem, while the dual attracts only mathematical interest. Interest is restricted to the $\theta_1/\theta_0 < 1$ case.

Expanding the cosine term η(t, θ), recall (8.2.1), it is

$$\eta(x, \theta) = \theta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 \tag{8.2.2}$$

$$\text{with } \beta_1 = \theta_1 \cos\theta_2, \quad \beta_2 = -\theta_1 \sin\theta_2, \quad x_0 = 1, \quad x_1 = \cos 2\pi t, \quad x_2 = \sin 2\pi t. \tag{8.2.2a}$$

Therefore model (8.2.1) can be written as

$$y(t) = W^T(t)\beta + \varepsilon, \quad \beta = (\theta_0, \beta_1, \beta_2), \quad W^T(t) = (x_0, x_1, x_2) \tag{8.2.3}$$

When the model (8.2.2) is fitted estimates for θ_1 and θ_2 can be obtained through:

$$\hat{\theta}_1 = \sqrt{(\hat{\beta}_1^2 + \hat{\beta}_2^2)}, \; \hat{\theta}_2 = \hat{\omega} + \kappa \tag{8.2.4}$$

where $\hat{\omega} = \arctan\left|\hat{\beta}_2/\hat{\beta}_1\right|$ and κ is an appropriate constant. The value $2\theta_1$ is the peak to the trough estimate and ω is the estimate of the phase of the rhythm i.e. the time of the computed acrophase. For different values of $\hat{\beta}_1, \hat{\beta}_2$ it is:

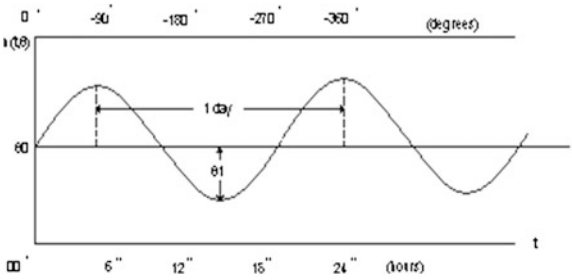

Fig. 8.1 A typical cosine function: $\eta(t, \theta) = \theta_0 + \theta_1 \cos(2\pi t + \theta_2)$

$$
\begin{array}{lll}
\hat{\beta}_1, \hat{\beta}_2 > 0 & \text{then} & \hat{\theta}_2 = -\hat{\omega}(\kappa = 0) \\
\hat{\beta}_1 < 0, \hat{\beta}_2 > 0 & \text{then} & \hat{\theta}_2 = -\pi + \hat{\omega}(\kappa = -\pi) \\
\hat{\beta}_2 < 0, \hat{\beta}_1 < 0 & \text{then} & \hat{\theta}_2 = -\pi - \hat{\omega}(\kappa = -\pi) \\
\hat{\beta}_1 > 0, \hat{\beta}_2 < 0 & \text{then} & \hat{\theta}_2 = -2\pi + \hat{\omega}(\kappa = -2\pi)
\end{array}
$$

In the sequel the design problem will be discussed for a nonlinear function of the parameters of the linearized model (8.2.3).

8.3 D-Optimal Design

For the model (8.2.1) it is known that in clinical or engineering practice an interest is mainly in estimating efficiently the relative stability i.e. the ratio

$$g = g(\theta_0, \theta_1) = \theta_1/\theta_0 \tag{8.3.1}$$

is asked to be estimated as well as possible. Therefore the optimum designs for estimation of g were considered. For the model (8.2.2) the design space, X say, is a circle, defined by

$$x_0 = 1, \quad x_1^2 + x_2^2 = 1. \tag{8.3.2}$$

The centre of the circle is on the x_0 axis at point (1, 0, 0); see Fig. 8.2. With its dual design space [cycle with center (−1, 0, 0)] form a cylinder.

Then it follows (Fedorov 1972, p.75) that the points of the D-optimal design must lie on the given circle. Moreover, any equally-weighted design whose support coincides with the vertices of any regular polygon, inscribed in the circle, is a D-optimal one. For instance a four point equally spaced and equally weighted design will be a D-optimal design. Consider that, in contrast, under c-optimality a two point unequal weighted design will be produced in Sect. 8.4, see for details Kitsos et al. (1988).

For the model (8.2.2) the average per observation information matrix was evaluated for this four point design, mentioned above. It equals (recall (2.5.3) for the discrete case)

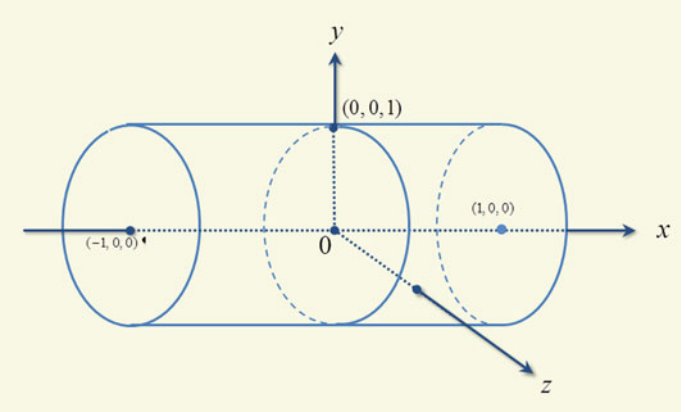

Fig. 8.2 The design space X in 3-dimensions and the −X

$$nM(\xi) = \begin{pmatrix} n & \sum\cos(2\pi t_i) & \sum\sin(2\pi t_i) \\ \sum\cos(2\pi t_i) & \sum\cos^2(2\pi t_i) & \sum\cos(2\pi t_i)\sin(2\pi t_i) \\ \sum\sin(2\pi t_i) & \sum\cos(2\pi t_i)\sin(2\pi t_i) & \sum\sin^2(2\pi t_i) \end{pmatrix} \quad (8.3.3)$$

Take the 4 points to be: t, $t + 1/2$, $t + 1/2$, $t + 3/4$
i.e. in angles correspond to : $2\pi t$, $2\pi t + \pi/2$, $2\pi t + \pi$, $2\pi t + 3\pi/2$.
Let $\delta_i = 2\pi t + T_i, \mathrm{T_i} = 0, \pi/2, \pi, 3\pi/2$. It is easy to see that

$$\sum\cos\delta_i = \sum\sin\delta_i = 0, \quad \sum\cos^2\delta_i = \sum\sin^2\delta_i = 2. \quad (8.3.4)$$

Thus for n observations obtained in n/4 days, (8.3.3) is reduced to

$$nM = \begin{pmatrix} n & 0 & 0 \\ 0 & 2(n/4) & 0 \\ 0 & 0 & 2(n/4) \end{pmatrix} = n \cdot diag\left(1, 1/2, 1/2\right) \quad (8.3.5)$$

Interest is in estimating (8.3.1), written as

$$g = \left(\sqrt{(\beta_1^2 + \beta_2^2)}\right)/\theta_0. \quad (8.3.6)$$

Thus the approximate variance of g is :

$$nVar(\hat{g}) \cong \sigma^2(\nabla g)^T M^{-1}(\nabla g) \quad (8.3.7)$$

where ∇g is the vector of partial derivatives of g(.) and equals to

$$\begin{aligned}(\nabla g)^T &= \left(-\left(\sqrt{(\beta_1^2 + \beta_2^2)}\right)/\theta_0^2, \beta_1/\left(\theta_0\sqrt{(\beta_1^2 + \beta_2^2)}\right), \beta_2/\left(\theta_0\sqrt{(\beta_1^2 + \beta_2^2)}\right)\right) \\ &= \theta_0^{-1}(-\theta_1/\theta_0, \beta_1/\theta_1, \beta_2/\theta_1)\end{aligned} \quad (8.3.8)$$

Substituting (8.3.8) to (8.3.7) the approximate variance, V_4 say, is obtained for an equally-spaced, equally-weighted, 4-point design

$$V_4 = (\sigma^2/(n\theta_0^2))[(\theta_1/\theta_0)^2 + 2]. \tag{8.3.9}$$

Note that from (8.3.7) the problem is approximately equivalent to a locally c-optimal design where "c" is given by (8.3.8).

8.4 c-Optimal Design

For given θ_0, θ_1, θ_2 and therefore β_1, β_2 the locally c-optimal design problem is to

$$\min\{c^T M^{-1}(\xi)c, \quad \xi \in \Xi\} \tag{8.4.1}$$

With $c = (c_0, c_1, c_2)^T = \nabla g$ as in (8.3.8) and $nM(\xi)$ as in (8.3.3) i.e. ξ imposes a measure which puts weight 1/n at $t_1, t_2, \ldots, t_n \in [0,1)$. In principle it is required an optimal design measure ξ on [0, 1) to solve (8.4.1). Elfving (1952) developed a geometrical approach to obtain c-optimal designs, see Sect. 3.7 and Fig. 3.2. Due to this theorem considering the reflection $-X$ of the design space X, then a cylinder is formed through the connection X and $-X$, with the x_0-axis as the x-axis and "direction" or "generation" the circle X, see Fig. 8.3.
The equation of the cylinder C, as in Fig. 8.3, is

$$C = \{(x_0, x_1, x_2): \quad -1 \le x_0 \le 1, x_1^2 + x_2^2 = 1\}. \tag{8.4.2}$$

Moreover any point on the cylinder C is either

(1) On the curved surface (ray r_1, point R_1 in Fig. 8.3) or
(2) On one of the side (ray r_2, point R_2 in Fig. 8.3).

Any ray R can be written

$$R = \{(\Delta c_0, \Delta c_1, \Delta c_2), \quad \Delta > 0\} \tag{8.4.3}$$

for some c_0, c_1, c_2. In particular, the case where c_0, c_1, c_2 is as in (8.3.8) is considered. The ray hits the axis $x_o = 1$ at $\Delta = 1/c_0 = -\theta_0/\theta_1$ therefore the point of intersection is $(1, c_1/c_0, c_2/c_0)$. Two cases (1) and (2) were distinguished as follows:

$$\begin{array}{lll} \text{if} & (c_1^2 + c_2^2)^2/c_0^2 > 1 & \text{then} \quad r_1 \text{ is considered} \\ & < 1 & \text{then} \quad r_2 \text{ is considered} \end{array} \tag{8.4.4}$$

It is easy to verify that

$$(c_1^2 + c_2^2)^2/c_0^2 = (\theta_0/\theta_1)^2. \tag{8.4.5}$$

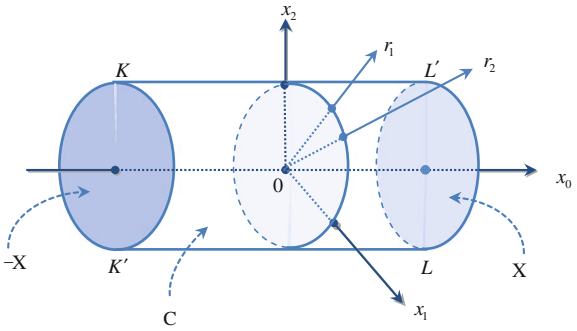

Fig. 8.3 The design space X and its reflection −X form the cylinder C for the (8.2.2)

The geometry of the problem suggests Elfving's theorem in order to tackle the two cases described above.

Case I Consider points such as R_1, i.e. $\theta_1/\theta_0 < 1$.

Thus for $\theta_1/\theta_0 < 1$ it can be proved, see Kitsos et al. (1988), that it should be allocated

$$\begin{aligned} \xi &= 0.5(1-\theta_1/\theta_0) \text{ percentage of observations at point } -\theta_2/2\pi \\ 1-\xi &= 0.5(1+\theta_1/\theta_0) \text{ percentage of observations at point } \pi-\theta_2/2\pi \end{aligned} \tag{8.4.6}$$

For the two point design the corresponding 3×3 average per observation information matrix $M = M(\theta, \xi)$ is singular with rank $r(M) = 2$. Considering the general form of $M(\theta, \xi)$ in (8.3.3), for this particular case, it is easy to verify that under (8.4.6)

$$M(\theta,\xi) = \begin{pmatrix} 1 & (2\xi-1)\cos\theta_2 & -(2\xi-1)\sin\theta_2 \\ (2\xi-1)\cos\theta_2 & \cos^2\theta_2 & -\cos\theta_2\sin\theta_2 \\ -(2\xi-1)\sin\theta_2 & -\cos\theta_2\sin\theta_2 & \sin^2\theta_2 \end{pmatrix} \tag{8.4.7}$$

Substituting ξ in (8.4.7) from (8.4.6), for the optimal design measure $\xi = \xi_2$ say, the average per observation information matrix is obtained as

$$M(\theta,\xi_2) = \begin{pmatrix} 1 & -(\theta_1/\theta_0)\cos\theta_2 & (\theta_1/\theta_0)\sin\theta_2 \\ -(\theta_1/\theta_0)\cos\theta_2 & \cos^2\theta_2 & -\cos\theta_2\sin\theta_2 \\ (\theta_1/\theta_0)\sin\theta_2 & -\cos\theta_2\sin\theta_2 & \sin^2\theta_2 \end{pmatrix}$$

To solve (8.4.1) the generalized inverse $M^-(\theta, \xi_2)$ is needed. Using a matrix result (Appendix 3, (v)) it is for $\left(\theta_2 \neq \frac{\pi}{2}\right)$

$$M^C(\theta,\xi_2) = \begin{pmatrix} \cos^2\theta_2 & (\theta_1/\theta_0)\cos\theta_2 & 0 \\ (\theta_1/\theta_0)\cos\theta_2 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \frac{1}{\cos^2\theta_2\left(1-\frac{\theta_1{}^2}{\theta_0^2}\right)} \tag{8.4.8}$$

Hence for $\xi = \xi_2$ and

$$c^T = (\nabla g)^T = (1/\theta_0)(-\theta_1/\theta_0, \beta_1/\theta_1, \beta_2/\theta_1) = (1/\theta_0)(-\theta_1/\theta_0, \cos\theta_2, -\sin\theta_2)$$

it is

$$c^{\mathrm{T}}\mathbf{M}^{-}(\theta, \xi)c = (1/\theta_0)^2 \tag{8.4.9}$$

Therefore the approximate variance $V_{2,}$ for the two point c-optimal design is

$$V_2 = Var(c^T\hat{\theta}) = (1/\theta_0)^2\sigma^2/n, \quad \theta_1/\theta_0 < 1 \tag{8.4.10}$$

Case II Consider points such as R_2. i.e. $\theta_1/\theta_0 > 1$.

This case has a limited practical application and has been investigated for theoretical fulfillment, see Kitsos et al. (1988).

8.5 Restricted Design Space

One important practical difficulty related with the optimal designs of Sect. 8.4 is that they require measurements to be made when the response function is maximum and minimum. The latter typically occurs, at least in biological rhythms, in the early hours of the morning.

It might be desirable for the design to be restricted to more social hours, i.e. avoid taking measurements during the night. The restriction of the design is considered to a portion $1 - T$, say of the day, where T is the length of the nighttime period, e.g. 11 pm till 7 am, see Fig. 8.4. Moreover it is assumed that the minimum of the response function occurs at the middle of T and the maximum, in $1 - T$, occurs at the middle of this interval. Any design depends on θ_1/θ_0 and θ_2. Moreover the restriction on time means that the new design space, X_N, say, is no longer a circle and hence the idea of a full cylinder is no any longer available. The cylinder will be "truncated", as shown in Fig. 8.5.

After some algebra, see Kitsos et al. (1988) and through the quantity q, given bellow

$$q = (1/3\theta_0)^2\left[(4\kappa+1)^2(1-\xi) + (2-\kappa)^2\xi\right]/[\xi(1-\xi)\Big], \quad \kappa = \theta_0/\theta_1 \tag{8.5.1}$$

the approximate variance for the three point optimally-weighted truncated design is

$$V_3 = Var(c^T\hat{\theta}) = q\sigma^2/n \tag{8.5.2}$$

The corresponding design weight ξ_3 can be evaluated with $T = 1/3$ as

$$\xi_3 = [0.5(\kappa+1)]/[0.25+\kappa] \tag{8.5.3}$$

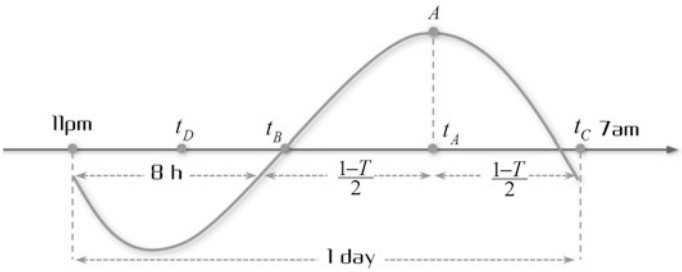

Fig. 8.4 A typical situation when the rhythm is "going down" during the night

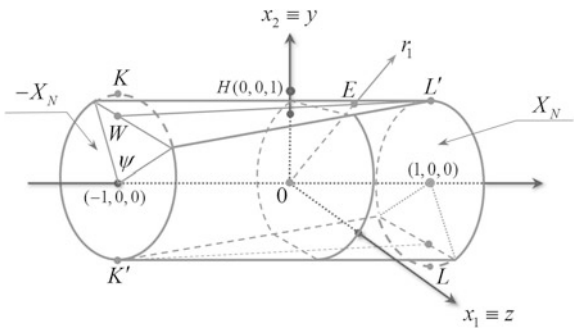

Fig. 8.5 The truncated space X_N and its corresponded truncated cylinder

Thus the design measure still depends on the ratio $\theta_1\theta_0$, the unknown quantity one is trying to estimate. If it is desired to construct an equally-weighted, 3 point design, in this truncated case, the corresponding approximate variance it would be

$$V_3^* = (2/3\theta_o^2)\left[2(\kappa+1)^2 + (2-\kappa)^2\right]\sigma^2/n \tag{8.5.4}$$

In the next paragraph a synopsis is presented of the above results and the efficiencies of the designs are evaluated.

8.6 Synopsis

Let us summarize the results obtained in previous sections for the optimal design for estimating the function $g = \theta_o/\theta_1$ of the parameters θ_o, θ_1 of the nonlinear model $n(t,\theta) = \theta_o + \theta_1\cos(2\pi t + \theta_2)$. The efficiency results are tabulated in Table 8.1. For the results evaluated in Sects. 8.3–8.5 efficiencies can be obtained. For the un-truncated case, the four point and the two point designs are compared

$$E_{2,4} = V_2/V_4 = 1/[(\theta_1/\theta_o)^2 + 2] \tag{8.6.1}$$

The truncated design compared to the un-truncated unequally weighted design gives efficiency equals to

Table 8.1 Efficiencies when $\theta_1/\theta_o < 1$ for the proposed designs

θ_1/θ_o	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$E_{2,4}$	0.50	0.49	0.48	0.46	0.44	0.42	0.40	0.37	0.35
$E_{2,3}$	0.53	0.51	0.48	0.46	0.44	0.42	0.40	0.39	0.37
$E_{2,3}$	0.49	0.49	0.48	0.46	0.44	0.42	0.40	0.37	0.35

$$E_{2,3} = V_2/V_3 = [9\xi_3(1-\xi_3)]/[\kappa_1(1-\xi_3) + \kappa_2\xi_3] \tag{8.6.2}$$

With: $\kappa_1 = (\theta_1/\theta_o + 1)^2$, $\kappa_2 = (2 - \theta_1/\theta_o)^2$, ξ_3 as in (8.5.3)

For the equally weighted truncated design compared with the two point design it can be proved that

$$E^*_{2,3} = V_2/V^*_3 = 3\left[2(\theta_1/\theta_o)^2 + (2-\theta_1/\theta_o)^2\right]^{-1} \tag{8.6.3}$$

The efficiencies and the design measures have been evaluated for different values of $\theta_1/\theta_o < 1$, given in Tables 8.1 and 8.2. It is interesting to notice that truncation does not greatly influence the nature of the design. Thus, an equally weighted truncated design can be recommended. Based on previous experience in biorhythms, the ratio θ_1/θ_o does not exceed 0.3. In principle in the nonlinear case, an optimal design depends on the parameters it is planning to estimate. In Table 8.1 this dependence is reflected. However, for small values of θ_1/θ_o, there is a little difference between the optimal designs. Thus it is recommended to choose D- or c-optimality, and provide a guess for θ_1/θ_o, and then continue the performance of the design as in Chap. 4.

8.7 Engineering Application

The cosinor optimal experimental design was suitably adopted for the Robotic Total Stations (RTS). Operating in a tracking mode, Zarikas et al. (2010) applied the statistical model discussed above. RTS or Robotic Theodolite is a new generation topographic instrument, which is capable of recording the coordinates of a moving target (reflector) through 3D space. In recent years, with the advent of technology, modern RTS are used complementary to other systems (such as GPS and accelerometers), for measuring the displacements of civil engineering structures that exhibit a cyclic pattern of motion.

Optimal Experimental Design techniques were adopted and demonstrated suitably the c- or D- optimal design. Therefore a fully exploit of the potential RTS as far as the computation of the amplitude of displacements when a sinusoidal fit was assumed. The experimental design was employed for the generation of a series of fully controlled sinusoidal oscillations of the known modal characteristics. The comparisons between two different, but equal sample sizes, was mainly investigated. The first sample consists in data measurements consecutively distributed

Table 8.2 Evaluating the design measures when $\theta_1/\theta_o < 1$ for D- and c-optimal designs

Design	θ_1/θ_o	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1.	ξ_i	0.25 for all i = 1, 2, 3, 4 and for every θ_1/θ_o								
2.	ξ_1	0.45	0.40	0.35	0.30	0.25	0.20	0.15	0.10	0.05
	ξ_2	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95

along the design space. The second sample consists in data following the optimal design criteria. The proposed c- or D-optimal design provides the estimation method of the amplitude of sine or cosine waves, with the minimum possible dispersion. The problems are

(a) Where is the optimal design point located, during a certain period and
(b) What is the number of observations allocating to the optimal design points.

Given the sinusoidal form of oscillation adopted in the experiments, the model of motion reads

$$Y(t) = A_0 + A_1 \sin(2\pi ft) + e \text{ or } Y(t) = A_0 + A_1 \cos(2\pi ft - \pi/2) \tag{8.7.1}$$

In Eq. (8.7.1), A_0, A_1 and f denote the mean value of Y(t), the amplitude and the frequency of oscillation respectively. Following the discussed above method, $A_1 < A_0$ and therefore, the c-optimal design criterion suggests that N measurements should be obtained with the following weights,

$$\begin{array}{ll} \frac{N}{2}\left(1 - \frac{A_1}{A_0}\right) & \text{Measurements at} \quad t_1 = \frac{\pi}{2} \\ \frac{N}{2}\left(1 + \frac{A_1}{A_0}\right) & \text{Measurements at} \quad t_2 = \frac{\pi}{2} + \pi \end{array} \tag{8.7.2}$$

Alternatively, one might choose the optimal design points at time locations $t_1 = \theta$ and $t_2 = \theta + \pi$. Equivalently, for convenience reasons one may perform for antipodal observations every cycle.

Two estimates of the variance of the effective amplitude A_1/A_0 were produced based on two equally sized data samples. As it is mentioned the first data sample was built up to adhere the optimal design criteria, whereas the second one was set up to contain a continuous stream of consecutively measurements of the raw data. In order to facilitate comparisons, the non linear regression model (8.7.1) was adopted. At an implementation stage any statistical or mathematical software package can be used.

Analysis of the results reveals a reduction (of the order of 3–10 %) in the variance of the effective amplitude if the optimal design method is used. More importantly, this is irrespective of the dataset used. From the analysis of the results a number of points become evident. The degree of reduction in the amplitude variance is practically insusceptible to changes in the amplitude of oscillation.

In contrast, the degree of reduction in the amplitude variance becomes more evident as the oscillation frequency increases. This finding is considered to be important, as at higher oscillation frequencies a smaller number of measurement

recordings contribute per oscillation cycle, and therefore, the benefit of using the proposed optimal design it becomes more apparent. Furthermore, it can be concluded that the significance of applying the optimal design is fully exploited in situations were functioning an analytical instrument to its operational limits and for datasets with large dispersion values.

References

Elfving G (1952) Optimum allocation in linear regression theory. Ann Math Stat 23:255–262

Fedorov VV (1972) Theory of optimal experiments. Academic Press, New York

Hetzel MR, Clark TJH (1980) Comparison of normal and asthmatic circadian rhythms in peak expiratory flow rate. Thorax 35:732–738

Kitsos CP, Titterington DM, Torsney B (1988) An optimal design problem in rhythmometry. Biometrics 44:657–671

Nelson W, Tong YL, Lee JK, Halberg F (1979) Methods for cosinor-rhythmometry. Chronobiologia 6:305–323

Zarikas V, Gikas V, Kitsos CP (2010) Evaluation of the optimal design cosinor model for enhancing the potential of robotic theodolite kinematic observation. Measurement 43(10): 1416–1424

Chapter 9
Discussion

Abstract An extensive discussion of the provided methodology and applications is provided, to help the reader to realize the importance of this particular field and encourage him to adopt it in practice.

This monograph aims to develop a solid and compact theoretical insight of the Nonlinear Optimal Experimental Design problem and to provide food for thought on how to apply the discussed theory on real life cases. That is why a number of examples were extensively discussed. Relevant chemical kinetic models were studied, avoiding chemical and statistical technicalities. The three simulations were completely analyzed and discussed. The objective was to construct a design that eventually estimates the unknown parameter vector θ, as adequate as possible. Thus there is no reference to designs for discriminating two rival nonlinear-models, see Atkinson and Federov (1975), or to any other kind of designs.

Needless to say, the theoretical insight of the nonlinear case is broader than for the linear one. The theoretical approaches developed over time do not focus on reduction from nonlinear to linear. A solid statistical background for the linear case, either continuous or discrete, has been created. This theoretical insight was used in the NLED application to extend the nonlinear case. The NLED is heavily dependent on (a number of) the involved parameters.

Notice than the Up and Down (UD) and Stochastic Approximation (SA) methods tackled in Sects. 5.4 and 5.5 have the following common characteristics:

1. They deal with the fully sequential method of design.
2. They have a nonparametric flavor.
3. They were developed to estimate the parameter of interest, usually a single one and not a subset of several parameters.

The fully sequential nature is obvious as one observation is made at each stage. Moreover both the UD and SA methods are "Markovian" in the sense that the choice of each run depends only on the current situation. The nonparametric feature of the discrete (binary) methods is based on the fact that a functional of the unknown response $T(.)$ was estimated, usually the p-th percentile L_p.

C. P. Kitsos, *Optimal Experimental Design for Non-Linear Models*,
SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-45287-1_9,

The assumption made about the cumulative distribution function (cdf) $T(.)$ is that it is either normal or logistic, leading to probit or logit analysis. The virtue of the logit model lies in its simplicity and approximation to the normal distribution in the range $p \in [0.2, 0.8]$. Moreover the canonical form discussed regarding the binary models is essential. This can be helpful when working within a group of transformations: movement from orbit to orbit can keep invariance under Affine Geometry principles.

Geometry can be really very helpful to understand the optimality criteria: D-optimality minimizes the volume of the confidence ellipsoidal, and remains invariant to linear transformation. This is not the case for A-optimality, since the confidence ellipsoidal volume might remain constant, but not the axes of the ellipsoidal. The c-optimality criterion is based on Elfving's geometrical-oriented theorem.

In the linear case, the sequential nature of the design is based on the augmentation of the data rather than an approach to estimate the parameters. The optimal design point are "well defined" in the linear case. In the nonlinear case the optimal design points depend on the parameters we want to estimate, and therefore are not well defined. Therefore in such a case the locally optimal design is defined, declaring this parameter dependence.

As far as the simulations are concerned, the first one referred, eventually, to a nonlinear problem as a ratio estimate was discussed through the first order autoregressive model, so by its "nature" was sequential. The dilution assessment was a different type of problem—binary or discrete—which could be faced either sequentially or as a static design. Various approaches were discussed, to provide evidence that "a design of the experiment" is necessary. Consequently, a strategy has to be followed and not just to perform an experiment many times "to see how the results turn out". There is a need to investigate an adequate statistical procedure, providing the best estimate of the parameter under investigation. In fact, the discrete case of the linear case is a rather complicated theoretical approach, which has much in common with Number Theory and Combinatorics.

The last simulation was about a two-parameter nonlinear model, from the continuous case. As in almost all cases, initial information is not needed for all the parameters involved. The fully sequential approach was developed based on two observations, as the number of the parameters at each stage. Truncation was applied and might prove helpful in real situations.

The rhythmometry problem shows that an application might have a strong mathematical insight: Geometry and Analysis were used to evaluate the D- and c-optimal design points for this particular problem. Although it was first considered as a medical application, it was also applied to engineering research, proving that the lines of thought for fields that might be considered disjoined are not that different!

All the practical investigations were based on the developed theory, but actually the sample size did not reach infinity (!). Even with small sample sizes approximated confidence intervals were obtained. Moreover the models found in the bibliography were mainly focused on no more than four parameters; five

parameters is the worst case scenario. Certainly the theoretical development for "any p-term vector of parameters" goes beyond the real situation.

The theoretical approach makes researchers hesitate to follow it, but practice shows that one can adopt the theory. The only problem seems to be the lack of communication; neither the experimenter nor the statistician invites anyone to "his own cloud"!

Acknowledgments I would like to thank Assistant Prof. V. Zarikas (Lamia), and part time Assistant Profs. T. Toulias, Y. Hatzikian and K. Kolovos (Athens), for useful discussions during the preparation of this monograph. I would like to thank the reviewers for their time working with the text. They made very sharp and useful comments, which are very much appreciated. The encouragement of Springer, BioMathematics/Statistics, is very much appreciated.

Reference

Atkinson AC, Fedorov VV (1975) The design of experiments for discriminating between two rival models. Biometrika 62:57–70

Appendix 1

A1.1 Caratheodory's Theorem

Definition 1
The set S is called convex if all points $s \in S$ of the form

$$s = \alpha \cdot s_1 + (1 - \alpha) \cdot s_2, \quad s_1, s_2 \in S, \quad \alpha \in [0, 1]$$

are elements of S.

Definition 2
The set of points, S^* say, with elements

$$s^* = \sum a_i s_i, \quad \sum a_i = 1, \quad a_i \in [0, 1], \quad s_i \in S$$

is a convex set. S^* is called the convex hull of the set S.

Theorem (Caratheodory)

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

$$s^* = \sum_{1}^{n+1} a_i s_i, \quad a_i \geq 0, \quad s_i \in s, \quad \sum_{1}^{n+1} a_i = 1$$

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

C. P. Kitsos, *Optimal Experimental Design for Non-Linear Models*,
SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-45287-1,

Appendix 2

A2.1 Silvapulle's Theorem

Silvapulle (1981) stated and proved the conditions under which the existence of the MLE in binary problems is guaranteed.

Let $u_1, u_2, \ldots, u_r$ be the design points corresponding to responses $y_i = 1$, $i = 1, \ldots, r$ and $u_{r+1}, \ldots, u_n$ corresponding to responses $y_i = 0$, $i = r+1, \ldots, n$. Consider the convex cones:

$$S = \{\Sigma \kappa_i u_i, \kappa_i \geq 0. \forall i = 1, \ldots, r\}$$
$$F = \{\Sigma \kappa_j u_j, \kappa_j \geq 0. \forall j = r+1, \ldots, n\}$$

Then the following theorem holds.

Theorem
Let the condition (L) be defined by

$$(L)\, S \cap F \neq \varnothing \text{ or one of S or F is } R^P \supseteq \Theta$$

Then for the binomial response model:

$$\text{Prob}(y_i = 1) = T(u_i^T \theta)$$

1. The MLE $\hat{\theta}$ of θ exists and the minimum set $\{\theta\}$ is bounded only when (L) is satisfied.
2. Suppose that $\ell(\theta) = -\Sigma \log T(u_i \theta) - \Sigma \log(1 - T(u_i^T \theta))$ is a proper closed convex function on R^P. Then the MLE $\hat{\theta}$ exists and the minimum set $\{\hat{\theta}\}$ is bounded if and only if (L) is satisfied.
3. Suppose that $-\log T$ and $\log(1 - T)$ are convex and $u_{1i} = 1$ for every i. Then $\hat{\theta}$ exists and the minimum set $\{\hat{\theta}\}$ is bounded if and only if $S \cap F \neq \varnothing$. Let us further assume that T is strictly increasing at every t satisfying

   $0 \langle T(t) \langle 1$. Then θ is uniquely defined if and only if $S \cap F = \varnothing$.

C. P. Kitsos, *Optimal Experimental Design for Non-Linear Models*,
SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-45287-1,

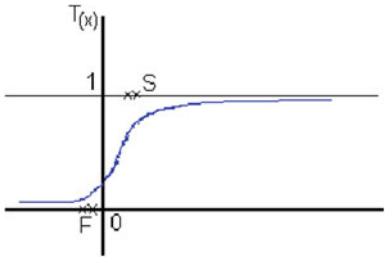

Fig. A2.1 No intersection between the set S and the set F

An example where the MLE does not exist is considered in Figure A2.1, where there is no "inter-blocking" condition between the sets S (set of successes) and F (set of failures).

Appendix 3

A3.1 Conditional Inverse

Definition
Let $A \in Mat(m, n)$. A^c is conditional inverse iff $AA^cA=A$

- The generalized inverse is also a conditional inverse. The opposite not necessarily true. The main Properties are:

1. $rank(A^c) \geq rank(A)$
2. $rank(A^c\ A) = rank(AA^c) = rank(A) = tr(AA^c) = tr(A^cA)$
3. AA^c, A^cA are idempotent matrices
4. $A^cA = I \Rightarrow rank(A) = n$, $AA^c = I \Rightarrow rank(A) = m$
5. If A of rank r is partitioned as

$$A = \begin{pmatrix} B & C \\ D & E \end{pmatrix} \quad \text{then} \quad A^c = \begin{pmatrix} B^{-1} & 0 \\ 0 & 0 \end{pmatrix}, \text{ with B of rank r.}$$

C. P. Kitsos, *Optimal Experimental Design for Non-Linear Models*,
SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-45287-1,

Appendix 4

A4.1 Simulation I Results

Tables A4.1, A4.2, A4.3, A4.4, A4.5

Table A4.1 Simulation study I, procedure P1

N	θ_{st}	ECP	$\overline{\theta}$	S	K	EMSE
1,000	2.196	0.953	3.18	0.23	2.97	0.018
	3.18	0.954	3.19	0.02	2.97	0.015
	7.15	0.958	3.18	0.05	2.93	0.019
100	2.196	0.950	3.24	0.59	3.98	0.204
	3.18	0.952	3.2	0.53	3.47	0.172
	7.15	0.952	3.0	0.17	3.29	0.201
50	2.196	0.977	3.31	1.52	5.09	0.454*
	3.18	0.958	3.28	0.56	3.43	0.365
	7.15	0.935	3.22	0.38	3.27	0.462

*Two "outliers" were not considered, therefore N=998.

Table A4.2 Simulation study I, procedure P2

N	θ_{st}	r	ECP	$\overline{\theta}$	S	K	EMSE
100	2.196	5	0.937	3.23	0.55	3.86	0.20
		25	0.946	3.24	0.66	3.85	0.20
		50	0.945	3.25	1.64	8.46	0.22*
	3.18	5	0.946	3.26	0.51	3.63	0.16
		25	0.950	3.19	0.45	3.72	0.16
		50	0.950	3.20	0.42	3.35	0.15
	7.15	5	0.955	3.23	0.36	2.83	0.17
		25	0.946	3.23	0.36	3.09	0.19
		50	0.955	3.24	0.36	2.96	0.20
50	2.196	5	0.954	3.28	0.83	4.32	0.43
		25	0.962	3.46	3.12	12.16	1.14#
	3.18	5	0.952	3.26	0.77	3.78	0.39

(continued)

C. P. Kitsos, *Optimal Experimental Design for Non-Linear Models*, SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-45287-1,

Table A4.2 (continued)

N	θ_{st}	r	ECP	$\overline{\theta}$	S	K	EMSE
		25	0.948	3.24	0.99	4.32	0.35[#]
	7.15	5	0.947	3.29	0.68	3.64	0.42
		25	0.940	3.28	0.57	3.46	0.43

*Two "outliners" were not considered, therefore N=998 (see Appendix A2.II).
[#] Three "outliners" were not considered, therefore N=997.(see Appendix A2.II)

Table A4.3 Simulation study I, procedure P3.

N	θ_{st}	r	ECP	$\overline{\theta}$	S	K	EMSE
100	2.196	5*	0.937	3.23	0.55	3.86	0.20
		25	0.944	3.21	0.51	3.43	0.19
		50	0.948	3.23	0.75	3.97	0.20
	3.18	5*	0.946	3.20	0.51	3.63	0.16
		25	0.947	3.20	0.51	3.28	0.17
		50	0.961	3.20	0.46	3.31	0.15
	7.15	5*	0.955	3.23	0.36	2.83	0.17
		25	0.947	3.20	0.44	3.36	0.17
		50	0.949	3.19	0.36	3.21	0.18
50	2.196	5*	0.954	3.28	0.83	4.32	0.43
		25	0.942	3.21	1.09	1.28	0.51
	3.18	5*	0.952	3.26	0.77	3.78	0.39
		25	0.953	3.26	0.88	5.43	0.38
	7,15	5*	0.947	3.29	0.68	3.64	0.42
		25	0.948	3.26	0.43	2.98	0.39

*From Table A4.2 (r=r'=5).

Table A4.4 Simulation study I, procedure P4.

N	θ_{st}	r	ECP	$\overline{\theta}$	S	K	EMSE
100	2.196	5	0.955	3.21	0.45	3.53	0.16
		25	0.959	3.23	0.48	3.21	0.18
		50	0.957	3.23	0.60	3.83	0.19
	3.18	5	0.960	3.21	0.46	3.33	0.15
		25	0.952	3.21	0.75	4.37	0.16
		50	0.962	3.20	0.59	3.84	0.15
	7.15	5	0.953	3.24	0.32	2.77	0.17
		25	0.956	3.24	0.41	3.23	0.19
		50	0.946	3.22	0.37	3.25	0.20
75	2.196	5	0.943	3.22	0.64	3.74	0.24

(continued)

Table A4.4 (continued)

N	θ_{st}	r	ECP	$\overline{\theta}$	S	K	EMSE
		25	0.951	3.24	0.57	3.22	0.25
		50	0.954	3.25	0.96	5.51	0.28
	3.18	5	0.948	3.23	0.60	3.65	0.23
		25	0.945	3.23	0.82	4.45	0.23
		50	0.955	3.23	0.68	3.80	0.22
	7.15	5	0.958	3.25	0.38	3.05	0.23
		25	0.953	3.26	0.49	3.38	0.26
		50	0.941	3.21	0.43	3.43	0.28
50	2.196	5	0.956	3.26	0.67	3.53	0.37
		25	0.948	3.30	0.92	4.28	0.46
	3.18	5	0.949	3.24	0.89	3.89	0.33[#]
		25	0.946	3.24	0.98	4.04	0.34*
	7.15	5	0.948	3.26	0.69	4.02	0.38
		25	0.948	3.29	0.46	3.14	0.44

[#] N = 997, *N = 998

Table A4.5 Simulation Study I, Procedure P5.

N	θ_{st}	r	ECP	$\overline{\theta}$	S	K	EMSE
100	2.196	5	0.961	3.24	0.58	3.90	0.17
		25	0.946	3.23	0.44	3.17	0.18
		50	0.956	3.23	0.41	3.43	0.17
	3.18	5	0.967	3.23	0.37	3.17	0.15
		25	0.958	3.21	0.32	3.17	0.15
		50	0.954	3.21	0.44	3.55	0.16
	7.15	5	0.958	3.28	0.58	3.47	0.17
		25	0.952	3.23	0.57	4.05	0.18
		50	0.954	3.21	0.20	3.06	0.18
75	2.196	5	0.955	3.27	0.64	3.78	0.25
		25	0.947	3.24	0.55	3.46	0.24
		50	0.976	3.33	0.86	3.90	0.46*
	3.18	5	0.961	3.24	0.59	3.64	0.22
		25	0.955	3.22	0.54	3.66	0.21
		50	0.955	3.25	0.82	4.43	0.37
	7.15	5	0.951	3.24	0.65	3.80	0.23
		25	0.951	3.25	0.60	3.86	0.25
		50	0.951	3.22	0.24	3.15	0.23
50	2.196	5	0.959	3.31	0.70	3.65	0.40
		25	0.957	3.28	0.82	4.61	0.41
	3.18	5	0.950	3.26	0.78	4.11	0.37
		25	0.947	3.25	0.73	4.24	0.36

(continued)

Table A4.5 (continued)

N	θ_{st}	r	ECP	$\overline{\theta}$	S	K	EMSE
	7.15	5	0.955	3.28	0.58	3.47	0.37
		25	0.958	3.28	0.82	4.52	0.41

*N = 999

Appendix 5

A5.1 Simulation II Results

Tables A5.1, A5.2, A5.3, A5.4, A5.5, A5.6, A5.7, A5.8

Table A5.1 Simulation study II, procedure Π1

θ_{2T}	θ_{2F}	Coverage probabilities			EMSE		Ln det M_0
		θ_1 and θ_2	θ_1	θ_2	θ_1	θ_2	
1.0	1.0	0.943	0.961	0.954	$50^{(.)}$	$57^{(*)}$	5.470
	3.0	0.954	0.955	0.953	138	168	5.096
2.0	2.0	0.942	0.955	0.954	26	28	6.606
	4.0	0.945	0.952	0.946	50	56	6.437
3.0	1.0	0.947	0.958	0.958	46	46	7.201
	3.0	0.961	0.956	0.954	8.4	9	7.991
	5.0	0.943	0.942	0.941	12	13	7.895
4.0	2.0	0.957	0.968	0.967	3.4	3	9.211
	4.0	0.947	0.940	0.940	2.2	2.3	9.478
	6.0	0.953	0.953	0.954	2.7	2.3	9.415

$^{(.)}$ All values $\times 10^{-3}$ $^{(*)}$ All values $\times 10^{-5}$

Table A5.2 Simulation study II, procedure Π2

θ_{2T}	θ_{2F}	Coverage probabilities			EMSE		ln det M_0
		θ_1 and θ_2	θ_1	θ_2	θ_1	θ_2	
1.0	1.0	0.955	0.962	0.959	$47^{(.)}$	$53^{(*)}$	5.470
	3.0	0.945	0.952	0.954	73	88	5.355
2.0	2.0	0.943	0.939	0.942	30	32	6.606
	4.0	0.945	0.950	0.950	35	40	6.541
3.0	1.0	0.953	0.949	0.949	14	14	7.758
	3.0	0.946	0.954	0.947	9	9.5	7.990
	5.0	0.954	0.959	0.960	10	10	7.950

(continued)

C. P. Kitsos, *Optimal Experimental Design for Non-Linear Models*, SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-45287-1,

Table A5.2 (continued)

θ_{2T}	θ_{2F}	Coverage probabilities			EMSE		ln det M_0
		θ_1 and θ_2	θ_1	θ_2	θ_1	θ_2	
4.0	2.0	0.957	0.968	0.971	2.4	2	9.367
	4.0	0.952	0.943	0.942	2.2	2.3	9.478
	6.0	0.949	0.968	0.969	2.2	2	9.451

(.) All values $\times 10^{-3}$ (*) All values $\times 10^{-5}$

Table A5.3 Simulation study II, procedure Π3

θ_{2T}	θ_{2F}	Coverage probabilities			EMSE		ln det M_0
		θ_1 and θ_2	θ_1	θ_2	θ_1	θ_2	
1.0	1.0	0.945	0.949	0.952	50(.)	55(*)	5.470
	3.0	0.944	0.949	0.950	56	65	5.435
2.0	2.0	0.953	0.956	0.956	27	29	6.605
	4.0	0.948	0.944	0.946	31	34	6.584
3.0	1.0	0.941	0.945	0.943	10	10	7.913
	3.0	0.950	0.938	0.942	9	9.5	7.991
	5.0	0.946	0.945	0.944	9.4	10	7.977
4.0	2.0	0.947	0.943	0.945	2.3	2.3	9.438
	4.0	0.957	0.960	0.960	2	2.1	9.478
	6.0	0.951	0.953	0.954	2.1	2.4	9.468

(.) All values $\times 10^{-3}$ (*) All values $\times 10^{-5}$

Table A5.4 Simulation study II, procedure Π4

θ_{2T}	θ_{2F}	Coverage probabilities			EMSE		ln det M_0
		θ_1 and θ_2	θ_1	θ_2	θ_1	θ_2	
1.0	1.0	0.950	0.941	0.945	52(.)	60(*)	5.470
	3.0	0.948	0.938	0.939	56	63	5.453
2.0	2.0	0.955	0.962	0.960	26	28	6.606
	4.0	0.942	0.943	0.943	30	32	6.595
3.0	1.0	0.948	0.941	0.946	10	10	7.953
	3.0	0.942	0.940	0.938	10	10	7.990
	5.0	0.957	0.964	0.965	8	8	7.984
4.0	2.0	0.953	0.957	0.957	2	2	9.458
	4.0	0.950	0.951	0.950	2	2.2	9.478
	6.0	0.955	0.958	0.960	2	2	9.473

(.) All values $\times 10^{-3}$ (*) All values $\times 10^{-5}$

Table A5.5 Simulation study II, procedure Π5

θ_{2T}	θ_{2T}	Coverage probabilities			EMSE		ln det M_0
		θ_1 and θ_2	θ_1	θ_2	θ_1	θ_2	
1.0	1.0	0.944	0.955	0.953	$46^{(.)}$	$53^{(*)}$	5.471
	3.0	0.954	0.950	0.955	52	59	5.463
2.0	2.0	0.936	0.931	0.935	30	32	6.605
	4.0	0.946	0.954	0.951	27	30	6.601
3.0	1.0	0.946	0.944	0.946	9.5	10	7.973
	3.0	0.956	0.951	0.950	8.7	9.4	7.990
	5.0	0.946	0.944	0.946	9	9	7.987
4.0	2.0	0.957	0.947	0.945	2.2	2	9.468
	4.0	0.941	0.944	0.940	2.4	2.5	9.478
	6.0	0.958	0.970	0.971	1.9	2	9.476

$^{(.)}$ All values $\times 10^{-3}$ $^{(*)}$ All values $\times 10^{-5}$

Table A5.6 Simulation study II, skewness and kurtosis for parameter θ_2

Procedure		Π1		Π2		Π3		Π4		Π5	
θ_{2T}	θ_{2F}	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
1.0	1.0	−0.001	2.91	0.015	2.67	0.074	2.99	0.042	2.88	0.020	3.16
	3.0	−0.002	3.01	0.092	3.05	0.125	3.29	−0.066	3.00	0.092	2.87
2.0	2.0	−0.012	2.98	−0.023	3.17	0.005	2.60	−0.008	2.83	0.083	3.24
	4.0	0.001	3.16	0.014	2.83	−0.101	2.85	0.019	2.58	0.020	3.17
3.0	1.0	0.097	3.40	0.020	2.96	−0.135	2.91	0.101	2.85	0.020	2.86
	3.0	−0.052	2.97	−0.039	2.94	−0.045	2.84	−0.043	3.09	0.019	2.69
	5.0	0.044	2.74	0.014	3.04	0.067	2.86	0.009	2.90	0.035	3.06
4.0	2.0	−0.034	2.88	0.034	2.76	0.038	2.94	−0.195	2.99	0.089	2.88
	4.0	−0.013	2.90	−0.001	3.20	0.064	2.66	0.054	3.13	0.010	2.97
	6.0	0.019	3.09	−0.064	2.99	–0.005	2.89	−0.194	2.98	−0.109	2.94

(1) Skewness θ_2, (2) Kurtosis θ_2

Table A5.7 Simulation study II, evaluating efficiencies for θ_1

Procedure		Π_1	Π_2	Π_3	Π_4	Π_5
θ_{2T}	θ_{2F}					
1.0	1.0	1.0	1.06	1.0	0.96	1.08
	3.0	0.36	0.68	0.89	0.89	0.96
2.0	2.0	1.0	0.86	0.96	1.0	0.81
	4.0	0.5	0.74	0.83	0.86	0.96
3.0	1.0	0.20	0.60	0.84	0.84	0.88
	3.0	1.0	0.93	0.93	0.84	0.96
	5.0	0.75	0.84	0.89	1.05	0.93
4.0	2.0	0.64	0.91	0.95	1.10	1.0
	4.0	1.0	1.0	1.10	1.10	0.91
	6.0	0.81	1.0	1.04	1.10	1.15

Table A5.8 Simulation study II, efficiencies for θ_2 and $\theta = (\theta_1, \theta_2)$

Procedure		Π1		Π2		Π3		Π4		Π5	
θ_{2T}	θ_{2F}	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
1.0	1.0	1.0	1.0	1.0	1.0	1.03	1.0	0.95	1.0	1.07	1.0
	3.0	0.34	0.93	0.64	0.98	0.88	0.99	0.90	0.99	0.97	0.99
2.0	2.0	1.0	1.0	0.87	1.0	0.96	1.0	1.0	1.0	0.87	1.0
	4.0	0.50	0.97	0.70	0.99	0.82	0.99	0.88	0.98	0.93	0.99
3.0	1.0	0.20	0.90	0.64	0.97	0.90	0.99	0.90	0.99	0.90	0.99
	3.0	1.0	1.0	0.95	1.0	0.95	1.0	0.90	1.0	0.96	1.0
	5.0	0.69	0.99	0.90	0.99	0.90	1.0	1.12	0.99	1.0	0.99
4.0	2.0	0.76	0.97	1.15	0.99	1.0	0.99	1.15	0.99	1.15	0.99
	4.0	1.0	1.0	1.0	1.0	1.09	1.0	1.04	1.0	0.92	1.0
	6.0	1.0	0.99	1.15	1.0	0.96	0.99	1.15	0.99	1.15	1.0

(1) Efficiency for θ_2 recall (7.7.5a), (2) Efficiency for θ_2 recall (7.7.5b)

Reference

Silvapulle MJ (1981) On the existence of the maximum likelihood estimators for the Binomial response models. J Roy Stat Soc 43:310–313