University of Alberta

Designs for Nonlinear Regression With a Prior on the Parameters

by

Md. Jamil Hasan Karami

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science

in

Statistics

Department of Mathematical and Statistical Sciences

Fall 2011
Edmonton, Alberta

Abstract

This thesis deals with finding design points for nonlinear regression models with the possibility that the fitted model is incorrect. The information matrix depends on the parameter in nonlinear situations. We have assumed a range of values of the parameter and have specified a prior on the parameter space. A loss function has been developed and then a minimax approach has been adopted to achieve our goal. We have found an explicit expression for the maximized loss and a numerical minimization of it has been done by a genetic algorithm (GA). The whole approach has been implemented by considering some well-known nonlinear functions. We see that changing the values of the parameter of the prior density have effect on design points. However, changing the tuning constants of GA does not alter the design points noticeably. This indicates that we have obtained the minimizing design.

Acknowledgements

All praises be to Allah, the most merciful, the most forgiving. From the bottom of my heart I express my gratitude to Him for enabling me to carry out my studies without any major interruptions.

I am grateful to my supervisor professor Douglas P. Wiens. In the Spring and Summer of 2010, Dr. Wiens conducted a series of seminars on the Robust methods in Statistics, in which I participated. This helped a lot in understanding and writing my thesis. Moreover, I have benefitted from his hands-on teaching on many issues. He expertly guided me through the problems I struggled with in a way that led me to find the solutions for myself. His continuous co-operation, suggestions and wise comments made the difficult task of writing a thesis easier than it was thought at the outset. Actually this thesis would not have been possible without his scholarly guidance.

I would like to thank Dr. Byron Schmuland, Dr. Pengfei Li and Dr. Giseon Heo for serving on the examination committee.

I am also grateful to my parents and brother whose continuous encouragement was instrumental in my successful completion of this program at the University of

Alberta. Finally, I am thankful to my beloved wife, Shabnam Sohani, who inspired me to work with zeal.

Contents

List of Figures

Chapter 1: Introduction

1.1 Nonlinear Regression

In regression analysis, the initial purpose is to find the relationship between the response and the covariates. This relationship is expressed through some models. The general framework of regression model is given below.

$$\text{Observed response} = \text{function of covariates} + \text{random error}.$$

So, the random response is presumably dependent on the values of covariates $\mathbf{x} = (x_1, x_2 \cdot \cdot \cdot, x_p)^T$. Usually the mean function $E[Y|\mathbf{x}] = f(\mathbf{x}; \boldsymbol{\theta})$ has a known form, but may depend on unknown parameters $\boldsymbol{\theta}$. If we assume that $f(\mathbf{x}; \boldsymbol{\theta})$ is a linear function of $\boldsymbol{\theta}$ then we have the linear regression model. On the other hand, if we assume that $f(\mathbf{x}; \boldsymbol{\theta})$ is a nonlinear function of $\boldsymbol{\theta}$ then it is the nonlinear regression model. There are different types of regression but here we will consider nonlinear regression. Nonlinear regression models are often encountered in chemical reactions, in biology, clinical trials (Begg and Kalish 1984), reliability and life testing (Maxim, Hendrickson, and Cullen 1977; Meeker and Hahn 1978; Meeker 1984). Also, the nonlinear models are engendered from dynamic systems where the design problem is to choose an input process in such a way that it can provide optimal identification of the system.

1.2 Examples and Model setup

Count Rumford performed an experiment in 1798 and obtained data on the amount of heat generated by friction. In this experiment an object was allowed to cool from an initial temperature of 130°F to an ambient temperature of 60°F. Data from this experiment are available in Bates and Watts (1988). It is evident, from Figure 1.1, that the linear relationship assumption is not reasonable here and the plot seems to exhibit exponential decay. A model based on Newton's law of cooling was proposed as

$$f(x;\theta) = 60 + 70\ e^{-\theta x},$$

where f is predicted temperature and x is time. Differentiating this function with respect to θ we get $-70xe^{-\theta x}$, which clearly depends on θ . Therefore, this model is indeed nonlinear.

The Michaelis-Menten model is believed to be appropriate in pharmacology and other fields where the output Y of a chemical reaction (e.g., initial 'velocity' of an enzymatic reaction) may depend on the input x (e.g., substrate concentration). The following is a Michaelis-Menten model:

$$Y = \frac{\theta_1 x}{\theta_2 + x} + \varepsilon,$$

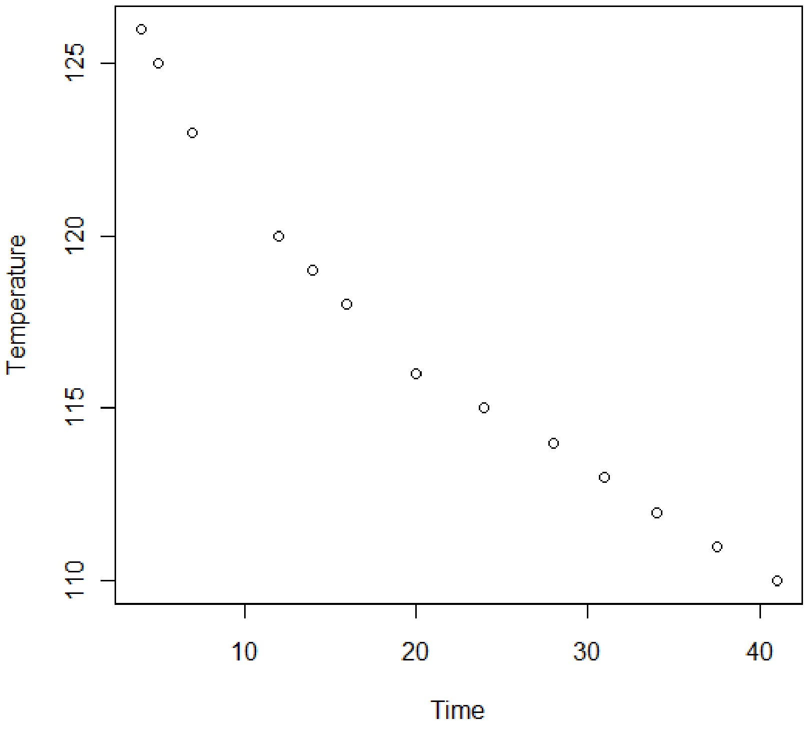

Figure 1.1: Rumford data.

where θ_1, θ_2 are parameters and ε is the random error term. We can observe pairs $(x_i,\ y_i)\,; i = 1, 2, \cdots, n$ and from these parameters can be estimated. Symbolically,

$$Y_i = f(\boldsymbol{\theta}; x_i) + \varepsilon_i,\ i = 1, 2, \cdots, n.$$

The function $f(\boldsymbol{\theta}; x) = \frac{\theta_1 x}{\theta_2 + x}$ where $\boldsymbol{\theta} = (\theta_1,\ \theta_2)$, is a nonlinear function of $\boldsymbol{\theta}$. If we differentiate the function with respect to θ_1 and θ_2 we get $\frac{x}{\theta_2 + x}$ and $\frac{-\theta_1 x}{(\theta_2 + x)^2}$ respectively. Therefore, it is a nonlinear regression model since the derivative involves at least one of the parameters. With

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix},\ \boldsymbol{\eta}(\boldsymbol{\theta}) = \begin{pmatrix} f(\boldsymbol{\theta}; x_1) \\ f(\boldsymbol{\theta}; x_2) \\ \vdots \\ f(\boldsymbol{\theta}; x_n) \end{pmatrix},\ \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

we have the model $\mathbf{Y} = \boldsymbol{\eta}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon}$. There are some important assumptions regarding the error terms. The mean error is zero, i.e. $E(\varepsilon_i) = 0$. Thus

$$E(\boldsymbol{\varepsilon}) = \begin{pmatrix} E(\varepsilon_1) \\ E(\varepsilon_2) \\ \vdots \\ E(\varepsilon_n) \end{pmatrix} = \mathbf{0}.$$

Moreover, errors are uncorrelated on different trials but have constant variance i.e.,

$$Cov(\varepsilon_i, \varepsilon_j) = \begin{cases} 0, & \text{if } i \neq j; \\ \sigma_\varepsilon^2, & \text{if } i = j. \end{cases}$$

Thus the dispersion matrix is obtained:

$$\begin{aligned} Cov(\boldsymbol{\varepsilon}) &= E\left[(\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon}))(\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon}))^T\right] \qquad (1.1) \\ &= \sigma_\varepsilon^2 \mathbf{I}_n. \end{aligned}$$

For nonlinear regression models, likelihood based inferences are usually recommended. The likelihood equations are solved using iterative methods, such as the Gauss-Newton method or Fisher's scoring method. Details of the model fitting and inference procedures can be found in Bates and Watts (1988) and also in Seber and Wild (1989).

1.3 Classical Linear Regression Design

In regression analysis, we obtain information on a response variable Y that depends on a (possibly vector valued) variable X. An experimenter might be able to choose the values of X where it is best to observe the values of Y. In an optimal design problem, the objective is to find the levels of X and to allocate observations at those X's so that the unknown parameters are estimated in an optimal fashion. For linear regression, the experimenter can assume a form given below:

$$Y_i = \mathbf{z}^T(\mathbf{x}_i)\boldsymbol{\theta} + \varepsilon_i \ ; \ i = 1, \ 2, \ \cdots , \ n$$

where $\mathbf{x}_i$ are the values of covariates, $\mathbf{z}(\mathbf{x}_i)$ are regressors, $\boldsymbol{\theta}$ is the vector of unknown parameters and ε_i are random errors. If the model is assumed to be correct, then the least squares estimates are unbiased and we seek to minimize variance. We have $Cov(\widehat{\boldsymbol{\theta}}) = \sigma_\varepsilon^2(\mathbf{Z}^T\mathbf{Z})^{-1}$ where $\mathbf{Z} = (\mathbf{z}(\mathbf{x}_1), \ \mathbf{z}(\mathbf{x}_2), \ \cdots , \ \mathbf{z}(\mathbf{x}_n))^T$. The information matrix $\mathbf{M} = \mathbf{Z}^T\mathbf{Z}$ depends on the design vector $\mathbf{x}$ through the matrix $\mathbf{Z}$. The matrix $\mathbf{Z}$ is known as the design matrix. We are to choose $\mathbf{x}_i$; $i = 1, 2, \cdots , n$ from a design space S so that some scalar valued function of $\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}$ is minimized. Then the exact design will be a discrete probability measure ξ on S with weights which are multiples of $\frac{1}{n}$. Also, these exact designs can be embedded in a suitable class Ξ of distributions. Thus we can express the information matrix $\mathbf{M}(\xi)$ as

$$\mathbf{M}(\xi) = \int\limits_S \mathbf{z}(\mathbf{x})\mathbf{z}^T(\mathbf{x})d\xi.$$

Here we assume that $\mathbf{M}(\xi)$ is a positive definite matrix. In order to get an optimal design, we optimize the scalar valued function of $\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}$ i.e., function of $\mathbf{M}(\xi)$. The most popular design criterion is the D-optimality criterion. This criterion consists of minimizing the determinant of $\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}$ or equivalently, maximizing the determinant of $\left(\mathbf{Z}^T\mathbf{Z}\right)$. That is,

$$\xi = \arg\max_{\xi \in \Xi} \det\left[\mathbf{M}\left(\xi\right)\right].$$

Another design criterion is the A-optimality criterion where the trace of $\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}$ is minimized, i.e.,

$$\xi = \arg\min_{\xi\in\Xi} tr\left[\mathbf{M}^{-1}\left(\xi\right)\right].$$

In the E-optimality design criterion, the maximum eigenvalue of $\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}$ is minimized, i.e.,

$$\xi = \arg\min_{\xi\in\Xi} Ch_{\max}\left[\mathbf{M}^{-1}(\xi)\right].$$

Another important design criterion is the I-optimality criterion where the integrated (or average) variance of the estimated response over the design space is minimized, i.e.,

$$\xi = \arg\min_{\xi\in\Xi} \int_S \mathbf{z}^T(\mathbf{x})\mathbf{M}^{-1}(\xi)\mathbf{z}(\mathbf{x})d\mathbf{x}.$$

In the literature, other criteria such as G-optimality and c-optimality have been studied. The G-optimality criterion seeks the design that minimizes the maximum (over the design space) variance of the predicted response (Kiefer and Wolfowitz, 1960). That is,

$$\xi = \arg\min_{\xi\in\Xi} \max_{\mathbf{x}\epsilon S}\left\{\mathbf{z}^T(\mathbf{x})\mathbf{M}^{-1}(\xi)\mathbf{z}(\mathbf{x})\right\}.$$

On the other hand, in the c-optimality criterion the variance of a given linear combination of parameters is minimized. That is,

$$\xi = \arg\min_{\xi\in\Xi}\left\{\mathbf{c}^T\mathbf{M}^{-1}(\xi)\mathbf{c}\right\},$$

where $\mathbf{c}$ is a fixed vector.

1.4 Robustness of Design

The assumed model, in most applications, is a reasonable approximation to the true model. The classical regression designs perform well if the assumed model is exactly correct. Let us consider an example to illustrate the perils of acting as if the fitted model is necessarily correct. Suppose an experimenter measures the water purity (y) as a function of an input variable chlorine (x). Also, he sets the design space $S = [-1,\ 1]$. Now for the straight line regression

$$y_i = \theta_0 + \theta_1 x_i + \varepsilon_i, \qquad -1 \leqslant x_i \leqslant 1,$$

we find that

$$nVar(\widehat{\theta}_1) = \frac{\sigma_\varepsilon^2}{S_X^2}, \tag{1.2}$$

$$nVar(\widehat{\theta}_0) = \sigma_\varepsilon^2 \left(1 + \frac{\overline{x}^2}{S_X^2}\right), \tag{1.3}$$

where $S_X^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 / n$, $\overline{x} = \sum_{i=1}^{n} x_i / n$ and σ_ε^2 is the error variance. It is easily seen that both the expressions (1.2) and (1.3) are minimized by putting half of the x's at each of -1 and $+1$ since then $\overline{x}^2 = 0$ and S_X^2 is the maximum. The classically optimal design measure, in this case, $\xi(x) =$ fraction of design points placed at x has the form

$$\xi(-1) = \xi(1) = \frac{1}{2}.$$

Usually, the experimenter wants to have $Cov\left(\widehat{\boldsymbol{\theta}}\right) = \sigma_\varepsilon^2 \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}$ small, i.e. he chooses the design so that the determinant, or trace, or maximum eigenvalue of $\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}$ is minimized. If we consider the quadratic regression

$$y_i = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \varepsilon_i, \qquad -1 \leqslant x_i \leqslant 1,$$

the D-optimal design measure has the form

$$\xi\,(-1) = \xi\,(0) = \xi\,(1) = \frac{1}{3},$$

i.e. $\frac{1}{3}$ of the observations on Y are placed at each of -1, 0 and 1. Note that in both of the cases (linear and quadratic regression) the estimates are unbiased, i.e. $E(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$ if the fitted model is correct. So the usual interest is on minimizing the variance.

Box and Draper (1959) carried out research on designs for polynomial fits when the true response is different from the one fitted. In their seminal paper, they compared some designs and concluded '$\cdots$ the optimal design in typical situations in which both variance and bias occur is very nearly the same as would be obtained if variance were ignored completely and the experiment designed so as to minimize bias alone.' But we have to be cautious that by fitting an incorrect model we may induce much larger bias than the gains attained by using an optimal design. Let us consider another example. Suppose that an experimenter fits a straight line using the classically optimal design on the design space $[-1,\ 1]$ when in fact $E(Y) = \phi_0 + \phi_1 x + \phi_2 x^2$. He fits $E(Y) =$

$\mathbf{z}^T(\mathbf{x})\boldsymbol{\theta}$ with $\boldsymbol{\theta} = (\theta_0,\ \theta_1)^T$. Then the least squares estimates are

$$\begin{aligned}\widehat{\boldsymbol{\theta}} &= \begin{pmatrix}\widehat{\theta}_0\\ \widehat{\theta}_1\end{pmatrix}\\ &= \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{y},\ where\ \mathbf{Z} = \left(\mathbf{1}_n \vdots \mathbf{x}\right).\end{aligned}$$

The $k-th$ sample moment about zero is $\tau_k = \sum_{i=1}^{n} x_i^k/n$. Let us assume that the design is symmetric and then $\tau_k = 0$ for odd k. Now

$$\begin{aligned}Cov(\widehat{\boldsymbol{\theta}}) &= \sigma_\varepsilon^2\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\\ &= \frac{\sigma_\varepsilon^2}{n}\begin{pmatrix}1 & 0\\ 0 & \tau_2^{-1}\end{pmatrix},\end{aligned}$$

$$\begin{aligned}E(\widehat{\boldsymbol{\theta}}) &= \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T E(\mathbf{y})\\ &= \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\left\{\mathbf{Z}\begin{pmatrix}\phi_0\\ \phi_1\end{pmatrix} + \phi_2\begin{pmatrix}x_1^2\\ x_2^2\\ \vdots\\ x_n^2\end{pmatrix}\right\}\\ &= \begin{pmatrix}\phi_0 + \phi_2\tau_2\\ \phi_1\end{pmatrix}.\end{aligned}$$

Therefore, the predictions $\widehat{Y}(x)$ have

$$\begin{aligned} E\left[\widehat{Y}(x)\right] &= (1,\ x)\, E\left(\widehat{\boldsymbol{\theta}}\right) \\ &= (1,\ x) \begin{pmatrix} \phi_0 + \phi_2\tau_2 \\ \phi_1 \end{pmatrix} \\ &= \phi_0 + \phi_2\tau_2 + \phi_1 x \\ &= \phi_0 + \phi_1 x + \phi_2 x^2 + \phi_2\tau_2 - \phi_2 x^2 \\ &= E\left[Y(x)\right] + \phi_2(\tau_2 - x^2). \end{aligned}$$

Here, the estimate $\widehat{Y}(x)$ of $E\left[Y(x)\right]$ has the bias $b(x) = \phi_2(\tau_2 - x^2)$, which dominates the mean squared error for all sufficiently large n:

$$\begin{aligned} MSE\left[\widehat{Y}(x)\right] &= E\left[\left\{\widehat{Y}(x) - E[Y(x)]\right\}^2\right] \\ &= (1,\ x)\ Cov(\widehat{\boldsymbol{\theta}}) \begin{pmatrix} 1 \\ x \end{pmatrix} + b^2(x) \\ &= (1,\ x)\ \frac{\sigma_\varepsilon^2}{n} \begin{pmatrix} 1 & 0 \\ 0 & \tau_2^{-1} \end{pmatrix} \begin{pmatrix} 1 \\ x \end{pmatrix} + b^2(x) \\ &= \frac{\sigma_\varepsilon^2}{n}\left(1 + \frac{x^2}{\tau_2}\right) + \left[\phi_2(\tau_2 - x^2)\right]^2. \end{aligned}$$

The Integrated Mean Squared Error (IMSE) of the fitted response is often used as a design criterion. So we can use it here as a measure of performance. We have

$$\begin{aligned}
IMSE &= \int\limits_S MSE\left[\widehat{Y}(x)\right] dx \\
&= \frac{\sigma_\varepsilon^2}{n} \int\limits_{-1}^{1} \left(1 + \frac{x^2}{\tau_2}\right) dx + \phi_2^2 \int\limits_{-1}^{1} \left(\tau_2 - x^2\right)^2 dx \\
&= 2\frac{\sigma_\varepsilon^2}{n} \left(1 + \frac{1}{3\tau_2}\right) + 2\phi_2^2 \left[\left(\tau_2 - \frac{1}{3}\right)^2 + \frac{4}{45}\right].
\end{aligned}$$

The integrated variance, i.e. the first term in the IMSE, is minimized by the classically optimal design. But, on the other hand, the integrated squared bias, i.e. the second term, dominates the first term for sufficiently large n. It is, however, minimized when $\tau_2 = \frac{1}{3}$, which is the second moment of the continuous uniform distribution on $[-1,\ 1]$. This can be approximated by an equally spaced design

$$x_i = -1 + 2\left(\frac{i-1}{n-1}\right), \quad i = 1,\ 2,\ \cdots,\ n$$

with $\tau_2 = \frac{1}{3} + \frac{2}{3(n-1)} = \frac{1}{3} + O(\frac{1}{n})$. Thus minimizing the bias alone may result in very close to minimizing the IMSE. It is also notable that we cannot test the lack of fit with the classical design.

There is another situation where the experimenter fits $E(Y) = \mathbf{z}^T(\mathbf{x})\boldsymbol{\theta}$ when in fact, for some unknown function f,

$$E(Y) = \mathbf{z}^T(\mathbf{x})\boldsymbol{\phi} + f(\mathbf{x}). \tag{1.4}$$

The presence of f in (1.4) indicates that $\boldsymbol{\phi}$ may differ from the 'true' regression parameter when f is absent. Some authors such as Marcus and Sacks (1976), Sacks and Ylvisaker (1978), Pesotchinsky (1982), Li and Notz (1982) and Li (1984) have considered f to be a member of

$$\{f|\ |f(\mathbf{x})| \leqslant \psi(\mathbf{x})\},$$

where $\psi(\mathbf{x})$ may be constant or some other function of $\mathbf{x}$. Minimizing some function of the MSE of $\widehat{Y}(x)$ we can get designs that are robust. However, the designs are sensitive to the choice of ψ. Moreover, they tend to concentrate all mass at extreme points of the design space. So, there is no scope of exploring its interior.

Huber (1975) considered approximate straight line regression. The true model was defined as

$$E\left[Y(x)\right] = \mathbf{z}^T(x)\boldsymbol{\theta} + f(x),$$

where the vector of regressor $\mathbf{z}(x) = (1,\ x)^T$, $x \in S = \left[-\frac{1}{2},\ \frac{1}{2}\right]$ and the contamination function $f \in \mathcal{F}$, which is an infinite dimensional space of functions, such that

$$\mathcal{F} = \left\{ f : \int\limits_S f^2(x)dx \leqslant \eta^2,\ \int\limits_S \mathbf{z}(x)f(x)dx = 0 \right\}. \tag{1.5}$$

Huber used the integrated mean squared error

$$IMSE = \int\limits_S E\left[\left(\mathbf{z}^T(x)\widehat{\boldsymbol{\theta}} - E[Y(x)]\right)^2\right] dx,$$

as the design criterion and found the design by solving $\min_{\xi \in \Xi} \max_{f \in \mathcal{F}} IMSE(\xi, f)$. He termed the obtained design as a minimax design, which is optimal for the worst possible contamination function $f \in \mathcal{F}$. This implies that it is a robust design. Later, Wiens (1990) extended Huber's work from simple linear regression to multiple linear regression where $\mathbf{z}(\mathbf{x}) = (1,\ x_1,\ x_2,\ \cdots,\ x_p)^T$, p is the number of regressors in the model. The robust regression design under a finite design space was considered by Fang and Wiens (2000). They used the average (over the design space) mean squared error of $\widehat{Y}(x) = \mathbf{z}^T(x)\widehat{\boldsymbol{\theta}}$ as the loss function. For example, suppose S is a finite design space with design points $\mathbf{x}_i,\ i = 1,\ 2,\ \cdots,\ N$. Now we have to allocate $n_i \geqslant 0$ observations to $\mathbf{x}_i$, with $\sum_{i=1}^{N} n_i = n$, the total number of observations. But the design problem is to choose the n_i's optimally. The loss function, according to them, is

$$I = \frac{1}{N} \sum_{i=1}^{N} E\left[\left(\widehat{Y}(\mathbf{x}_i) - E[Y(\mathbf{x}_i)]\right)^2\right].$$

After some calculation it takes the following form:

$$I = \frac{1}{N}\mathbf{d}^T\mathbf{Z}^T\mathbf{Z}\mathbf{d} + \frac{1}{N}\ tr\left(\mathbf{Z}\mathbf{C}\mathbf{Z}^T\right) + \frac{1}{N}\mathbf{f}^T\mathbf{f}, \tag{1.6}$$

where $\mathbf{Z} = (\mathbf{z}(\mathbf{x}_1),\ \mathbf{z}(\mathbf{x}_2),\ \cdots,\ \mathbf{z}(\mathbf{x}_N))^T$, $\mathbf{d}$ is the bias vector $E(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$, $\mathbf{C}$ is the covariance matrix $Cov(\widehat{\boldsymbol{\theta}})$ and $\mathbf{f} = (f(\mathbf{x}_1),\ f(\mathbf{x}_2),\ \cdots,\ f(\mathbf{x}_N))^T$. Now assume $\mathbf{Z}$ is of full rank p and the singular value decomposition of $\mathbf{Z}$ is $\mathbf{Z} = \mathbf{U}_{N\times p}\boldsymbol{\Lambda}_{p\times p}\mathbf{V}^T_{p\times p}$ with $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_p$ and $\boldsymbol{\Lambda}$ is diagonal and invertible. The matrix $\mathbf{U}$ is augmented by $\widetilde{\mathbf{U}}_{N\times(N-p)}$ so that $\left[\mathbf{U} \vdots \widetilde{\mathbf{U}}\right]_{N\times N}$ is orthogonal. Now using the conditions of the class of contamination function $\mathcal{F}$ as in (1.5) with summation instead of integration, an $(N-p) \times 1$ vector $\mathbf{c}$, with $\|\mathbf{c}\| \leqslant 1$, satisfying $\mathbf{f} = \eta\sqrt{N}\widetilde{\mathbf{U}}\mathbf{c}$ is found. By using matrix algebra, (1.6) can then be maximized over f by maximizing over c. Then the allocation of n_i observations to $\mathbf{x}_i$ is obtained by minimizing that maximized loss function subject to some restrictions. A simulated annealing algorithm was adopted by Fang and Wiens for numerical minimization of the maximized loss function. In this thesis, we will use minimax design criteria, but the minimization will be carried out through a Genetic Algorithm.

1.5 Designs for Nonlinear Regression

Suppose we have a response variable Y and a model of the form $\mathbf{Y} = \boldsymbol{\eta}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon}$, where $\boldsymbol{\eta}(\boldsymbol{\theta})$ has elements $f(\boldsymbol{\theta}; x_i)$ and derivative $\mathbf{Z}(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{\eta}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$. The information matrix for parameters is given by

$$\begin{aligned} \mathbf{M}(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \frac{n_i}{n} \mathbf{z}(\mathbf{x}_i; \boldsymbol{\theta}) \mathbf{z}^T(\mathbf{x}_i; \boldsymbol{\theta}) \\ &= \mathbf{Z}^T(\boldsymbol{\theta}) \mathbf{D}_\xi \mathbf{Z}(\boldsymbol{\theta}), \end{aligned}$$

where $\mathbf{z}(\mathbf{x}_i; \boldsymbol{\theta}) = \frac{\partial f(\boldsymbol{\theta}; x_i)}{\partial \boldsymbol{\theta}}$, $\mathbf{Z}(\boldsymbol{\theta}) = (\mathbf{z}(\mathbf{x}_1; \boldsymbol{\theta}),\ \mathbf{z}(\mathbf{x}_2; \boldsymbol{\theta}),\ \cdots,\ \mathbf{z}(\mathbf{x}_n; \boldsymbol{\theta}))^T$, $\mathbf{D}_\xi$ is the diagonal matrix with diagonal elements $\{n_i/n\}$ and n_i is the number of observations made at x_i. The information matrix can also be written as

$$\mathbf{M}(\boldsymbol{\theta}, \xi) = \int\limits_S \mathbf{z}(\mathbf{x}) \mathbf{z}^T(\mathbf{x}) d\xi,$$

where ξ belongs to a suitable class Ξ of distributions. Now we may want to minimize some function of the information matrix, say, $\phi(\mathbf{M}(\boldsymbol{\theta}, \xi))$. This function, in the nonlinear problem, depends on $\boldsymbol{\theta}$ in such a way as to affect the choice of design. Thus we encounter a peculiar situation: 'The theory of optimal designs for linear models has been extensively developed but the adaptation to nonlinear models is more difficult, as the criteria involve the unknown parameters $\boldsymbol{\theta}$ as well as the design points x_i' (Seber and Wild 1989). Designing an efficient experiment for the estimation of parameters will require knowledge of the parameters. So the optimal design depends on the proper values of the parameters. This dependency has been dealt with using different approaches such as locally optimal, sequential, minimax or Bayesian.

Chernoff (1953) implemented the earliest locally optimal approach where an experimenter makes an initial guess of values for the vector of unknown parameters. The performance of this approach depends on how close these initial values are to the true parameter values. So, based on it, the efficiency of such designs can be high or low. That is, the locally optimal design may be very sensitive to small perturbations in the parameter value. This problem can be alleviated by adopting a sequential design. In the sequential design approach, parameters are estimated at every stage. But in each stage, the estimates obtained from the previous stage are used as best guesses of the parameters in order to obtain design points in the current stage. Abdelbasit and Plackett (1983), Chaudhuri and Mykland (1993), Sinha and Wiens (2002) have used the sequential design approach to nonlinear design problems.

The minimax is another important approach where a range of possible parameter values are considered, i.e., $\boldsymbol{\theta} \in \Theta$, where Θ denotes the plausible parameter values.

The minimax optimal design minimizes the maximum of a criterion and the following form is taken into account:

$$\min_{\xi}\ \max_{\theta\in\Theta}\ \phi\left\{\mathbf{M}\left(\boldsymbol{\theta},\xi\right)\right\}.$$

It is notable that the minimax design approach is robust because it gives the design corresponding to the worst possible values of Θ. King and Wong (2000), Biedermann, Dette and Pepelyshev (2006), Braess and Dette (2007) and some other authors have considered the minimax (or maximin) design approach for nonlinear regression models.

If we have prior knowledge of the parameters we can take advantage of it. In the Bayesian approach, a prior distribution $\pi(\cdot)$ is assumed on the unknown parameters. Then by maximizing the expectation of the criterion over the assumed prior distribution we can find the optimal Bayesian design. That is,

$$E_\theta\phi(\mathbf{M}(\xi,\theta)) = \int_\theta \phi(\mathbf{M}(\xi,\theta))\pi(\theta)d\theta.$$

Chaloner and Larntz (1989), Chaloner and Verdinelli (1995), Dette and Wong (1996), and Matthews and Allcock (2004) have worked on Bayesian designs.

1.6 Literature Review

There are some authors who have used the design of nonlinear problems in their research. Among them, as a primer, Box and Lucas (1959), Fedorov (1972), Cochran

(1973), St. John and Draper (1975), Silvey (1980), and Steinberg and Hunter (1984) are noteworthy. Box and Lucas (1959) carried out a study on the design of experiments in nonlinear situations. Their objective was to obtain a programme of trials that can be used for estimating the parameters with high accuracy. To do this, a set of preliminary values of parameters was assumed known. Then the design points were chosen by maximizing the determinant of Fisher information or, equivalently, by minimizing the asymptotic formula for Wilks' generalized variance of the maximum likelihood estimates of the parameters. They used a model from chemical reaction and some standard nonlinear regression models as examples in order to illustrate their research.

Draper and Hunter (1967) discussed the use of prior distributions in the design of experiments for parameter estimation in nonlinear situations. They took an example with the solution obtained from Box and Lucas (1959), and showed how the positioning of the design points change based on the availability of prior information on the parameters.

In his expository article, Cochran (1973) made some candid comments on nonlinear design problems. We need to know the values of parameters before we start finding the design points in order to estimate the parameters. 'This is a standard feature that distinguishes nonlinear from linear problems', Cochran said while commenting on the dilution series experiment. He also claimed: 'You tell me the value of $\boldsymbol{\theta}$ and I promise to design the best experiment for estimating $\boldsymbol{\theta}$'. In that article, Cochran reviewed some works previously done on experiments for nonlinear functions. He mentioned the sequential and nonsequential approaches for estimating the parameters and also commented on model adequacy and discrimination between models and model building.

White (1973) extended the general equivalence theorem, due to Kiefer and Wolfowitz (1960), to nonlinear models. He developed the information matrix and a variance function and then used these to prove the equivalence theorem for the nonlinear model.

St. John and Draper (1975) reviewed some major results on the theory of design and the way the design criterion has been extended to nonlinear models. They discussed the algorithm for obtaining the D-optimal design. The general algorithm was developed by Fedorov and his coworkers (Fedorov 1972), Fedorov and Dubova (1968). Later it was studied and modified by Atwood (1973) and St. John (1973). However, Silvey and Titterington (1973) outlined a slightly different algorithm than Fedorov for obtaining a D-optimal design.

Abdelbasit and Plackett (1981) worked on designs for categorized data. They mentioned that the asymptotic dispersion matrix of the estimators for models of categorized data usually contains some unknown parameters. So they considered design criteria that suit the nonlinear models. Moreover, Abdelbasit and Plackett (1983) worked on experimental designs for binary data. Since the information matrix in this situation depends on the unknown parameters they have used the initial point estimates and applied the sequential methods. Also they discussed the criterion of constant information for models with one or two parameters.

Atkinson (1982) wrote an article highlighting the developments in the design of experiment based on literature mostly from 1976 to 1980. He cited the comments of

Box (1979): ‘by invention of the concept of experimental design, Fisher promoted the statistician from a curator of dusty relics to a valued member of a scientific team’. In that article, he summarized the development of designs for nonlinear models.

Steinberg and Hunter (1984) presented an article reviewing and commenting on the major developments in the design of experiments. They put important guidelines and recommendations for experimenters, statisticians, practitioners and researchers who are jointly exploring new frontiers. One of the important topics in that paper was nonlinear models. Under this topic they described some work done before 1983 on designs for nonlinear models. Mentioning the relatively few studies of experimental design for nonlinear models as compared to linear models, they have suggested to enhance research on designing experiments for nonlinear models. Also they suggested some studies regarding the design of experiments for nonlinear models. One is the design of experiments for nonlinear models that are proposed as tentative empirical approximations and the other is finding a link between empirical models and underlying nonlinear mechanisms.

In an article Ford, Titterington and Kitsos (1989) have summarized some work in optimal experimental design in nonlinear problems. They have also described some design approaches such as static and sequential design schemes with application to nonlinear models. Moreover, they have expressed concern about the misspecification of the model itself. They state, ‘Indeed, if the model is seriously in doubt, the forms of design that we have considered may be completely inappropriate.’ They, however, concluded that for a reliable nonlinear model an acceptable design can be obtained if we have reasonably good prior information about the parameters and also by using sequential design approach we can get acceptable designs.

Chaudhuri & Mykland (1993) carried out research on nonlinear experiments with the objective in mind of getting design points so that the parameters can be estimated efficiently. They have used an initial static design at first and then a fully adaptive sequential design. Under some assumptions on the regression model, they showed that the asymptotic distribution of the maximum likelihood estimates of the parameters is normal. Moreover, they obtained asymptotically D-optimal design by using their scheme for choosing design points sequentially.

Sinha & Wiens (2002) have introduced the notion of approximately specified nonlinear regression model. Allowing for the possibility of an incorrect model, the sequential design approaches have been used. It is notable that the authors have developed the loss functions in their research and estimated the components of the loss functions. Moreover, small-sample simulation studies have shown that their obtained new designs can be very successful with respect to mean squared error. The authors wanted to have a robust design so that the loss is the least even when the model is approximately true.

This thesis will focus on finding designs for the nonlinear regression problem when the model is a misspecified one with response contamination function. In order to achieve our goal, we will be using the minimax criterion where a range of values of the parameter is assumed. Therefore, we can specify a prior on the region of the parameter space where the design will be robust.

Chapter 2: The Loss Function for Nonlinear Models

2.7 The Approximate Model

In regression analysis, it is usually assumed that the form of the model under consideration is exactly correct. But we will allow the possibility that the fitted model is incorrect. Suppose an underlying regression model is $Y = E\left(Y|\mathbf{x}\right) + \varepsilon,\ \mathbf{x} \in S$, where $S{=}\left\{\mathbf{x}_i\right\}_{i=1}^{N}$ is the design space and ε is the random error term. Now an experimenter acts according to the belief that

$$E\left(Y|\mathbf{x}\right) \approx f\left(\mathbf{x}; \boldsymbol{\theta}_0\right), \tag{2.1}$$

where $\boldsymbol{\theta}_0$ is a $p \times 1$ vector of parameters. Suppose n_i observations at $\mathbf{x}_i$ are taken and $\sum_{i=1}^{N} n_i = n$. Here N is a finite number and $n_i \geqslant 0$; there is no requirement that observations be made at every $\mathbf{x}_i$. If we are given the data $(\mathbf{x}_1,\ y_1),\ (\mathbf{x}_2,\ y_2),\ \cdots,\ (\mathbf{x}_n,\ y_n)$, parameters can be estimated by least squares estimate:

$$\widehat{\boldsymbol{\theta}}_n = \arg\ \min \sum_{i=1}^{n} \left[y_i - f(\mathbf{x}_i; \boldsymbol{\theta})\right]^2 .$$

The estimates that are obtained from the above result in

$$\sum_{i=1}^{n} \mathbf{z}(\mathbf{x}_i; \widehat{\boldsymbol{\theta}}_n) r_i(\widehat{\boldsymbol{\theta}}_n) = \mathbf{0},$$

where $\mathbf{z}(\mathbf{x}; \boldsymbol{\theta}) = \partial f(\mathbf{x}; \boldsymbol{\theta})/\partial \boldsymbol{\theta}$ is the $p \times 1$ gradient vector and $r_i(\widehat{\boldsymbol{\theta}}_n) = y_i - f(\mathbf{x}_i; \widehat{\boldsymbol{\theta}}_n)$ is the residual.

We can define $d(\mathbf{x}; \boldsymbol{\theta}_0) = E\,(Y|\mathbf{x}) - f\,(\mathbf{x}; \boldsymbol{\theta}_0)$, which gives the exact but only approximately specified model

$$Y = f\,(\mathbf{x}; \boldsymbol{\theta}_0) + d(\mathbf{x}; \boldsymbol{\theta}_0) + \varepsilon,$$

where the true $\boldsymbol{\theta}_0$ is defined in such a way that (2.1) becomes most accurate; in other words, the sum of squared discrepancy is minimized:

$$\boldsymbol{\theta}_0 = \arg\ \min \sum_{i=1}^{N} d^2(\mathbf{x}_i; \boldsymbol{\theta}). \tag{2.2}$$

Thus $f\,(\mathbf{x}; \boldsymbol{\theta}_0)$ is, on an average, the best predictor of $E\,(Y|\mathbf{x})$. We may think of $d(\mathbf{x}; \boldsymbol{\theta}_0)$ due to the model misspecification or unestimated curvature in the model. Thus its presence in a model increases the bias of $\boldsymbol{\theta}_0$ and of $E\,(Y|\mathbf{x})$. Moreover, by dint of (2.2) as in Sinha & Wiens (2002), we have that

$$\sum_{i=1}^{N} \mathbf{z}(\mathbf{x}_i;\boldsymbol{\theta}_0)d(\mathbf{x}_i;\boldsymbol{\theta}_0) = \mathbf{0}. \tag{2.3}$$

2.8 Developing the Loss Function

Let us assume a finite design space $S = \{\mathbf{x}_1,\ \mathbf{x}_2,\ \cdots,\ \mathbf{x}_N\}$. Suppose we have an n point design. Define $\mathbf{Z}(\boldsymbol{\theta})$ to be the $N \times p$ matrix whose rows are the $\mathbf{z}^T(\mathbf{x}_i;\boldsymbol{\theta})$, $\mathbf{D}_\xi$ the diagonal matrix with diagonal elements $\{\xi_i\} = \{n_i/n\}$, where the design allocates $n_i \geqslant 0$ observations to the locations $\mathbf{x}_i$ and $\sum_{i=1}^{N} n_i = n$. Put

$$\begin{aligned}\mathbf{A}(\boldsymbol{\theta}) &= \frac{1}{N}\sum_{i=1}^{N} \mathbf{z}(\mathbf{x}_i;\boldsymbol{\theta})\mathbf{z}^T(\mathbf{x}_i;\boldsymbol{\theta}) \\ &= \frac{1}{N}\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{Z}(\boldsymbol{\theta}).\end{aligned}$$

Define $\mathbf{d}(\boldsymbol{\theta}) = (d(\mathbf{x}_1;\boldsymbol{\theta}),\ d(\mathbf{x}_2;\boldsymbol{\theta}),\ \cdots,\ d(\mathbf{x}_N;\boldsymbol{\theta}))^T$ and let $\boldsymbol{\Sigma}$ be the $N \times N$ diagonal matrix with diagonal elements $\{\sigma^2(\mathbf{x}_i)\}$. Then the asymptotic mean squared error matrix, following Sinha and Wiens (2002), is

$$\mathbf{MSE}(\boldsymbol{\theta}) = \mathbf{M}^{-1}(\boldsymbol{\theta})\left\{\mathbf{Q}(\boldsymbol{\theta}) + \mathbf{b}(\boldsymbol{\theta})\mathbf{b}^T(\boldsymbol{\theta})\right\}\mathbf{M}^{-1}(\boldsymbol{\theta}),$$

where

$$\begin{aligned}\mathbf{M}(\boldsymbol{\theta})_{p\times p} &= n\sum_{i=1}^{N}\mathbf{z}(\mathbf{x}_i;\boldsymbol{\theta})\mathbf{z}^T(\mathbf{x}_i;\boldsymbol{\theta})\xi_i \\ &= n\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{D}_\xi\mathbf{Z}(\boldsymbol{\theta}),\end{aligned}$$

$$\begin{aligned}\mathbf{Q}(\boldsymbol{\theta})_{p\times p} &= n\sum_{i=1}^{N}\mathbf{z}(\mathbf{x}_i;\boldsymbol{\theta})\sigma^2(\mathbf{x}_i)\mathbf{z}^T(\mathbf{x}_i;\boldsymbol{\theta})\xi_i \\ &= n\mathbf{Z}^T(\boldsymbol{\theta})\boldsymbol{\Sigma}\mathbf{D}_\xi\mathbf{Z}(\boldsymbol{\theta}),\end{aligned}$$

$$\begin{aligned}\mathbf{b}(\boldsymbol{\theta})_{p\times 1} &= n\sum_{i=1}^{N}\mathbf{z}(\mathbf{x}_i;\boldsymbol{\theta})d(\mathbf{x}_i;\boldsymbol{\theta})\xi_i \\ &= n\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{D}_\xi\mathbf{d}(\boldsymbol{\theta}).\end{aligned}$$

The initial purpose of nonlinear regression is typically response estimation or prediction. Therefore, we should choose a design so that the average error is minimized when $E(Y|\mathbf{x})$ is predicted by $f(\mathbf{x},\widehat{\boldsymbol{\theta}}_n)$. Here we can mention an asymptotic property of the estimate $\widehat{\boldsymbol{\theta}}_n$. From the asymptotic theory of Gallant (1987), and by using (1.1), the asymptotic normal approximation is

$$\begin{aligned}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0) &\sim N\left(\mathbf{M}^{-1}(\boldsymbol{\theta}_0)\mathbf{b}(\boldsymbol{\theta}_0),\ \mathbf{M}^{-1}(\boldsymbol{\theta}_0)\mathbf{Q}(\boldsymbol{\theta}_0)\mathbf{M}^{-1}(\boldsymbol{\theta}_0)\right) \\ &= N\left(\mathbf{M}^{-1}(\boldsymbol{\theta}_0)\mathbf{b}(\boldsymbol{\theta}_0),\ \sigma^2\mathbf{M}^{-1}(\boldsymbol{\theta}_0)\right),\end{aligned}$$

where the information matrix $\mathbf{M}(\boldsymbol{\theta}_0)$ arises from the concept of the Fisher information. Though the Fisher information is built under a normal likelihood, it is used in asymptotic even when the original likelihood is not necessarily normal. In nonlinear situations, estimates can be obtained as if they are from normal likelihood.

If $\boldsymbol{\theta}$ is the 'true' value, the first order approximation of the error is

$$\begin{aligned} f(\mathbf{x};\widehat{\boldsymbol{\theta}}_n) - E(Y|\mathbf{x}) &\approx f(\mathbf{x};\boldsymbol{\theta}) + \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right)^T [\partial f(\mathbf{x};\boldsymbol{\theta})/\partial\boldsymbol{\theta}] - E(Y|\mathbf{x}) \\ &= \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right)^T \mathbf{z}(\mathbf{x};\boldsymbol{\theta}) - d(\mathbf{x};\boldsymbol{\theta}). \end{aligned}$$

Then the loss is

$$\begin{aligned} & \frac{1}{N}\sum_{i=1}^{N} E\left[\left\{f(\mathbf{x}_i;\widehat{\boldsymbol{\theta}}_n) - E(Y|\mathbf{x}_i)\right\}^2\right] \\ \approx\; & \frac{1}{N}\sum_{i=1}^{N} E\left[\left\{\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right)^T \mathbf{z}(\mathbf{x}_i;\boldsymbol{\theta}) - d(\mathbf{x}_i;\boldsymbol{\theta})\right\}^2\right] \\ =\; & tr\left[\mathbf{MSE}(\boldsymbol{\theta}) \cdot \mathbf{A}(\boldsymbol{\theta})\right] + \frac{1}{N}\|\mathbf{d}(\boldsymbol{\theta})\|^2 \\ =\; & tr\mathbf{M}^{-1}(\boldsymbol{\theta})\mathbf{Q}(\boldsymbol{\theta})\mathbf{M}^{-1}(\boldsymbol{\theta})\mathbf{A}(\boldsymbol{\theta}) + \mathbf{b}^T(\boldsymbol{\theta})\mathbf{M}^{-1}(\boldsymbol{\theta})\mathbf{A}(\boldsymbol{\theta})\mathbf{M}^{-1}(\boldsymbol{\theta})\mathbf{b}(\boldsymbol{\theta}) \\ & + \frac{1}{N}\|\mathbf{d}(\boldsymbol{\theta})\|^2. \end{aligned}$$

As in (2.3), $\boldsymbol{\theta}_0$ is the 'true' value that leads to the orthogonality condition

$$\frac{1}{N}\sum_{i=1}^{N} d(\mathbf{x}_i;\boldsymbol{\theta}_0)\mathbf{z}(\mathbf{x}_i;\boldsymbol{\theta}_0) = \frac{1}{N}\mathbf{Z}^T(\boldsymbol{\theta}_0)\mathbf{d}(\boldsymbol{\theta}_0) = \mathbf{0},$$

and a bound

$$\sum_{i=1}^{N} d^2(\mathbf{x}_i; \boldsymbol{\theta}_0) \leqslant \frac{\eta^2}{n},$$

where η is a fixed quantity chosen by the designer.

We are considering the nonlinear regression for which an optimal design depends on the proper values of parameters that are unknown. As we mentioned earlier there are some approaches to handle this problem. For example, the locally optimal design approach can be adopted. But the initial guess of the unknown parameter vector $\boldsymbol{\theta}_0$ should have to be as close as possible to the true parameter vector since the efficiency of such designs depends on the quality of initial approximation. To alleviate this problem of locally optimal design, a sequential design approach can be considered, where estimates of parameters are updated at each stage. On the other hand, a minimax design approach can be implemented, where minimization is done after maximizing the loss over a neighbourhood of $\boldsymbol{\theta}_0$.

In this thesis we aim to minimize the maximum, over $d(\cdot)$, of

$$\begin{aligned} \mathcal{L}_I &= \mathcal{L}_I(\xi, d) \\ &= N \int\limits_{\Theta} \left\{ tr \left[\mathbf{MSE}(\boldsymbol{\theta}) \cdot \mathbf{A}(\boldsymbol{\theta}) \right] + \frac{1}{N} \|\mathbf{d}(\boldsymbol{\theta})\|^2 \right\} p\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta}, \end{aligned}$$

where $p(\boldsymbol{\theta})$ is some density on the parameter space Θ. The maximization is to be done subject to the conditions that, for each $\boldsymbol{\theta}$,

$$\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{d}(\boldsymbol{\theta}) = \mathbf{0}, \tag{2.4}$$

$$\|\mathbf{d}(\boldsymbol{\theta})\|^2 \leqslant \eta^2/n. \tag{2.5}$$

We are considering n point design and the discrepancy $d(\mathbf{x}_i;\boldsymbol{\theta}_0)$ would not exceed the order of error, which is $1/\sqrt{n}$.

Since the loss no longer depends on the parameters, it is not necessary to design sequentially. We could instead look for minimax designs, with the maximization being done over $d(\cdot)$, subject to (2.4) and (2.5).

2.9 Minimax Designs

In a minimax design problem we try to find a design measure

$$\xi(\mathbf{x}) = \ fraction\ of\ observations\ made\ at\ x,$$

which results in small values of the maximum, over departures from the fitted model, of the integrated mean squared error. The minimax design can be obtained by minimizing (over a class of designs) the maximum (over d) value of the expected loss function. In minimax approach, we minimize the loss function in the worst case, i.e.

$$\min_{\xi}\ \max_{d}\ \mathcal{L}_I(\xi, d).$$

Now we shall first maximize the loss function over d. The maximization part of the minimax problem can be solved completely, i.e. the maximum loss can be written down as an explicit function of the design.

2.10 Maximizing the Loss Function

Let $\mathbf{U}(\boldsymbol{\theta})$ be an $N \times p$ matrix whose columns form an orthogonal basis for the column space of $\mathbf{Z}(\boldsymbol{\theta})$. Augment $\mathbf{U}(\boldsymbol{\theta})$ by $\widetilde{\mathbf{U}}(\boldsymbol{\theta})$, an $N \times N - p$ matrix whose columns form an orthonormal basis of the orthogonal complement to the column space of $\mathbf{Z}(\boldsymbol{\theta})$. Then $\left[\mathbf{U}(\boldsymbol{\theta}) \vdots \widetilde{\mathbf{U}}(\boldsymbol{\theta})\right]$ is an $N \times N$ orthogonal matrix. Assuming that $\mathbf{Z}(\boldsymbol{\theta})$ has full column rank, we can write

$$\mathbf{Z}(\boldsymbol{\theta}) = \mathbf{U}(\boldsymbol{\theta})\mathbf{R}(\boldsymbol{\theta}),$$

for some $p \times p$ non-singular matrix $\mathbf{R}(\boldsymbol{\theta})$. Using this QR-decomposition, $\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{d}(\boldsymbol{\theta}) = \mathbf{0}$ can be written as $\mathbf{R}^T(\boldsymbol{\theta})\mathbf{U}^T(\boldsymbol{\theta})\mathbf{d}(\boldsymbol{\theta}) = \mathbf{0}$. Since $\mathbf{R}(\boldsymbol{\theta})$ is non-singular, we have $\mathbf{U}^T(\boldsymbol{\theta})\mathbf{d}(\boldsymbol{\theta}) = \mathbf{0}$, i.e. $\mathbf{d}(\boldsymbol{\theta})$ is orthogonal to the columns of $\mathbf{U}(\boldsymbol{\theta})$. This says that $\mathbf{d}(\boldsymbol{\theta})$ lies in the orthogonal complement to the column space of $\mathbf{U}(\boldsymbol{\theta})$, hence is a linear combination of the vectors in a basis for this space. The columns of $\widetilde{\mathbf{U}}(\boldsymbol{\theta})$ form such a basis. Therefore, the condition (2.4) is equivalent to

$$\mathbf{d}(\boldsymbol{\theta}) = \widetilde{\mathbf{U}}(\boldsymbol{\theta})\mathbf{c}(\boldsymbol{\theta}),$$

for some $\mathbf{c}(\boldsymbol{\theta}) : N - p \times 1$. Then since the columns of $\widetilde{\mathbf{U}}(\boldsymbol{\theta})$ are orthonormal, $\|\mathbf{d}(\boldsymbol{\theta})\| = \|\mathbf{c}(\boldsymbol{\theta})\|$ and so the condition (2.5) is

$$\|\mathbf{c}(\boldsymbol{\theta})\|^2 \leqslant \eta^2/n. \tag{2.6}$$

The maximization problem is now to maximize

$$
\begin{aligned}
\mathcal{L}_I &= N \int\limits_{\Theta} \left\{ tr\left[\mathbf{MSE}(\boldsymbol{\theta}) \cdot \mathbf{A}(\boldsymbol{\theta})\right] + \frac{1}{N} \|\mathbf{d}(\boldsymbol{\theta})\|^2 \right\} p\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta} \\
&= N \int\limits_{\Theta} \left\{ tr\mathbf{M}^{-1}(\boldsymbol{\theta})\mathbf{Q}(\boldsymbol{\theta})\mathbf{M}^{-1}(\boldsymbol{\theta})\mathbf{A}(\boldsymbol{\theta}) \right\} p(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&\quad + N \int\limits_{\Theta} \left\{ \mathbf{b}^T(\boldsymbol{\theta})\mathbf{M}^{-1}(\boldsymbol{\theta})\mathbf{A}(\boldsymbol{\theta})\mathbf{M}^{-1}(\boldsymbol{\theta})\mathbf{b}(\boldsymbol{\theta}) \right\} p(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&\quad + \int\limits_{\Theta} \|\mathbf{d}(\boldsymbol{\theta})\|^2 p(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \frac{1}{n} \int\limits_{\Theta} tr \left\{ \begin{array}{c} \left[\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{D}_\xi\mathbf{Z}(\boldsymbol{\theta})\right]^{-1} \left[\mathbf{Z}^T(\boldsymbol{\theta})\boldsymbol{\Sigma}\mathbf{D}_\xi\mathbf{Z}(\boldsymbol{\theta})\right] \cdot \\ \left[\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{D}_\xi\mathbf{Z}(\boldsymbol{\theta})\right]^{-1} \left[\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{Z}(\boldsymbol{\theta})\right] \end{array} \right\} p(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&\quad + \int\limits_{\Theta} \left\{ \begin{array}{c} \mathbf{c}^T(\boldsymbol{\theta})\widetilde{\mathbf{U}}^T(\boldsymbol{\theta})\mathbf{D}_\xi\mathbf{Z}(\boldsymbol{\theta}) \ \left[\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{D}_\xi\mathbf{Z}(\boldsymbol{\theta})\right]^{-1} \mathbf{Z}^T(\boldsymbol{\theta}) \cdot \\ \mathbf{Z}(\boldsymbol{\theta}) \left[\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{D}_\xi\mathbf{Z}(\boldsymbol{\theta})\right]^{-1} \mathbf{Z}^T(\boldsymbol{\theta})\mathbf{D}_\xi\widetilde{\mathbf{U}}(\boldsymbol{\theta})\mathbf{c}(\boldsymbol{\theta}) \end{array} \right\} \cdot p(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&\quad + \int\limits_{\Theta} \|\mathbf{c}(\boldsymbol{\theta})\|^2 p(\boldsymbol{\theta}) d\boldsymbol{\theta},
\end{aligned}
$$

subject to (2.6). The maximum value of $\|\mathbf{c}(\boldsymbol{\theta})\|^2$ is η^2/n by (2.6), $\int\limits_{\Theta} p(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is unity and from matrix algebra we know that if $\mathbf{W}$ is a symmetric matrix of order $n \times n$ and $\mathbf{v}$ is a column vector, then we have $\max\limits_{\mathbf{v} \neq \mathbf{0}} \left[\frac{\mathbf{v}^T\mathbf{W}\mathbf{v}}{\mathbf{v}^T\mathbf{v}}\right] = ch_{\max}\left\{\mathbf{W}\right\}$, where $ch_{\max}\left\{\mathbf{W}\right\}$ denotes the largest eigenvalue of $\mathbf{W}$. Using these results in the above expression of

$\mathcal{L}_I$, the maximum loss is

$$\begin{aligned}\max \mathcal{L}_I \;=\; & \frac{1}{n}\int\limits_{\Theta} tr \left\{ \begin{array}{c} \left[\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{D}_\xi\mathbf{Z}(\boldsymbol{\theta})\right]^{-1}\left[\mathbf{Z}^T(\boldsymbol{\theta})\boldsymbol{\Sigma}\mathbf{D}_\xi\mathbf{Z}(\boldsymbol{\theta})\right]\;\cdot \\ \left[\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{D}_\xi\mathbf{Z}(\boldsymbol{\theta})\right]^{-1}\left[\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{Z}(\boldsymbol{\theta})\right] \end{array} \right\} p(\boldsymbol{\theta})d\boldsymbol{\theta} \qquad (2.7) \\ & + \frac{\eta^2}{n}\left[\begin{array}{c} \displaystyle\int\limits_{\Theta} ch_{\max}\left\{ \begin{array}{c} \widetilde{\mathbf{U}}^T(\boldsymbol{\theta})\mathbf{D}_\xi\mathbf{Z}(\boldsymbol{\theta})\;\left[\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{D}_\xi\mathbf{Z}(\boldsymbol{\theta})\right]^{-1}\mathbf{Z}^T(\boldsymbol{\theta})\cdot \\ \mathbf{Z}(\boldsymbol{\theta})\left[\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{D}_\xi\mathbf{Z}(\boldsymbol{\theta})\right]^{-1}\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{D}_\xi\widetilde{\mathbf{U}}(\boldsymbol{\theta}) \end{array} \right\} p(\boldsymbol{\theta})d\boldsymbol{\theta} \\ +\,1 \end{array} \right].\end{aligned}$$

We repeatedly use the fact that the non-zero eigenvalues of a product $\mathbf{AB}$ are the same as those of $\mathbf{BA}$. Moreover, $\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T = \mathbf{I} - \mathbf{U}\mathbf{U}^T$. Applying these to the second term of (2.7) we have

$$\begin{aligned} & ch_{\max}\left\{ \begin{array}{c} \mathbf{U}(\boldsymbol{\theta})\left[\mathbf{U}^T(\boldsymbol{\theta})\mathbf{D}_\xi\mathbf{U}(\boldsymbol{\theta})\right]^{-1}\mathbf{U}^T(\boldsymbol{\theta})\mathbf{D}_\xi^2\mathbf{U}(\boldsymbol{\theta}) \\ \left[\mathbf{U}^T(\boldsymbol{\theta})\mathbf{D}_\xi\mathbf{U}(\boldsymbol{\theta})\right]^{-1}\mathbf{U}^T(\boldsymbol{\theta}) - \mathbf{I} \end{array} \right\} \\ = \;& ch_{\max}\left\{ \begin{array}{c} \left[\mathbf{U}^T(\boldsymbol{\theta})\mathbf{D}_\xi\mathbf{U}(\boldsymbol{\theta})\right]^{-1}\mathbf{U}^T(\boldsymbol{\theta})\mathbf{D}_\xi^2\mathbf{U}(\boldsymbol{\theta}) \\ \left[\mathbf{U}^T(\boldsymbol{\theta})\mathbf{D}_\xi\mathbf{U}(\boldsymbol{\theta})\right]^{-1} \end{array} \right\} - 1. \end{aligned}$$

Now for the purpose of doing the computations, it is simpler to express all in terms of $\mathbf{Z}(\boldsymbol{\theta})$. Thus final term above becomes

$$ch_{\max}\left\{ \begin{array}{c} \left[\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{D}_\xi\mathbf{Z}(\boldsymbol{\theta})\right]^{-1}\left[\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{D}_\xi^2\mathbf{Z}(\boldsymbol{\theta})\right] \\ \left[\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{D}_\xi\mathbf{Z}(\boldsymbol{\theta})\right]^{-1}\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{Z}(\boldsymbol{\theta}) \end{array} \right\} - 1.$$

Substituting into (2.7) and considering the covariance matrix $\mathbf{\Sigma}$ of (1.1) we obtain

$$\max \mathcal{L}_I = \left(\frac{\sigma_\varepsilon^2}{n} + \frac{\eta^2}{n}\right) \Re_\pi(\xi),$$

where

$$\pi = \frac{\eta^2}{(\sigma^2 + \eta^2)} \in [0,\ 1]$$

and

$$\Re_\pi(\xi) = \int_\Theta \left[\begin{array}{c} (1-\pi)\ tr\left\{ \left[\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{D}_\xi \mathbf{Z}(\boldsymbol{\theta})\right]^{-1}\ \left[\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{Z}(\boldsymbol{\theta})\right] \right\} \\ +\pi ch_{\max} \left\{ \begin{array}{c} \left[\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{D}_\xi \mathbf{Z}(\boldsymbol{\theta})\right]^{-1} \left[\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{D}_\xi^2 \mathbf{Z}(\boldsymbol{\theta})\right] \\ \left[\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{D}_\xi \mathbf{Z}(\boldsymbol{\theta})\right]^{-1} \left[\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{Z}(\boldsymbol{\theta})\right] \end{array} \right\} \end{array} \right] p(\boldsymbol{\theta}) d\boldsymbol{\theta}. \qquad (2.8)$$

Now we can look at minimizing $\Re_\pi(\xi)$ over the design, for fixed values of π and $\mathbf{\Sigma}$. To do this, a numerical minimization will be done by the Genetic Algorithm (Coley 1999) in the next chapter.

Chapter 3: The Genetic Algorithm and Loss Minimization

3.11 Introduction

Genetic Algorithms (GAs), (Coley 1999) which are a family of computational models, have been developed by using the notion of evolution theory. The implementation of a GA starts with some inputs as a set of solutions, called a population, to a particular problem. A metric called a fitness function is defined to evaluate each candidate quantitatively. These candidates are usually generated at random. Only the promising candidates are kept and allowed to reproduce. It is interesting to note that random changes occur during reproduction. A new pool of candidate solutions is obtained after reproduction. Again their fitnesses are evaluated and promising candidates are selected. These winning candidates are copied over into the next generation possibly with random changes and the process repeats. It is expected that the average fitness of the population will increase each round. So we can hope to find very good solutions to a problem by repeating this process for hundreds of rounds. There are other computerized problem-solving techniques such as Hill-climbing, Simulated annealing etc., which are in some ways similar to genetic algorithms. But by using GAs we can solve many large complex problems while other methods have experienced difficulties.

Before we begin to work with a GA, it requires to encode the potential solutions to a specific problem. We can use binary strings: sequences of 1's and 0's, where they represent the value of some aspect of the solution. Besides this, real valued numbers and strings of letters are also used for encoding. However, we will follow the first approach for our problem. Another important issue of GA is the way we select individuals to be copied over into the next generation. There are different methods such as Elitist selection, Roulette-wheel selection, Fitness-proportionate selection, Tournament selection etc.

3.12 Parameters of the GA

Crossover and mutation are the most important parts of the genetic algorithm. Actually, the performance of a GA is greatly influenced by these two operators. While crossover selects characteristics from parents to create a new offspring, mutation changes the characteristics of the new offspring randomly.

The two basic parameters of GA are crossover probability and mutation probability. The crossover probability indicates how often crossover will be performed. If there is no crossover, the offspring is an exact copy of the parents. However, if there is a crossover, offspring are made up of characteristics taken from both of parents. Crossover is made in hopes of getting new offsprings that will be better inputs on the way to obtaining a good solution to a specific problem. Typical values of crossover probability are 0.4 to 0.9 (Coley 1999). On the other hand, mutation probability indicates how often the characteristics of offspring will be changed. If there is no mutation, an offspring is the same as is obtained after crossover. Mutation prevents GA from falling into a local extreme, but it should not occur very often. The mutation probability is typically of the order 0.001 (Coley 1999). However, the correct setting for it will be problem dependent. Another important parameter of GA is population size that indicates the number of members that will be considered in one generation in a population. GA will not work efficiently if the population size is too large or too small. In our case, a population consists of the set of designs.

3.13 The GA for Our Problem

For a nonlinear regression model we want to find design points so that (2.8) is minimized. To do this, the following genetic algorithm, as in Welsh and Wiens (2011), will be implemented. Note that we will often choose equally spaced design points $x_1, x_2, \cdots, x_N$.

1. Start by randomly generating a first 'generation' of n_g designs.

2. For the current generation of designs, compute the loss: $loss_k = \mathcal{L}_I(\xi_k)$ for each design $\xi_k,\ k = 1, 2, \cdots, n_g$, and the corresponding 'fitness levels'

$$fitness_k = \frac{1}{(loss_k - 0.99\ loss_{\min})^2},\ k = 1, 2, \cdots, n_g,$$

where $loss_{\min}$ is the minimum value of the loss in the current population. Scale the fitness levels $\{fitness_k\}_{k=1}^{n_g}$ to form a probability distribution

$$\psi_k = \frac{fitness_k}{\sum_{j=1}^{n_g} fitness_j},\ k = 1, 2, \cdots, n_g.$$

3. Form a new generation of n_g designs to replace the current generation in the following way.

(a) Include the fittest $N_{elite} = n_g P_{elite}$ of the current generation; they are an 'elite' group which survives through to the next generation. The remaining $n_g - N_{elite}$ members are formed by crossover and mutation.

(b) Crossover proceeds as follows:

- Choose two members of the current population to be parents with probability proportional to their fitness level: If $\zeta_1, \zeta_2 \sim independent\ Uniform\ (0, 1)$, then choose to be parents the current generation members i_1^* and i_2^*, where

  $$i_1^* = \min\left\{i : \sum_{j=1}^{i} \psi_j \geqslant \zeta_1\right\}\ and\ i_2^* = \min\left\{i : \sum_{j=1}^{i} \psi_j \geqslant \zeta_2\right\}.$$

  (The same parent can be chosen twice without posing difficulties for the algorithm.)

- With probability $1 - P_{crossover}$, the child is identical to the fittest parent.

- With probability $P_{crossover}$, the parents both contribute towards the child, in the following manner. Each member of the current generation can be represented by its vector $n\boldsymbol{\xi}$ of allocations. The two vectors of allocations arising from the parents are averaged, and any fractional allocations are rounded down. This results in a vector with integer elements, with sum s possibly less than n. If $s < n$ then $n - s$ design points are randomly chosen from S (with replacement) and added to the design. The child formed in this way is added to the new generation.

(c) Mutation is applied independently to each child - regardless of how the child is formed - as follows. With probability $P_{mutation}$, k elements of the vector $\boldsymbol{\xi}$ defining the child are randomly chosen, and permuted. The value of k is chosen by the user; we typically use $2 \leqslant k \leqslant 6$. With probability $1 - P_{mutation}$ we do nothing.

4. Step 3 is carried out until the next generation has been formed. Then its fitness levels are computed and the process is repeated from Step 2. The loss is guaranteed to decrease (weakly) in each generation, because of the inclusion of the elite members. We run the algorithm until the best design has not changed in G consecutive generations.

3.14 Minimization of the Loss

We will do a numerical minimization of (2.8). Let us consider a nonlinear function of the form

$$f(x;\theta) = \exp(-\theta x);\ 0 < \theta < 1,\ x \geqslant 0.$$

Now we have to assume a prior distribution on θ. There are some ways of choosing it. If there are information about the desired prior in the literature, it can be helpful for finding a suitable prior distribution. However, often we adopt subjective method for choosing it. We can look at the values that θ can take on and choose the prior accordingly from our own belief. Suppose, for the above exponential model, the prior distribution of θ is uniform:

$$p(\theta) = 1;\ 0 < \theta < 1.$$

Then assuming $\pi = \frac{1}{2}$ and using (2.8), we have

$$\Re_\pi(\xi) = \frac{1}{2}\int_0^1 \left\{ \frac{f(\theta;\xi_i)}{g(\theta;\xi_i)} + \frac{h(\theta;\xi_i) f(\theta;\xi_i)}{\left(g(\theta;\xi_i)\right)^2} \right\} d\theta, \tag{3.1}$$

where $f(\theta;\xi_i) = \sum_{i=1}^N x_i^2 \exp(-2\theta x_i)$, $g(\theta;\xi_i) = \sum_{i=1}^N \frac{n_i}{n} x_i^2 \exp(-2\theta x_i)$, $h(\theta;\xi_i) = \sum_{i=1}^N \frac{n_i^2}{n^2} x_i^2 \exp(-2\theta x_i)$.

Now suppose we have generated 40 equally spaced points ranging from 0 to 10 of the covariate x. We would like to get a 30 point design minimizing the expression (3.1). We use generations of size $n_g = 40$, and vary $P_{mutation}$ linearly from 0 to 0.5. We also consider $P_{crossover} = 0.95$, $P_{elite} = 0.1$, $G = 200$. The integration was carried out by Simpson's rule, using a 101-point quadrature for the one-dimensional integration. Based on the genetic algorithm mentioned in the previous section, an execution of R codes gives the output as in the Figure 3.2, where we can see that the loss sharply drops at the beginning and then it decreases slowly and remains fixed after certain

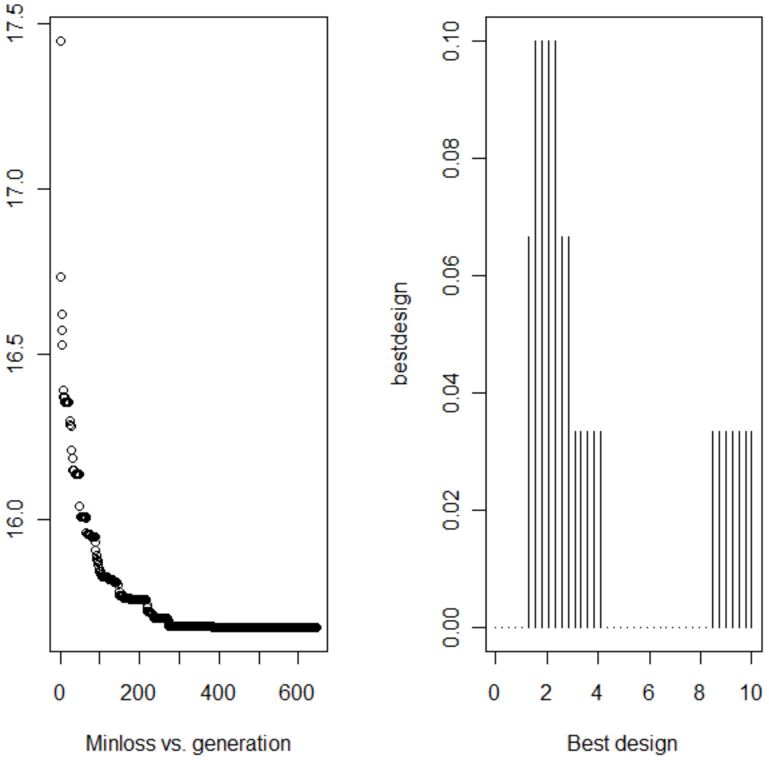

Figure 3.2: Minimax design with $n = 30$, $N = 40$ and $n_g = 40$; $loss$ =15.67.

number of generations. Moreover, the design points have been shown in the plot. It is observed that larger masses are made at around the site of $x = 2$ and $x = 3$.

It is interesting to say that our methods are also valid for linear models. In order to check how our methods are performing, let us consider a linear model, say cubic regression on $[-1,\ 1]$. Both the designs obtained by our methods and by Fang and Wiens (2000) are shown in the Figure 3.3. They can be compared based on the loss. It is clear that the loss obtained in the present study is much less than the loss obtained by Fang and Wiens (2000). So, an improvement over loss minimization has been achieved.

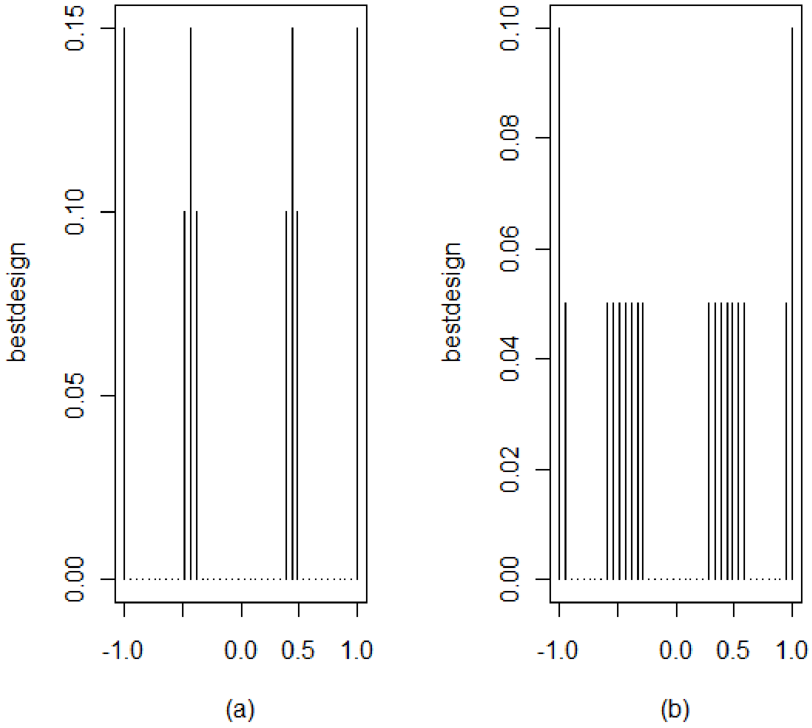

Figure 3.3: Minimax design for cubic regression on $[-1, 1]$ with $n = 20$, $N = 40$. (a) obtained in current study by using GA with $loss = 113.09$; (b) obtained in Fang and Wiens (2000) by simulated annealing with $loss = 116.52$.

Chapter 4: Implementation

4.15 Designs for a model with Rumford's data

We mentioned in section 1.2 that Count Rumford, in 1798, conducted an experiment where an object was allowed to cool from an initial temperature of 130^0F to an ambient temperature of 60^0F. A model based on Newton's law of cooling was fitted, using $f(x|\theta) = 60 + 70e^{-\theta x}$. Here the covariate x represents time and f is predicted temperature. From Bates and Watts (1988), we have the following values of x:

$$4,\ 5,\ 7,\ 12,\ 14,\ 16,\ 20,\ 24,\ 28,\ 31,\ 34,\ 37.5,\ 41.$$

We redesign this experiment using the same values of x. In order to get the minimax design, we use $n = 20$, $n_g = 40$, $P_{crossover} = 0.95$, $P_{elite} = 0.1$, $G = 200$. We also assume that $p(\theta)$ is the uniform density on $[0, 1]$, $\pi = 0.5$ and $P_{mutation}$ will vary linearly from 0 to 0.5. The minimax designs for Rumford's experiment are shown in Figure 4.4.

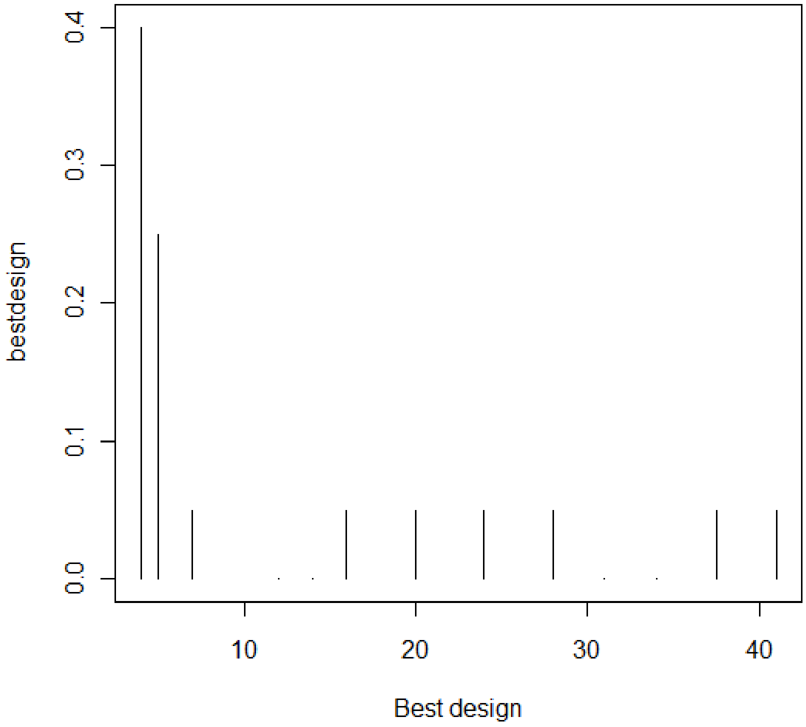

Figure 4.4: Minimax design for the Rumford experiment; $loss = 3.423$.

4.16 Designs for Exponential response model

Let us consider the following approximate exponential response model

$$f\left(x|\theta\right) = e^{-\theta x};\ x \geqslant 0,\ 0 < \theta < 1.$$

We assume that the prior density of θ is $Beta(p, q)$ i.e. beta density with parameters p and q. Considering different combinations of values of p and q, we shall construct minimax designs for the above model. The design space, in this example, will consist of 40 equally spaced points ranging from 0 to 10. We take $n = 30$, $\pi = 0.5$, $n_g = 40$. The other tuning parameters of the GA are same as in the example of Rumford design of the previous section. The minimax designs for this example are shown in Figure 4.5 and Figure 4.6.

Now we will look at minimax designs for exponential model with different combinations of the tuning constants of GA. To do this, we will consider 20 equally spaced points of x ranging from 0 to 10, $\pi = .5$, $n = 20$ and the fixed values of parameters of the prior density $(p, q) = (3, 3)$. See Figure 4.7, where it is evident that even if the GA parameters are changed, the designs are not changing; they remain identical with the same loss.

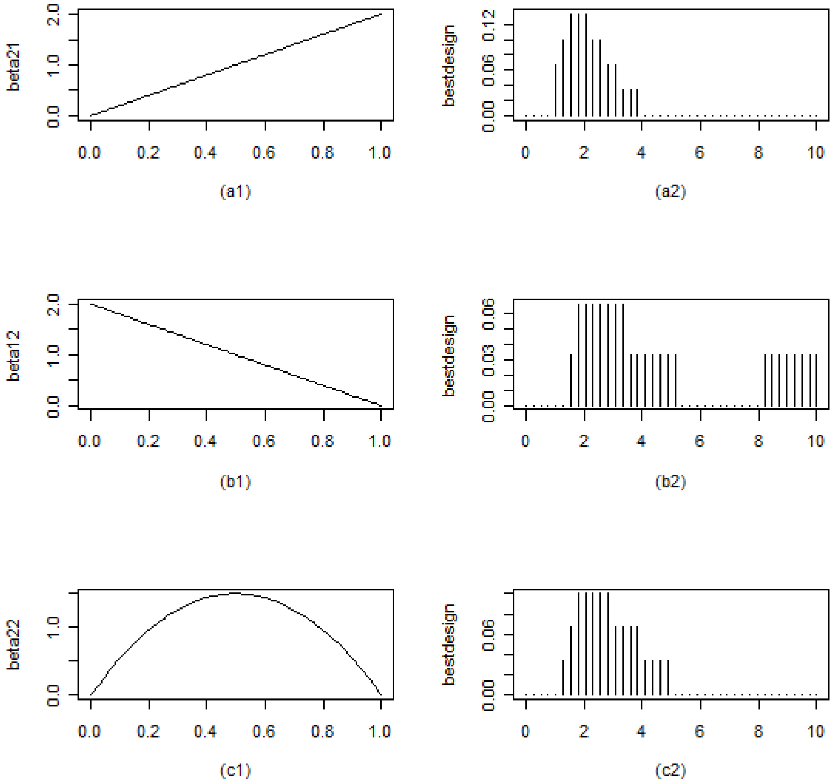

Figure 4.5: Minimax designs for exponential response model and various beta priors. (a1) beta density with $(p,\ q) = (2, 1)$ and (a2) best design with $loss = 10.39$; (b1) beta density with $(p, q) = (1, 2)$ and (b2) best design with $loss = 17.97$; (c1) beta density with $(p, q) = (2, 2)$ and (c2) best design with $loss = 13.35$.

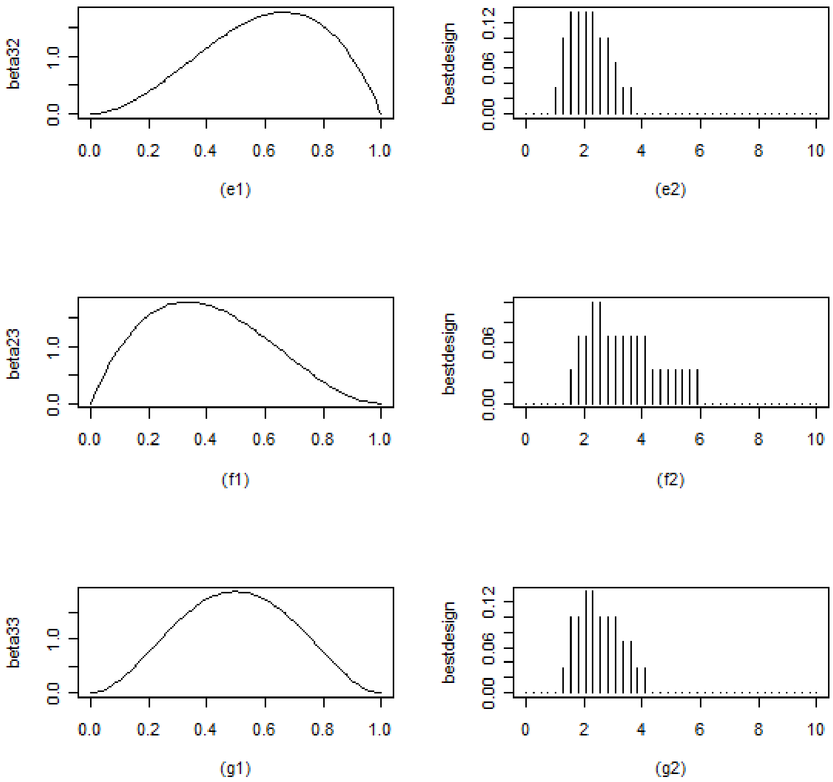

Figure 4.6: Minimax designs for exponential response model and various beta priors. (e1) beta density with $(p,\ q) = (3, 2)$ and (e2) best design with $loss = 10.02$; (f1) beta density with $(p, q) = (2, 3)$ and (f2) best design with $loss = 15.20$; (g1) beta density with $(p, q) = (3, 3)$ and (g2) best design with $loss = 11.81$.

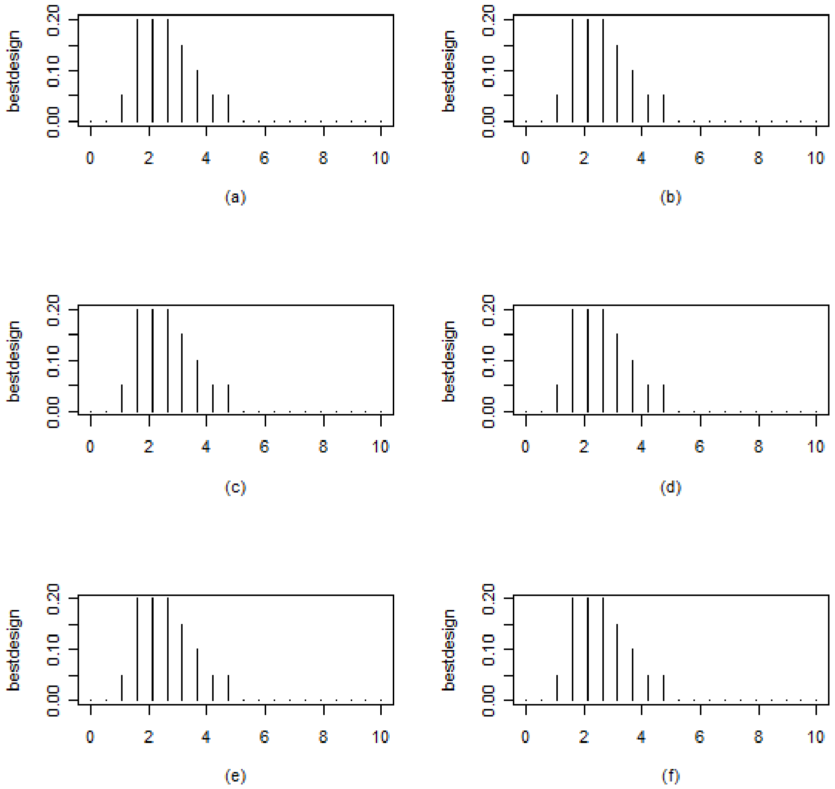

Figure 4.7: Minimax designs for different combinations of tuning constants of GA, namely (G, n_g, P_c, $P_{int.m}$, $P_{final.m}$, P_{elite}). (a) (200, 20, .9, 0, .4, .1); (b) (210, 30, .91, 0, .45, 1/15); (c) (220, 25, .92, .05, .49, 2/25); (d) (230, 45, .94, .09, .5, 2/45); (e) (240, 50, .89, .07, .52, 2/50); (f) (250, 40, .95, 0, .5, 2/45); $loss = 6.275$.

4.17 Designs for Michaelis-Menten model

Let us consider an approximate, two-parameter Michaelis-Menten model, with

$$f\left(x|\boldsymbol{\theta}\right) = \frac{\theta_1 x}{\theta_2 + x}.$$

Also assume that the covariate x takes on $N = 11$ equally spaced values spanning $[0,1]$. We develop this example based on the Puromycin experiment from Bates and Watts (1988), where estimates

$$\begin{aligned} \widehat{\boldsymbol{\theta}} &= (195.8, 0.0484) \\ &\approx (200, 0.05) \end{aligned}$$

were obtained by linearizing $1/f$ and carrying out a preliminary linear regression. Thus we may introduce two random variables $\psi_1, \psi_2 \in [0,1]$, defined by

$$\begin{aligned} \theta_1 &= 200(\psi_1 + 0.5) \in [100, 300]\,, \\ \theta_2 &= \frac{1}{20}(\psi_2 + 0.5) \in [0.025, 0.075]\,. \end{aligned}$$

By (2.8), we have

$$\Re_\pi(\xi) = \int_0^1 \int_0^1 \phi\left(200(\psi_1 + 0.5),\ \frac{(\psi_2 + 0.5)}{20},\ \xi\right) p(\psi_1) p(\psi_2) d\psi_1 d\psi_2,$$

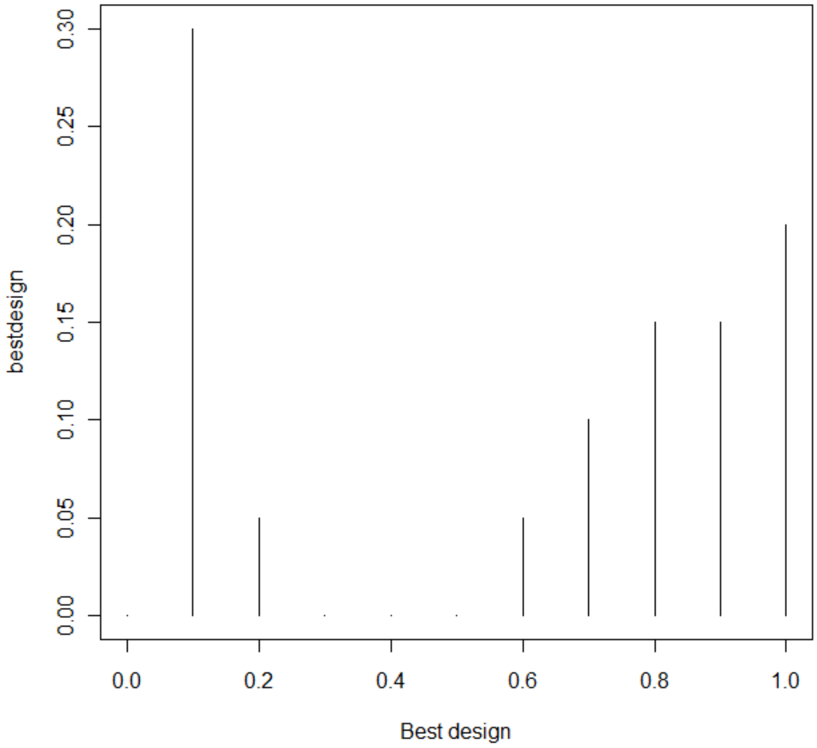

Figure 4.8: Designs for approximate Michaelis-Menten model with loss $= 8.51$.

where $\phi(\cdot)$ represents MSE, $p(\psi_1)$ and $p(\psi_2)$ are identical $Beta(p, q)$ densities. We take $\pi = .5$, $n = 20$, $n_g = 20$ and obtain the design shown in Figure 4.8 for the combination $(p, q) = (20, 20)$. The integration was carried out by Simpson's rule, using a 51-point quadrature on each axis for two-dimensional integration.

4.18 Discussion

In obtaining an optimal design, it is helpful to run the GA several times with seeding the best design found to date into the initial generation. It is realized from the Figure 4.4 that in order to get a minimax design for Rumford's experiment, most of the observations were made at the time points 4 and 5. The minimax design for this experiment shows that there are some time points where no observations were made at all.

For the plots of Figure 4.5 and Figure 4.6 of minimax designs for exponential response model with simulated values of the covariate, it is seen that larger masses are placed at around the point $x = 2$ of the design space. We can realize from these plots that the designs change to some extent with change in the values of the parameters of the prior density. However, the minimax designs do not spread throughout the whole design space. On the other hand, minimax designs for exponential response model based on fixed values of parameters of the prior density with six different combinations of tuning constants of GA have been shown in Figure 4.7. We see that the designs do not change.

The minimax design of the approximate Michaelis-Menten model in Figure 4.8 shows that most of the mass is placed at 0.1, and then the weight of masses increases after the site 0.6.

Bibliography

Abdelbasit, K.M., and Plackett, R.L. (1981), "Experimental Design for Categorized Data," International Statistical Review, 49, 111-126.

Abdelbasit, K. M., and Plackett, R. L. (1983), "Experimental Design for Binary Data," Journal of the American Statistical Association, 78, 90-98.

Atkinson, A.C. (1982), "Developments in the Design of Experiments," International Statistical Review, 50, 161-177.

Atwood, C.L. (1973), "Sequences converging to D-optimal designs of experiment," The Annals of Statistics, 1, 342-352.

Bates, D.M., and Watts D.G. (1988), Nonlinear Regression Analysis and Its Applications, New York: John Wiley & Sons.

Begg, C.B., and Kalish, L.A. (1984), "Treatment Allocation for Nonlinear Models in Clinical Trials: The Logistic Model," Biometrics, 40, 409-420.

Biedermann, S., Dette, H., and Pepelyshev, A. (2006), "Some Robust Design Strategies for Percentile Estimation in Binary Response Models," The Canadian Journal of Statistics, 34, 603-622.

Box, G.E.P. (1979), “Some Problems of Statistics and Everyday Life,” Journal of the American Statistical Association, 74, 1-4.

Box, G.E.P., and Draper, N.R. (1959), “A Basis for the Selection of a Response Surface Design,” Journal of the American Statistical Association, 54, 622-654.

Box, G.E.P., and Lucas, H.L. (1959), “Design of Experiments in Nonlinear Situations,” Biometrika, 49, 77-90.

Braess, D., and Dette, H. (2007), “On the Number of Support Points of Maximin and Bayesian Optimal Designs,” The Annals of Statistics, 35, 772-792.

Chaloner, K., and Larntz, K. (1989), “Optimal Bayesian Design Applied to Logistic Regression Experiments,” Journal of Statistical Planning and Inference, 21, 191-208.

Chaloner, K., and Verdinelli, I. (1995), “Bayesian Experimental Design: A Review,” Statistical Science, 10, 273-304.

Chaudhuri, P., and Mykland, P.A. (1993),“Nonlinear Experiments: Optimal Design and Inference Based on Likelihood,” Journal of the American Statistical Association, 88, 538-546.

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

Cochran, W.G. (1973), “Experiments for Nonlinear Functions,” Journal of the American Statistical Association, 68, 771-781.

Coley, D.A. (1999), An Introduction to Genetic Algorithms for Scientists and Engineers, Singapore: World Scientific Publishing Co.

Dette, H., and Wong, W.K. (1996), “Optimal Bayesian Designs for Models with Partially Specified Heteroscedastic Structure,” The Annals of Statistics, 24, 2108-2127.

Draper, N.R., and Hunter, W.G. (1967), “The Use of Prior Distributions in the Design of Experiments for Parameter Estimation in Non-Linear Situations,” Biometrika, 54, 147-153.

Fang, Z., and Wiens, D.P. (2000), “Integer-valued, Minimax Robust Designs for Estimation and Extrapolation in Heteroscedastic, Approximately Linear Models,” Journal of the American Statistical Association, 95, 807-818.

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

Fedorov, V. V., and Dubova, I. S. (1968). Methods for constructing optimal designs in regression experiments. Preprint No. 4 LSM, Izd-vo Moscow State University, Moscow, USSR.

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

Gallant, A. R. (1987), Nonlinear Statistical Models, New york: Wiley.

Huber, P.J. (1975), “Robustness and Designs,” in: A Survey of Statistical Design and Linear Models, ed. J.N. Srivastava, North Holland, Amsterdam, pp. 287-303.

Kiefer, J., and Wolfowitz, J. (1960), "The Equivalence of Two Extremum Problems," Canadian Journal of Mathematics, 12, 363-366.

King, J., and Wong, W.K. (2000), "Minimax D-Optimal Designs for the Logistic Model," Biometrics, 56, 1263-1267.

Li, K.C. (1984), "Robust Regression Designs when the Design Space Consists of Finitely Many Points," The Annals of Statistics, 12, 269-282.

Li, K.C., and Notz, W. (1982), "Robust Designs for Nearly Linear Regression," Journal of Statistical Planning and Inference, 6, 135-151.

Marcus, M.B., and Sacks, J. (1976), "Robust designs for regression problems," in: Statistical Decision Theory and Related Topics II, ed. S. S. Gupta and D. S. Moore, New York: Academic press, pp. 245-268.

Matthews, J.N.S., and Allcock, G.C. (2004), "Optimal designs for Michaelis–Menten kinetic studies," Statistics in Medicine, 23, 477–491.

Maxim, L.D., Hendrickson, A.D., and Cullen, D.E. (1977),"Experimental Design for Sensitivity Testing: The Weibull Model," Technometrics, 19, 405-412.

Meeker, W.Q. (1984),"A Comparison of Accelerated Life Test Plans for Weibull and Lognormal Distributions and Type I Censoring," Technometrics, 26, 157-171.

Meeker, W.Q., and Hahn, G.J. (1978), "A Comparison of Accelerated Test Plans to Estimate the Survival Probability at a Design Stress," Technometrics, 20, 245-247.

Pesotchinsky, L. (1982), “Optimal Robust Designs: Linear Regression in R^k,” The Annals of Statistics, 10, 511-525.

Sacks, J., and Ylvisaker, D. (1978), “Linear Estimation for Approximately Linear Models,” The Annals of Statistics, 6, 1122-1137.

Seber, G.A.F., and Wild, C.J. (1989), Nonlinear Regression, New York: John Wiley & Sons.

Silvey, S.D. (1980), Optimal design : an introduction to the theory for parameter estimation, London:Chapman and Hall.

Silvey, S.D., and Titterington, D.M. (1973),“A geometric approach to optimal design theory,” Biometrika, 60, 21-32.

Sinha, S., and Wiens, D.P. (2002), “Robust sequential designs for nonlinear regression,” The Canadian Journal of Statistics, 30, 601-618.

Steinberg, D.M., and Hunter, W.G. (1984), “Experimental Design: Review and Comment,” Technometrics, 26, 71-97.

St. John, R. C. (1973). Models and designs for experiments with mixtures. Ph.D. Thesis, Department of Statist., University of Wisconsin, Madison, Wisconsin.

St. John, R.C., and Draper, N. R. (1975), “D-Optimality for Regression Designs: A Review,” Technometrics, 17, 15-23.

Welsh, A. H., and Wiens D. P. (2011), “Robust Model-based Sampling Designs,” preprint.

White, L.V. (1973), “An Extension of the General Equivalence Theorem to Nonlinear Models,” Biometrika, 60, 345-348.

Wiens, D.P. (1990), “Robust, Minimax Designs for Multiple Linear Regression,” Linear Algebra and Its Applications, Second Special Issue on Linear Algebra and Statistics, 127, 327-340.

Appendix

A sample R codes are given below.

```
DESIGN <- function(n, unchangedLimit, popSize, crossoverProb,
                   initMutationProb, finalMutationProb, eliteProp,
                   a, b, c, d, p, q, nu) {
tic <- proc.time()
x=seq(0, 1, by=.1)
N = length(x)
#initial generation
lossvec=vector(length=popSize)
genmat=NULL
generation=1
unchanged=0
#simulate initial population of designs, an N by popSize matrix
#whose columns are the allocation vectors for the first generation
population = rmultinom(popSize, n, prob = rep(1,N))
#Compute loss
```

```
for (count in 1:popSize) { lossvec[count]=LOSS(population[ ,count], n, N, x, a, b,c,d, p, q, nu) }
# The vector of psi_k values, summing to 1
psivec=FITNESS(lossvec)
#Create subsequent generations
while (unchanged<unchangedLimit)
{
newPopulation=matrix(ncol=popSize,nrow=N)
#Force elites (the best of the current population) to survive
numberElites=floor(popSize*eliteProp)
I=order(psivec,decreasing=TRUE)
newPopulation[,1:numberElites]=population[,I[1:numberElites]]
lossvec[1:numberElites]=lossvec[I[1:numberElites]]
newPopSize=numberElites
while(newPopSize < popSize) {
# Choose fit parents:
cumfit <- cumsum(psivec)
popIndices <- 1:ncol(population)
choices = vector(length = 2)
rr <- runif(2)
for (i in 1:2) choices[i] <- min(popIndices[cumfit >= rr[i]])
# Create a child from 2 randomly chosen members of
#the current population
mutationProb = initMutationProb +
(finalMutationProb-initMutationProb)*unchanged/unchangedLimit
newchild = child(allocation1 = population[ ,choices[1]], allocation2 = population[ ,choices[2]],
psi1 = psivec[choices[1]], psi2 = psivec[choices[2]], n, N, crossoverProb, mutationProb)
newPopSize = newPopSize + 1
newPopulation[ , newPopSize] = newchild
}
population = newPopulation
for (count in 1:popSize) {
lossvec[count]=LOSS(population[ ,count], n, N, x, a, b, c,d,p, q, nu) }
psivec=FITNESS(lossvec)
# Summarize this generation:
bestdesign <- population[, lossvec == min(lossvec)]
if (is.matrix(bestdesign) && ncol(bestdesign) > 1) bestdesign <- bestdesign[, 1]
genmat <- cbind(genmat, c(bestdesign, min(lossvec)))
if (generation > 1 && identical(genmat[1:N,generation], genmat[1:N, generation - 1])) {
unchanged <- unchanged + 1
} else {
unchanged <- 0
}
```

```
psivec <- FITNESS(lossvec) # fitness of all designs in generation
cat("generation =", generation, "unchanged =", unchanged,
"mutation prob =", round(mutationProb,3), "min loss =", round(min(lossvec),5), "\n")
generation <- generation + 1
} # End of outer "while"
# Summarize and plot the output
bestDesignInGenerations <- genmat[1:N, ] # best designs; one for #each generation
bestLossInGenerations <- genmat[N+1,] # best loss per generation
# best overall is in the last column of bestDesignInGenerations:
bestdesign <- bestDesignInGenerations[1:N, ncol(genmat)]/n
PLOTS(x, bestLossInGenerations, bestdesign)
cat("Number of generations =", ncol(genmat), "\n")
toc <- proc.time()
cat("Time used =",round((toc[3]-tic[3])/60,1),"minutes","\n")
list(DES = cbind(x, bestdesign), MINLOSS = genmat[N+1, ncol(genmat)])
}
#Simpson's rule for double integral
simp2=function(mattheta12,a,b,c,d)
{
mtheta1=nrow(mattheta12)-1
htheta1=(b-a)/mtheta1
```

```
dtheta1=c(1,rep(c(4,2),mtheta1/2-1),4,1)

mtheta2=ncol(mattheta12)-1

htheta2=(d-c)/mtheta2

dtheta2=c(1,rep(c(4,2),mtheta2/2-1),4,1)

(htheta1*htheta2*sum(outer(dtheta1,dtheta2)*mattheta12))/9

}

#Loss function

LOSS=function(allocations, n, N, x, a, b,c,d, p, q, nu) {

Dshi = diag(allocations)/n

theta1=seq(a,b,length=51)

theta2=seq(c,d,length=51)

mattheta12=matrix(0,nrow=length(theta1), ncol=length(theta2))

betadist1=c()

betadist2=c()

for (i in 1:length(theta1))

{

for      (j in 1:length(theta2))

{

Z= cbind(x/(x+theta2[j]), -theta1[i]*x/(x+theta2[j])^2)

betadist1[i]=(1/beta(p,q))*((theta1[i]/200-0.5)^(p-1))*

((1-theta1[i]/200+.5)^(q-1))

betadist2[j]=(1/beta(p,q))*((20*theta2[j]-.5)^(p-1))*

((1-20*theta2[j]+.5)^(q-1))
#Trace function for the 1st term in Loss
tr = function(mat) {
sum(diag(mat))
}
mat1=(solve(t(Z)%*%Dshi%*%Z))%*%(t(Z)%*%Z)
tr(mat1)
# Maximum eigenvalue of the 2nd term
mat2=(solve(t(Z)%*%Dshi%*%Z))%*%(t(Z)%*%(Dshi^2)%*%Z)%*%
(solve(t(Z)%*%Dshi%*%Z))%*%(t(Z)%*%Z)
chmax=eigen(mat2, only.values = T)$values[1]
mattheta12[i,j]=((1-nu)*tr(mat1)+(nu)*chmax)*(1/10)*betadist1[i]*betadist2[j]
}
}
simp2(mattheta12,a,b,c,d)
}
#Fitness function
FITNESS=function(lossvec){
fit=1/(lossvec-.99*min(lossvec))^2
fit/sum(fit)
}
#Child function
child = function(allocation1, allocation2, psi1, psi2, n,
N, crossoverProb, mutationProb) {
```

```
U = runif(2)
# As in the second bullet of Step 3(b):
if (U[1] > crossoverProb) {if (psi1 > psi2) child = allocation1 else child = allocation2}
# As in the third bullet:
if (U[1] <= crossoverProb) {
child = floor((allocation1 + allocation2)/2) # The average of the #two allocations, rounded down
deficiency = n - sum(child)
if (deficiency > 0) {
# Choose the indices of the additional design points:
newpoints = sample(N, size = deficiency, replace = TRUE)
# Assign these to the child:
for (i in 1:deficiency) { child[newpoints[i]] = child[newpoints[i]] + 1 }
}
}
Q1=length(child)
Q2=sample(Q1,3)
# Mutate, with prob. Pmutation
if (U[2] <= mutationProb) child[Q2] = sample(child[Q2])
child # This is the ’child’ allocation vector
}
#Plot function
```

```
PLOTS <- function(x, minloss, bestdesign) {

des <- bestdesign

par(mfrow = c(1,2))

plot(minloss, xlab = "Minloss vs. generation", ylab = "")

plot(x, bestdesign, type = ’h’, xlab = "Best design", main = NULL)

}

output = DESIGN( n = 20, unchangedLimit = 200, popSize = 20,

crossoverProb = .95, initMutationProb = 0,

finalMutationProb = .5, eliteProp = .1, a =100, b = 300,

c=.025, d=.075, p = 20, q =20, nu = .5)

output
```