Optimal Design of Experiments

by

Linwei Hu

(Under the direction of Professor John Stufken)

Abstract

In design of experiments, optimal designs are designs that can glean the maximal amount of information from a study. Therefore, an optimal design can reduce the number of experimental units needed and saving the cost of study. However, the research in designing optimal experiments has not kept up with the increasingly complicated structure of data and models; especially for correlated data and multiple-covariate models, finding optimal designs is very difficult.

In a series of papers by Yang and Stufken, the complete class approach has been revitalized by applying it to the optimal design problem with great success. Their inspirational idea has spawned my research, which includes three projects for three different topics.

In the first project, we develop a general approach to find optimal designs for independent data with a single covariate. There has been lots of research under this topic, but most of the work is done on a case by case basis. So we propose a unified way of finding optimal designs for a class of models under general optimality criteria. In the second project, we consider correlated data with a single covariate. There are very few results under this topic. To bridge the gap, we extend the result from independent data to correlated data. Finally, we consider multiple-covariate models under independent data. We are unable to find closed-

form solutions for optimal designs, but we give a complete class result that can save the computational resources by orders of magnitude.

INDEX WORDS: Chebyshev system, complete class, concavity, locally optimal design, Loewner ordering, robustness

Optimal Design of Experiments

by

Linwei Hu

B.A., Zhejiang University, 2009

M.A., University of Missouri, 2011

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2014

Optimal Design of Experiments

by

Linwei Hu

Approved:

Major Professor: John Stufken

Committee: Abhyuday Mandal
Jaxk Reeves
Pengsheng Ji
Xiangrong Yin

Electronic Version Approved:

Julie Coffield
Interim Dean of the Graduate School
The University of Georgia
December 2014

Optimal Design of Experiments

Linwei Hu

December 4, 2014

Acknowledgments

This PhD thesis would not have been possible without the help of many people, whom I would like to thank.

First, I would like to thank my supervisor Professor John Stufken, for his continued support, encouragement and guidance. Despite his heavy duties as a department head and journal editor, he spent a great amount of time reading my reports, discussing with me and providing insightful comments over the past three years. Besides, he is a very kind and fun person, it's been my great pleasure working with him.

Second, I would like to thank my supervisor in the University of Missouri, Professor Min Yang. He led me into the design of experiments world, and helped me immensely in getting started on research. My first research project was done under the guidance of Professor Yang, and my later research projects also benefited greatly from his continued input.

Third, I also want to thank my committee members, Professor Abhyuday Mandal, Professor Pengsheng Ji, Professor Xiangrong Yin and Professor Jaxk Reeves. They offered me valuable input for my dissertation as well as other help.

Finally, I am much obliged to my family for their long-standing support and encouragement, especially my wife Hui, who has always been supportive and made it all possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In statistics, an experiment is an information gathering procedure carried out in order to understand the relationship between certain inputs and outputs. By manipulating the input factors and observing what happens to the output, data are collected from the experiment, and statistical analysis can be performed to study the relationship between the input and output variables.

Experimentations have a very broad application across all the natural and social sciences as well as engineering. For example, a pharmaceutical company that hopes to understand the pharmacokinetics of a new drug can conduct an experiment, whereby the experimental units, patients, are each injected with one of several doses of the drug and their drug concentrations are measured over time. As another example, a company that decides to launch an email campaign wants to know which of the following two calls of action is more effective: "Offer ends this Saturday!" or "Offer ends soon!". An experiment can be conducted by sending emails to 1000 people using the first call of action and another 1000 people with the second call of action, and compare the response rates.

Design of experiment deals with planning and conducting a better experiment so that a better understanding of the relationship between input and output is achieved. Modern

experimental design theory dates back to the seminal work of R.A. Fisher, in his innovative books: The Arrangement of Field Experiments (1926) and The Design of Experiments (1935), when he was at the Rothamsted Agricultural Experimental Station in UK. Back at that time, design work was motivated by problems in agriculture, and the main purpose was to design experiments that allow for treatment comparisons. The important concepts of randomization, blocking and replication were developed at that time. During the last several decades, statistics has revolutionized numerous disciplines and as a result, design of experiments has also been adopted in different areas, such as psychology, medicine, business, etc. Meanwhile, the goal of experimental designs has also evolved from treatment comparison to variable screening, response surface exploration, system optimization, and so on. In this dissertation, we will focus specifically on optimal experimental design.

1.1 Optimal design

The goal of optimal design is to design the best experiment that gives us the most information about the unknown parameters. Because an optimal design is more efficient, it requires fewer experimental runs to achieve the same result as a non-optimal design, hence saves resources.

A design is usually denoted as $\xi = \{(x_k, \omega_k)\}_{k=1}^{t}$, where x_k's are the design points and ω_k's are the weights put on each design point. The total number of experimental units is fixed to be n, thus $\sum_{k=1}^{t} \omega_k = n$. In an exact design, ω_k's have to be integers. However this creates difficulty in finding the optimal designs due to discrete optimization. Hence *approximate designs* are usually used in which the weights can be any non-negative numbers (Kiefer 1959). An efficient exact design can be found by rounding the weights in an optimal approximate design.

The theory of optimal design can be described as given a model, find the design that can maximize the information matrix (or minimize the variance-covariance matrix) under a

certain criterion. For example, for univariate linear/nonlinear models

$$y = \eta(x, \boldsymbol{\beta}) + \epsilon,$$

the (scaled) information matrix under ξ can be calculated as (assuming observations are independent)

$$\mathbf{M}(\xi) = \sum_{k=1}^{t} \omega_k \mathbf{M}(x_k) = \sum_{k=1}^{t} \omega_k \frac{\partial \eta(x_k, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \left(\frac{\partial \eta(x_k, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right)^T.$$

Because matrices cannot be maximized directly, different optimality criteria have been proposed, for example,

1. D-optimality maximizes $|\mathbf{M}(\xi)|$. Thus a D-optimal design minimizes the expected volume of the confidence ellipsoid for the parameters.

2. A-optimality maximizes $-\text{tr}(\mathbf{M}^{-1}(\xi))$. Thus an A-optimal design minimizes the average variance of all parameter estimates.

3. E-optimality maximizes the minimal eigenvalue of $\mathbf{M}(\xi)$. Thus an E-optimal design minimizes the largest variance of estimating $\mathbf{c}^T\boldsymbol{\beta}$ over all unit length vector $\mathbf{c}$'s.

4. c-optimality maximizes $-\mathbf{c}^T\mathbf{M}^{-1}(\xi)\mathbf{c}$. Thus a c-optimal design minimizes the variance of estimating $\mathbf{c}^T\boldsymbol{\beta}$ for some specific vector $\mathbf{c}$.

For nonlinear models, the information matrix depends on the unknown parameter $\boldsymbol{\beta}$. To circumvent this problem, an initial guess of $\boldsymbol{\beta}$ is assumed to be known, and the optimal design found based on this guess is called *locally optimal design*.

A review of the literature indicates that modern optimal design theory can be traced back to Kirstine Smith (1878-1939), who wrote a thesis that later became a precursor to modern optimal design theory, published in 1918 Biometrika. In Smith (1918), she calculated G-optimal designs for polynomial regression models of order up to 6. Other designs were also

considered in her paper, for example, uniform designs. However, she did not extend her work on optimal experimental designs; in fact, most work in this area started to appear in the 50s and 60s, with the most significant contribution from Jack Kiefer (1924-1981). The foundations behind optimal design are laid out in Kiefer (1959), in which different optimality criteria and objectives were examined and methods of deriving optimal designs are given on a sound mathematical basis. After this, a major development on optimal design is the series of equivalence theorems. For example, Kiefer and Wolfowitz (1960) proved the equivalence between G-optimality and D-optimality; Kiefer (1974) gave a general equivalence theorem for general convex optimality criteria. Such an equivalence theorem provides means to both construct an optimal design as well as check the optimality of a design. Some other developments in optimal design include Bayesian optimal designs, optimal model discrimination designs, multi-objective designs, etc.

However, over all, the research in optimal experimental design has not kept up with the increasingly complicated structure of data and data analysis methods. Therefore, great efforts need to be devoted to this area. My research tries to address three optimal design problems.

First, explicit solutions for optimal designs under specific model and specific optimality criteria are given in a few papers (eg, Dette et al. 2008, 2010), but such solutions don't exist in general. So theoretical guidance is lacking in general. Algorithms have been developed to fill in the gap, however, they usually don't provide much insight into the problem. Moreover, the bulk of them find the optimal designs on a grid of the design domain, and it has three disadvantages. First, the optimal design found on the grid is inferior to the optimal design on the whole continuous design domain. Second, a finer grid is preferred, but it requires more computational resources, and the computational cost also increases with the sample size N. Third, the majority of algorithms focus on D-optimality, for general optimality criteria, it is harder to search and a standard algorithm may not exist.

Second, how to find optimal designs for correlated data? Compared to the independent data case, results on correlated data are much more scarce. The reason is that the information matrix becomes so complicated. For nonlinear mixed effects models, the likelihood function even doesn't even have an explicit form in general.

Third, how to find optimal designs for multiple-covariate models? For multiple-covariate models (for example, logistic regression with $m > 1$ covariates), results derived under single covariate models, such as complete class results in Yang and Stufken (2013), cannot be generalized, and there are few theoretical results available. Even computationally it is very expensive to find the optimal designs, because the grid size increases exponentially.

In the remainder of this dissertation, Chapter 2 gives implicit solutions of optimal designs for a class of models under general optimality criteria, and the uniqueness of optimal designs is proved under mild conditions. Chapter 3 generalizes this approach to correlated data under mixed effects models, and studies the robustness against mis-specification of the covariance structure. Chapter 4 gives preliminary results for multiple-covariate generalized linear regression models, these results can be used to reduce the computational burden of finding an optimal design. Chapter 5 discusses future research.

Chapter 2

One covariate, independent data[1]

2.1 Introduction

In this chapter, we consider optimal design problem under independent data for a class of models with a single covariate. In particular, we give implicit solutions for optimal designs in a general case and the solution can be solved explicitly in special cases. Moreover, we show the uniqueness of optimal designs under mild conditions. It is true that there are available algorithms that can solve the optimal design problem numerically, but a theoretical solution is always preferred over a numerical one due to the following two reasons.

First, a theoretical solution is concise, readily available and provides more insights into the problem. Obviously, if an explicit solution exists, then finding a numerical solution is a waste of time. If an explicit solution doesn't exists because of the complexity of design problems, an implicit solution is the best we can get, and it is still concise because it can be described as a solution to an equation; it is readily available because a simple Newton's algorithm can solve it instantly. Moreover, our theoretical results show the uniqueness of optimal designs, a phenomenon that has been observed in several papers.

[1]This chapter is taken from Hu et al. (2015), with minor changes.

Second, there are some concerns about the numerical algorithms. The bulk of the algorithms need to discretize the design domain into a grid of points, and construct the optimal design on the grid. So it is inferior to the optimal design obtained on the whole continuous design domain. While a finer grid leads to a better approximation of the optimal design, it comes at a cost of more computational resources. Besides the speed issue, convergence is not guaranteed for some algorithms.

In order to avoid intricacies caused by the discreteness of the problem, we will work with approximate designs (see Section 2.2). Because the information matrix usually depends on the unknown parameters, we consider locally optimal designs obtained by plugging in values for the parameters in the information matrix. This gives good designs when prior knowledge of the parameters is available, and it also provides a benchmark for evaluating other designs. For the sake of simplicity, we omit the word locally hereafter.

Finding optimal designs is difficult. There are two challenges, the first challenge is to determine the number of support points in an optimal design. By Caratheodory's theorem, the number of support points in an optimal design can be taken as no more than $1+d(d+1)/2$, where d is the number of parameters. However, this bound is often weak. To reduce the bound, Kiefer's equivalence theorem [Kiefer and Wolfowitz (1960)] and Elfving's Geometric method [Elfving (1952)] have been used. These two methods have proven to be successful, and the optimal designs are found to be saturated (a saturated design is a design that has d support points, where d is the number of parameters) in many cases [see for example Dette et al. (2010)]. However, these methods can only be used on a case-by-case basis, and usually only for D- and c-optimality. So Yang and Stufken (2009), Yang (2010), Dette and Melas (2011), Yang and Stufken (2012) and Dette and Schorning (2013) studied a general method called complete class approach. Based on this approach, optimal designs can be found in a small class of designs called the complete class, and in many cases, this complete class only contains designs with at most d design points. This solves the first challenge.

The second challenge is to find an optimal design given the number of support points. Based on the optimality criterion, we can formulate the objective function to be maximized. Many optimality criteria have been proposed, and most are quite complicated, which also makes the objective function complicated. Moreover, the required number of support points increases as the model becomes more complex, also making the objective function more complex. For these reasons, explicit solutions for optimal designs are only given for a few models under particular optimality criteria. In addition, the uniqueness of optimal designs has been shown in only a few situations [see for example Dette, Melas and Wong (2006)], and has not been systematically studied either.

The goal of this project is to solve the second challenge based on the complete class results. While an explicit solution does not exist in general, we provide an implicit solution which can be solved easily using Newton's algorithm, and we prove the uniqueness of these optimal designs under mild conditions. Since our approach is based on the complete class results, it applies to a variety of models under general optimality criteria.

2.2 Theory

2.2.1 Complete class results

The models under consideration include polynomial regression models, nonlinear regression models and generalized linear models, with a univariate response y and a single covariate x which belongs to the interval $[L, U]$ (L or U could be $-\infty$ or ∞ respectively, with $[L, U]$ being half open or open). The unknown parameter is a $d \times 1$ vector denoted as $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T$. To be specific, for polynomial regression models and nonlinear models $\boldsymbol{\beta}$ is the unknown parameter in the mean response $\eta(x, \boldsymbol{\beta}) = \mathrm{E}(y)$ and we take the constant variance to be 1 since it doesn't affect the optimal design; for generalized linear models, $\boldsymbol{\beta}$ is the unknown parameter in the linear predictor $\eta(x, \boldsymbol{\beta}) = h(\mathrm{E}(y))$, where h is the link function.

In approximate design context, a design ξ with at most t design points can be written as $\xi = \{(x_k, \omega_k)\}_{k=1}^t$, where $x_k \in [L, U], \omega_k \geq 0, k = 1, \ldots, t$, x_k's and ω_k's are the design points and corresponding design weights, and $\sum_{k=1}^t \omega_k = 1$. An optimal design is a design that maximizes the Fisher information matrix $\mathbf{M}(\xi) = \sum_{k=1}^t \omega_k \mathbf{M}(x_k)$ under a certain criterion Φ. In this project we focus on a general class of optimality criteria. Specifically, let NND(d) be the set of all $d \times d$ nonnegative definite matrices, PD(d) be the set of all $d \times d$ positive definite matrices, and Φ be any nonnegative, nonconstant function defined on NND(d) that satisfies the following assumption [see Pukelsheim (1993), p. 115]

(1) it is concave, i.e., $\Phi(\alpha\mathbf{M}_1 + (1-\alpha)\mathbf{M}_2) \geq \alpha\Phi(\mathbf{M}_1) + (1-\alpha)\Phi(\mathbf{M}_2)$, where $\alpha \in (0,1), \mathbf{M}_1, \mathbf{M}_2 \in$ NND(d).

(2) it is isotonic, i.e., $\Phi(\mathbf{M}_1) \geq \Phi(\mathbf{M}_2)$ if $\mathbf{M}_1 \geq \mathbf{M}_2$ under the Loewner ordering, $\mathbf{M}_1, \mathbf{M}_2 \in$ NND(d).

(3) it is smooth on PD(d), i.e., when interpreted as a function of the $d(d+1)/2$-dimensional vector of elements in the upper triangle of $\mathbf{M} \in$PD(d), Φ is differentiable and the partial derivatives are continuous.

A design ξ^* is Φ-optimal if it maximizes $\Phi(\mathbf{M}(\xi))$ with respect to ξ.

This class of optimality criteria is very broad and includes, for example, the well known Φ_p-optimality criteria when $-\infty < p \leq 1$, which are defined as follows. Suppose we are interested in estimating a smooth function of $\boldsymbol{\beta}$, say $g(\boldsymbol{\beta}) : \mathbb{R}^d \to \mathbb{R}^v$, where $v \leq d$ and $\mathbf{K} = (\partial g(\boldsymbol{\beta})/\partial\boldsymbol{\beta})^T$ has full column rank v. It can be estimated as long as the columns of $\mathbf{K}$ are contained in the range of $\mathbf{M}(\xi)$. The information matrix for $g(\boldsymbol{\beta})$ under design ξ is defined as $\mathbf{I}(\xi) = (\mathbf{K}^T\mathbf{M}(\xi)^-\mathbf{K})^{-1}$, where $\mathbf{M}^-(\xi)$ is a generalized inverse if $\mathbf{M}(\xi)$ is singular. Then a Φ_p-optimal design for $g(\boldsymbol{\beta})$ is defined to maximize

$$\Phi(\mathbf{M}(\xi)) = \Phi_p(\mathbf{I}(\xi)) = \left(\frac{1}{v}\ \text{trace}\ (\mathbf{I}^p(\xi))\right)^{1/p}, \quad p \in (-\infty, 1].$$

In addition to the Φ_p-optimality criteria, our general Φ-optimality criteria also include compound optimality criteria, criteria for evaluating a mixture of information matrices obtained from nested models [see Pukelsheim (1993), Chap. 11] and so on.

Now we briefly introduce a fundamental theorem from Yang and Stufken (2012) for our later use. Decompose the Fisher information matrix in the following way:

$$\mathbf{M}(\xi) = \mathbf{P}\mathbf{C}(\xi)\mathbf{P}^T, \quad \mathbf{C}(\xi) = \left(\sum_{k=1}^{t} \omega_k \mathbf{C}(u_k)\right), \tag{2.2.1}$$

where $\mathbf{C}(u)$ is a $d \times d$ symmetric matrix,

$$\mathbf{C}(u) = \begin{pmatrix} \Psi_{11}(u) & & & \\ \Psi_{21}(u) & \Psi_{22}(u) & & \\ \vdots & \vdots & \ddots & \\ \Psi_{d1}(u) & \Psi_{d2}(u) & \cdots & \Psi_{dd}(u) \end{pmatrix},$$

$\mathbf{P}$ is a $d \times d$ nonsingular matrix that only depends on $\boldsymbol{\beta}$, and $u \in [A, B]$ is a smooth monotonic transformation of x that depends on $\boldsymbol{\beta}$.

For some $\tilde{d}$, $1 \leq \tilde{d} < d$, define $\mathbf{C}_{22}(u)$ as the lower $\tilde{d} \times \tilde{d}$ principal submatrix of $\mathbf{C}(u)$. Choose a maximal set of linearly independent non-constant functions from the first $d - \tilde{d}$ columns of the matrix $\mathbf{C}(u)$, and rename them as $\Psi_\ell(u)$, $\ell = 1, \ldots, q-1$. Let $\Psi_q(u) = \mathbf{C}_{22}(u)$ and define the functions $h_{r,s}(u)$, $1 \leq r \leq s \leq q$, to be

$$\begin{pmatrix} h_{1,1} = \Psi_1' & & & & \\ h_{2,1} = \Psi_2' & h_{2,2} = \left(\frac{h_{2,1}}{h_{1,1}}\right)' & & & \\ h_{3,1} = \Psi_3' & h_{3,2} = \left(\frac{h_{3,1}}{h_{1,1}}\right)' & h_{3,3} = \left(\frac{h_{3,2}}{h_{2,2}}\right)' & & \\ \vdots & \vdots & \vdots & \ddots & \\ h_{q,1} = \Psi_q' & h_{q,2} = \left(\frac{h_{q,1}}{h_{1,1}}\right)' & h_{q,3} = \left(\frac{h_{q,2}}{h_{2,2}}\right)' & \cdots & h_{q,q} = \left(\frac{h_{q,q-1}}{h_{q-1,q-1}}\right)' \end{pmatrix},$$

where the entries on the last row are matrices, and the derivatives of matrices are element-wise derivatives (assuming all derivatives exist). Define matrix $\mathbf{H}(u) = \prod_{\ell=1}^{q} h_{\ell,\ell}(u)$. Then the following theorem due to Yang and Stufken (2012) is available [see also Dette and Schorning (2013), Theorem 3.1].

Theorem 2.2.1 (Yang and Stufken (2012))**.** *For a regression model with a single covariate, suppose that either $\boldsymbol{H}(u)$ or $-\boldsymbol{H}(u)$ is positive definite for all $u \in [A, B]$. Then the following results hold:*

(a) If $q = 2a-1$ is odd and $\boldsymbol{H}(u) < 0$, then designs with at most a support points, including point A, form a complete class Ξ.

(b) If $q = 2a-1$ is odd and $\boldsymbol{H}(u) > 0$, then designs with at most a support points, including point B, form a complete class Ξ.

(c) If $q = 2a$ is even and $\boldsymbol{H}(u) < 0$, then designs with at most a support points, form a complete class Ξ.

(d) If $q = 2a - 2$ is even and $\boldsymbol{H}(u) > 0$, then designs with at most a support points, including both A and B, form a complete class Ξ.

It is helpful to sketch how Theorem 2.2.1 is proved. For some carefully chosen $\tilde{d}$ (see example below) where one of the conditions in Theorem 2.2.1 holds, it can be proved that for any design $\xi \notin \Xi$, we can find a design $\tilde{\xi} \in \Xi$ such that $\mathbf{C}(\tilde{\xi}) \geq \mathbf{C}(\xi)$ under the Loewner ordering, hence $\mathbf{M}(\tilde{\xi}) \geq \mathbf{M}(\xi)$. To be specific, $\mathbf{C}(\tilde{\xi}) - \mathbf{C}(\xi)$ has a *positive definite* lower $\tilde{d} \times \tilde{d}$ principal submatrix, and is 0 everywhere else. So the search for optimal designs can be restricted within Ξ.

Theorem 2.2.1 also applies to generalized linear models. Besides, while it is stated in terms of the "transformed design point" u, the result can be easily translated back into x using the relationship between them, and we will state results in x unless otherwise specified.

In Theorem 2.2.1, there are four different types of complete classes, the difference being whether one or both of the endpoints are fixed design points (note however a fixed design point can have weight 0 so that it need not be a support point). To make it easier to distinguish, let $fix(\Xi)$ denote the set of fixed design points for the designs in the complete class Ξ. For example, $fix(\Xi) = \emptyset$ and $\{L, U\}$ refers to the complete classes in Theorem 2.2.1(c) and Theorem 2.2.1(d), respectively.

Applications of Theorem 2.2.1 can be found in Yang and Stufken (2009), Yang (2010) and Yang and Stufken (2012). Obviously, $a \geq d$, however, in many applications we actually find $a = d$. Take the LINEXP model from Yang and Stufken (2012) as an example.

The LINEXP model is used to characterize tumor growth delay and regrowth. The natural logarithm of tumor volume is modeled using a nonlinear regression model with mean

$$\eta(x, \boldsymbol{\beta}) = \beta_1 + \beta_2 e^{\beta_3 x} + \beta_4 x, \tag{2.2.2}$$

where $x \in [L, U]$ is the time, $\beta_1 + \beta_2$ is the logarithm of initial tumor volume, $\beta_3 < 0$ is the rate at which killed cells are eliminated, $\beta_4 > 0$ is the final growth rate.

The information matrix for $\boldsymbol{\beta}$ can be written in the form of (2.2.1) with

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \frac{\beta_2}{\beta_3} \\ 0 & 0 & \frac{1}{\beta_3} & 0 \end{pmatrix}, \quad \mathbf{c}(u) = \begin{pmatrix} 1 & & & \\ e^u & e^{2u} & & \\ u & ue^u & u^2 & \\ ue^u & ue^{2u} & u^2e^u & u^2e^{2u} \end{pmatrix},$$

where $u = \beta_3 x \in [A, B] = [\beta_3 U, \beta_3 L]$. Let $\tilde{d} = 2$, $\mathbf{C}_{22}(u)$ be the lower 2×2 principal submatrix of $\mathbf{C}(u)$, and $\Psi_1(u) = u, \Psi_2(u) = e^u, \Psi_3(u) = ue^u \Psi_4(u) = e^{2u}, \Psi_5(u) = ue^{2u}$ be the set of linearly independent non-constant functions from the first two columns of $\mathbf{C}(u)$.

Then $q = 6$, $h_{1,1} = 1, h_{2,2} = e^u, h_{3,3} = 1, h_{4,4} = 4e^u, h_{5,5} = 1$, and

$$h_{6,6}(u) = \begin{pmatrix} 2e^{-2u} & e^{-u}/2 \\ e^{-u}/2 & 2 \end{pmatrix}, \quad \mathbf{H}(u) = \prod_{\ell=1}^{6} h_{\ell,\ell}(u) = \begin{pmatrix} 8 & 2e^u \\ 2e^u & 8e^{2u} \end{pmatrix}.$$

Because $\mathbf{H}(u) > 0$, Theorem 2.2.1(d) can be applied with $a = 4 = d$, and Ξ consists of designs with at most four design points including both endpoints, thus $fix(\Xi) = \{L, U\}$.

2.2.2 Identifying the optimal design

If one of the cases in Theorem 2.2.1 holds, an optimal design exists of the form $\xi = \{(x_k, \omega_k)\}_{k=1}^{a}$, where x_k's are strictly increasing, with x_1 or x_a possibly fixed to be L or U, respectively; ω_k's are nonnegative, and $\omega_1 = 1 - \sum_{k=2}^{a} \omega_k$. Let $\boldsymbol{Z}$ be the vector of unknown design points (i.e., exclude x_1 or x_a if fixed to be the endpoint) and $a - 1$ unknown weights $\omega_2, \ldots, \omega_a$. Thus we can use $\boldsymbol{Z}$ to represent the design ξ. Now the objective function $\Phi(\mathbf{M}(\xi)$ is a function of $\boldsymbol{Z}$, denoted as $\tilde{\Phi}(\boldsymbol{Z})$, and it is smooth by the smoothness of Φ. To find an optimal design, we need to maximize $\tilde{\Phi}(\boldsymbol{Z})$ with respect to $\boldsymbol{Z}$. The simplest way is to find the critical points. Theorem 2.2.2 states the conditions when a critical point gives an optimal design, where the condition "a valid design with support size a" means all a weights are positive, all a design points are distinct and within $[L, U]$.

Theorem 2.2.2. *Assume one of the cases in Theorem 2.2.1 holds. If there exists a vector $\boldsymbol{Z}^c$ such that it is a critical point of the objective function $\tilde{\Phi}(\boldsymbol{Z})$ and the corresponding design ξ^c is a valid design with support size a, then ξ^c is a Φ-optimal design.*

Proof. See Appendix □

Theorem 2.2.2 gives an implicit solution of an optimal design if there exists a critical point corresponding to a valid design with support size a, in which case we will call it a *feasible*

critical point. Such a point can be given explicitly in special situations, but not in general due to the complexity of the objective function. Nevertheless, we have an implicit solution and it can be easily solved using Newton's algorithm. However, we need to guarantee the existence of a feasible critical point in the first place. Theorem 2.2.3 gives some sufficient conditions that a feasible critical point exists.

Theorem 2.2.3. *Suppose one of the cases in Theorem 2.2.1 holds and any* Φ*-optimal design has at least a support points. Further assume the end points* L *and* U *are either a fixed design point of* Ξ *or the information matrix at which is 0, then a feasible critical point of* $\tilde{\Phi}(\boldsymbol{Z})$ *must exist, and by Theorem 2.2.2, the corresponding design is a* Φ*-optimal design.*

Proof. Let $\xi^* \in \Xi$ be a Φ-optimal design, then ξ^* has at least m support points. By Theorem 2.2.1, designs in the complete class have at most m support points, hence ξ^* has exactly m support points. Let $\boldsymbol{Z}^*$ be the vector corresponding to ξ^* according to the definition in the beginning of Section 2.2.2. For each of conditions (a)$\sim$(d), we know the design points in $\boldsymbol{Z}^*$ don't include any of the endpoints (recall the fixed design points are excluded in $\boldsymbol{Z}^*$), hence they all belong to the open interval (L, U). The weights in $\boldsymbol{Z}^*$ are all positive, hence all belong to the open interval $(0, 1)$, so $\boldsymbol{Z}^*$ is not on the boundary and must be a critical point of $\tilde{\Phi}(\boldsymbol{Z})$. This proves the existence. □

As we have stated, for many models, the complete class given by Theorem 2.2.1 only consists of designs with at most $a = d$ support points. So the condition that any Φ-optimal design has at least a support points is met for many optimality criteria. The condition (d) is found to be satisfied for several models, as we will see in Section 2.3. For condition (a), usually the information matrix becomes 0 at U only when $U = \infty$, so the condition fails if we are interested in a finite design region, and so do conditions (b) and (c). This issue will be addressed later in Theorem 2.2.6.

For the most commonly used Φ_p-optimality criteria, Corollary 2.2.4 gives some useful applications of Theorem 2.2.3.

Corollary 2.2.4. *Suppose that one of the cases in Theorem 2.2.1 holds with $a = d$, and one of the four conditions in Theorem 2.2.3 is met. Consider Φ_p-optimal design for $g(\boldsymbol{\beta})$ where $g(\boldsymbol{\beta})$ satisfies either case (i) or (ii) below,*

(i) $g(\boldsymbol{\beta}) = \boldsymbol{\beta}$ or a reparametrization of $\boldsymbol{\beta}$.

(ii) $g(\boldsymbol{\beta}) = \boldsymbol{c}^T\boldsymbol{\beta}, \boldsymbol{c} = (c_1, \ldots, c_d)^T$ is a $d \times 1$ vector and $\boldsymbol{c}^T\boldsymbol{\beta}$ can only be estimated with at least d support points.

Then a feasible critical point of $\tilde{\Phi}(\boldsymbol{Z})$ exists, and the corresponding design is a Φ_p-optimal design for $g(\boldsymbol{\beta})$.

Remark 2.2.1. In Corollary 2.2.4(i), a special case of a reparameterization is $g(\boldsymbol{\beta}) = \mathbf{W}\boldsymbol{\beta}$, $\mathbf{W}$ is a diagonal matrix with positive diagonal elements. This will result in a rescaling of the covariance matrix for $\hat{\boldsymbol{\beta}}$, and it makes sense when the magnitudes of $\text{var}(\hat{\beta}_i)$'s differ by orders. For example, in Dette (1997), the author proposed "standardized" optimality criteria, where the matrix $\mathbf{W}$ has diagonal elements $W_{ii} = \sqrt{1/(\mathbf{M}^{-1}(\xi_i^*))_{ii}}$, ξ_i^* is the c-optimal design for estimating β_i alone, $i = 1, \ldots, d$. Under the conditions of Corollary 2.2.4(i), finding such optimal designs is easy after we find ξ_i^*'s.

Remark 2.2.2. Corollary 2.2.4(ii) considers c-optimality. The set of all vectors $\boldsymbol{c}$ such that $\boldsymbol{c}^T\boldsymbol{\beta}$ is only estimable with at least d design points is denoted as A^* in Kiefer and Wolf (1965). When $\boldsymbol{c} \in A^*$, the c-optimal design is supported at the full set of Chebyshev points in many cases [see Studden (1968)], but our method gives another way of finding c-optimal designs. When $\boldsymbol{c} \notin A^*$, sometimes a feasible critical point still exists, and it still gives an optimal design. However, if there is no such critical point, then c-optimal designs must be supported at fewer points, which may not be the Chebyshev points, and this problem

becomes harder. Nevertheless, we can approximate such c-optimal designs. Suppose $c_1 \neq 0$, consider $g_\epsilon(\boldsymbol{\beta}) = (\boldsymbol{c}^T\boldsymbol{\beta}, \epsilon\beta_2, \ldots, \epsilon\beta_d)^T, \epsilon > 0$. A Φ_p-optimal design for $g_\epsilon(\boldsymbol{\beta})$ can be found easily by Corollary 2.2.4(i). Let $\epsilon \to 0$, it can be shown that these Φ_p-optimal designs will eventually converge to the c-optimal design for $\boldsymbol{c}^T\boldsymbol{\beta}$ (i.e., the efficiencies of these Φ_p-optimal designs under c-optimality will converge to 1), for any $p \leq -1$. Some examples are provided in Section 2.3.

To verify the condition $\boldsymbol{c} \in A^*$, let $\mathbf{f}(x, \boldsymbol{\beta}) = (f_1(x, \boldsymbol{\beta}), \ldots, f_d(x, \boldsymbol{\beta}))^T = \partial\eta(x, \boldsymbol{\beta})/\partial\boldsymbol{\beta}$. The condition $\boldsymbol{c} \in A^*$ is equivalent to

$$\begin{vmatrix} f_1(x_1, \boldsymbol{\beta}) & \cdots & f_1(x_{d-1}, \boldsymbol{\beta}) & c_1 \\ f_2(x_1, \boldsymbol{\beta}) & \cdots & f_2(x_{d-1}, \boldsymbol{\beta}) & c_2 \\ \vdots & \ddots & \vdots & \vdots \\ f_d(x_1, \boldsymbol{\beta}) & \cdots & f_d(x_{d-1}, \boldsymbol{\beta}) & c_d \end{vmatrix} \neq 0 \tag{2.2.3}$$

for all $L \leq x_1 < x_2 < \ldots < x_{d-1} \leq U$ (this is also true for generalized linear models). In particular, if we are interested in estimating the individual parameter β_i, i.e., $\boldsymbol{c} = \boldsymbol{e}_i$ where $\boldsymbol{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ denotes the ith unit vector, then $\boldsymbol{e}_i \in A^*$ is equivalent to $\mathbf{f}_{-i} = \{f_j | j \in \{1, \ldots, d\} \backslash \{i\}\}$ being a Chebyshev system [see Karlin and Studden (1966)]. Here the traditional definition of a Chebyshev system is used, which only requires the determinant in (2.2.3) to be nonzero instead of positive.

Next, the uniqueness of optimal designs can also be established under mild conditions. We first introduce some additional terminologies. A criterion Φ is called *strictly isotonic* on PD(d) if $\Phi(\mathbf{M}_1) > \Phi(\mathbf{M}_2)$ for any $\mathbf{M}_1 \geq \mathbf{M}_2$ but $\mathbf{M}_1 \neq \mathbf{M}_2$, where $\mathbf{M}_1, \mathbf{M}_2 \in$ PD(d). It is called *strictly concave* on PD(d) if $\Phi(\alpha\mathbf{M}_1 + (1-\alpha)\mathbf{M}_2) > \alpha\Phi(\mathbf{M}_1) + (1-\alpha)\Phi(\mathbf{M}_2)$ for any $\alpha \in (0, 1), \mathbf{M}_1 > 0, \mathbf{M}_2 \geq 0$ and $\mathbf{M}_2 \not\propto \mathbf{M}_1$. For example, Φ_p-optimality criteria are both strictly isotonic and strictly concave on PD(d) when $g(\boldsymbol{\beta})$ is $\boldsymbol{\beta}$ or a reparameterization of $\boldsymbol{\beta}$ and $p \in (-\infty, 1)$ [see Pukelsheim (1993), p. 151]. Moreover, a compound optimality criterion

which involves a strictly isotonic and strictly concave criterion is also strictly isotonic and strictly concave. For these criteria, we have Theorem 2.2.5.

Theorem 2.2.5. *Assume that one of the cases in Theorem 2.2.1 holds. If Φ is both strictly isotonic and strictly concave on PD(d) and there exists a Φ-optimal design ξ^* which has at least d support points, then ξ^* is the unique Φ-optimal design. In particular, the Φ_p-optimal design under Corollary 2.2.4(i) is unique for $p \in (-\infty, 1)$.*

Proof. See Appendix □

Remark 2.2.3. The c-optimality criterion with $g(\boldsymbol{\beta}) = \boldsymbol{c}^T\boldsymbol{\beta}$ maybe neither strictly concave nor strictly isotonic on PD(d). However, if $\boldsymbol{c} \in A^*$ and $\mathbf{f}(x, \boldsymbol{\beta})$ is a Chebyshev system, the uniqueness is proved in Studden (1968).

The uniqueness is not only of interest in itself, but also has implications for finding optimal designs. As we have stated earlier, conditions (a), (b) and (c) in Theorem 2.2.3 may only hold on a large design region, call it the "full design region". Let ξ^{**} be a Φ-optimal design on the full design region with smallest support point x^{**}_{min} and largest support point x^{**}_{max}. Then for a smaller design region $[L, U]$, under the same optimality criterion Φ, we have Theorem 2.2.6.

Theorem 2.2.6. *Assume that one of the cases in Theorem 2.2.1 holds for the full design region, and both Φ-optimal designs on $[L, U]$ and the full design region are unique with support size a, then we have*

(a) under $fix(\Xi) = \{L\}$, if $U < x^{**}_{max}$, then the Φ-optimal design on $[L, U]$ has both L and U as support points; otherwise, the optimal design is ξ^{**}.

(b) under $fix(\Xi) = \{U\}$, if $x^{**}_{min} < L$, then the Φ-optimal design on $[L, U]$ has both L and U as support points; otherwise, the optimal design is ξ^{**}.

(c) under $fix(\Xi) = \emptyset$*, if* $x^{**}_{min} < L$ *or* $U < x^{**}_{max}$*, then the* Φ*-optimal design on* $[L, U]$ *has at least one end point as a support point; otherwise, the optimal design is* ξ^{**}*.*

Proof. We only give the proof for case (a), others being similar. When $U \geq x^{**}_{max}$, the design ξ^{**} is still a feasible design on the region $[L, U]$, and it is optimal because it is optimal on the full design region. When $U < x^{**}_{max}$, ξ^{**} is no longer a feasible design, let ξ^* be the optimal design on $[L, U]$. A complete class of the same type exists for design region $[L, U]$ because, e.g., $\mathbf{F}(c) > 0$ on the full design region implies $\mathbf{F}(c) > 0$ on the smaller design region. So $x^*_1 = L$. If the largest support point $x^*_m < U$, then $\boldsymbol{Z}^* = (x^*_2, \ldots, x^*_m, \omega^*_2, \ldots, \omega^*_m)^T$ must be a critical point of $\tilde{\Phi}(\boldsymbol{Z})$. Now if we consider the optimal design problem on the full design region again, $\boldsymbol{Z}^*$ is a feasible critical point and, by Theorem 2.2.2, ξ^* must be an optimal design on the full design region. However, $\xi^* \neq \xi^{**}$, this contradicts the uniqueness assumption. □

2.3 Application

The theorems we have established can be used to find optimal designs for many models. In Sections 2.3.1 through 2.3.3, we consider Φ_p-optimal designs for models with two, three and four or six parameters, respectively. In Section 2.3.4, we consider polynomial regression models with arbitrary d parameters under more general optimality criteria.

2.3.1 Models with two parameters

Yang and Stufken (2009) considered complete class results for two-parameter models, including logistic/probit regression model, Poisson regression model and Michaelis–Menten model. The theorems we have established can be used to find the optimal designs. Take the Poisson regression model as an example (the applications to other models are similar). It has the

following form:

$$\eta(x, \boldsymbol{\beta}) = \log(\mathrm{E}(y)) = \beta_1 + \beta_2 x, \quad x \in [L, U].$$

Theorem 2.2.1(b) can be applied to this model, and a complete class consists of designs with at most 2 design points including one boundary point [see Yang and Stufken (2009), Theorem 4]. Specifically, when $\beta_2 > 0$, U is a fixed design point, and $\mathbf{M}(-\infty) = 0$ (since $\mathbf{M}(x) = e^{\beta_1+\beta_2 x}(1, x)^T(1, x)$); when $\beta_2 < 0$, L is a fixed design point, and $\mathbf{M}(\infty) = 0$. Thus on any one-sided restricted region $(-\infty, U]$ (when $\beta_2 > 0$) or $[L, \infty)$ (when $\beta_2 < 0$), Φ_p-optimal designs for $\boldsymbol{\beta}$ can be found by solving for the critical points, according to Corollary 2.2.4(i). For c-optimality, recall $\mathbf{f}(x, \boldsymbol{\beta}) = \partial\eta(x, \boldsymbol{\beta})/\partial\boldsymbol{\beta} = (1, x)$, thus $\mathbf{f}_{-2} = \{1\}$ is a Chebyshev system, which means β_2 can only be estimated with at least $d = 2$ support points. Therefore, according to Corollary 2.2.4(ii), an $\boldsymbol{e}_2$-optimal design (c-optimal design for β_2) can also be found by solving for the critical points.

In particular, D- and $\boldsymbol{e}_2$-optimal designs can be found analytically through symbolic computation software (for example, by using the solve function in Matlab) and are listed in (2.3.1) and (2.3.2). Note that they do not depend on β_1 since e^{β_1} is merely a multiplicative factor in $\mathbf{M}(x)$.

$$\xi_D^* = \begin{cases} \{(U - 2/\beta_2, 1/2),\ (U, 1/2)\}, & \beta_2 > 0, \\ \{(L - 2/\beta_2, 1/2),\ (L, 1/2)\}, & \beta_2 < 0. \end{cases} \tag{2.3.1}$$

$$\xi_{\boldsymbol{e}_2}^* = \begin{cases} \{(U - 2.557/\beta_2, 0.782),\ (U, 0.218)\}, & \beta_2 > 0, \\ \{(L - 2.557/\beta_2, 0.782),\ (L, 0.218)\}, & \beta_2 < 0. \end{cases} \tag{2.3.2}$$

However, A-optimal designs do not have explicit forms. Nevertheless the solutions can be found easily using Newton's algorithm. For the case of $\beta_2 < 0$, some examples are listed in Table 2.1 (again the optimal designs do not depend on β_1).

In addition, the Φ_p-optimal design for $\boldsymbol{\beta}$ and $\boldsymbol{e}_2$-optimal design are unique, due to Theorem 2.2.5. For finite design regions, Theorem 2.2.6 can be applied. For example, the A-optimal design for $\boldsymbol{\beta} = (1, -1)^T$ on $[0, U]$ when $U \geq 2.261$ is $\{(0, 0.444), (2.261, 0.556)\}$; when $U < 2.261$, the optimal design is supported at exactly two points 0 and U, and the weights can be determined easily.

Table 2.1: A-optimal designs for Poisson regression model on $[0, \infty)$

	A-optimal	
β_2	(x_1, x_2)	(ω_1, ω_2)
-1	(0, 2.261)	(0.444, 0.556)
-2	(0, 1.193)	(0.320, 0.680)

2.3.2 Models with three parameters

Dette et al. (2008) and Dette et al. (2010) considered optimal designs for the Emax and log-linear models. These models, often used to model dose-response curves, are nonlinear regression models with means

$$\eta(x, \boldsymbol{\beta}) = \begin{cases} \beta_1 + \beta_2 x/(x + \beta_3), & \text{Emax}, \\ \beta_1 + \beta_2 \log(x + \beta_3), & \text{log-linear}. \end{cases}$$

Here $x \in [L, U] \subseteq (0, \infty)$ is the dose range, $\beta_2 > 0$ and $\beta_3 > 0$. Theorem 2.2.1(d) can be applied to both models, and a complete class consists of designs with at most 3 design points including *both* endpoints [Yang (2010), Theorem 3]. Hence Corollary 2.2.4 is applicable on design space $[L, U]$. In particular, D-optimal designs can be computed explicitly and are listed in (2.3.3). They are consistent with the results in Dette et al. (2010).

$$\xi_D^* = \begin{cases} \{(L, 1/3), (x_E^*, 1/3), (U, 1/3)\}, & \text{Emax}, \\ \{(L, 1/3), (x_l^*, 1/3), (U, 1/3)\}, & \text{log-linear}, \end{cases} \tag{2.3.3}$$

where

$$x_E^* = \frac{L(U+\beta_3)+U(L+\beta_3)}{L+U+2\beta_3}, \quad x_l^* = \frac{(L+\beta_3)(U+\beta_3)}{U-L}\log\left(\frac{U+\beta_3}{L+\beta_3}\right)-\beta_3. \tag{2.3.4}$$

For A-optimality, numerical solutions can be obtained easily by Newton's algorithm. Table 2.2 gives some examples for the Emax model using parameter settings in Dette et al. (2008) (the optimal designs don't depend on β_1 since it is not involved in the information matrix; and although it seems that the optimal weights are constant, they do change gradually with β_2 and β_3).

Table 2.2: A-optimal designs for the Emax model on [0,150]

β_2	β_3	(x_1, x_2, x_3)	$(\omega_1, \omega_2, \omega_3)$
7/15	15	(0, 12.50, 150)	(0.250, 0.500, 0.250)
7/15	25	(0, 18.75, 150)	(0.250, 0.500, 0.250)
10/15	25	(0, 18.75, 150)	(0.250, 0.500, 0.250)

For c-optimality, Dette et al. (2010) gave explicit solutions for ED_p-optimal designs, where an ED_p-optimal design is a design that is optimal for estimating the dose that achieves $100p\%$ of the maximum effect in dose range $[L, U]$, $0 < p < 1$. In fact, ED_p-optimality is equivalent to $\boldsymbol{e}_3$-optimality regardless of p, and we can find the optimal designs using our method. First, we have

$$\mathbf{f}(x, \boldsymbol{\beta}) = \begin{cases} (1, x/(x+\beta_3), -\beta_2 x/(x+\beta_3)^2), & \text{Emax}, \\ (1, \log(x+\beta_3), \beta_2/(x+\beta_3)), & \text{log-linear}. \end{cases}$$

It is easy to prove for both the Emax and log-linear models that $\mathbf{f}_{-3}$ is a Chebyshev system, which means that β_3 is only estimable with at least $d = 3$ support points. So $\boldsymbol{e}_3$-optimal designs can be found by solving for the critical points, by Corollary 2.2.4(ii). The solutions are listed in (2.3.5). They are consistent with the results in Dette et al. (2010).

$$\xi^*_{\boldsymbol{e}_3} = \xi^*_{ED_p} = \begin{cases} \{(L, 1/4), (x^*_E, 1/2), (U, 1/4)\}, & \text{Emax}, \\ \{(L, \omega^*_l), (x^*_l, 1/2), (U, 1/2 - \omega^*_l)\}, & \text{log-linear}, \end{cases} \tag{2.3.5}$$

where x^*_E and x^*_l are the same as in (2.3.4), and

$$\omega^*_l = \frac{\log(x^*_l + \beta_3) - \log(U + \beta_3)}{2(\log(L + \beta_3) - \log(U + \beta_3))}.$$

Regarding $\mathbf{f}_{-2}$, it can be shown that it is always a Chebyshev system for the log-linear model, and it is a Chebyshev system for the Emax model if $\beta_3 \notin (L, U)$. In such cases, $\boldsymbol{e}_2$-optimal designs can be found according to Corollary 2.2.4(ii), and the solutions can be derived analytically as shown in (2.3.6).

$$\xi^*_{\boldsymbol{e}_2} = \begin{cases} \{(L, \frac{1}{4} - \frac{(U-L)\beta_3}{8(\beta_3^2 - LU)}), (x^*_E, \frac{1}{2}), (U, \frac{1}{4} + \frac{(U-L)\beta_3}{8(\beta_3^2 - LU)})\}, & \text{Emax}, \beta_3 \notin (L, U), \\ \{(L, \frac{(U-x^*_l)(L+\beta_2)}{2(U-L)(x^*_l+\beta_2)}), (x^*_l, \frac{1}{2}), (U, \frac{(x^*_l-L)(U+\beta_2)}{2(U-L)(x^*_l+\beta_2)})\}, & \text{log-linear}. \end{cases} \tag{2.3.6}$$

When $\beta_3 \in (L, U)$, $\mathbf{f}_{-2}$ is no longer a Chebyshev system for the Emax model. However, if $|(U-L)\beta_3| < |2(\beta_3^2 - LU)|$, the weights of $\xi^*_{\boldsymbol{e}_2}$ in (2.3.6) are still positive, and the design is still $\boldsymbol{e}_2$-optimal; otherwise, the optimal design is supported at fewer than 3 points, which may not be the Chebyshev points. Nevertheless, we can approach the optimal design using the method in Remark 2.2.2. To show this, consider the setting where the dose range is [0,150], $\beta_2 = 7/15$ and $\beta_3 = 25$. The exact $\boldsymbol{e}_2$-optimal design can be found to be $\xi^*_{\boldsymbol{e}_2} = \{(\beta_3^2/U, 0.5), (U, 0.5)\} = \{(25/6, 0.5), (150, 0.5)\}$ using Elfving's method. Now let $\epsilon = 10^{-5}, 10^{-6}, 10^{-7}$; the Φ_p-optimal designs for estimating $g_\epsilon(\boldsymbol{\beta}) = (\epsilon\beta_1, \beta_2, \epsilon\beta_3)^T$ can be found by Corollary 2.2.4(i) and are used to approximate the $\boldsymbol{e}_2$-optimal design. Table 2.3 shows the errors and $1-$efficiencies of the approximation for $p = -1$ and -3. As we can see, the error gets sufficiently small after a few iterations, especially when $|p|$ is larger; however, due to singularity issues, the error cannot be made arbitrary small.

Table 2.3: Approximating two point $\boldsymbol{e}_2$-optimal design using three point designs

p	ϵ	$\lvert x_2 - \frac{25}{6}\rvert/\frac{25}{6}$	$\lvert\omega_1\rvert$	$\lvert\omega_2 - 0.5\rvert$	$\lvert\omega_3 - 0.5\rvert$	$1 - \text{eff}$
-1	10^{-5}	10^{-2}	$4 \cdot 10^{-3}$	$4 \cdot 10^{-3}$	$2 \cdot 10^{-4}$	$6 \cdot 10^{-4}$
	10^{-6}	10^{-3}	$4 \cdot 10^{-4}$	$4 \cdot 10^{-4}$	$2 \cdot 10^{-5}$	$6 \cdot 10^{-5}$
	10^{-7}	10^{-4}	$4 \cdot 10^{-5}$	$4 \cdot 10^{-5}$	$2 \cdot 10^{-6}$	$6 \cdot 10^{-6}$
-3	10^{-5}	$3 \cdot 10^{-4}$	$1 \cdot 10^{-4}$	$9 \cdot 10^{-5}$	$4 \cdot 10^{-6}$	$2 \cdot 10^{-5}$
	10^{-6}	$8 \cdot 10^{-6}$	$3 \cdot 10^{-6}$	$3 \cdot 10^{-6}$	$5 \cdot 10^{-8}$	$5 \cdot 10^{-7}$
	10^{-7}	$7 \cdot 10^{-7}$	$3 \cdot 10^{-7}$	$2 \cdot 10^{-7}$	$1 \cdot 10^{-8}$	$4 \cdot 10^{-8}$

2.3.3 Models with four or six parameters

Demidenko (2004) used a double exponential model to characterize the regrowth of tumor after radiation. The natural logarithm of tumor volume can be modeled using a nonlinear regression model with mean

$$\eta(x, \boldsymbol{\beta}) = \beta_1 + \log(\beta_2 e^{\beta_3 x} + (1 - \beta_2) e^{-\beta_4 x}),$$

where $0 \leq x \in [L, U]$ is the time, β_1 is the logarithm of the initial tumor volume, $0 < \beta_2 < 1$ is the proportional contribution of the first compartment, and $\beta_3, \beta_4 > 0$ are cell proliferation and death rates.

Demidenko (2006) used the LINEXP model to characterize tumor growth delay and regrowth. The model was described in Section 2.2.1 and re-presented below:

$$\eta(x, \boldsymbol{\beta}) = \beta_1 + \beta_2 e^{\beta_3 x} + \beta_4 x,$$

Li and Balakrishnan (2011) considered D- and c-optimal designs for these two models, but our approach yields more general results. For both models, Theorem 2.2.1(d) can be applied, and a complete class consists of designs with at most four design points including both endpoints [see Yang and Stufken (2012)]. Thus Corollary 2.2.4 can again be applied

on the design space $[L, U]$, and Φ_p-optimal designs for $\boldsymbol{\beta}$ and certain c-optimal designs can be found by solving for the critical points. In particular, $\mathbf{f}_{-3}$ and $\mathbf{f}_{-4}$ are Chebyshev systems under both models [see Li and Balakrishnan (2011)], thus $\boldsymbol{e}_3$- and $\boldsymbol{e}_4$-optimal designs for both models can be found by solving for the critical points.

There is no explicit solution for the optimal designs, but numerical solutions can be easily found using Newton's algorithm. Here we give some A-optimal designs for the LINEXP model in Table 2.4 (the optimal designs for the LINEXP model do not depend on β_1 and β_4 since they are not involved in the information matrix). For D- and c-optimality, our approach gives the same results as in Li and Balakrishnan (2011).

Table 2.4: A-optimal designs for the LINEXP model on [0,1]

β_2	β_3	(x_1, x_2, x_3, x_4)	$(\omega_1, \omega_2, \omega_3, \omega_4)$
0.5	-1	(0, 0.220, 0.717, 1)	(0.156, 0.324, 0.344, 0.176)
1	-1	(0, 0.220, 0.717, 1)	(0.151, 0.319, 0.349, 0.181)
1	-2	(0, 0.195, 0.681, 1)	(0.146, 0.315, 0.355, 0.184)

Consider one more example. Dette, Melas and Wong (2006) studied D-optimal designs for biexponential regression models, which are nonlinear regression models with mean

$$\eta(x, \boldsymbol{\beta}) = \sum_{s=1}^{S} \beta_{2s-1} e^{-\beta_{2s} x}, \quad 0 \le x \in [L, U], \tag{2.3.7}$$

where $\beta_{2s-1} \neq 0, s = 1, \ldots, S, 0 < \beta_2 < \ldots < \beta_{2S}$. When $S = 2$ and $\beta_4/\beta_2 < 61.98$ or $S = 3, 2\beta_4 = \beta_2 + \beta_6$ and $\beta_4/\beta_2 < 23.72$, Theorem 2.2.1(b) can be applied, and a complete class consists of designs with at most $2S$ design points including the lower endpoint L [see Yang and Stufken (2012), Theorem 3 and Theorem 4]. Moreover, it is easy to see that the information matrix $\mathbf{M}(x)$ goes to 0 when x approaches infinity, thus Corollary 2.2.4 can be applied on any design region $[L, \infty)$. Table 2.5 gives some A-optimal designs when $S = 2, \beta_1 = \beta_2 = 1$.

Table 2.5: A- and $\boldsymbol{e}_2$-optimal designs for biexponential regression model on $[0, \infty)$

Criterion	β_3	β_4	(x_1, x_2, x_3, x_4)	$(\omega_1, \omega_2, \omega_3, \omega_4)$
A-optimality	1	2	$(0, 0.275, 1.196, 3.416)$	(0.078, 0.178, 0.251, 0.493)
	1	4	$(0, 0.170, 0.768, 2.472)$	(0.118, 0.261, 0.287, 0.334)
	3	4	$(0, 0.172, 0.760, 2.450)$	(0.083, 0.199, 0.296, 0.422)
$\boldsymbol{e}_2$-optimality	1	2	$(0, 0.273, 1.197, 3.425)$	(0.054, 0.124, 0.200, 0.623)
	1	4	$(0, 0.168, 0.769, 2.492)$	(0.033, 0.082, 0.201, 0.683)
	3	4	$(0, 0.168, 0.769, 2.492)$	(0.033, 0.082, 0.201, 0.683)

For c-optimality, first we have

$$\mathbf{f}(x, \boldsymbol{\beta}) = \begin{cases} (e^{-\beta_2 x}, -\beta_1 x e^{-\beta_2 x}, e^{-\beta_4 x}, -\beta_3 x e^{-\beta_4 x}), & S = 2, \\ (e^{-\beta_2 x}, -\beta_1 x e^{-\beta_2 x}, e^{-\beta_4 x}, -\beta_3 x e^{-\beta_4 x}, e^{-\beta_6 x}, -\beta_5 x e^{-\beta_6 x}), & S = 3. \end{cases}$$

Both are Chebyshev systems. In addition, we can show that $\mathbf{f}_{-2s}, s = 1, \ldots, S$ are Chebyshev systems for $S = 2$ and $S = 3$, so the c-optimal designs for $\beta_{2s}, s = 1, \ldots, S$ on $[L, \infty)$ can be found by solving for the critical points. Table 2.5 gives some $\boldsymbol{e}_2$-optimal designs when $S = 2, \beta_1 = \beta_2 = 1$.

Moreover, the Φ_p-optimal designs for $\boldsymbol{\beta}$ and c-optimal design for β_{2s}'s are unique by Theorem 2.2.5. For a finite design region, Theorem 2.2.6 can be applied. For example, the A-optimal design for $\boldsymbol{\beta} = (1, 1, 1, 2)^T$ on $[0, U]$ when $U \geq 3.416$ is the same as in Table 2.5; when $U < 3.416$, the optimal design is supported at 4 design points including both 0 and U.

2.3.4 Polynomial regression model with d parameters

Yang (2010) considered the general $(d-1)$th degree polynomial regression model P_{d-1} with variance $\sigma^2/\lambda(x)$ and mean

$$\eta(x, \boldsymbol{\beta}) = \beta_1 + \sum_{i=2}^{d} \beta_i x^{i-1}. \tag{2.3.8}$$

For different choices of the efficiency function $\lambda(x)$, Theorem 2.2.1 gives the following complete class results [see Yang (2010), Theorem 9]:

(a) When (i) $\lambda(x) = 1 - x, x \in [-1, 1]$ or (ii) $\lambda(x) = e^{-x}, x \in [0, \infty)$, a complete class consists of designs with at most d design points including the left endpoint. Moreover, the information matrix $\mathbf{M}(U) = 0$.

(b) When $\lambda(x) = 1 + x$, $x \in [-1, 1]$, a complete class consists of designs with at most d design points including the right endpoint. Moreover, the information matrix $\mathbf{M}(L) = 0$.

(c) When (i) $\lambda(x) = (1 - x)^{u+1}(1 + x)^{v+1}, x \in [-1, 1], u + 1 > 0, v + 1 > 0$ or (ii) $\lambda(x) = x^{u+1}e^{-x}, x \in [0, \infty), u + 1 > 0$ or (iii) $\lambda(x) = e^{-x^2}, x \in (-\infty, \infty)$ or (iv) $\lambda(x) = (1 + x^2)^{-t}, x \in (-\infty, \infty), d \leq t$, a complete class consists of designs with at most d design points. Moreover, the information matrices $\mathbf{M}(L) = \mathbf{M}(U) = 0$.

(d) When $\lambda(x) \equiv 1, x \in [L, U]$, a complete class consists of designs with at most d design points including both endpoints.

Corollary 2.2.4 can be applied to the above models on the respective (full) design regions, thus Φ_p-optimal designs for $\boldsymbol{\beta}$ and c-optimal designs for β_d can be found by solving for the critical points. Furthermore, those designs are unique, so Theorem 2.2.6 can be used when the design regions are small.

Finally, we apply our theorems to more general optimality criteria. Dette and Studden (1995) considered optimal designs under nested polynomial regression models. To be specific, suppose the degree of the polynomial regression model is an unknown integer between 1 and $d-1$. The D-optimal design ξ_D^ℓ under a given model P_ℓ, $1 \leq \ell \leq d-1$, may not be efficient under another model with a different degree. To take this uncertainty into consideration,

the authors proposed the following weighted optimality criteria $\Phi_{p',\boldsymbol{\beta}}$:

$$\Phi_{p',\boldsymbol{\beta}}(\mathbf{M}_\xi) = \left[\sum_{\ell=1}^{d-1} \beta_\ell (\text{eff}_D^\ell(\xi))^{p'}\right]^{1/p'}, \tag{2.3.9}$$

where $p' \in [-\infty, 1]$, $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_{d-1}\}$ is a prior on the set $\{1, \ldots, d-1\}$ with $\beta_{d-1} > 0$,

$$\text{eff}_D^\ell(\xi) = \left(\frac{\det \mathbf{M}_\xi^\ell}{\det \mathbf{M}_{\xi_D^\ell}^\ell}\right)^{\frac{1}{\ell+1}}, \quad \ell = 1, \ldots, d-1,$$

$\mathbf{M}_\xi^\ell$ is the information matrix of ξ under model P_ℓ, and $\text{eff}_D^\ell(\xi)$ is the D-efficiency of ξ under model P_ℓ.

Dette and Studden (1995) gave the solution of $\Phi_{p',\boldsymbol{\beta}}$-optimal design for $\lambda(x) \equiv 1, x \in [-1, 1]$. The solution is rather complicated, and it requires knowledge of canonical moments. An alternative way is to use Theorem 2.2.3, and it can be applied to more general settings.

First, the D-efficiency in the definition of $\Phi_{p',\boldsymbol{\beta}}$ can be generalized to any Φ_p-efficiency, $p \in (-\infty, 1]$ (e.g., A-efficiency when $p = -1$), and we denote the resulting optimality criteria as $\Phi_{p,p',\boldsymbol{\beta}}$. Second, the efficiency function $\lambda(x)$ can be generalized to any function in cases (a)$\sim$(d) in this subsection, where x belongs to the respective (full) design regions.

Under this general setting, $\Phi_{p,p',\boldsymbol{\beta}}$ always satisfies our assumption about optimality criteria in Section 2.2.1 [see Pukelsheim (1993), p. 285]. Moreover, while this optimality criterion is defined on a mixture of different models, these models are nested within the largest model P_{d-1}, thus our complete class result for P_{d-1} can be applied to $\Phi_{p,p',\boldsymbol{\beta}}$. Finally, to use Theorem 2.2.3, any $\Phi_{p,p',\boldsymbol{\beta}}$-optimal design must have at least d support points. This requirement is reasonable since otherwise the optimal design will not be able to estimate the model P_{d-1}, which may be the true model. To meet the requirement, it is sufficient to restrict ourselves to $p, p' \in (-\infty, 0]$, since any singular matrix will result in $\Phi_{p,p',\boldsymbol{\beta}}$ to be 0. So by Theorem 2.2.3, $\Phi_{p,p',\boldsymbol{\beta}}$-optimal designs for models in cases (a)$\sim$(d) of this subsection can be

found by solving for the critical points. Some examples are given in Table 2.6 for the case $\lambda(x) = 1 - x^2, x \in [-1, 1], p = -1$ (i.e., for A-efficiency in (2.3.9)), $d = 4$ and $\boldsymbol{\beta}$ a uniform prior.

Table 2.6: $\Phi_{p,p',\boldsymbol{\beta}}$-optimal designs for polynomial regression models

p'	(x_1, x_2, x_3, x_4)	$(\omega_1, \omega_2, \omega_3, \omega_4)$	$(\text{eff}_A^1(\xi), \text{eff}_A^2(\xi), \text{eff}_A^3(\xi))$
0	(−0.860, −0.346, 0.346, 0.860)	(0.263, 0.237, 0.237, 0.263)	(0.692, 0.745, 0.902)
−1	(−0.854, −0.343, 0.343, 0.854)	(0.268, 0.232, 0.232, 0.268)	(0.701, 0.753, 0.879)
-3	(−0.846, −0.339, 0.339, 0.846)	(0.273, 0.227, 0.227, 0.273)	(0.714, 0.759, 0.846)

In addition, $\Phi_{p,p',\boldsymbol{\beta}}$-optimality is strictly isotonic and strictly concave on $\text{PD}(d)$ since $\beta_{d-1} > 0$ and the Φ_p-efficiency under model P_{d-1} is strictly isotonic and strictly concave on $\text{PD}(d)$ for $p \in (-\infty, 0]$. Hence by Theorem 2.2.5, the optimal designs are unique. However, for smaller design regions, the optimality criterion $\Phi_{p,p',\boldsymbol{\beta}}$ changes as the design region changes. For example, when $p = 0$, the design ξ_D^ℓ changes when the design region changes, which causes $\Phi_{p,p',\boldsymbol{\beta}}$ to change. So the optimal design on the full design region cannot be used to obtain the optimal design on a smaller region as we did in Theorem 2.2.6.

2.4 Computational advantages

Although it is not the main motivation, our method does provide computational advantages over other algorithms, as Newton's algorithm is well studied, easy to program and fast. For comparison, we choose the optimal weight exchange algorithm (OWEA) proposed in Yang, Biedermann and Tang (2013), which is among the most general and fastest algorithms.

OWEA algorithm starts with an initial design on a grid of the design space, then iterates between optimizing the weights for the current set of support points and adding a new grid point to the current support points, until the condition for optimality in general equivalence theorem is satisfied. The computing time increases as the grid size κ becomes larger. So to reduce the computing time, the authors proposed a modified algorithm. The modified algorithm starts with a coarse grid and finds the optimal design on the coarse grid. Based on

that, the grid near the support points of the optimal design is refined and a more accurate optimal design is found on the finer grid. We refer to their original and modified algorithm as OWEA I and OWEA II, respectively. All algorithms are coded using SAS IML and run on a Dell Desktop (2.5GHz and 4Gb RAM). Comparisons are made for different grid sizes, different models, and under both A- and D-optimality criterion.

Table 2.7: Computation time (seconds) for the LINEXP model

	A-optimal			D-optimal		
	$\kappa = 100$	$\kappa = 1000$	$\kappa = 10000$	$\kappa = 100$	$\kappa = 1000$	$\kappa = 10000$
Newton's	0.08	0.08	0.08	0.08	0.08	0.08
OWEA I	0.19	0.22	0.63	0.24	0.37	1.28
OWEA II	0.17	0.18	0.21	0.20	0.23	0.29

First, we consider the LINEXP model given in (2.2.2). The parameters are set to be $\boldsymbol{\beta} = (1, 0.5, -1, 1)^T$, and the design space is $[0, 1]$. Three different grid sizes, $\kappa = 100, 1000$ and 10000, are used for OWEA I and II; and for OWEA II, the initial coarse grid sizes are chosen to be 10, 100, and 100 respectively. The computing times are shown in Table 2.7. Note the grid size κ is irrelevant for the speed of Newton's algorithm.

Table 2.8: Computation time (seconds) for polynomial regression model

	A-optimal			D-optimal		
	$\kappa = 100$	$\kappa = 1000$	$\kappa = 10000$	$\kappa = 100$	$\kappa = 1000$	$\kappa = 10000$
Newton's	0.17	0.17	0.17	0.14	0.14	0.14
OWEA I	0.33	0.61	3.49	0.48	1.09	4.83
OWEA II	0.34	0.48	0.89	0.44	0.72	1.35

From Table 2.7, we can see all three algorithms are very efficient in finding optimal designs. Newton's algorithm is at least twice as fast as the other two algorithms. The speed gain is more prominent when comparing to OWEA I, especially when the grid size κ is large.

Second, we consider a polynomial regression model given in (2.3.8) with $d = 6$ and $\lambda(x) = 1 - x^2, x \in [-1, 1]$. It has more parameters than the previous example so finding optimal designs takes longer. The results are shown in Table 2.8. Again, we can see all three

algorithms are very efficient in finding optimal designs. Newton's algorithm is at least twice as fast as the other two algorithms. The speed gain is more prominent when comparing to OWEA I, especially when the grid size κ is large.

Chapter 3

One covariate, correlated data

3.1 Introduction

Correlation may arise when there is a cluster structure to the data. For example, in a longitudinal study, repeated measurements taken from the same subject may be correlated. Observations from different subjects, for example, patients from the same physician or hospital, can also be correlated. Despite the ubiquity of correlated data, there are very few results on optimal designs for this type of data. These few results can be classified into two approaches, depending on how the correlation is modeled.

The first approach, which composes the majority of the existing literature, usually assumes the parameters are fixed and models the correlated errors using a covariance kernel function, for example, an exponential covariance function. This approach has been applied mainly in the linear models context. Sacks and Ylvisaker (1966, 1968) are among the very first papers to consider optimal design problems for correlated data. They considered a linear stochastic process where the parameters were fixed and the covariance between errors was modeled using a covariance kernel function. The covariance kernel function made the variance-covariance matrix of the weighted least squares estimator more complicated and the

design problem became much harder. So the authors proposed *asymptotic designs*, which are approximate optimal designs when the sample size goes to infinity, and are usually represented by density functions on the design space. This concept was adopted and extended by other authors. For example, Bickel and Herzberg (1979) and Bickel et al. (1981) developed an asymptotic theory for ordinary least squares estimators and a correlation function that was assumed to depend on the sample size.

Asymptotic designs are difficult to find, and for a long time, results only existed for location-scale models. Recently, Dette et al. (2013) provided results for certain more general linear regression models. While a significant step forward, asymptotic designs are only known for few combinations of covariance kernel function and regression function. To the best of our knowledge, for nonlinear models, there are no theoretical results using this approach, and even numerical results are rare.

Part of the difficulty is that many tools established for independent data do not hold for dependent data. Firstly, the convexity property of information matrices with respect to the designs no longer holds (see Dette et al., 2013). This means that only the necessity part of Kiefer's equivalence theorem (Kiefer and Wolfowitz, 1960) can be proved. Secondly, the powerful complete class results established in a series of papers, for example, Yang and Stufken (2012) and Dette and Schorning (2013), cannot be generalized either.

The second approach assumes a mixed effects model, i.e., the parameters are random and the observational errors are independent. Mixed effects models are widely used for analyzing correlated data, but research on optimal design is very underdeveloped, and there are even fewer theoretical results available compared to the first approach. However, we will adopt this second approach and show that it is well-suited for generalizing classical results for independent data to correlated data.

The rest of this Chapter is organized as follows. Section 3.2 generalizes results under independent data to mixed effects models. These results are used to find optimal designs

for a pharmacokinetics study in Section 3.3. The robustness of design efficiency against miss-specifying the covariance structure of random effects is investigated in Section 3.4.

3.2 Theory

Mixed effects models are similar to models in Chapter 2 but with some parameters being random. Suppose $\mathbf{y}_i = (y_{i1}, ..., y_{im_i})^T$ is the m_i-dimensional vector of responses from subject (cluster) i, where y_{ij} is the jth observation given by the following mixed effects model:

$$y_{ij} = \eta(x_{ij}, \boldsymbol{\beta}_i) + \epsilon_{ij}, \quad i = 1, ..., n; j = 1, ..., m_i. \tag{3.2.1}$$

Here $\eta(x_{ij}, \boldsymbol{\beta}_i)$ is the conditional mean given $\boldsymbol{\beta}_i = (\beta_{i1}, ..., \beta_{id})^T$ and $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{b}_i$ is the parameter vector for the ith individual, $\boldsymbol{\beta} = (\beta_1, ..., \beta_d)$ is the population parameter vector and $\mathbf{b}_i \sim N(0, \sigma^2\Sigma)$ is the random effects vector, Σ is the scaled variance-covariance matrix and Σ can be singular if some parameters are fixed. The observational errors ϵ_{ij}'s are independent, normally distributed with constant variance σ^2. The random effects $\mathbf{b}_i$'s are independent of each other and are independent of the observational errors.

As usual, our main interest lies in estimating the regression parameter $\boldsymbol{\beta}$. It can be estimated using maximum likelihood estimator, but the likelihood function doesn't always have a closed form. To begin with, let's consider a simpler case where the random effects enter the model linearly. Split $\boldsymbol{\beta}_i$ into two parts, $\boldsymbol{\beta}_i = (\boldsymbol{\beta}_i^{(1)T}, \boldsymbol{\beta}^{(2)T})^T$, where $\boldsymbol{\beta}_i^{(1)}$ is the d_1-dimensional random parameter vector and $\boldsymbol{\beta}^{(2)}$ is the $(d - d_1)$-dimensional fixed parameter vector. Assume the regression function can also be decomposed into two parts, a linear part and a nonlinear part,

$$\eta(x_{ij}, \boldsymbol{\beta}_i) = \mathbf{f}(x_{ij})^T \boldsymbol{\beta}_i^{(1)} + \tilde{f}(x_{ij}, \boldsymbol{\beta}^{(2)}).$$

In particular, when $d_1 = d$, this is a linear mixed effects model; when $d_1 = 1$ and $\mathbf{f}(x) \equiv 1$, this is a random intercept model.

For the above model, the likelihood function of $\mathbf{y}_i$ can be derived explicitly. Let $\boldsymbol{\eta}(\mathbf{x}_i, \boldsymbol{\beta}) = (\eta(x_{i1}, \boldsymbol{\beta}), ..., \eta(x_{im_i}, \boldsymbol{\beta}))^T$, $\mathbf{F}_i = \partial\boldsymbol{\eta}(\mathbf{x}_i, \boldsymbol{\beta})/\partial\boldsymbol{\beta}$ and $\mathbf{V}_i = \mathbf{I}_{m_i} + \mathbf{F}_i\Sigma\mathbf{F}_i^T$, the marginal distribution of $\mathbf{y}_i$ can be written as

$$\mathbf{y}_i \sim N(\boldsymbol{\eta}(\mathbf{x}_i, \boldsymbol{\beta}), \sigma^2\mathbf{V}_i).$$

The maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ is asymptotically normally distributed with mean $\boldsymbol{\beta}$ and variance-covariance matrix

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\sum_{i=1}^{n} \mathbf{F}_i^T \mathbf{V}_i^{-1} \mathbf{F}_i \right)^{-1}, \quad \mathbf{F}_i = \frac{\partial\boldsymbol{\eta}(\mathbf{x}_i, \boldsymbol{\beta})}{\partial\boldsymbol{\beta}}, \quad \mathbf{V}_i = \mathbf{I}_{m_i} + \mathbf{F}_i\Sigma\mathbf{F}_i^T. \tag{3.2.2}$$

However, in general, the likelihood of $\mathbf{y}_i$ in model (3.2.1) requires an integration over the random effects, which usually doesn't have a closed form. So there is no analytical expression for the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$. In the literature, a first order linearization of the model is often used, see Retout et al. (2001), Schmelter (2007b) or Dette et al. (2010), among the others. To be specific, expand η around $\boldsymbol{\beta}$ (in the locally optimal design context, we assume $\boldsymbol{\beta}$ is known or a good guess is available), we get

$$\eta(x_{ij}, \boldsymbol{\beta}_i) \approx \eta(x_{ij}, \boldsymbol{\beta}) + \left(\frac{\partial\eta(x_{ij}, \boldsymbol{\beta})}{\partial\boldsymbol{\beta}} \right)^T (\boldsymbol{\beta}_i - \boldsymbol{\beta}).$$

Therefore, the model is linearized, and (3.2.2) can be used as an approximation of the intractable variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ under nonlinear mixed models.

Now we formulate the locally optimal design problem. Assume the number of subjects (clusters) is fixed to be n, the total number of measurements is fixed to be N. The number of measurements taken from each subject and the way of distributing these measurements

within the subject need to be designed. To avoid the difficulty of discrete optimization, we consider approximate designs; efficient exact designs can usually be found by rounding.

According to Schmelter (2007a), the optimal design should have the same number of measurements taken from each subject and the same design within each subject. Hence $m = N/n$ is the number of measurements taken from each subject. Let $\xi_i \equiv \xi = \{(x_k, \omega_k)\}_{k=1}^{t}$ be the common design for all subjects, where $\sum_{k=1}^{t} \omega_k = m$, and $\mathbf{F}_i \equiv \mathbf{F}, \mathbf{V}_i \equiv \mathbf{V}, i = 1, ..., n$. The covariance matrix in (3.2.2) now becomes $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(n\mathbf{F}^T\mathbf{V}^{-1}\mathbf{F})^{-1}$. Define $\mathfrak{M}(\xi) = \mathbf{F}^T\mathbf{V}^{-1}\mathbf{F}$ as the information matrix for each subject, $n\mathfrak{M}(\xi)$ as the total information matrix, and our goal is to maximize $\mathfrak{M}(\xi)$ with respect to ξ.

We first generalize the complete class results in Chapter 2 to mixed models. To do this, we present a lemma which relates the information matrix under independence, denoted as $\mathbf{M}(\xi)$ in Chapter 2, to the information matrix under mixed model. The lemma and its proof can be found in Schmelter (2007a).

Lemma 3.2.1 (Schmelter (2007a))**.** *The information matrix $\mathfrak{M}(\xi)$ can be rewritten as*

$$\mathfrak{M}(\xi) = \left(\boldsymbol{M}^{+}(\xi) + \boldsymbol{M}^{+}(\xi)\boldsymbol{M}(\xi)\Sigma\boldsymbol{M}(\xi)\boldsymbol{M}^{+}(\xi)\right)^{+},$$

where $\boldsymbol{M}^{+}(\xi)$ is the Moore-Penrose inverse of $\boldsymbol{M}(\xi)$. If $\boldsymbol{M}(\xi)$ is non-singular the right hand simplifies to

$$(\boldsymbol{M}^{-1}(\xi) + \Sigma)^{-1}.$$

For singular $\boldsymbol{M}(\xi)$ it holds that

$$\lim_{\delta \to 0} \left((\boldsymbol{M}(\xi) + \delta \boldsymbol{I}_d)^{-1} + \Sigma\right)^{-1} = \mathfrak{M}(\xi)$$

Lemma 1 connects the information matrix under independent data with its counterpart under correlated data in a surprisingly nice way, it allows results established under indepen-

dent data to be generalized. For example, an immediate observation here is, $\mathfrak{M}(\xi)$ and $\mathbf{M}(\xi)$ have the same rank and the same column space. So a design ξ is feasible for estimating $\boldsymbol{\beta}$ (under mixed model) if and only if it has at least d support points; it is feasible for estimating $\mathbf{c}^T\boldsymbol{\beta}$ if and only if it is feasible under independent data. More similarities will be revealed along the way.

Using Lemma 3.2.1, we can generalize the complete class results to mixed models easily.

Theorem 3.2.2. *The complete class Ξ derived under independent model in Chapter 2 is also the complete class under mixed model (3.2.1).*

Proof. We only need to show for any design $\xi \notin \Xi$, we can find a design $\tilde{\xi} \in \Xi$ such that $\mathfrak{M}(\tilde{\xi}) \geq \mathfrak{M}(\xi)$. Using results in Chapter 2, there exists a design $\tilde{\xi} \in \Xi$ such that $\mathbf{M}(\tilde{\xi}) \geq \mathbf{M}(\xi)$.

Case 1, $\mathbf{M}(\xi)$ is nonsingular, by Lemma 3.2.1 we have

$$\mathfrak{M}(\tilde{\xi}) = (\mathbf{M}^{-1}(\tilde{\xi}) + \Sigma)^{-1} \geq (\mathbf{M}^{-1}(\xi) + \Sigma)^{-1} = \mathfrak{M}(\xi),$$

Case 2, $\mathbf{M}(\xi)$ is singular. Still we have

$$\Big((\mathbf{M}(\tilde{\xi}) + \delta\mathbf{I}_d)^{-1} + \Sigma\Big)^{-1} \geq \big((\mathbf{M}(\xi) + \delta\mathbf{I}_d)^{-1} + \Sigma\big)^{-1}$$

Take the limit as $\delta \to 0$, we have $\mathfrak{M}(\tilde{\xi}) \geq \mathfrak{M}(\xi)$. □

Now the search of optimal designs can be restricted within Ξ, in particular, we focus on Φ_p-optimal designs. The Φ_p-optimal design for a smooth function of the parameters, $g(\boldsymbol{\beta}) : \mathbb{R}^d \to \mathbb{R}^v$ where $v \leq d$, is defined to maximize the following objective function with respect to ξ:

$$\Phi_p(\mathfrak{I}(\xi)) = \left(\frac{1}{v}\ \mathrm{tr}(\mathfrak{I}^p(\xi))\right)^{1/p}, \quad p \in (-\infty, 1],$$

where $\mathfrak{I}(\xi) = (\mathbf{K}^T\mathfrak{M}^-(\xi)\mathbf{K})^{-1} = (\mathbf{K}^T(\mathbf{M}^-(\xi) + \Sigma)\mathbf{K})^{-1}$ is the (scaled) information matrix for $g(\boldsymbol{\beta})$, $\mathbf{K} = (\partial g(\boldsymbol{\beta})/\partial\boldsymbol{\beta})^T$ has full column rank v and $\mathbf{M}^-(\xi)$ is the generalized inverse if $\mathbf{M}(\xi)$ is singular.

Before finding the optimal designs, it is important to point out the concavity of the information matrix mapping $\mathfrak{I}(\xi)$. Under independent data, for any two feasible designs ξ_1, ξ_2, let $\xi_\alpha = \alpha\xi_1 + (1-\alpha)\xi_2, \alpha \in (0,1)$, we have

$$(\mathbf{K}^T\mathbf{M}^-(\xi_\alpha)\mathbf{K})^{-1} \geq \alpha(\mathbf{K}^T\mathbf{M}^-(\xi_1)\mathbf{K})^{-1} + (1-\alpha)(\mathbf{K}^T\mathbf{M}^-(\xi_2)\mathbf{K})^{-1}.$$

Same concavity holds under mixed models. To see that, Schmelter (2007a) proved $\mathfrak{M}(\xi_\alpha) \geq \alpha\mathfrak{M}(\xi_1) + (1-\alpha)\mathfrak{M}(\xi_2)$, thus $\mathfrak{I}(\xi_\alpha) \geq (\mathbf{K}^T(\alpha\mathfrak{M}(\xi_1) + (1-\alpha)\mathfrak{M}(\xi_2))^-\mathbf{K})^{-1} \geq \alpha\mathfrak{I}(\xi_1) + (1-\alpha)\mathfrak{I}(\xi_2)$.

To find the Φ_p-optimal designs, Theorem 2.2.3 can be used in many situations. Rewrite $\Phi_p(\mathfrak{I}(\xi))$ as $\Phi(\mathbf{M}(\xi))$. It is easy to see Φ satisfies assumptions (2) $\sim$ (3); for assumption (1), since we have established the concavity of the information mapping $\mathfrak{I}(\xi)$, this assumption holds as well. So Theorem 2.2.3 can be applied which gives Theorem 3.2.3.

Theorem 3.2.3. *Suppose the complete class Ξ satisfies (i) $a = d$ (ii) the end points L and U are either a fixed design point of Ξ or the information matrix at which is 0, then a feasible critical point of $\tilde{\Phi}(\boldsymbol{Z})$ must exist, and any such point gives a Φ_p-optimal design for $\boldsymbol{\beta}$.*

Examples will be given for the biexponential regression model in Section 3.3. Note that for c-optimality, Theorem 3.2.3 also applies if $\mathbf{c}^T\boldsymbol{\beta}$ can only be estimated with at least d support points. However, we decided not to go into much details here.

Next, Theorem 3.2.4 and Theorem 3.2.5 gives some interesting properties of the optimal designs under the commonly used A-, c-, and D-optimality.

Theorem 3.2.4. *When $p = -1$, a Φ_{-1}-optimal design for $g(\boldsymbol{\beta})$ under a fixed effects model is also a Φ_{-1}-optimal design under its corresponding mixed effects model, for any Σ.*

Proof. When $p = -1$, the optimality criterion reduces to

$$\Phi_{-1}(\mathfrak{I}(\xi)) = v\left(\mathrm{tr}(\mathfrak{I}^{-1}(\xi))\right)^{-1} = v\left(\mathrm{tr}(\mathbf{K}^T(\mathbf{M}^-(\xi)+\Sigma)\mathbf{K})\right)^{-1} = v\left(\mathrm{tr}(\mathbf{K}^T\mathbf{M}^-(\xi)\mathbf{K}) + \mathrm{tr}(\mathbf{K}^T\Sigma\mathbf{K})\right)^{-1}.$$

Since $\mathrm{tr}(\mathbf{K}^T\Sigma\mathbf{K})$ is a constant scalar, the optimal design obtained under independent data, which maximizes $\Phi_{-1}(\mathfrak{I}(\xi))$ under $\Sigma = 0$, must also maximizes $\Phi_{-1}(\mathfrak{I}(\xi))$ under any non-negative definite Σ. □

By Theorem 3.2.4, the A-optimal designs and c-optimal designs obtained under independent data are also optimal under mixed models. On the other hand, for D-optimality and other optimality criteria where $p \neq -1$, the optimal design depends on Σ in general. However, for random intercept models, the D-optimal design actually doesn't depend on Σ either.

Theorem 3.2.5. *A D-optimal design for $\boldsymbol{\beta}$ under a fixed effects model is also a D-optimal design under its corresponding random intercept model, for any Σ.*

Proof. The D-optimal design maximizes $\det(\mathfrak{M}(\xi))$. From Lemma 3.2.1, we can assume $\mathbf{M}(\xi)$ is nonsingular since otherwise $\det(\mathfrak{M}(\xi)) = 0$. We only need to prove $\det(\mathfrak{M}(\xi)) = \det(\mathbf{M}(\xi))/(1+m\sigma_{11})$, where σ_{11} is the $(1,1)$th element of Σ.

For a random intercept model, $\mathbf{F} = (\mathbf{F}^{(1)}, \mathbf{F}^{(2)})$ where $\mathbf{F}^{(1)}$ is the m-dimensional vector of all 1's, and Σ is a sparse matrix with all 0's except $\sigma_{11} > 0$. Thus

$$\mathbf{M}(\xi) = \begin{pmatrix} m & \mathbf{F}^{(1)T}\mathbf{F}^{(2)} \\ \mathbf{F}^{(2)T}\mathbf{F}^{(1)} & \mathbf{F}^{(2)T}\mathbf{F}^{(2)} \end{pmatrix}, \quad \mathbf{M}(\xi)\Sigma = \begin{pmatrix} \sigma_{11}m & 0 \\ \sigma_{11}\mathbf{F}^{(2)T}\mathbf{F}^{(1)} & 0 \end{pmatrix},$$

$$\det(\mathfrak{M}(\xi)) = \det(\mathbf{M}^{-1}(\xi)+\Sigma)^{-1} = \det(\mathbf{M}(\xi)) \times \det(\mathbf{I}_d + \mathbf{M}(\xi)\Sigma)^{-1} = \det(\mathbf{M}(\xi))/(1+m\sigma_{11}).$$

□

A nice property about D-optimal designs in independent data case is that they are invariant to re-parameterizations. This property also holds under mixed data. Suppose $\tilde{\boldsymbol{\beta}} = g(\boldsymbol{\beta})$ is a reparameterization of $\boldsymbol{\beta}$, then $\widetilde{\mathbf{M}}(\xi) = \mathbf{K}^{-1}\mathbf{M}(\xi)\mathbf{K}^{-T}$, where $\mathbf{K} = (\partial g(\boldsymbol{\beta})/\partial\boldsymbol{\beta})^T$ is nonsingular. Moreover, using Delta method, the variance covariance matrix of $\tilde{\boldsymbol{\beta}}_i$'s is $\sigma^2\widetilde{\Sigma}$, where $\widetilde{\Sigma} = \mathbf{K}^T\Sigma\mathbf{K}$. Therefore, $\widetilde{\mathfrak{M}}(\xi) = (\widetilde{\mathbf{M}}^{-1}(\xi) + \widetilde{\Sigma})^{-1} = \mathbf{K}^{-1}\mathfrak{M}(\xi)\mathbf{K}^{-T}$, which means $\det \widetilde{\mathfrak{M}}(\xi) \propto \det \mathfrak{M}(\xi)$, and the D-optimal designs are invariant.

Finally, Kiefer's equivalence theorem also holds because of the concavity of information mapping $\mathfrak{I}(\xi)$.

Theorem 3.2.6. *A design ξ^* is a Φ_p-optimal design for estimating $g(\boldsymbol{\beta})$ if and only if*

$$tr(\mathfrak{I}^{p+1}(\xi^*)\boldsymbol{K}^T\boldsymbol{M}^-(\xi^*)\boldsymbol{M}(\xi_x)\boldsymbol{M}^-(\xi^*)\boldsymbol{K}) \leq tr(\mathfrak{I}^{p+1}(\xi^*)\boldsymbol{K}^T\boldsymbol{M}^-(\xi^*)\boldsymbol{K}) \tag{3.2.3}$$

for all $x \in [L, U]$, where $\boldsymbol{M}^-(\xi^*)$ is some G-inverse of $\boldsymbol{M}(\xi^*)$, $\xi_x = \{x, m\}$ is the one-point design. Moreover, the equality holds at the support points of ξ^*.

3.3 Application

Before going to an application, there is another practical issue that needs to be addressed. The optimal designs suggest replications on each support point. However, in reality, it may occur that observations are non-repeatable within each subject, for example, when the covariate x is time. In such situation, an efficient design can be found by replacing each (x_k, ω_k) with a set of distinct x's that are close to x_k.

Applications of our method can be presented for multiple families of models, for example, models in Yang (2010) and Yang and Stufken (2012). We will take the mixed effects version of the biexponential regression model as an example. To make it meaningful, the experimental settings are taken from a real-life study described in Pinheiro and Bates (2000). It is a study

about the pharmacokinetics of the drug indomethicin. In this study, there are six human subjects, each subject received the same dose of indomethicin and his/her drug concentration was measured 11 times between 0.25 hours and 8 hours after injection. In Pinheiro and Bates (2000), the jth measurement from subject i was modeled as

$$y_{ij} = \beta_{i1}\exp(-e^{\beta_{i2}}x_{ij}) + \beta_{i3}\exp(-e^{\beta_{i4}}x_{ij}) + \epsilon_{ij}, \quad i = 1, ..., 6; j = 1, ..., 11,$$

where $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{i2}, \beta_{i3}, \beta_{i4})^T \sim N(\boldsymbol{\beta}, \sigma^2\Sigma)$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^T$ and $\beta_2 > \beta_4$. Different covariance structure of the random effects were proposed and compared, and finally, the authors used the following block-diagonal structure for Σ:

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & 0 & 0 \\ \sigma_{21} & \sigma_{22} & 0 & 0 \\ 0 & 0 & \sigma_{33} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Based on this covariance model, the parameters are estimated and given in Table 3.1 (Pinheiro and Bates 2000, p.287).

Table 3.1: Parameter estimates for indomethicin data

Parameter	β_1	β_2	β_3	β_4	σ_{11}	σ_{12}	σ_{22}	σ_{33}	σ
Estimate	2.8045	0.8502	0.5887	-1.1029	78.93	17.54	4.76	5.50	0.0782

We will design our locally optimal designs based on the estimated parameters $\hat{\boldsymbol{\beta}}$ and $\widehat{\Sigma}$ in Table 3.1. The design space is chosen to be $[0.25, 8]$.

Because $e^{\hat{\beta}_2}/e^{\hat{\beta}_4} = e^{0.8502+1.1029} < 61.98$, by Yang and Stufken (2012), designs with at most 4 support points including L form a complete class. Now we are going to find optimal designs within Ξ, specifically, we will focus on A and D optimality.

If the design space were $[0.25, \infty)$, Theorem 3.2.3 can be used to find the optimal designs, and the results are listed in Table 3.2. Because the largest support points are 6.018 and 5.153

for A-optimal design and D-optimal design, respectively, which are below 8, we can conclude that they are also the optimal designs on the restricted design space of $[0.25, 8]$.

Table 3.2: A- and D-optimal designs for indomethicin study

A-optimal		D-optimal	
(x_1, x_2, x_3, x_4)	(w_1, w_2, w_3, w_4)	(x_1, x_2, x_3, x_4)	(w_1, w_2, w_3, w_4)
(0.250, 0.619, 1.655, 6.018)	(1.417, 1.770, 2.770, 5.043)	(0.250 0.601 1.882 5.487)	(0.774 2.117 2.956 5.153)

Now we compare the optimal designs with other designs. We consider four alternatives, the design used in the original study ξ_0 , the uniform design ξ_u and the "practical" optimal designs ξ^*_{p1} and ξ^*_{p2}. The original design takes observations at $0.25, 0.5, 0.75, 1, 1.25, 2, 3, 4, 5, 6, 8$, the uniform design takes observations at 11 equally spaced design points, $0.25, 1.025, 1.8, ..., 8$, the "practical" optimal designs are constructed from the locally optimal design by rounding and distributing the weight on each support point to distinct design points. Specifically, round the weights in ξ^* to integers first, say $(\bar{w}_1, \bar{w}_2, \bar{w}_3, \bar{w}_4)$; then for any x_k that has a weight $\bar{w}_k > 1$, we will replace $(x_k, \bar{w}_k)$ with $\bar{w}_k$ distinct design points that are close to x_k. In practice, we may not be able to choose the $\bar{w}_k$ new design points to be arbitrarily close to x_k, so we consider two situations here. First, assume we can only take observations at a multiple of 0.25 hours, and the practical optimal design obtained under this constraint is denoted as ξ^*_{p1}. Second, assume we can take observations at any time as long as they are at least 0.1 hours apart, denote the practical optimal design obtained under this constraint as ξ^*_{p2}. Therefore, the practical optimal designs can be chosen as (they don't have to be unique)

$$\xi^*_{p1} = \begin{cases} (0.25, 0.5, 0.75, 1.5, 1.75, 2, 5.5, 5.75, 6, 6.25, 6.5) & \text{A-optimality} \\ (0.25, 0.5, 0.75, 1.75, 2, 2.25, 5, 5.25, 5.5, 5.75, 6) & \text{D-optimality} \end{cases}$$

and

$$\xi^*_{p2} = \begin{cases} (0.25, 0.57, 0.67, 1.56, 1.66, 1.76, 5.8, 5.9, 6, 6.1, 6.2) & \text{A-optimality} \\ (0.25, 0.55, 0.65, 1.80, 1.90, 2.0, 5.3, 5.4, 5.5, 5.6, 5.7) & \text{D-optimality} \end{cases}$$

Now we compare the efficiencies of these four alternatives with ξ^* under A- and D-optimality criteria. The efficiency of any design ξ compared to the optimal design ξ^* is defined as

$$\text{eff}(\xi) = \begin{cases} \frac{\text{tr}(\mathfrak{M}^{-1}(\xi^*))}{\text{tr}(\mathfrak{M}^{-1}(\xi))}, & \text{A-optimality} \\ \left(\frac{\det(\mathfrak{M}^{-1}(\xi^*))}{\det(\mathfrak{M}^{-1}(\xi))}\right)^{1/d}, & \text{D-optimality} \end{cases},$$

here $d = 4$.

Efficiency is directly related with sample size. For example, an efficiency of 0.8 means the optimal design only needs 80% of the sample size ξ needed to achieve the same accuracy of parameter estimation. The efficiencies of the four alternative designs are listed in Table 3.3.

Table 3.3: Efficiencies for the alternative designs

Optimality	ξ_0	ξ_u	ξ^*_{p1}	ξ^*_{p2}
A	0.855	0.776	0.976	0.990
D	0.870	0.765	0.973	0.994

From Table 3.3, we can see the uniform design has the worst efficiency under both optimality criteria. Its efficiencies are 20% lower than the optimal designs and the practical optimal designs. The practical optimal designs have only a negligible loss of efficiency, and it is not surprising to see ξ^*_{p2} performs slightly better than ξ^*_{p1}. The original design has a decent efficiency, however, it is still 10% less than the practical optimal designs.

3.4 Robustness

Because the covariance information is hard to obtain, we consider the robustness of design efficiency with respect to miss-specifying the covariance matrix Σ. Since A-optimal designs are invariant to Σ, we only consider D-optimal designs. It would be too time consuming and lengthy to consider all different models and different parameter settings, so we will restrict considerations to the biexponential model and the following parameter settings. The

regression parameters are fixed to be $\hat{\boldsymbol{\beta}}$, and we consider two parallel scenarios about the structure of Σ. One is that Σ is singular, specifically we consider the case where $\sigma_{44} = 0$. The other scenario is that Σ is positive definite. We will use $\widetilde{\Sigma}$ and $\widetilde{R}$ to denote the mis-specified covariance and correlation matrices, respectively. We will find D-optimal designs under the mis-specified covariance matrix $\widetilde{\Sigma}$, and compute their efficiencies with respect to optimal designs under the true covariance matrix Σ. The design space is kept as $[0{\cdot}25,\ 8]$, and n as 11. Under this setting, all D-optimal designs below can be found using Theorem 3.2.3.

First, assume the variances are known exactly and study the effect of the miss-specified correlation matrix. Each time, we sample a random correlation matrix using `rcorrmatrix` function in the R package `clusterGeneration`, and treat it as the true correlation. The miss-specified correlation matrix is fixed to be diagonal (no correlation). The variances are assumed to be known exactly, but to make the result more general, we draw the variances from a distribution instead of fixing them at certain values.

To be specific, in the first scenario, let $\sigma_{ii} = \hat{\sigma}_{ii} \times r_i, i = 1, 2, 3, 4$ (note $\hat{\sigma}_{44} = 0$), where each r_i has a discrete uniform distribution on the set of ratios $\{1/\bar{r}, 2/\bar{r}, ..., (\bar{r}-1)/\bar{r}, 1, 2, ..., \bar{r}\}$, $\bar{r}$ is a positive integer. In the second scenario we consider a positive definite Σ. We can still let $\sigma_{ii} = \hat{\sigma}_{ii} \times r_i, i = 1, 2, 3, 4$, except replacing $\hat{\sigma}_{44}$ with a positive value. Because the scale of $\hat{\beta}_4$ is about one half the scale of $\hat{\beta}_1$ and two time the scale of $\hat{\beta}_3$, 20 seems to be a reasonable value to replace $\hat{\sigma}_{44}$. We consider three values for $\bar{r}$, $\bar{r} = 1, 10, 100$, thus allowing the variances to vary by a maximum of four orders of magnitude. For each value of $\bar{r}$ in each scenario, 500 optimal designs are generated based on the miss-specified diagonal covariance matrices, also 500 optimal designs are generated based on the true covariance matrices, and the efficiencies of the former compared to the latter are summarized in Table 3.4.

Table 3.4 shows that the design efficiencies are robust against mis-specified correlation matrix. For scenario one, the minimum efficiency goes below 90%, however the 5% quantile shows that 95% of the efficiencies are above 97%, and the P(eff$\geq$0.95) column shows more

Table 3.4: Efficiencies of D-optimal designs under mis-specified correlation

	Scenario 1			Scenario 2		
$\bar{r}$	min	5%	P(eff≥0.95)	min	5%	P(eff≥0.95)
1	94.2%	97.2%	99.8%	95.1%	97.7%	100%
10	89.5%	97.4%	99.2%	95.3%	98.1%	100%
100	89.0%	98.8%	99.2%	94.2%	99.1%	99.6%

than 99% of the efficiencies are above 95%. For scenario two, the results look even better. Therefore, we can conclude that when the variances are known exactly, the design efficiencies are robust against mis-specification in the correlation matrix.

Second, we allow mis-specification in both variances and correlations, but we assume there are some prior knowledge about Σ, for example, an estimate. This naturally leads us to sampling $\widetilde{\Sigma}$ from the Wishart distribution. Specifically, sample a random matrix $\mathbf{W}$ from Wishart distribution $W_4(\Sigma, \nu)$ with degrees of freedom ν, and let $\widetilde{\Sigma} = \mathbf{W}/\nu$ be the mis-specified covariance matrix. Again we consider two scenarios here. In the first scenario, we let $\Sigma = \widehat{\Sigma}$. Note due to the singular structure of $\widehat{\Sigma}$, $\widetilde{\Sigma}$ is also rank 3 with the last column 0. In the second scenario, we let the variances be $\sigma_{ii} = \hat{\sigma}_{ii}, i = 1, 2, 3$ and $\sigma_{44} = 20$, and the correlation structure be exchangeable with $\rho = 0.5$.

Table 3.5: Efficiencies of D-optimal designs under mis-specified Wishart Σ's

		Scenario 1			Scenario 2		
ν	c_v	min	5%	P(eff≥0.95)	min	5%	P(eff≥0.95)
4	0.707	93.4%	97.2%	99.8%	95.3%	97.6%	100%
7	0.535	96.7%	98.7%	100%	97.1%	98.6%	100%
10	0.447	98.0%	99.2%	100%	97.9%	98.9%	100%

We know $\mathrm{E}(\widetilde{\Sigma}) = \Sigma$ and $\mathrm{Var}(\tilde{\sigma}_{ij}) = (\sigma_{ii}^2 + \sigma_{ii}\sigma_{jj})/\nu$, $i, j = 1, 2, 3, 4$. Specially $\mathrm{Var}(\tilde{\sigma}_{ii}) = 2\sigma_{ii}^2/\nu$, so the coefficient of variation for any $\tilde{\sigma}_{ii}$ is $c_v = \sqrt{2\sigma_{ii}^2/\nu}/\sigma_{ii} = \sqrt{2/\nu}$. Therefore, we can manipulate ν to control the error between the mis-specified covariance and the true one. We consider $\nu = 4, 7, 10$. For each value of ν in each scenario, 500 optimal designs

are constructed based on 500 mis-specified $\widetilde{\Sigma}$'s, and their efficiencies under the true Σ are evaluated and given in Table 3.5.

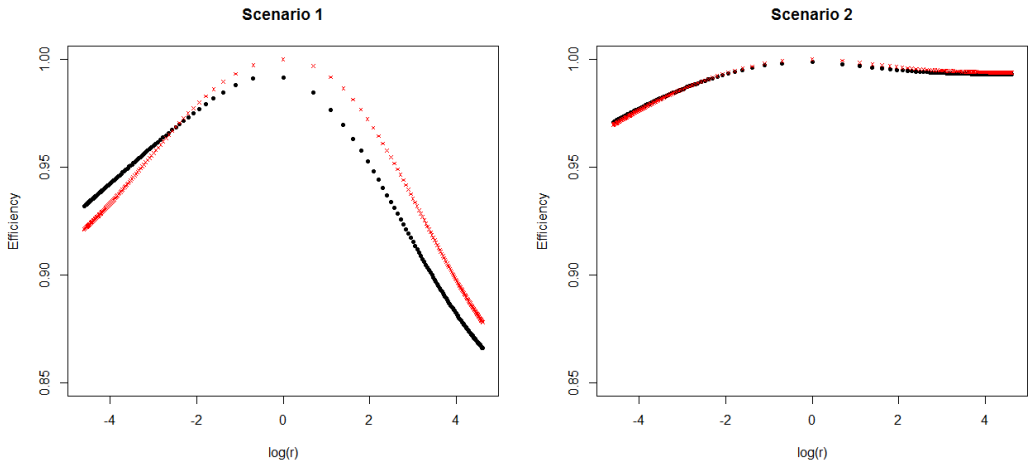

Figure 3.1: Efficiency plots for scenario one and two, the third study

From Table 3.5, we can see the design efficiencies are quite robust given we have certain amount of knowledge about Σ from previous experience. The worse case happens when $\nu = 4$, and the minimum efficiency goes below 95% (93.4%) in scenario one. However, 95% of the time, the efficiency is above 97.2%, and 99.8% of the time, the efficiency is above 95%. For Scenario two, the efficiencies are above 95% all the time.

Third, we don't assume any prior knowledge about the structure of Σ, so there will be larger errors in specifying the covariance matrix. For the true covariance matrix, let them to be the same as in the second study. For the mis-specified covariance, we consider two cases in each scenario. The first case, assume a diagonal covariance matrix with mis-specified variances $\tilde{\sigma}_{ii} = \sigma_{ii} \times r, i = 1, 2, 3, 4$. Note unlike the first study, the ratio r is kept the same for all four variance components. This means the variances will be over estimated ($r > 1$) or under estimated ($r < 1$) simultaneously. The second case, we use the same mis-specified variance components as in the first case, but adopting the true correlation structure. This is an extension of the first study, to see the robustness against mis-specification in the

correlation when the variances are also mis-specified. Take $r = 1/100, 1/99, ..., 99, 100$, the efficiencies of optimal designs derived under mis-specified covariance matrices are given in Figure 3.1, where red crosses show design efficiencies under diagonal correlation structure and black dots show design efficiencies under true correlation structure. Note the x-axis is $\log(r)$ instead of r.

In Figure 3.1, it is not surprising to see that the efficiency decreases as the mis-specification in variances aggravates (r moves further away from 1). However, the pattern and pace of the decline are very different between the two scenarios. The decline is much faster in the first scenario than in the second scenario; and while the efficiencies are worse for over estimated variances than under estimated variances in the first scenario, it is the contrary in the second scenario. In both scenarios, we find that the efficiency differences between case one (red crosses) and case two (black dots) are minor, which suggests miss-specifying correlation structures doesn't affect design efficiencies much.

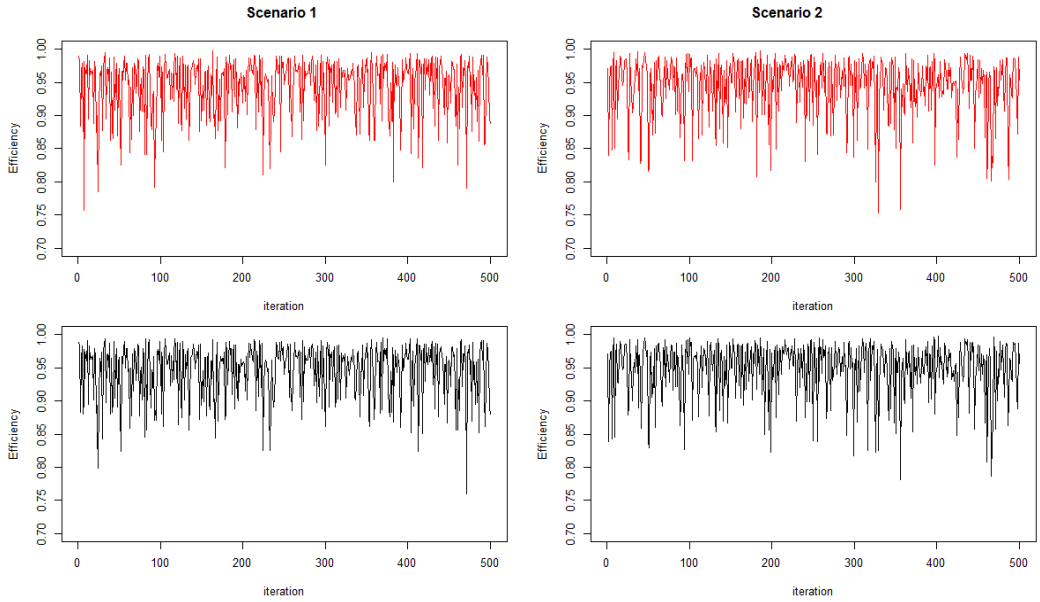

Figure 3.2: Efficiency plots for scenario one and two, the fourth study

Finally, the fourth study is similar to the third study but more general. In the third study, the variances are over estimated or under estimated simultaneously, but in prac-

tice this is hardly the case. Moreover, only one true correlation structure is considered in each scenario. So to make a more general study, the mis-specified variances are set to be $\tilde{\sigma}_{ii} = \sigma_{ii} \times r_i, i = 1, 2, 3, 4$, where r_i has a discrete uniform distribution on the set of ratios $\{1/\bar{r}, 2/\bar{r}, ..., 1, 2, ..., \bar{r}\}$. The true correlation structure is sampled randomly using rcorrmatrix function. Still we consider two cases in each scenario, as did in the third study. Set $\bar{r} = 100$ and 500 iterations are run for each case, the results are given in Figure 3.2, where upper panel shows design efficiencies under diagonal correlation structure and lower panel shows design efficiencies under true correlation structure. Note the lines in Figure 3.2 are jagged because different correlation matrix and variances are sampled at each iteration.

Comparing upper panel with lower panel, we can see the efficiency plots are very similar. So this again confirms our observation that design efficiencies are generally robust to mis-specified correlation structures, and it also means the imperfect design efficiencies are caused mainly by mis-specified variances. Table 3.6 summarizes the efficiencies. The minimum efficiency for either case under either scenario is only slightly above 75%, the 5% quantile is around 85% and the proportion of efficiencies that are at least 95% is merely 60%. The efficiencies aren't poor, but it is not that robust either. Note that unlike in the third study, the efficiencies between the two scenarios are more comparable now.

Table 3.6: Efficiencies of D-optimal designs under mis-specified covariance matrices

	Scenario 1			Scenario 2		
Case	min	5%	P(eff≥0.95)	min	5%	P(eff≥0.95)
One	75.7%	86.2%	63.4%	75.3%	84.4%	58.4%
Two	76.0%	86.8%	62.8%	78.2%	85.8%	61.2%

Summarizing the results, design efficiencies are generally very robust against mis-specification in the correlation structure, but not so robust against mis-specification in the variance components unless some prior knowledge is known.

Chapter 4

Multiple-covariate generalized linear models, independent data

4.1 Introduction

The models we have considered so far all contain only one covariate. For models with multiple covariates, particularly generalized linear models, research is far behind. Most efforts in the literature rely solely on limited computational tools like simulated annealing or Fedorov exchange algorithm. However, as the number of covariates increases, the optimization problem can get very time-consuming. As pointed out in Khuri et al. (2006), "The situations where one has several covariates (control variables) or multiple responses corresponding to each subject demand extensive work to evaluate optimal or at least efficient designs". Therefore, theoretical guidance is desired in this multiple-covariate setting.

However, methodological research is even more scarce than numerical studies. There are only a couple of papers that provided theoretical results. For count data, Russell et al. (2009) gave explicit formula for D-optimal designs under a Poisson regression model, when the design space is relatively large. For binary data, Sitter and Torsney (1995) considered c-

and D-optimal designs for logistic/probit regression models with two covariates. Haines et al. (2007) also studied D-optimal designs for the logistic regression model with two covariates, but under a slightly different set-up. Yang, Zhang and Huang (2011) derived closed-form solutions of optimal designs for logistic/probit regression models with an arbitrary number of covariates, where the design space is unbounded for exactly one covariate and bounded for all other covariates. Specifically, a small class of designs in which the optimal designs can be found was identified, and explicit forms of D-, A- and E-optimal designs were given. However, none of these papers considered any interaction effects between the covariates, which usually makes the problem much harder.

In this chapter, we will extend the methods in Yang, Zhang and Huang (2011) to more general models and more general design spaces. We will focus on loglinear models for count data and logistic/probit regression models for binary data, with an arbitrary number of covariates and possible interactions between the covariates. The design space can be any hypercube. We are not able to give explicit forms of optimal designs in such a general setting, but we will use the complete class approach to identify a relatively small class of designs in which the optimal designs can be found, thus greatly simplifying the computational cost of the numerical search.

4.2 Theory

Assume there are d unrelated continuous covariates, $x_1, \ldots, x_d$, and the response Y can be either binary or count. First, consider models without interactions. The loglinear, logistic and probit regression models can be written in the following unifying form:

$$E(Y_i) = P(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_d x_{id}), \quad x_{ij} \in [L_j, U_j], i = 1, ..., n, j = 1, ..., d, \qquad (4.2.1)$$

where n is the number of subjects, x_{ij} is the value of covariate x_j for subject i, x_{ij} belongs to the design space $[L_j, U_j]$, and $P(u) = e^u$ for loglinear models, $P(u) = e^u/(1+e^u)$ for logistic models and $P(u) = \Phi(u)$ for probit models, where $\Phi(u)$ is the cumulative distribution function of the standard normal distribution.

Let $\mathbf{x}_q = (x_{q1}, ..., x_{qd})^T$. An approximate design $\xi = \{(\boldsymbol{x}_q, w_q)\}_{q=1}^t$ puts nw_q observations at the design point $\mathbf{x}_q$, where $\sum_{q=1}^t w_q = 1$ and the nw_q's don't have to be integers. The (scaled) information matrix under ξ can be written as

$$\mathbf{M}(\xi) = \sum_{q=1}^{t} w_q \mathbf{M}(\boldsymbol{x}_q) = \sum_{q=1}^{t} w_q \Psi(u_q) \mathbf{z}_q \mathbf{z}_q^T, \tag{4.2.2}$$

where $u_q = \beta_0 + \beta_1 x_{q1} + \cdots + \beta_d x_{qd}, \mathbf{z}_q = (1, \mathbf{x}_q^T)^T, \Psi(u) = e^u$ for loglinear models and $\Psi(u) = [P'(u)]^2/[P(u)(1-P(u))]$ for logistic/probit models.

Let $\chi = \prod_{j=1}^d [L_j, U_j]$, define $S_j = \{(x_1, ..., x_d) \in \chi | x_\ell = L_\ell \text{ or } U_\ell \text{ for any } \ell \neq j\}$ as the union of all parallel 1-dimensional edges of the design hypercube.

Theorem 4.2.1. *Under model (4.2.1), for any design ξ, there exists another design $\tilde{\xi} = \{(\tilde{\boldsymbol{x}}_s, \tilde{w}_s)\}_{s=1}^{t'}$, such that*
(1) $\boldsymbol{M}(\tilde{\xi}) \geq \boldsymbol{M}(\xi)$ under Loewner ordering.
(2) $\tilde{\boldsymbol{x}}_s \in \bigcup_{j=1}^d S_j, s = 1, ..., t'$.

Proof. We will construct $\tilde{\xi}$ by replacing each design point that's not on the edge with a few design points on the edge and meanwhile improve the information matrix. Without loss of generality, assume $L_j < x_{qj} < U_j$ for $j = d-1$ and d, for design point $\boldsymbol{x}_q$. Fix $\boldsymbol{x}_q^{(1)} = (x_{q1}, ..., x_{q,d-2})^T$, consider the constant logit line that passes through $\boldsymbol{x}_q^{(2)} = (x_{q,d-1}, x_{q,d})^T$,

$$\beta_{d-1} z_1 + \beta_d z_2 = \beta_{d-1} x_{q,d-1} + \beta_d x_{q,d}, \quad z_1 \in [L_{d-1}, U_{d-1}],\ z_2 \in [L_d, U_d].$$

Let $\bar{\boldsymbol{x}}_q^{(2)} = (\bar{x}_{q,d-1}, \bar{x}_{q,d})^T$ and $\underline{\boldsymbol{x}}_q^{(2)} = (\underline{x}_{q,d-1}, \underline{x}_{q,d})^T$ be the two points on the perimeter of the design rectangle $[L_{d-1}, U_{d-1}] \times [L_d, U_d]$ where this line intersects at. There exists a λ such that $\boldsymbol{x}_q^{(2)} = \lambda \bar{\boldsymbol{x}}_q^{(2)} + (1-\lambda)\underline{\boldsymbol{x}}_q^{(2)}$, $0 < \lambda < 1$, and by convexity,

$$\lambda \bar{\boldsymbol{x}}_q^{(2)} \bar{\boldsymbol{x}}_q^{(2)T} + (1-\lambda)\underline{\boldsymbol{x}}_q^{(2)} \underline{\boldsymbol{x}}_q^{(2)T} \geq \boldsymbol{x}_q^{(2)} \boldsymbol{x}_q^{(2)T}. \tag{4.2.3}$$

Replace $\boldsymbol{x}_q$ with $\bar{\boldsymbol{x}}_q = (\boldsymbol{x}_q^{(1)T}, \bar{\boldsymbol{x}}_q^{(2)T})^T$ and $\underline{\boldsymbol{x}}_q = (\boldsymbol{x}_q^{(1)T}, \underline{\boldsymbol{x}}_q^{(2)T})^T$, with weight λw_q and $(1-\lambda)w_q$, respectively. Because of (4.2.3),

$$\begin{aligned}
\mathbf{M}(\boldsymbol{x}_q) &= \Psi(u_q) \begin{pmatrix} 1 & \boldsymbol{x}_q^{(1)T} & \boldsymbol{x}_q^{(2)T} \\ \boldsymbol{x}_q^{(1)} & \boldsymbol{x}_q^{(1)} \boldsymbol{x}_q^{(1)T} & \boldsymbol{x}_q^{(1)} \boldsymbol{x}_q^{(2)T} \\ \boldsymbol{x}_q^{(2)} & \boldsymbol{x}_q^{(2)} \boldsymbol{x}_q^{(1)T} & \boldsymbol{x}_q^{(2)} \boldsymbol{x}_q^{(2)T} \end{pmatrix} \\
&\leq \Psi(u_q) \begin{pmatrix} 1 & \boldsymbol{x}_q^{(1)T} & \boldsymbol{x}_q^{(2)T} \\ \boldsymbol{x}_q^{(1)} & \boldsymbol{x}_q^{(1)} \boldsymbol{x}_q^{(1)T} & \boldsymbol{x}_q^{(1)} \boldsymbol{x}_q^{(2)T} \\ \boldsymbol{x}_q^{(1)} & \boldsymbol{x}_q^{(2)} \boldsymbol{x}_q^{(1)T} & \lambda \bar{\boldsymbol{x}}_q^{(2)} \bar{\boldsymbol{x}}_q^{(2)T} + (1-\lambda)\underline{\boldsymbol{x}}_q^{(2)} \underline{\boldsymbol{x}}_q^{(2)T} \end{pmatrix} \\
&= \lambda \mathbf{M}(\bar{\boldsymbol{x}}_q) + (1-\lambda)\mathbf{M}(\underline{\boldsymbol{x}}_q).
\end{aligned}$$

Thus the above replacement produces a information matrix that is at least as large. By continuing to do this for $\bar{\boldsymbol{x}}_q$, $\underline{\boldsymbol{x}}_q$ and the other $\boldsymbol{x}_q$'s until all designs points are located at the edges, we end up with a design $\tilde{\xi}$ that satisfies (1) and (2). □

With Theorem 4.2.1, the search of optimal designs can be greatly simplified. Suppose we are going to use Fedorov's algorithm to find the D-optimal designs. To form the candidate set, κ grid points are chosen from each dimension $[L_j, U_j]$, thus the size of the candidate set $N = \kappa^d$. However, with Theorem 4.2.1, we only need the design points on the edges, which reduces N from κ^d to $\kappa d 2^{d-1}$. For example, when $\kappa = 100$ and $d = 3$, N reduces from 1 million to 1200. As d gets larger, the reduction will only become more substantial.

Now let's look at models with two-way interactions. Let $I_2 = \{(j,k)|1 \le j < k \le d\}$ be the set of two-way interactions we are interested in, which need not be all two-way interactions. Then the model can be written as

$$E(Y_i) = P(\beta_0 + \sum_{j=1}^{d} \beta_j x_{ij} + \sum_{(j,k)\in I_2} \beta_{jk} x_{ij} x_{ik}), \quad x_{ij} \in [L_j, U_j],\ i = 1, ..., n, \tag{4.2.4}$$

The information matrix under design ξ can be written in the same format as in 4.2.2 with $\mathbf{z}_q = (1, x_{q1}, ..., x_{qd}, ..., x_{qj}x_{qk}, ...)$.

Define $S_{j,k} = \{(x_1, ..., x_d)^T | x_\ell = L_\ell \text{ or } U_\ell \text{ for any } \ell \ne j, k\}$, so $S_{j,k}$ is actually a set of parallel 2-dimensional faces of the design hypercube χ.

Theorem 4.2.2. *If there is no $j < k < l$ such that all three two-way interactions between x_j, x_k, x_l exist in (4.2.4), then for any design ξ, there exists another design $\tilde{\xi} = \{(\tilde{\boldsymbol{x}}_s, \tilde{w}_s)\}_{s=1}^{t'}$, such that*

(1) $\boldsymbol{M}(\tilde{\xi}) \ge \boldsymbol{M}(\xi)$ under Loewner ordering.

(2) $\tilde{\boldsymbol{x}}_s \in (\bigcup_{j=1}^{d} S_j) \bigcup (\bigcup_{(j,k)\in I_2} S_{j,k})$, $s = 1, ..., t'$..

Proof. First, suppose there exists some q and $j < k < l$ such that $L_\ell < x_{q\ell} < U_\ell$ for $\ell = j, k, l$. Since not all three interactions between x_j, x_k, x_l exist in (4.2.4), we can assume $(j,k) \notin I_2$. Fix $x_{q\ell}$ for $\ell \ne j, k$, so that (4.2.4) becomes a model with only two covariates and no interaction term. So by Theorem 4.2.1, (x_{qj}, x_{qk}) can be replaced by design points (with suitable weights) on the perimeter of the design rectangle. Apply the same replacement to $\boldsymbol{x}_q$. Keep on doing this until all design points are on the faces of the design hypercube. Meanwhile, condition (1) is satisfied.

Now that all design points are on the faces, assume $\boldsymbol{x}_q \in S_{j,k}$ but for some $(j,k) \notin I_2$. Fix all covariates other than x_j, x_k, so that (4.2.4) again reduces to a model with only two covariates and no interaction. Therefore $\boldsymbol{x}_q$ can be replaced by design points on the edges of the design hypercube while keeping the information matrix at least as large. Repeat this

procedure to any $\boldsymbol{x}_q$ that doesn't satisfy (2), we finally get a better design $\tilde{\xi}$ satisfying both Conditions (1) and (2). □

Remark 4.2.1. The condition that not all three two-way interactions exist is necessary. For example, for the following logistic regression model with 3 covariates and all two way interactions:

$$logit(Prob(Y_i = 1)) = \beta_0 + \sum_{j=1}^{3} \beta_j x_{ij} + \sum_{1 \le j < k \le 3} \beta_{jk} x_{ij} x_{ik}, \quad x_{ij} \in [0, 1], i = 1, ..., n, j = 1, 2, 3, \tag{4.2.5}$$

the c-optimal design with $\mathbf{c} = (1, 0.5, 0.5, 0.5, 0.25, 0.25, 0.25)^T$ can be found to be the single point design which puts all weights on the point $(0.5, 0.5, 0.5)$. Moreover, this design is better than any design found on the faces of the design cube. Therefore, Theorem 4.2.2 does not hold in this case.

4.3 Computational savings

Theorem 4.2.1 and 4.2.2 greatly reduce the searching space for the optimal designs. This can in turn save substantial computational resources.

Consider a logistic regression model with 3 covariates and no interactions:

$$logit(Prob(Y_i = 1)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}, \quad x_{ij} \in [0, 1], i = 1, ..., n, j = 1, 2, 3.$$

We use the optimal weight exchange algorithm (OWEA) proposed in Yang, Biedermann and Tang (2013) to find the optimal designs. We consider two different candidate sets

1. the full candidate set with all κ^d grid points, with grid size κ;
2. the reduced candidate set with $d2^{d-1}\kappa$ grid points from the edges only, with the same grid size κ.

Table 4.1: Computation time (seconds) for A- and D-optimal designs, $\boldsymbol{\beta} = (-4, 1, 1, 1)^T$

	A-optimal			D-optimal		
candidate	$\kappa = 50$	$\kappa = 100$	$\kappa = 200$	$\kappa = 50$	$\kappa = 100$	$\kappa = 200$
full	4.62	34.58	270.17	4.64	35.10	271.14
reduced	0.24	0.28	0.31	0.24	0.28	0.29

Table 4.1 shows the computation times for finding A- and D-optimal designs. We can see that computation time using the reduced candidate set is much less, and the saving increases exponentially as the grid size κ increases. In particular, when κ doubles, the computation time using the full candidate set increases by approximately a factor of $2^3 = 8$; this is consistent with the increment in the size of the full candidate set. However, the increment is marginal if the reduced candidate set is used. Moreover, the saving increases exponentially as the number of covariates d increases.

4.4 Discussion

For some specific optimality criteria, the optimal designs may be found on a smaller space than the one given in Theorem 4.2.2. For example, Table 4.2 shows that the support points of A- and D-optimal designs for Model 4.2.5 are all located on the faces, where the searching space given in Theorem 4.2.2 has to be the whole design cubic. However, the reasons behind this are unclear and needs further investigation.

Table 4.2: Support points in A- and D-optimal designs, for $\boldsymbol{\beta} = (-4, 1, 1, 1, 1, 1, 1)^T$

D-optimality	x_1	0	1	0	0	1	1	0	1	1	0.9
	x_2	0	0	1	0	1	0	1	1	0.9	1
	x_3	0	0	0	1	0	1	1	0.9	1	1
A-optimality	x_1	0	1	0	0	1	1	0	1		
	x_2	0	0	1	0	1	0	1	1		
	x_3	0	0	0	1	0	1	1	1		

Chapter 5

Discussion

This dissertation covers three different topics of the optimal design area: single covariate linear/nonlinear/generalized linear models, single covariate linear/nonlinear mixed effects models and multiple-covariate generalized linear models. The underlying idea is nevertheless the same; by using the complete class approach demonstrated in Yang and Stufken (2009), Yang (2010) and Yang and Stufken (2012), general theoretical results can be obtained for a class of models and optimality criteria. However, the problems are not completely solved.

First, for single covariate fixed effects models, we point out some models that cannot be accommodated. One case is that when the complete class given by Theorem 2.2.1 is not small enough. For example, in Dette et al. (2010), D-optimal designs for a nonlinear model with mean $\eta(x, \boldsymbol{\beta}) = \beta_1 + \beta_2 \exp(x/\beta_3), x \in [L, U]$ are found to be 3-point designs with both endpoints, whereas a complete class consists of designs with at most 3 design points including only the upper endpoint as a fixed design point [Yang (2010), Theorem 3]. So the D-optimal designs are actually on the boundary of the $\mathbf{Z}$-space, hence no feasible critical points can be found, and Theorem 2.2.3 does not apply.

Another case is that when the model contains multiple covariates. In general, theoretical results are very hard to obtain for multiple-covariate models, and only a couple of papers

have provided some theoretical guidance. Specific to our approach, complete class results similar to Theorem 2.2.1 are not available. The reason is that complete class results are built upon Chebyshev systems. However, there is no satisfactory multi-dimensional generalization of the Chebyshev system yet. While Yang, Zhang and Huang (2011) gave complete class results for logistic and probit models with multiple covariates, the complete classes are not derived using multidimensional Chebyshev systems, and they are not small enough for our method to be applied.

Second, for mixed effects models, we haven't considered generalized linear mixed effects models. Similar to the nonlinear mixed models, the marginal log likelihood in a generalized linear mixed model also does not have a closed form. A first order linearization can be used, as we did for nonlinear mixed models; however, unlike nonlinear mixed models, the responses are not normal for generalized linear mixed models, thus a normal approximation of the response is needed, and the approximation can be poor for binary outcomes. In addition to that, the conditional variance (conditioning on the random effects) depends on the conditional mean. This is different from nonlinear mixed models, where the conditional variances are homogeneous. This creates another complication.

Third, for multiple-covariate generalized linear models, we know that optimal designs may not be found on the 2-dimensional faces if for example, all three two-way interactions exist in a logistic regression model with three covariates. However, our numerical study shows that the Φ_p-optimal designs for $\boldsymbol{\beta}$ can still be found on the faces, but it is unclear how to prove it. Besides, it is also not intuitively clear why the complete class result does not hold in this case.

Appendix A

Proofs

We will prove Theorem 2.2.2 and Theorem 2.2.5. Before proving Theorem 2.2.2, we first provide a lemma. This lemma is easier stated in terms of u, but it can be translated into x. Recall that Theorem 2.2.1 gives the form of a complete class. For any design ξ, we can find a design $\tilde{\xi} = \{(\tilde{u}_j, \tilde{\omega}_j)\}_{j=1}^m$ in the complete class that is non-inferior ($\mathbf{M}(\tilde{\xi}) \geq \mathbf{M}(\xi)$).

In particular, for ξ specified in Lemma A.0.1, let $\Psi_0(u) \equiv 1$, a design $\tilde{\xi}$ can be found by solving the following nonlinear equation system [see Yang and Stufken (2012) and Dette and Schorning (2013)]:

$$\sum_i \omega_i \Psi_\ell(u_i) = \sum_j \tilde{\omega}_j \Psi_\ell(\tilde{u}_j), \quad \ell = 0, 1, \ldots, q-1 \tag{A.0.1}$$

where $\tilde{u}_1$ and $\tilde{u}_a$ may be fixed to be boundary points (see Lemma A.0.1). Multiply both sides of (A.0.1) by a positive constant, the equation system still holds, so we can remove the constraint of $\sum_i w_i = 1$ for ξ and allow $\sum_i w_i$ to be any positive number in the following Lemma A.0.1; similarly for $\tilde{\xi}$ (but we still refer to them as designs for convenience). Let $\boldsymbol{X} = (\mathbf{u}^T, \boldsymbol{\omega}^T)^T$ be the vector of all u_i's and ω_i's in ξ. Let $\mathbf{S}_1$ and $\mathbf{S}_2$ be the sets of all possible vectors $\boldsymbol{X}$ corresponding to designs in cases (1a)$\sim$(1d) and (2) of Lemma A.0.1,

respectively. Further let $\boldsymbol{Y}$ be the vector of all $\tilde{u}_j$'s except those fixed as boundary points (if any) and all $\tilde{\omega}_j$'s in design $\tilde{\xi}$ given in Lemma A.0.1. We will define function H, $H(\boldsymbol{X}) = Y$, where $\boldsymbol{X} \in \mathbf{S} = \mathbf{S}_1 \bigcup \mathbf{S}_2$, and show this function is smooth on $\mathbf{S}$ under certain conditions.

Lemma A.0.1. *Suppose one of the conditions in Theorem 2.2.1 holds.*

(1a) If $q = 2a - 1$ and $\boldsymbol{H}(u) < 0$, then for any design $\xi = \{(u_i, \omega_i)\}_{i=1}^{a}, A < u_1 < \ldots < u_a \leq B, \omega_i > 0$ for $i \geq 1$, there exists a non-inferior design $\tilde{\xi} = \{(\tilde{u}_j, \tilde{\omega}_j)\}_{j=1}^{a}$, where $\tilde{u}_1 = A, \tilde{\omega}_j > 0$ for $j \geq 1$, that solves (A.0.1).

(1b) If $q = 2a - 1$ and $\boldsymbol{H}(u) > 0$, then for any design $\xi = \{(u_i, \omega_i)\}_{i=1}^{a}, A \leq u_1 < \ldots < u_a < B, \omega_i > 0$ for $i \geq 1$, there exists a non-inferior design $\tilde{\xi} = \{(\tilde{u}_j, \tilde{\omega}_j)\}_{j=1}^{a}$, where $\tilde{u}_a = B, \tilde{\omega}_j > 0$ for $j \geq 1$, that solves (A.0.1).

(1c) If $q = 2a$ and $\boldsymbol{H}(u) < 0$, then for any design $\xi = \{(u_i, \omega_i)\}_{i=1}^{a+1}, A \leq u_1 < \ldots < u_{a+1} \leq B, \omega_i > 0$ for $i \geq 1$, there exists a non-inferior design $\tilde{\xi} = \{(\tilde{u}_j, \tilde{\omega}_j)\}_{j=1}^{a}$, where $\tilde{\omega}_j > 0$ for $j \geq 1$, that solves (A.0.1).

(1d) If $q = 2a - 2$ and $\boldsymbol{H}(u) > 0$, then for any design $\xi = \{(u_i, \omega_i)\}_{i=1}^{a-1}, A < u_1 < \ldots < u_{a-1} < B, \omega_i > 0$ for $i \geq 1$, there exists a non-inferior design $\tilde{\xi} = \{(\tilde{u}_j, \tilde{\omega}_j)\}_{j=1}^{a}$, where $\tilde{u}_1 = A, \tilde{u}_a = B, \tilde{\omega}_j > 0$ for $j \geq 1$, that solves (A.0.1).

Such solution is unique under each case, hence H is well defined on $\boldsymbol{S}_1$.

(2) For each case of (1a)$\sim$(1d), let ξ be similarly defined as above except that there is exactly one 0 weight and all other weights are positive. Then rewriting ξ in the form of $\tilde{\xi}$ in each corresponding case solves (A.0.1) and defines H on $\boldsymbol{S}_2$. Moreover, H is smooth on $\boldsymbol{S} = \boldsymbol{S}_1 \bigcup \boldsymbol{S}_2$.

proof of Lemma A.0.1. We only prove for case (a), others being similar. First, let's consider (1a). From Lemma 1 in Yang (2010) [see also Dette and Schorning (2013), Theorem 3.1], we

know that a solution to (A.0.1) exists with $\tilde{u}_1 = A, \tilde{\omega}_j > 0, j \geq 1$. Moreover, $\mathbf{H}(u) < 0$ implies that $\{\Psi_0, \Psi_1, \ldots, \Psi_{2a-2}\}$ is a Chebyshev system [see Yang and Stufken (2012), Proposition 4], thus such solution is unique. So H is well defined on $\mathbf{S}_1$. Now we show the smoothness on $\mathbf{S}_1$.

We have $\boldsymbol{X} = (u_1, \ldots, u_a, \omega_1, \ldots, \omega_a)^T, \boldsymbol{Y} = (\tilde{u}_2, \ldots, \tilde{u}_a, \tilde{\omega}_1, \ldots, \tilde{\omega}_m)^T$ by definition ($\tilde{u}_1$ is excluded in $\boldsymbol{Y}$ since it is fixed to be A). Subtract the left hand side from the right hand side in (A.0.1), we get an equation system $G(\boldsymbol{X}, \boldsymbol{Y}) = 0$, where G is smooth. So $\boldsymbol{Y} = H(\boldsymbol{X})$ is the implicit function defined by $G(\boldsymbol{X}, \boldsymbol{Y}) = 0$. By implicit function theorem, to ensure H to be smooth, we only need the Jacobian matrix $G_{\boldsymbol{Y}}(\boldsymbol{X}, \boldsymbol{Y}) = \partial G(\boldsymbol{X}, \boldsymbol{Y})/\partial \boldsymbol{Y}$ to be nonsingular, i.e.,

$$
\begin{aligned}
& \det G_{\boldsymbol{Y}}(\boldsymbol{X}, \boldsymbol{Y}) \\
= & \begin{vmatrix}
0 & \cdots & 0 & 1 & \cdots & 1 \\
\tilde{\omega}_2 \Psi_1'(\tilde{u}_2) & \cdots & \tilde{\omega}_a \Psi_1'(\tilde{u}_a) & \Psi_1(A) & \cdots & \Psi_1(\tilde{u}_a) \\
\vdots & \ddots & \vdots & \vdots & \vdots & \ddots \\
\tilde{\omega}_2 \Psi_{2a-2}'(\tilde{u}_2) & \cdots & \tilde{\omega}_a \Psi_{2a-2}'(\tilde{u}_a) & \Psi_{2a-2}(A) & \cdots & \Psi_{2a-2}(\tilde{u}_a)
\end{vmatrix} \\
= & (\prod_{j=2}^{a} \tilde{w}_j) d(\tilde{\mathbf{u}}) \neq 0,
\end{aligned}
$$

where

$$
d(\tilde{\mathbf{u}}) = \begin{vmatrix}
1 & \cdots & 1 & 0 & \cdots & 0 \\
\Psi_1(A) & \cdots & \Psi_1(\tilde{u}_a) & \Psi_1'(\tilde{u}_2) & \cdots & \Psi_1'(\tilde{u}_a) \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\Psi_{2a-2}(A) & \cdots & \Psi_{2a-2}(\tilde{u}_a) & \Psi_{2a-2}'(\tilde{u}_2) & \cdots & \Psi_{2a-2}'(\tilde{u}_a)
\end{vmatrix}. \tag{A.0.2}
$$

Since $\tilde{w}_j > 0$ for all $1 \leq j \leq a$, we only need to show $d(\tilde{\mathbf{u}}) \neq 0$. We first do some column manipulations to the matrix in (A.0.2). Subtract the first column from the second to the ath

column, then for the resulting matrix, subtract the second column from the third to the ath column, continue doing this until finally subtract the $(a-1)$th column from the ath column. Because the determinant doesn't change during this process,

$$d(\tilde{\mathbf{u}}) = \begin{vmatrix} \Psi_1(\tilde{u}_2) - \Psi_1(A) & \cdots & \Psi_1(\tilde{u}_a) - \Psi_1(\tilde{u}_{a-1}) & \\ \vdots & \ddots & \vdots & \mathbf{D} \\ \Psi_{2a-2}(\tilde{u}_2) - \Psi_{2a-2}(A) & \cdots & \Psi_{2a-2}(\tilde{u}_a) - \Psi_{2a-2}(\tilde{u}_{a-1}) & \end{vmatrix}, \tag{A.0.3}$$

where $\mathbf{D}$ is the $(2a-2) \times (a-1)$ matrix,

$$\mathbf{D} = \begin{pmatrix} \Psi_1'(\tilde{u}_2) & \cdots & \Psi_1'(\tilde{u}_a) \\ \vdots & \ddots & \vdots \\ \Psi_{2a-2}'(\tilde{u}_2) & \cdots & \Psi_{2a-2}'(\tilde{u}_a) \end{pmatrix}.$$

Treat A in the first column of the matrix in (A.0.3) as a variable and fix everything else, then the determinant becomes a real-valued function of A. By the mean value theorem,

$$\begin{aligned} &d(\tilde{\mathbf{u}}) = (\tilde{u}_2 - A)\times \\ &\begin{vmatrix} \Psi_1'(\hat{u}_1) & \Psi_1(\tilde{u}_3) - \Psi_1(\tilde{u}_2) & \cdots & \Psi_1(\tilde{u}_a) - \Psi_1(\tilde{u}_{a-1}) & \\ \vdots & \vdots & \ddots & \vdots & \mathbf{D} \\ \Psi_{2a-2}'(\hat{u}_1) & \Psi_{2a-2}(\tilde{u}_3) - \Psi_{2a-2}(\tilde{u}_2) & \cdots & \Psi_{2a-2}(\tilde{u}_a) - \Psi_{2a-2}(\tilde{u}_{a-1}) & \end{vmatrix}, \end{aligned} \tag{A.0.4}$$

where $A < \hat{u}_1 < \tilde{u}_2$. Let $\varepsilon = \text{sign } d(\tilde{\mathbf{u}})$ be the sign of $d(\tilde{\mathbf{u}})$, treat $\tilde{u}_2$ in the second column of the matrix in (A.0.4) as a variable, and use the mean value theorem again to obtain

$$\varepsilon = \text{sign} \begin{vmatrix} \Psi_1'(\hat{u}_1) & \Psi_1'(\hat{u}_2) & \cdots & \Psi_1(\tilde{u}_a) - \Psi_1(\tilde{u}_{a-1}) & \\ \vdots & \vdots & \ddots & \vdots & \mathbf{D} \\ \Psi_{2a-2}'(\hat{u}_1) & \Psi_{2a-2}'(\hat{u}_2) & \cdots & \Psi_{2a-2}(\tilde{u}_a) - \Psi_{2a-2}(\tilde{u}_{a-1}) & \end{vmatrix},$$

where $\tilde{u}_2 < \hat{u}_2 < \tilde{u}_3$. Keep on doing this, and finally get

$$\varepsilon = \text{sign} \begin{vmatrix} \Psi_1'(\hat{u}_1) & \cdots & \Psi_1'(\hat{u}_{a-1}) & \Psi_1'(\tilde{u}_2) & \cdots & \Psi_1'(\tilde{u}_a) \\ \vdots & \ddots & \cdots & \cdots & \ddots & \cdots \\ \Psi_{2a-2}'(\hat{u}_1) & \cdots & \Psi_{2a-2}'(\hat{u}_{a-1}) & \Psi_{2a-2}'(\tilde{u}_2) & \cdots & \Psi_{2a-2}'(\tilde{u}_a) \end{vmatrix},$$

and $A = \tilde{u}_1 < \hat{u}_1 < \tilde{u}_2 < \hat{u}_2 < \ldots < \hat{u}_{a-1} < \tilde{u}_a$. Since $\{\Psi_1', \ldots, \Psi_{2a-2}'\}$ is a Chebyshev system, $\varepsilon \neq 0$. Hence the Jacobian matrix is invertible, and the function H is smooth on $\mathbf{S}_1$.

Turning to case (2), without loss of generality, assume $\omega_1 = 0, \omega_i > 0$ for $i \geq 2$. If we can show the function $H(\boldsymbol{X})$ is continuous on $\mathbf{S}_2$ and its partial derivatives can be extended continuously to $\mathbf{S}_2$, then it can be proved that $H(\boldsymbol{X})$ is also differentiable on $\mathbf{S}_2$. So first, we prove its continuity.

To show this, for any sequence $\boldsymbol{X}^n = (u_1^n, ..., u_a^n, \omega_1^n, ..., \omega_a^n)^T, n \geq 1$, $\boldsymbol{\omega}^n > 0$ and $\boldsymbol{X}^n$ approaching $\boldsymbol{X}^0 = (u_1, ..., u_a, 0, \omega_2, \ldots, \omega_a)^T$, we need to show $\boldsymbol{Y}^n = (\tilde{u}_2^n, ..., \tilde{u}_a^n, \tilde{\omega}_1^n, ..., \tilde{\omega}_a^n)^T$ approaches $\boldsymbol{Y}^0 = (u_2, ...u_a, 0, \omega_2..., \omega_a)^T$.

By definition we have

$$\sum_{i=1}^{a} \omega_i^n \Psi_\ell(u_i^n) = \sum_{j=1}^{a} \tilde{\omega}_j^n \Psi_\ell(\tilde{u}_j^n), \quad \ell = 0, ..., 2a-2. \tag{A.0.5}$$

Suppose we have $\boldsymbol{Y}_{j_1}^n$ doesn't converge to $\boldsymbol{Y}_{j_1}^0$ for some j_1, then because $\boldsymbol{Y}^n$ is a bounded sequence, there exists a subsequence $\{n_t | t = 1, 2, ...\}$ such that $\boldsymbol{Y}^{n_t}$ converges to some $\bar{\boldsymbol{Y}}^0 = (\bar{u}_2, ..., \bar{u}_a, \bar{\omega}_1, ..., \bar{\omega}_a)^T$ and $\bar{\boldsymbol{Y}}_{j_1}^0 \neq \boldsymbol{Y}_{j_1}^0$.

Now let $n_t \to \infty$, take the limit of (A.0.5) on both sides, we get:

$$\sum_{i=2}^{a} \omega_i \Psi_\ell(u_i) = \sum_{j=1}^{a} \bar{\omega}_j \Psi_\ell(\bar{u}_j), \quad \ell = 0, ..., 2a-2. \tag{A.0.6}$$

Since $\{\Psi_0, \ldots, \Psi_{2a-2}\}$ is a Chebyshev system and the maximum number of different support

points in (A.0.6) is $2a-1$, (A.0.6) only holds if $\bar{\omega}_1 = 0, \bar{\omega}_i = \omega_i, \bar{u}_i = u_i$ for $i \geq 2$, which means $\bar{\boldsymbol{Y}}^0 = \boldsymbol{Y}^0$, leading to a contradiction.

Next, we show the partial derivatives can be extended continuously to $\mathbf{S}_2$. Using the implicit function theorem, we know

$$\frac{\partial H(\boldsymbol{X})}{\partial \boldsymbol{X}} = -G_{\boldsymbol{Y}}^{-1}(\boldsymbol{X}, H(\boldsymbol{X}))G_{\boldsymbol{X}}(\boldsymbol{X}, H(\boldsymbol{X})), \quad G_{\boldsymbol{X}}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\partial G(\boldsymbol{X}, \boldsymbol{Y})}{\partial \boldsymbol{X}},$$

for $\boldsymbol{X} \in \mathbf{S}_1$. When $\boldsymbol{X} \to \boldsymbol{X}^0$, $H(\boldsymbol{X}) \to H(\boldsymbol{X}^0)$ by continuity, hence $G_{\boldsymbol{Y}}(\boldsymbol{X}, H(\boldsymbol{X})) \to G_{\boldsymbol{Y}}(\boldsymbol{X}^0, H(\boldsymbol{X}^0))$ since $G_{\boldsymbol{Y}}(\boldsymbol{X}, \boldsymbol{Y})$ is continuous. Furthermore, $G_{\boldsymbol{Y}}(\boldsymbol{X}^0, H(\boldsymbol{X}^0))$ is nonsingular by the similar argument as previously, therefore, $G_{\boldsymbol{Y}}^{-1}(\boldsymbol{X}, H(\boldsymbol{X})) \to G_{\boldsymbol{Y}}^{-1}(\boldsymbol{X}^0, H(\boldsymbol{X}^0))$. It is easy to see $G_{\boldsymbol{X}}(\boldsymbol{X}, H(\boldsymbol{X})) \to G_{\boldsymbol{X}}(\boldsymbol{X}^0, H(\boldsymbol{X}^0))$, therefore, the derivative $\partial H(\boldsymbol{X})/\partial \boldsymbol{X} \to -G_{\boldsymbol{Y}}^{-1}(\boldsymbol{X}^0, H(\boldsymbol{X}^0))G_{\boldsymbol{X}}(\boldsymbol{X}^0, H(\boldsymbol{X}^0))$, i.e., the derivative can be extended continuously to $\mathbf{S}_2$. So $H(\boldsymbol{X})$ is differentiable on $\mathbf{S}_2$ and the partial derivatives are continuous. □

Now we are ready to prove Theorem 2.2.2, the proof is stated in terms of x to be consistent with the theorem.

proof of Theorem 2.2.2. We only prove the case where the complete class consists of designs with at most a points including L, other cases being similar. Assume the design ξ^c given by a feasible critical point is not an optimal design, and an optimal design exists as $\xi^* = \{(L, 1-\sum_{i=2}^a \omega_i^*), \{(x_i^*, \omega_i^*)\}_{i=2}^a\}$, where $L < x_2^* < \ldots < x_a^*$ is a strictly increasing sequence (some of the weights ω_i^* may be 0 if the support size of ξ^* is less than a). We have $\Phi(\mathbf{M}(\xi^*)) > \Phi(\mathbf{M}(\xi^c))$. Consider the linear combination of the two designs, $\xi_\epsilon = \epsilon\xi^* + (1-\epsilon)\xi^c$, $0 \leq \epsilon \leq 1$, so

$$\xi_\epsilon = \{(L, 1-(1-\epsilon)\sum_{i=2}^a \omega_i^c - \epsilon\sum_{i=2}^a \omega_i^*), \{(x_i^c, (1-\epsilon)\omega_i^c)\}_{i=2}^a, \{(x_i^*, \epsilon\omega_i^*)\}_{i=2}^a\}.$$

By the concavity of the optimality criterion Φ, we have $\Phi(\mathbf{M}(\xi_\epsilon)) \geq (1-\epsilon)\Phi(\mathbf{M}(\xi^c)) +$

$\epsilon\Phi(\mathbf{M}(\xi^*))$, which gives

$$\frac{\Phi(\mathbf{M}(\xi_\epsilon)) - \Phi(\mathbf{M}(\xi^c))}{\epsilon} \geq \Phi(\mathbf{M}(\xi^*)) - \Phi(\mathbf{M}(\xi^c)) > 0. \tag{A.0.7}$$

Now, if we can find $\tilde{\xi}_\epsilon = \{(L, 1 - \sum_{i=2}^{a} \omega_{i,\epsilon}), \{(x_{i,\epsilon}, \omega_{i,\epsilon})\}_{i=2}^{a}\}$, $\epsilon \geq 0$ belongs to a neighborhood of 0, to be a set of designs with a support points such that

1. $\Phi(\mathbf{M}(\tilde{\xi}_\epsilon)) \geq \Phi(\mathbf{M}(\xi_\epsilon))$;

2. $\boldsymbol{Z}_\epsilon = (\boldsymbol{x}_\epsilon, \boldsymbol{\omega}_\epsilon)$ depends smoothly on ϵ, where $\boldsymbol{x}_\epsilon = (x_{2,\epsilon}, \ldots, x_{a,\epsilon})$, $\boldsymbol{\omega}_\epsilon = (\omega_{2,\epsilon}, \ldots, \omega_{a,\epsilon})$;

3. $\boldsymbol{Z}_0 = \boldsymbol{Z}^c = (\boldsymbol{x}^c, \boldsymbol{\omega}^c)$, thus $\tilde{\xi}_0 = \xi^c$.

Then, applying (A.0.7) to get

$$\frac{\Phi(\mathbf{M}(\tilde{\xi}_\epsilon)) - \Phi(\mathbf{M}(\tilde{\xi}_0))}{\epsilon} \geq \frac{\Phi(\mathbf{M}(\xi_\epsilon)) - \Phi(\mathbf{M}(\xi^c))}{\epsilon} \geq \Phi(\mathbf{M}(\xi^*)) - \Phi(\mathbf{M}(\xi^c)) > 0.$$

Because $\tilde{\xi}_\epsilon$ has $a \geq d$ support points, $\mathbf{M}(\tilde{\xi}_\epsilon)$ must belong to $\mathrm{PD}(d)$. By our smoothness assumption of Φ, $\Phi(\mathbf{M}(\tilde{\xi}_\epsilon))$ is a smooth function of ϵ. Take the limit as $\epsilon \to 0$, it gives

$$\left.\frac{\partial \Phi(\mathbf{M}(\tilde{\xi}_\epsilon))}{\partial \epsilon}\right|_{\epsilon=0} > 0. \tag{A.0.8}$$

On the other hand, by our definition, $\Phi(\mathbf{M}(\tilde{\xi}_\epsilon)) = \tilde{\Phi}(\boldsymbol{Z}_\epsilon)$. Applying the chain rule and using the fact that $\boldsymbol{Z}_0 = \boldsymbol{Z}^c$ is a critical point of $\tilde{\Phi}(\boldsymbol{Z})$, we can get

$$\left.\frac{\partial \Phi(\mathbf{M}(\tilde{\xi}_\epsilon))}{\partial \epsilon}\right|_{\epsilon=0} = \left.\frac{\partial \tilde{\Phi}(\boldsymbol{Z}_\epsilon)}{\partial \epsilon}\right|_{\epsilon=0} = \left.\frac{\partial \tilde{\Phi}(\boldsymbol{Z})}{\partial \boldsymbol{Z}}\right|_{\boldsymbol{Z}=\boldsymbol{Z}_0} \left.\frac{\partial \boldsymbol{Z}_\epsilon}{\partial \epsilon}\right|_{\epsilon=0} = 0.$$

This contradicts with (A.0.8). Hence ξ^c must be an optimal design.

To find such designs $\tilde{\xi}_\epsilon$, first, if the design ξ^* doesn't have new design points other than those in ξ^c, that is, $\forall 2 \leq i \leq a$, we have either $\omega_i^* = 0$ or $x_i^* \in \boldsymbol{x}^c$, then the design ξ_ϵ is itself

a design with a support points, therefore we can simply let $\tilde{\xi}_\epsilon = \xi_\epsilon$, and Conditions $1 \sim 3$ are satisfied.

Otherwise, suppose we have $r > 0$ new design points $x^*_{i_1}, ..., x^*_{i_r}$ introduced by ξ^*, with $\omega^*_{i_k} > 0, k = 1, ..., r$. Let $\delta_{ii'} = 1$ if $x^c_i = x^*_{i'}$ and 0 otherwise. Rewrite the design ξ_ϵ as

$$\begin{aligned}\xi_\epsilon &= \{(L, 1-(1-\epsilon)\sum_{i=2}^{a}\omega^c_i - \epsilon\sum_{i=2}^{a}\omega^*_i), \{(x^c_i, (1-\epsilon)\omega^c_i + \epsilon\sum_{i'=1}^{a}\omega^*_{i'}\delta_{ii'})\}_{i=2}^{a}\} \\ &\quad \cup\{(x^*_{i_k}, \epsilon\omega^*_{i_k})\}_{k=1}^{r} \\ &= \{(L, \omega^{(0)}_{1,\epsilon}), \{(x^{(0)}_{i,\epsilon}, \omega^{(0)}_{i,\epsilon})\}_{i=2}^{a}\} \cup \{(x^*_{i_k}, \epsilon\omega^*_{i_k})\}_{k=1}^{r},\end{aligned}$$

where the second equation simply renames the design points and design weights. It is easy to verify that conditions $2 \sim 3$ are satisfied for $\boldsymbol{Z}^{(0)}_\epsilon = (\boldsymbol{x}^{(0)}_\epsilon, \boldsymbol{\omega}^{(0)}_\epsilon) = (x^{(0)}_{2,\epsilon}, \ldots, x^{(0)}_{a,\epsilon}, \omega^{(0)}_{2,\epsilon}, \ldots, \omega^{(0)}_{a,\epsilon})$.

To find the desired m-point design $\tilde{\xi}_\epsilon$, we need to reduce the number of design points in a "smooth" way. We reduce one point at a time. First consider the design $\{(x^{(0)}_{i,\epsilon}, \omega^{(0)}_{i,\epsilon})\}_{i=2}^{a} \cup \{(x^*_{i_1}, \epsilon\omega^*_{i_1})\}$, all the weights are positive when $0 < \epsilon < 1$, and when $\epsilon = 0$, only one weight is 0. So applying Lemma A.0.1 to this design we can get a new design $\{(L, \omega^{(1)}_{1,\epsilon}), \{(x^{(1)}_{i,\epsilon}, \omega^{(1)}_{i,\epsilon})\}_{i=2}^{a}\}$ that is non-inferior, and conditions $2 \sim 3$ are satisfied for $\boldsymbol{Z}^{(1)}_\epsilon = (\boldsymbol{x}^{(1)}_\epsilon, \boldsymbol{\omega}^{(1)}_\epsilon)$, where $\boldsymbol{\omega}^{(1)}_\epsilon > 0$ for $0 \le \epsilon < 1$.

Next, we add point $x^*_{i_2}$ to $\{(x^{(1)}_{i,\epsilon}, \omega^{(1)}_{i,\epsilon})\}_{i=2}^{a}$ (we can always assume $x^*_{i_2}$ is a new point to $\boldsymbol{x}^{(1)}_\epsilon$ by taking ϵ small enough). Again, all the weights are positive when $\epsilon > 0$, and when $\epsilon = 0$, only one weight is 0. Use the same method to reduce one design point again. Keep on doing this until all r new points have been added and reduced, we get $\tilde{\xi}_\epsilon = \{(L, 1 - \sum_{i=2}^{a}\omega^{(r)}_{i,\epsilon}), \{(x^{(r)}_{i,\epsilon}, \omega^{(r)}_{i,\epsilon})\}_{i=2}^{a}\}$, that is not inferior to ξ_ϵ, with the conditions $1 \sim 3$ satisfied. □

Finally, we prove Theorem 2.2.5, the proof is stated in terms of u for convenience.

proof of Theorem 2.2.5. We only consider the case of Theorem 2.2.1(a). First, ξ^* must belong to the complete class. Otherwise, we can find a design $\tilde{\xi}^*$ with $\mathbf{M}(\tilde{\xi}^*) \ge \mathbf{M}(\xi^*)$ and

$\mathbf{M}(\tilde{\xi}^*) \neq \mathbf{M}(\xi^*)$. Because ξ^* has at least d support points, $\mathbf{M}(\xi^*)$ is positive definite. Since Φ is strictly isotonic on PD(d), we have $\Phi(\mathbf{M}(\tilde{\xi}^*)) > \Phi(\mathbf{M}(\xi^*))$, which is a contradiction.

Now suppose there is another optimal design $\tilde{\xi}^*$.

(i) if $\tilde{\xi}^*$ also has at least d support points, then it also belongs to the complete class by previous arguments, and we can write $\xi^* = \{(u_i^*, \omega_i^*)\}_{i=1}^a$, $\tilde{\xi}^* = \{(\tilde{u}_i^*, \tilde{\omega}_i^*)\}_{i=1}^a$, $u_1^* = \tilde{u}_1^* = A$. By strict concavity, we must have $\mathbf{M}(\xi^*) \propto \mathbf{M}(\tilde{\xi}^*)$ since otherwise $\Phi(\alpha\mathbf{M}(\xi^*) + (1-\alpha)\mathbf{M}(\tilde{\xi}^*)) > \alpha\Phi(\mathbf{M}(\xi^*)) + (1-\alpha)\Phi(\mathbf{M}(\tilde{\xi}^*)) = \Phi(\mathbf{M}(\xi^*))$ for all $\alpha \in (0,1)$. Let $\mathbf{M}(\xi^*) = \delta\mathbf{M}(\tilde{\xi}^*)$, then $\Phi(\delta\mathbf{M}(\tilde{\xi}^*)) = \Phi(\mathbf{M}(\tilde{\xi}^*))$. The strict isotonicity of Φ implies $\delta = 1$, hence $\mathbf{M}(\xi^*) = \mathbf{M}(\tilde{\xi}^*)$ and $\mathbf{C}(\xi^*) = \mathbf{C}(\tilde{\xi}^*)$. Then we have (A.0.1) holds. Because $\mathbf{H}(u) < 0$, $\{\Psi_0, \ldots, \Psi_{2a-2}\}$ is a Chebyshev system. The maximum number of different support points in (A.0.1) is $2a-1$, so (A.0.1) only holds if the design points and weights on two sides of the equations are equal, which means $\xi^* = \tilde{\xi}^*$.

(ii) if $\tilde{\xi}^*$ has less than d support points, let $\xi_\alpha = \alpha\xi^* + (1-\alpha)\tilde{\xi}^*, 0 < \alpha < 1$. By concavity, ξ_α is also an optimal design, moreover, it has at least d support points. Thus following the arguments in case (i), we have $\xi_\alpha = \xi^*$, which means $\xi^* = \tilde{\xi}^*$. This contradicts with the fact that $\tilde{\xi}^*$ has less than d support points. □

Bibliography

[1] Bickel, P. J. and Herzberg, A. M. (1979), "Robustness of design against autocorrelation in time I: Asymptotic theory, optimality for location and linear regression", *Annals of Statistics*, 7, 77-95.

[2] Bickel, P. J. and Herzberg, A. M. and Schilling, M. F. (1981), "Robustness of design against autocorrelation in time II: Optimality, theoretical and numerical results for the first-order autoregressive process", *Journal of the American Statistical Association*, 76, 870-877.

[3] de la Garza, A. (1954), "Spacing of information in polynomial regression", *Annals of Mathematical Statistics*, 25, 123–130.

[4] Demidenko, E. (2004), *Mixed Model: Theory and Applications*, John Wiley & Sons, Hoboken, New Jersey.

[5] Demidenko, E. (2006), "The assessment of tumor response to treatment", *Applied Statistics*, 55, 365-377.

[6] Dette, H. (1997), "Designing experiments with respect to 'standardized' optimality criteria", *J.R. Stat. Soc. Ser. B Stat. Methodol.*, 59, 97-110.

[7] Dette, H. and Bretz, F. and Pepelyshev, A. and Pinheiro, J. (2008), "Optimal designs for dose-finding studies", *Journal of the American Statistical Association*, 103, 1225-1237.

[8] Dette, H. and Kiss, C. and Bevanda, M. and Bretz, F. (2010), "Optimal designs for the Emax, log-linear and exponential Models", *Biometrika*, 97, 513-518.

[9] Dette, H. and Melas, V. B. and Wong, W. K. (2006), "Locally D-optimal designs for exponential regression models", *Statistica Sinica*, 16, 789-803.

[10] Dette, H. and Melas, V. B. (2011), "A note on the de la Garza phenomenon for locally optimal designs", *The Annals of Statistics*, 39, 1266-1281.

[11] Dette, H. and Pepelyshev, A. and Holland-Letz, T. (2010), "Optimal designs for random effect models with correlated errors with applications in population pharmacokinetics", *The Annals of Applied Statistics*, 4, 1430-1450.

[12] Dette, H. and Pepelyshev, A. and Zhigljavsky, A. (2013), "Optimal design for linear models with correlated observations", *The Annals of Statistics*, 41, 143-176.

[13] Dette, H. and Schorning, K. (2013), "Complete classes of designs for nonlinear regression models and principal representations of moment spaces", *The Annals of Statistics*, 41, 1260-1267.

[14] Dette, H. and Studden, W.J. (1995), "Optimal designs for polynomial regression when the degree is not known", *Statistica Sinica*, 5,459-473.

[15] Elfving, G. (1952), "Optimum allocation in linear regression theory", *Annals of Mathematical Statistics*, 23, 255-262.

[16] Fisher, R. A. (1926), "The arrangement of field experiments", *Journal of the Ministry of Agriculture of Great Britain*, 33, 503-513.

[17] Fisher, R. A. (1935), *The Design of Experiments*, Oliver and Boyd.

[18] Haines, L. M. and Kabera, G. and Ndlovu, P. and O'Brien, T. E. (2007), "D-optimal designs for logistic regression in two variables", *mODa 8-Advances in Model-Oriented Design and Analysis*, 91-98.

[19] Hu, L. and Yang, M. and Stufken, J. (2015), "Saturated locally optimal designs under differentiable optimality criteria", *The Annals of Statistics*, 43, 30-56.

[20] Karlin, S. and Studden, W.J. (1966), *Tchebycheff systems: with applications in analysis and statistics*, Interscience, New York.

[21] Kiefer, J. (1959), "Optimum experimental designs", *Journal of the Royal Statistical Society, Series B*, 21, 272-319.

[22] Kiefer, J. and Wolfowitz, J. (1960), "The equivalence of two extremum problems", *Canadian Journal of Mathematics*, 12, 363-366.

[23] Kiefer, J. and Wolfowitz, J. (1965), "On a theorem of Hoel and Levine on extrapolation", *Annals of Mathematical Statistics*, 36, 1627-1655.

[24] Khuri, A. and Mukherjee, B. and Sinha B. and Ghosh M. (2006), "Design Issues for Generalized Linear Models: A Review", *Statistical Science*, 21, 376-399.

[25] Li, G. and Balakrishnan, N. (2011), "Optimal designs for tumor regrowth models", *Journal of Statistical Planning and Inference*, 141, 644-654.

[26] Pukelsheim, F. (1993), *Optimal Design of Experiments*, John Wiley & Sons, New York.

[27] R Development Core Team (2011), R: A language and environment for statistical computing, Vienna, Austria, R Foundation for Statistical Computing.

[28] Retout, S. and Dufful, S. and Mentre, F. (2001), "Development and implementation of the population Fisher information matrix for the evaluation of population pharmacokinetic designs". *Comput. Methods Programs Biomed*, 65, 141-151.

[29] Russell, K. G. and Woods, D. C. and Lewis, S. M. and Eccleston, J. A. (2009), "D-optimal designs for Poisson regression models", *Statistica Sinica*, 19, 721-730.

[30] Sacks, J. and Ylvisaker, N. D. (1966), "Designs for regression problems with correlated errors", *Annals of Mathematical Statistics*, 37, 66-89.

[31] Sacks, J. and Ylvisaker, N. D. (1968), "Designs for regression problems with correlated errors; many parameters", *Annals of Mathematical Statistics*, 39, 49-69.

[32] Schmelter, T. (2007a), "The optimality of single-group designs for certain mixed models", *Metrika*, 65, 183-193.

[33] Schmelter, T. (2007b), "Considerations on group-wise identical designs for linear mixed models", *Journal of Statistical Planning and Inference*, 137, 4003-4010.

[34] Sitter, R. R. and Torsney, B. (1995), "Optimal designs for binary response experiments with two design variables", *Statistics Sinica*, 5, 405-419.

[35] Smith, K. (1918), "On the standard deviations 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.

[36] Studden, W.J. (1968), "Optimal designs on Tchebycheff points", *The Annals of Mathematical Statistics*, 39, 1435-1447.

[37] Yang, M. (2010), "On the de la Garza Phenomenon", *The Annals of Statistics*, 38, 2499-2524.

[38] Yang, M. and Biedermann, S. and Tang, E. (2013), "On Optimal Designs for Nonlinear Models: A General and Efficient Algorithm", *Journal of the American Statistical Association*, 108,1411-1420.

[39] Yang, M. and Stufken, J. (2009), "Support points of locally optimal designs for nonlinear models with two parameters", *The Annals of Statistics*, 37, 518-541.

[40] Yang, M. and Stufken, J. (2012), "Identifying locally optimal designs for nonlinear models: A simple extension with profound consequences", *The Annals of Statistics*, 40, 1665-1681.

[41] Yang, M. and Zhang, B. and Huang, S. (2011), "Optimal designs for generalized linear models with multiple design variables", *Statistica Sinica*, 21, 1415-1430.