Experimental Design for Mixed Models with Application to Population Pharmacokinetic Studies

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)

von Thomas Schmelter
geb. am 26.03.1977 in Mainz

genehmigt durch die Fakultät für Mathematik
der Otto-von-Guericke-Universität Magdeburg

Gutachter: Prof. Dr. Rainer Schwabe
Dr. Barbara Bogacka

eingereicht am: 22.12.2006
Verteidigung am: 25.05.2007

Acknowledgment

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

I would like to express my deep gratitude to my advisor Prof. Dr. Rainer Schwabe for all his support throughout my PhD project. I am very grateful for his encouragement, his helpful suggestions, and the fruitful discussions we had.

I would like to thank the whole "Institut für Mathematische Stochastik" for the nice atmosphere and for the cross-word puzzles during the coffee breaks.

Since this PhD thesis was a joint project together with Schering AG, I had the luck to have Dr. Norbert Benda (now working at Novartis) as an additional advisor, whom I would really like to thank for his motivation, enthusiam, and helpful discussions.
Special thanks also go to Marc Vandemeulebroecke. I could really benefit from the experiences he had gathered within his own "project". We had many enlightening discussions.
Furthermore, I would like to thank the whole Clinical Statistics Europe department at Schering AG, where I was immediately warmly welcomed. Especially I would like to thank Dr. Jürgen Dinger, Dr. Hermann Kulmann, and Dr. Christoph Gerlinger for giving me any support I needed.
In addition, I would like to thank two kineticists from Schering AG, Dr. Stefanie Reif and Dr. Gabriele Fliß, for giving me an insight into the practical aspects of pharmacokinetic studies and modeling.

I would like to thank all my friends for a life besides work. Besides this, special thanks go to Christina Wunder for a last-minute proofreading of this thesis and Tobias Dussa, whom I want to thank not only for solving my TEXnical problems.

I would like to express my gratitude to my parents for their love and support and for never pushing me.
Most of all I want to thank Sonja Schlauch. She did not only help me practically with this thesis, but was always there for me during the last years with love and (sometimes necessary) encouragement.

The financial support of this project by Schering AG is gratefully acknowledged.

Summary

In this thesis topics of optimal experimental design for linear and non-linear mixed models are considered. This is motivated by the field of application of population pharmacokinetics in drug development. Thus, examples from this area will be used throughout this thesis for illustration purposes.

First an introduction to the well-known topic of optimal experimental designs for the ordinary linear model is given. Then the linear and non-linear mixed models that are considered within this thesis are introduced. Focus is, on one hand, put on the so-called random coefficient regression model, which is a regression model with random parameters, and, on the other hand, on a more general mixed model, where additional factors that are not to be controlled by the investigator are included into the model.

After this, a general definition of designs in mixed models is given. The designs are defined in two stages: The first stage are the elementary designs, which specify the settings for single individuals, while the second stage are population designs, by which the settings for the whole sample population are defined. On both levels we allow approximate designs.

As designs are usually evaluated by using real-valued functions of the respective information matrices, we derive different representations of the information matrices for the introduced designs, which allow their calculation also for approximate designs.

Two extreme cases of classes of population designs are considered. These are, on one hand, the single-group designs, where all individuals are observed under the same experimental settings and, on the other hand, general population designs, where different settings are allowed for different individuals.

We show that the design optimization can be restricted to the class of single group designs if the mean number of observations per individual is prespecified and criteria are considered that are only based on the population parameter vector and not on the variance parameters. The larger class of general population designs then does not contain better designs. This result is extended to the considered general mixed models. In this case, however, one elementary design for each distinct value of the uncontrolled factor is necessary.

Besides this, equivalence theorems, similar to the ones known for the ordinary linear model, are derived for various situations. They allow to check the optimality of given designs.

The thesis closes with a discussion, in which also practical aspects of experimental designs for population pharmacokinetic studies are addressed.

Zusammenfassung

Diese Arbeit beschäftigt sich mit Themen der optimalen Versuchsplanung für lineare und nicht-lineare gemischte Modelle. Motiviert wird dieses Thema durch das Anwendungsgebiet der Populationspharmakokinetik in der Arzneimittelentwicklung. Beispiele aus diesem Feld werden die ganze Arbeit hindurch zur Illustration immer wieder aufgegriffen.

Zunächst wird in der Arbeit eine Einführung in das bekannte Thema der optimalen Versuchsplanung in gewöhnlichen linearen Modellen gegeben. Anschließend werden die im weiteren Verlauf der Arbeit verwendeten linearen und nicht-linearen gemischten Modelle eingeführt. Im Vordergrund stehen das sogenannte „Random-Coefficient-Regression-Modell", ein Regressionsmodell mit zufälligen Parametern, und ein allgemeineres gemischtes Modell, in dem weitere nicht-steuerbare, aber bekannte Einflussgrößen zusätzlich in das Modell aufgenommen werden.

Es folgt eine allgemeine Definition der Versuchspläne (Designs) für gemischte Modelle. Diese werden in zwei Stufen definiert: zum einen Elementar-Designs, die die Versuchsbedingungen für einzelne Individuen vorgeben, und zum anderen Populationsdesigns, durch die die Einstellungen für alle zu untersuchenden Individuen definiert werden. Auf beiden Ebenen werden approximative Designs zugelassen.
Da Designs üblicherweise durch reellwertige Funktionen der zugehörigen Informationsmatrizen beurteilt werden, werden für die betrachteten Designs verschiedene Darstellungsweisen der Informationsmatrizen hergeleitet, die unter anderem auch deren Berechnung im approximativen Fall zulassen.

Im weiteren Verlauf der Arbeit werden zwei Extremfälle von Populationsdesign-Klassen betrachtet, zum einen Ein-Gruppen-Designs, in denen alle Individuen unter dem gleichen Versuchsplan beobachtet werden, und zum anderen allgemeine Populationsdesigns, in denen verschiedene Elementardesigns für die einzelnen Individuen zugelassen sind.
Wir zeigen, dass die Optimierung auf die Klasse der Ein-Gruppen-Designs eingeschränkt werden kann, wenn die mittlere Anzahl von Beobachtungen pro Individuum fest vorgegeben wird und Kriterien betrachtet werden, die nur auf den Populationsparametern und nicht auf den Varianzparametern basieren. Die größere Menge der allgemeinen Populationsdesigns bietet in diesem Fall keine besseren Designs. Dieses Resultat wird auf die allgemeineren gemischten Modelle ausgeweitet. Allerdings ist in diesem Fall je ein Elementardesign pro unterschiedlicher Ausprägung der zusätzlichen nicht zu steuernden Einflussgröße nötig.
Desweiteren werden Äquivalenzsätze, analog zu den bekannten aus dem gewöhnlichen linearen Modell, für verschiedene Situationen hergeleitet, die die Überprüfung der Optimalität von Designs ermöglichen.

Die Arbeit schließt mit einer Diskussion, in der unter anderem auf praktische Belange der Versuchsplanung für Populationspharmakokinetik-Studien eingegangen wird.

Contents

1 Introduction

The history of the theory of experimental design goes back about one hundred years (see, e. g., the review paper "One hundred years of the design of experiments on and off the pages of Biometrika" by Atkinson and Bailey (2001)). One remarkable very early work is by Smith (1918). Most of the theory for optimal experimental design, however, has been developed in the second half of the 20th century. One important milestone was the development of convex design theory including a series of equivalence theorems, of which the first and most famous one is by Kiefer and Wolfowitz (1960) showing the equivalence of D- and G-optimality.

Concerning theoretical results for experimental design in mixed models, which will be the topic of this thesis, the amount of literature is still relatively sparse (especially for the nonlinear case). There are results by Gladitz and Pilz (1982) for individual predictions in a Bayesian framework, Fedorov and Hackl (1997) touch these models in their book and derive an equivalence theorem, Liski et al. (2002) and Luoma (2000) give special results for linear and quadratic regression.

In the last years the demand for optimal or at least efficient designs in mixed models has strongly increased, which was also caused by the introduction and acceptance of population pharmacokinetic modeling in drug development. Besides this, there are also other areas, in which mixed models gained popularity, like agriculture, psychology or market research.

Population pharmacokinetic modeling is a development of the recent thirty years, starting probably with a row of publications by Sheiner (Sheiner et al. (1972), Sheiner et al. (1977), Sheiner and Beal (1980), Sheiner and Beal (1981)). The population approach is an alternative to the ordinary pharmacokinetic evaluation of a drug, where the concentration profiles are individually assessed for each subject and a dense sampling scheme is needed. Those pharmacokinetic studies are usually phase I studies and the samples are obtained from few healthy volunteers. In contrast, in population pharmacokinetic studies, few blood samples from many subjects are evaluated together in one model assuming that the same regression functions can be used for all subjects, however, with different parameter sets for the different individuals modeled by random effects. A further difference is that population pharmacokinetic investigations are usually carried out in the target population for the drug to assess the variability of the pharmacokinetic parameters of the substance.

As in all controlled investigations, planning of the experiment is a very crucial thing. By an insufficiently planned clinical study unnecessary many subjects might be included or unnecessary many blood samples might be taken, which is both costly and unethical. In the worst case, the objectives of the study cannot be investigated with the collected data, and the study is worthless.

The popularity of population pharmacokinetic modeling has led to an increased number of publications on optimal design approaches for these studies. A large part of the literature has

appeared in pharmaceutical journals treating practical design issues for specific concrete studies (see Mentré et al. (2001),Retout et al. (2002), or Hennig et al. (2006) besides others).

In this thesis the design problem in mixed models is investigated from a more theoretical perspective. Theoretical results are given, which are mainly not based on a specific kinetic model or criterion. Nevertheless, examples from the practical context of pharmacokinetics are given wherever appropriate. The actual calculation of optimal designs plays a minor role in this thesis, since similar algorithms as known for ordinary regression models can be used. Other standard optimization approaches like simulated annealing (see Duffull et al. (2002)) or a simplex algorithm (see Retout and Mentré (2003)) have also already been discussed for this purpose in the literature.

The thesis is organized as follows.

Chapter 2 gives an introduction to the area of pharmacokinetics. The most relevant models, namely the compartment models, and the parameters commonly used therein are described. Furthermore, the idea of the population approach to pharmacokinetics is motivated. In Chapter 3 some of the existing theory of experimental design for ordinary linear models is summarized. The chapter serves as a starting point for later theoretical investigations in the mixed model situation. Chapters 4 and 5 give an introduction to the linear and non-linear models that will be used throughout this thesis. A special emphasis is laid on the random coefficient regression model, which is a special case of the mixed model commonly used in population pharmacokinetics. In Chapter 6 the term "design" is formally defined for the mixed model situation. The definition is based on a similar hierarchical structure as used in the definition of the models. Different representations of the information matrices of designs for mixed models are given. It is shortly discussed which criteria might be useful for evaluating designs in the mixed model case. In Chapter 7 special results are given for criteria only based on the population mean parameters. It is shown that the set of candidate designs can be drastically restricted without losing quality. In Chapter 8 the convex design theory described in Chapter 3 for the ordinary linear model is transfered to the mixed model case. Equivalence theorems are derived for different situations. In Chapter 9 issues are addressed that can arise when theoretically optimal designs are implemented in practical studies. The results of this thesis are discussed in particular under this perspective.

2 Introduction to Pharmacokinetic Modeling

In this chapter we want to give a brief introduction to the topic of pharmacokinetics. We do not go into any detail regarding the pharmacological or physiological background, but rather give a short description from the point of view of modeling. We introduce some of the most common models and terms, as we will need them in the later chapters for the examples and the motivation of the discussed problems. In the description of the models we mainly follow the book of Derendorf et al. (2002). In the last section of this chapter we describe the idea of the population approach to pharmacokinetic modeling.

We start with a definition of the term *pharmacokinetics* that can be found in the Internet encyclopedia Wikipedia (2006).

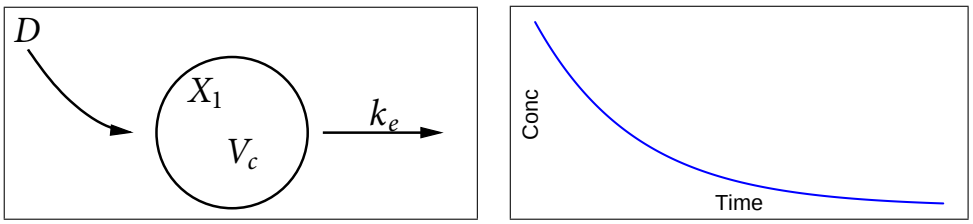
> **Pharmacokinetics** is a branch of pharmacology dedicated to the study of the time course of substances and their relationship with an organism or system. In practice, this discipline is applied mainly to drug substances, though in principle it concerns itself with all manner of compounds residing within an organism or system, such as nutrients, metabolites, endogenous hormones, toxins, etc. So, in basic terms, while pharmacodynamics explores what a drug does to the body, pharmacokinetics explores what the body does to the drug.

The aim of pharmacokinetic studies is, therefore, to collect data, for example from blood samples, to be able to describe and model the time course of the concentration of a substance in the body. One wants to gain quantitative information about the absorption, distribution, metabolism, and elimination of a drug in the body. A usual way to do this is by finding a structural model that describes the collected data appropriately and whose parameters are interpretable by the pharmacokineticist.

2.1 Models

The most commonly used models to describe the time course of a drug concentration are compartment models. Here, the body is thought to be subdivided into several compartments. By this it can be taken into account that the drug is not distributed to all parts of the body in the same way and with the same speed. The compartments, though, do not necessarily need to have a physiological interpretation. We concentrate on models with linear kinetics, where the concentrations are proportional to the dose and the rate of elimination of the drug is proportional to the concentration. In general much more complex models can be thought of.

The most simple one of these models is obviously the one-compartment model.



(a) (b)

Figure 2.1: *One-compartment model without absorption. (a): Structural model, (b): typical time course.*

2.1.1 One-compartment models

The organism is here seen as one system in which all body fluids are in a balance of flow. For these models one assumes that the distribution of the drug in the body happens within a negligible period of time, that is, the concentration of the drug in the organism is the same anywhere in the body at any time. Depending on the form of application, for example bolus injection, tablet, or infusion, the speed of the absorption of the drug in the body, however, can differ.

One-compartment model without absorption

In the case of an intravenous bolus injection we can assume that the whole amount of the drug is immediately absorbed into the compartment. So, in this case the absorption phase is negligible short. The elimination of the drug out of the body is usually modeled by a so-called first-order kinetics, where the speed of the elimination is proportional to the amount of the drug that is still left in the compartment.

The easiest way to describe this mathematically is by a simple initial value problem of the form

$$\begin{aligned} \frac{dX_1(t)}{dt} &= -k_e X_1(t), \\ X_1(0) &= D, \end{aligned}$$

where $X_1(t)$ is the amount of drug in the main (central) compartment, k_e is the elimination rate constant and D is the given dose. Figure 2.1(a) shows a graphical illustration of this process. The solution for $X_1(t)$ of this initial value problem is

$$X_1(t) = De^{-k_e t}$$

and, hence, the concentration $C(t)$ at time t can be modeled by

$$C(t) = \frac{X_1(t)}{V_c} = \frac{D}{V_c} e^{-k_e t},$$

where V_c is the volume (or rather the "volume of distribution", see Section 2.2) of the compartment. Figure 2.1(b) shows a typical time course of the concentration in this model.

The most common form of application, however, is the oral application in form of tablets. Here, the whole amount of the drug does not instantly reach the compartment, but it takes some time until the drug is absorbed. For the moment, however, we will still assume that there are no differences of the concentration within the body, that is, we still consider only one compartment.

One-compartment model with first-order absorption

Similar to the elimination process of the drug in the previous model we now assume that also the absorption process of the drug follows a first-order kinetics. Thus, the speed of the absorption is assumed to be proportional to the amount of drug that has not yet been absorbed. We can describe this process of the absorption and elimination by an initial value problem of the form

$$\begin{aligned} \frac{dX_1(t)}{dt} &= k_a X_2(t) - k_e X_1(t), \\ \frac{dX_2(t)}{dt} &= -k_a X_2(t) \end{aligned}$$

with initial conditions

$$X_2(0) = D \text{ and } X_1(0) = 0.$$

Here, $X_2(t)$ represents the amount of drug that has not yet been absorbed until time t, k_a is the absorption rate constant, and X_1 and k_e have the same meaning as before. Figure 2.2 gives a graphical illustration of the structure of this model.

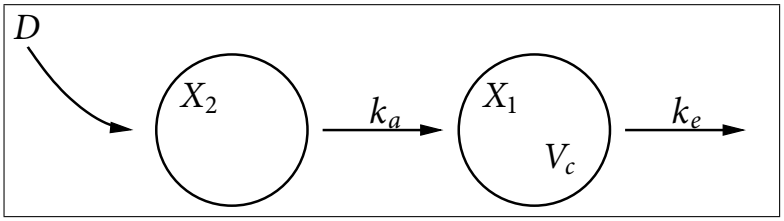


Figure 2.2: *Structure of a one-compartment model with absorption.*

The closed form solution for the concentration as a function of time is the so-called *Bateman function* given by

$$C(t) = \frac{X_1(t)}{V_c} = \frac{Dk_a}{V_c(k_a - k_e)} \left(e^{-k_e t} - e^{-k_a t} \right).$$

Often, depending on the form of application, not the whole amount of drug is absorbed. To model this a factor F is introduced in front of the dose D that gives the fraction of the amount of drug that is available to the organism. This constant F depends on the form of application

and is called *bioavailability*. That means the concentration is usually modeled by

$$C(t) = \frac{FDk_a}{V_c(k_a - k_e)}\left(e^{-k_e t} - e^{-k_a t}\right).$$

For a bolus injection this constant is 1 and could hence be omitted.

Figure 2.4(a) shows a typical time course of the concentration in this model on the semi-logarithmic scale.

2.1.2 Two-compartment models

Up to now we assumed that the absorbed amount of drug will be instantaneously distributed within the whole organism. That is, the process of the distribution of the drug within the body was not modeled. Often, however, it can take some time until the concentrations of the substance in all body liquids are in balance, making it necessary to consider more than just one compartment.

Exemplarily we consider the two-compartment model with first-order absorption.

Two-compartment model with first-order absorption

Compared to the respective one-compartment model, we now introduce an additional so-called peripheral compartment. The drug is absorbed and eliminated via the central compartment only. However, there is an exchange between the central and the peripheral compartment modeled by first-order kinetics. The process is illustrated in Figure 2.3.

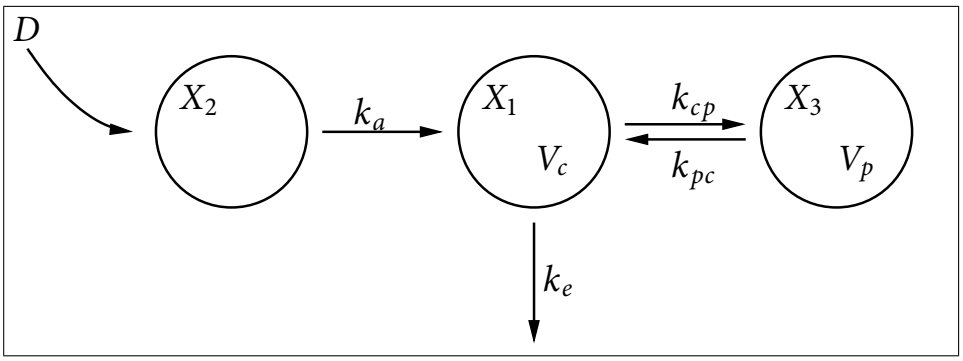


Figure 2.3: *Two-compartment model with first-order absorption*

Mathematically the model can be described by the following system of differential equations

$$\begin{aligned}
\frac{dX_1(t)}{dt} &= -(k_e + k_{cp})X_1(t) + k_a X_2(t) + k_{pc}X_3(t),\\
\frac{dX_3(t)}{dt} &= k_{cp}X_1(t) - k_{pc}X_3(t),\\
\frac{dX_2(t)}{dt} &= -k_a X_2(t)
\end{aligned}$$

with initial conditions

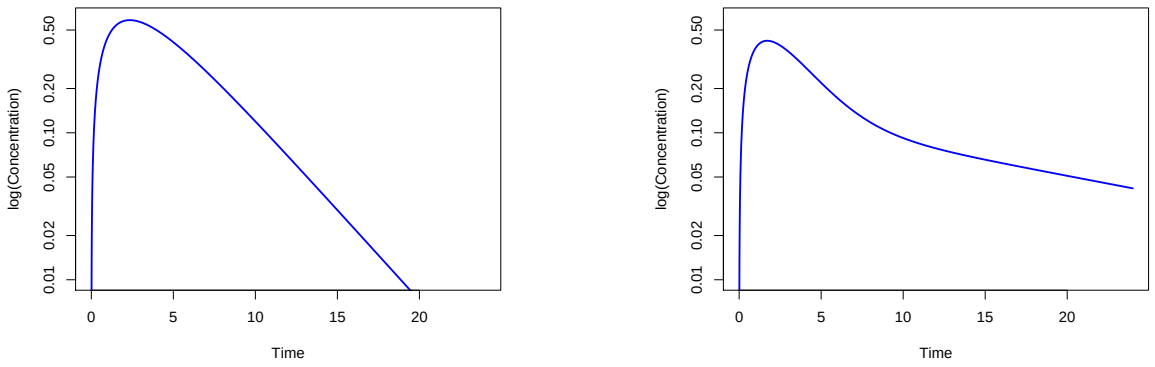
$$X_1(0) = 0, \quad X_2(0) = D, \quad \text{and} \quad X_3(0) = 0.$$

The parameters k_{cp} and k_{pc} are the rates of flow from the central to the peripheral compartment and vice versa. The solution for the concentration in the central compartment is a function of the form

$$C(t) = ae^{-\alpha t} + be^{-\beta t} - (a+b)e^{-k_a t},$$

where the constants a, b, α, and β can be expressed by k_e, k_{cp}, k_{pc}, and V_c. Note that usually only the concentrations in the central compartment are of interest, as there is often no physiological counterpart to the modeled peripheral compartment.

In the Figures 2.4(a) and 2.4(b) one can see the difference in the shape of the time courses of a one- and a two-compartment model on the log scale. In the one-compartment model two phases can be identified, an absorption phase, where the concentration increases very fast, and an elimination phase, where the concentration decreases exponentially (linear on the log scale). In the two-compartment model we can additionally see a distribution phase expressed by a strong decay of the concentration. In this phase the drug is distributed from the central compartment to the other compartment until the concentrations are balanced while already being eliminated. After a balance of the concentrations is reached, the elimination process outweighs.



(a) (b)

Figure 2.4: *Time course (on the log scale) of the concentration in (a) a one-compartment model with first-order absorption and (b) a two-compartment model with first-order absorption.*

2.1.3 Multiple-dose experiments

Often a drug is administered not only once but regularly in certain time intervals. For example, one could imagine tablets that are taken every day in the morning. When the drug has not yet been completely eliminated when the next dose is taken, the remaining amounts in the

compartments have to be taken into account as initial values in the differential equations for the modeling of the concentrations in the next dosing intervals. For the one-compartment model

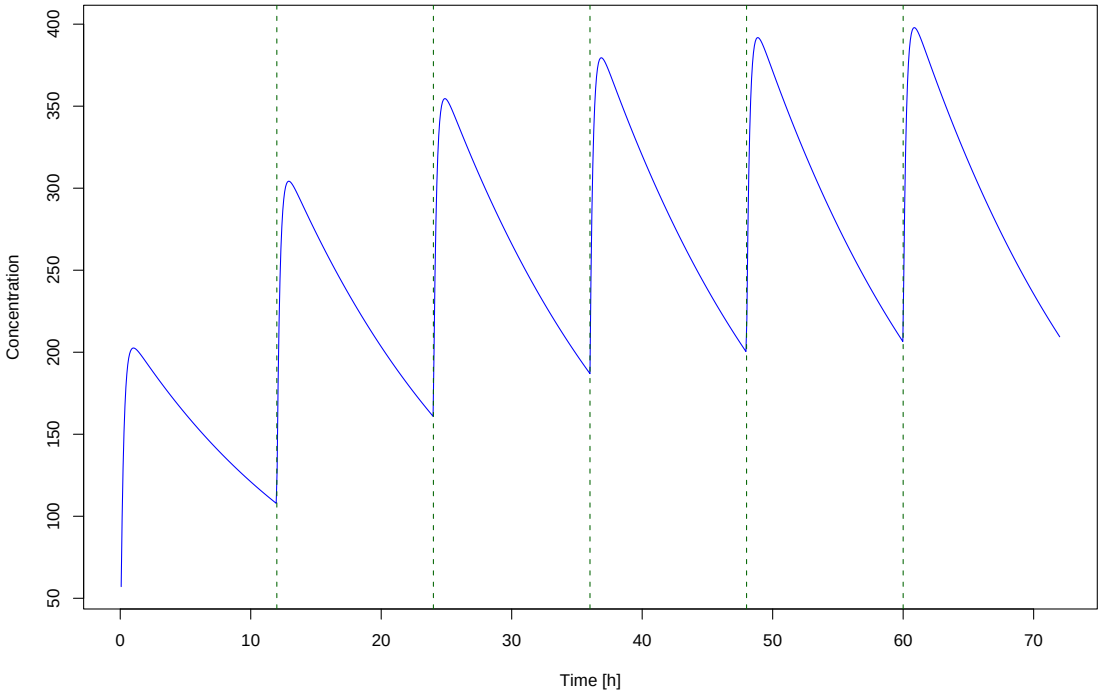


Figure 2.5: *Concentration in a one-compartment model with first-order absorption in a multiple-dose experiment.*

with zero-order absorption the concentration at time t after the n-th dose can be calculated by the following formula (see Derendorf et al. (2002), p. 55)

$$C(t;\tau,n) = \frac{FDk_a}{V_c(k_a - k_e)}\left(\frac{1-e^{-nk_e\tau}}{1-e^{-k_e\tau}}e^{-k_e t} - \frac{1-e^{-nk_a\tau}}{1-e^{-k_a\tau}}e^{-k_a t}\right),$$

where τ is the time interval between the administrations, t is the time since the last application, and n is the number of the repetition. After a certain time (approximately five times the half-life of the drug) the concentration curve reaches a steady-state where the concentration fluctuates between a constant maximal and a constant minimal level. Figure 2.5 shows an example of the time course of the concentration of a one-compartment model with first-order absorption, where the drug is taken every twelve hours. By letting n go to infinity in the previous formula we get the expression for the concentration curve in the steady-state

$$C(t;\tau) = \frac{FDk_a}{V_c(k_a - k_e)}\left(\frac{e^{-k_e t}}{1-e^{-k_e\tau}} - \frac{e^{-k_a t}}{1-e^{-k_a\tau}}\right).$$

For the two-compartment model with zero-order absorption we obtain similar formulas for the concentrations in multiple-dose applications. We get

$$C(t;\tau,n) = a\frac{1-e^{-n\alpha\tau}}{1-e^{-\alpha\tau}}e^{-\alpha t} + b\frac{1-e^{-n\beta\tau}}{1-e^{-\beta\tau}}e^{-\beta t} - (a+b)\frac{1-e^{-nk_a\tau}}{1-e^{-k_a\tau}}e^{-k_a t}$$

for the concentration at time t after the n-th drug intake and

$$C(t;\tau) = a\frac{e^{-\alpha t}}{1-e^{-\alpha\tau}} + b\frac{e^{-\beta t}}{1-e^{-\beta\tau}} - (a+b)\frac{e^{-k_a t}}{1-e^{-k_a\tau}}$$

for the steady-state concentrations (see Derendorf et al. (2002), p. 74).

2.2 Important pharmacokinetic parameters

In this section we want to list and explain several often used pharmacokinetic parameters. Some of them have already been used in the formulas in the previous section.

$t_{\max}$ The time of maximum concentration.

$c_{\max}$ The maximum of the concentration profile, i. e. $C(t_{\max})$.

AUC The integral (area) of the concentration curve. (AUC = Area under the Curve)

V_c (and) V_p The volumes of distribution. For apparent reasons the true volume of the blood (or another fluid) in the body can usually not be assessed. A theoretical volume, the so-called volume of distribution, however, can be easily calculated by dividing the amount of drug that is in the body by the concentration of the drug in the fluid

$$V_c = \frac{X}{C}.$$

Thus, the volume of distribution is a factor that relates the concentration to the total amount of drug present in the compartment. Usually the volume of distribution differs from the true volume of the compartment as often a part of the drug is bound in the tissue or the solution of the drug is not homogeneous. The volume of distribution corresponds to the volume of a homogeneous solution with the respective concentration that would be necessary to contain the total amount of the drug.

F The bioavailability describes the fraction of the administered dose that reaches the system. By definition, for a bolus injection the bioavailability is 1. For oral application the bioavailability is usually lower than 1, for example due to incomplete absorption. The absolute bioavailability of a non-intravenous administration of a drug is the AUC of the non-intravenous application divided by the AUC of the intravenous application.

Cl The clearance is defined as the amount of plasma that is cleared of the drug within unit time. The units of clearance are given in terms of $^{\text{volume}}/_{\text{time}}$. For the one-compartment

model the clearance is $k_e V_c$. For more-compartment models different types of clearance can be defined. However, we do not want to go into detail here.

With these parameters the models introduced in the previous sections can be described with different parameterizations. For example, the one-compartment model is often expressed with the parameters k_a, Cl, and V_c.

$$C(t) = \frac{FDk_a}{V_c k_a - Cl}\left(e^{-\frac{Cl}{V_c}t} - e^{-k_a t}\right). \tag{2.1}$$

Note that the model in Eq. (2.1) is overparameterized. The bioavailability parameter F is usually not estimated, if the model is fitted to data.

The two-compartment model can be parameterized by k_a, Cl, Q, V_c, and V_p. Here, Q is also a clearance parameter, for which $k_{cp} = {}^{Q}/_{V_c}$ and $k_{pc} = {}^{Q}/_{V_p}$ holds.

An extensive list with pharmacokinetic models and different parameterizations can be found in the documentation to the pharmacokinetics software package ADAPT II (D'Argenio and Schumitzky (1997)).

2.3 The population approach

The first investigations of the pharmacokinetic parameters of a drug are usually done in a non-parametric way. The drug is applied to healthy volunteers in a phase I study and blood samples are taken in a dense scheme so that parameters like C_{max} or t_{max} can be directly read out of the data or easily be calculated, like the AUC using the trapezoidal rule.

The data can then also be used to develop a model for the behavior of the drug in the body to be able to predict concentrations to be expected if different doses or multiple doses are given.

One aim of pharmacokinetic studies in later phases of the clinical development is to assess the variability of both the kinetic parameters as well as the concentrations within the target population of ill subjects the drug is designed for. This can be done with the so-called *population approach*, which is becoming more and more accepted by the authorities. A definition to this approach can be found in the FDA guidance for industry "Population Pharmacokinetics" (FDA (1999)):

> "Population pharmacokinetics is the study of the sources and correlates of variability in drug concentrations among individuals who are the target patient population receiving clinical relevant doses of a drug of interest."

To be able to assess the variability, the kinetic parameters and the time course of the concentrations has to be obtained for the whole sample population. In trials within the target population of ill patients it is usually not possible to get blood samples in a dense sampling scheme due to logistical, economical, and, most important, ethical reasons. It is hence not possible to fit independent concentration profiles for each single individual.

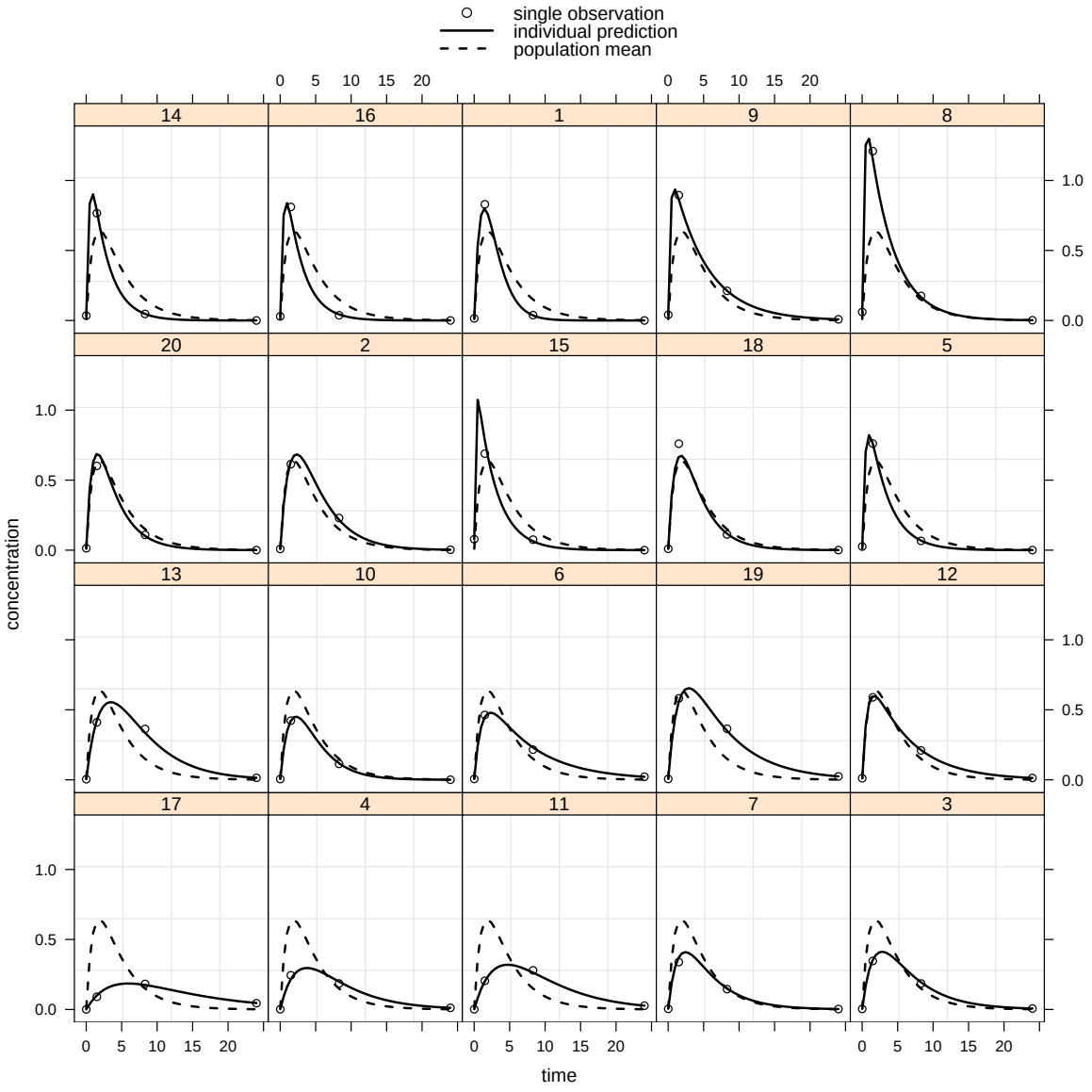


Figure 2.6: *Concentration profiles for different individuals in a population pharmacokinetic model. The dashed line denotes the estimated profile of the "mean" individual. The solid lines denote the predicted individual profiles obtained by using empirical-Bayes estimates for the individual parameters.*

The concentration profiles of the individuals look very similar with respect to the shape but can substantially differ one from another. We can, however, assume that the underlying mechanistic model that describes the absorption, distribution, and elimination of the drug is the same for all individuals. This justifies to assume that the same regression function(s) for all individuals can be used, but with different parameters for each individual. The standard approach in population pharmacokinetics is to model the variability of the profiles between the individuals by the use of random parameters. Thus, each individual has its own vector of individual pharmacokinetic parameters, which is the realization of a random vector. The mean parameter vector results in the profile of a typical individual. From a statistical point of view this approach is based on linear

or non-linear mixed models that we will describe in the Chapters 4 and 5. Figure 2.6 shows a typical scenario of a population pharmacokinetic study. Each of the panels in the plot describes the data for one individual. The circles denote the actual measurements, the dashed line shows the population mean curve, that is, the curve of a typical individual of the population. The solid line describes the prediction for the individual concentration profile, which are obtained by using the empirical-Bayes estimates for the individual parameters.

One of the pioneers in using and promoting this kind of modeling is L. B. Sheiner (see e. g. Sheiner and Beal (1980), Sheiner and Beal (1981), Sheiner and Wakefield (1999)), who is also one of the authors of the most often used software package for the analysis of population pharmacokinetic data NONMEM (see Beal and Sheiner (1989)).

The acceptance of this approach also by the regulatory side can be seen in two recent guidelines of the European authority EMEA. The draft "Guideline on reporting the results of population pharmacokinetic analysis" (EMEA-CHMP (2006a)) gives specific guidance what elements of the population pharmacokinetic analysis are considered important by the authorities and how they should be published. In the "Guideline on the role of pharmacokinetics in the development of medicinal products in the paediatric population" (EMEA-CHMP (2006b)) the use of the population approach is explicitly encouraged to obtain pharmacokinetic data for children of substances that have already been approved for adults.

> "Population pharmacokinetic analysis, using non-linear mixed effects models, is an appropriate methodology for obtaining pharmacokinetic information in paediatric trials both from a practical and ethical point of view. Mean and variances are estimated and information from all individuals is merged making it possible to use sparse sampling schemes." (EMEA-CHMP (2006b))

An example of a population pharmacokinetic study designed for a paediatric population is given in Mentré et al. (2001).

Although sparse sampling schemes are used in population pharmacokinetic studies it is still important to keep the number of blood samples and the number of patients as low as possible as each blood sample not only causes costs but also pain to the patient. It is therefore crucial to carefully plan the experiment and to select optimal sampling schemes that allow good parameter estimates, as it would be even more unethical if a study ends up with unreliable results. This careful planning is also recommended by the authorities:

> "The population approach may replace conventionally designed pharmacokinetic studies with rich sampling. Simulations or theoretical optimal design approaches, based on prior knowledge, should be considered as tools for the selection of sampling times and the number of subjects." (EMEA-CHMP (2006b))

The amount of literature on optimal design approaches for population pharmacokinetic studies has steadily increased since the 1990s. Starting with simulation based approaches as in Al-Banna et al. (1990) or Wang and Endrenyi (1992), it is up to now usually based on an approximation of the Fisher information matrix as given by Mentré et al. (1997) or Retout et al. (2001). Examples for studies with optimized designs are given in Green and Duffull (2003), Retout et al. (2002) besides others.

3 Experimental Design in the Ordinary Linear Model

This chapter is thought as a short introduction to the topic of optimal design of experiments. Important terms and definitions from the well-established theory of experimental design for ordinary linear models are introduced. Later we will use them as a basis to derive results for the mixed model case. The foundation for this chapter are the books by Silvey (1980), Atkinson and Donev (1996), and Fedorov and Hackl (1997), as well as the introductory chapter in Schwabe (1996). A deeper mathematical background to the theory of optimal design can be found in the books by Pázman (1986) and Pukelsheim (1993).

In Section 3.1 the considered ordinary linear model is described. Section 3.2 introduces all the important terms that are used in connection with designed (planned) experiments. In the remaining sections the concept of optimality criteria is introduced, which are used to evaluate the quality of the experimental designs, and some fundamental background on convex design theory is given.

3.1 Model

Throughout this chapter we consider a simple ordinary linear model of the form

$$Y_i = \mathbf{f}(x_i)^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, m, \tag{3.1}$$

where Y_i denotes the ith observation conducted under the experimental settings x_i and $\mathbf{f}(x_i) = (f_1(x_i), \ldots, f_p(x_i))^\top$ is a vector of known real-valued regression functions evaluated at x_i. For technical reasons we assume that the experimental settings x_i are elements of a compact set $\mathcal{X}$ called *design region* and that the regression functions are continuous on $\mathcal{X}$. This implies that the so-called *design locus* $\mathbf{f}(\mathcal{X}) = \{y | y = \mathbf{f}(x),\ x \in \mathcal{X}\}$ is also a compact set, which will later assure the existence of optimal designs. The p-dimensional parameter vector $\boldsymbol{\beta}$ is unknown and is of primary interest to be estimated. The observational errors ε_i, $i = 1, \ldots, m$ have zero mean, a constant variance σ^2 and are assumed to be uncorrelated. It is often useful to summarize all observations to $\mathbf{Y} = (Y_1, \ldots, Y_m)^\top$ and to use matrix/vector notation to express the model for all observations as

$$\mathbf{Y} = \mathbf{F}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{F} = (\mathbf{f}(x_1), \ldots, \mathbf{f}(x_m))^\top$ is called *design matrix* and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_m)^\top$ is the vector of observational errors. With the assumptions on the variances and correlations from above we get that $\mathrm{E}(\mathbf{Y}) = \mathbf{F}\boldsymbol{\beta}$ and $\mathrm{Cov}(\mathbf{Y}) = \mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_m$.

If $(\mathbf{F}^\top\mathbf{F})$ is regular then according to the well-known Gauss-Markov theorem (see e. g. Christensen (1987), p. 29) the best linear unbiased estimator (BLUE) for the parameter vector $\boldsymbol{\beta}$ is the least-squares estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top\mathbf{Y}.$$

The covariance matrix of the estimator $\hat{\boldsymbol{\beta}}$ is

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{F}^\top\mathbf{F})^{-1}.$$

In the case that $(\mathbf{F}^\top\mathbf{F})$ is singular, no linear unbiased estimator for $\boldsymbol{\beta}$ exists. However, frequently the interest lies only on a linear aspect φ of the parameter vector $\boldsymbol{\beta}$, i. e. on a function φ of the form $\varphi(\boldsymbol{\beta}) = \mathbf{L}_\varphi\boldsymbol{\beta}$. For the estimation of $\varphi(\boldsymbol{\beta})$ the assumption of the regularity of $\mathbf{F}^\top\mathbf{F}$ can be relaxed to just the requirement that φ has to be identifiable, that is, we only need that $\mathbf{L}_\varphi = \mathbf{QF}$ for some matrix $\mathbf{Q}$. In this case the best linear unbiased estimator for $\varphi(\boldsymbol{\beta})$ is given by

$$\hat{\varphi} = \mathbf{L}_\varphi(\mathbf{F}^\top\mathbf{F})^-\mathbf{F}^\top\mathbf{Y},$$

where $(\mathbf{F}^\top\mathbf{F})^-$ is an arbitrary generalized inverse of $\mathbf{F}^\top\mathbf{F}$. The covariance matrix of $\hat{\varphi}$ is then

$$\mathrm{Cov}(\hat{\varphi}) = \sigma^2\mathbf{L}_\varphi(\mathbf{F}^\top\mathbf{F})^-\mathbf{L}_\varphi^\top.$$

3.2 Designed experiments

As we just saw in the previous section the covariance matrices of $\hat{\boldsymbol{\beta}}$, and $\hat{\varphi}$ resp., depend on the design matrices $\mathbf{F}$ and with that on the experimental settings x_i. In many cases the settings $x_i \in \mathcal{X}$ at which the observations are to be taken can be chosen by the investigator. This makes it reasonable to try to "design" the experiment in a way that the estimates can be obtained with the highest precision possible. For this we first formally define the term *design*.

Definition 3.1 *An (exact) design d of size m is a vector $(x_1, \ldots, x_m)$ of possible experimental settings $x_i \in \mathcal{X}$, $i = 1, \ldots, m$.*

For a given design $d = (x_1, \ldots, x_m)$ the corresponding linear model has the form

$$\mathbf{Y}_{(d)} = \mathbf{F}_{(d)}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{F}_{(d)} = (\mathbf{f}(x_1), \ldots, \mathbf{f}(x_m))^\top$ is the corresponding design matrix.

If the experiment is conducted using design d, then $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{F}_{(d)}^\top\mathbf{F}_{(d)})^{-1}$ and $\mathrm{Cov}(\hat{\varphi}) = \sigma^2\mathbf{L}_\varphi(\mathbf{F}_{(d)}^\top\mathbf{F}_{(d)})^-\mathbf{L}_\varphi^\top$ respectively. As we can see, all the information provided by the design d is contained in $\mathbf{F}_{(d)}^\top\mathbf{F}_{(d)}$ justifying the term *information matrix*:

Definition 3.2 *Consider an ordinary linear model as defined in Section 3.1 and a design d of size m as defined in Definition 3.1. Then*

$$\mathbf{M}(d) = \frac{1}{m}\mathbf{F}_{(d)}^\top\mathbf{F}_{(d)}$$

is called the (normalized) information matrix of the design d.

If replications of the observations at the chosen experimental settings occur, a design d of size m can alternatively be expressed by its k different settings $x_1, \ldots, x_k$ and the corresponding numbers of replications $m_1, \ldots, m_k$ with $\sum m_j = m$, i. e.

$$d \sim \begin{pmatrix} x_1 & \ldots & x_k \\ m_1 & \ldots & m_k \end{pmatrix}.$$

The information Matrix $\mathbf{M}(d)$ can then be calculated by

$$\mathbf{M}(d) = \sum_{j=1}^{k} w_j \mathbf{f}(x_j)\mathbf{f}(x_j)^\top$$

with $w_j = \frac{m_j}{m}$, which is the same as

$$\mathbf{M}(d) = \mathbf{F}_{(d)}^\top \mathbf{W} \mathbf{F}_{(d)},$$

where $\mathbf{W} = \begin{pmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_k \end{pmatrix}$ and $\mathbf{F}_{(d)} = (\mathbf{f}(x_1), \ldots, \mathbf{f}(x_k))^\top$. Note, that although we used the same symbol $\mathbf{F}_{(d)}$ as before, the design matrix now contains each of the different settings x_i only once, i. e. $\mathbf{F}_{(d)} \in \mathbb{R}^{k \times p}$.

Note, that a linear aspect $\varphi(\boldsymbol{\beta}) = \mathbf{L}_\varphi \boldsymbol{\beta}$ is identifiable using the design d if there exists a matrix $\tilde{\mathbf{Q}}$ for which $\mathbf{L}_\varphi = \tilde{\mathbf{Q}}\mathbf{M}(d)$.

Due to the normalization the information matrix $\mathbf{M}(d)$ is independent of the actual number of observations m but only depends on the proportions w_i of observation to be taken at the different x_i.

Every design d can be associated with a *design measure* $\xi_{(d)}$ defined by

$$\xi_{(d)} = \sum_{j=1}^{k} w_j \delta_{\{x_j\}},$$

where $\delta_{\{x\}}$ denotes the one-point measure in x. The information matrix of d can then be calculated as an integral with respect to $\xi_{(d)}$:

$$\mathbf{M}(d) = \int \mathbf{f}(x)\mathbf{f}(x)^\top \xi_{(d)}(dx).$$

The natural aim is to find designs that allow the estimation of the parameter vector (or an aspect of the parameters) with the highest possible precision. As the set of exact designs of size m is a discrete and, hence, non-convex set, it is too sparse for the usual optimization processes (Kiefer and Wolfowitz (1959)). Therefore, the assumption that mw_j is an integer is often dropped and arbitrary weights $w_j \geq 0$, $j = 1, \ldots, k$ with $\sum w_j = 1$ are allowed, which leads to the class of *approximate designs*.

Definition 3.3 *ξ is called approximate design on $\mathcal{X}$ if $\xi = \sum_{j=1}^{k} w_j \delta_{\{x_j\}}$ for some $x_j \in \mathcal{X}$ and weights $w_j \geq 0$, $j = 1, \ldots, k$, $\sum w_j = 1$ and $k \in \mathbb{N}$.*

This set of approximate designs is convex and, hence, easier to handle. If the closure of this set is taken, the set of all probability measures on $\mathcal{X}$ is reached. This leads to a very general design definition.

Definition 3.4 *ξ is a design (design measure) on $\mathcal{X}$ if ξ is a probability measure on $\mathcal{X}$. The set of all designs on $\mathcal{X}$ will be denoted by Ξ. The set of designs under which a certain linear aspect φ is identifiable will be denoted by Ξ_φ. The set of designs under which the whole vector $\boldsymbol{\beta}$ is estimable will be denoted by Ξ_β. Designs, for which not the whole vector $\boldsymbol{\beta}$ is identifiable will be called singular or non-regular as then the corresponding information matrix is singular.*

The set of all possible information matrices

$$\mathcal{M} := \{\mathbf{M}(\xi); \xi \in \Xi\}$$

is convex as $\alpha\mathbf{M}(\xi_1) + (1-\alpha)\mathbf{M}(\xi_2) = \mathbf{M}(\alpha\xi_1 + (1-\alpha)\xi_2) \in \mathcal{M}$ holds due to the linearity of the integral. Moreover, $\mathcal{M}$ is the convex hull of the set of information matrices belonging to the one-point design measures.

From the point of view of information matrices it can be shown with Caratheodory's theorem (see Silvey (1980), p. 72) that it is sufficient to investigate only designs (design measures) with a finite number of support points, that is, approximate designs in the sense of Definition 3.3.

Theorem 3.5
For every design $\xi \in \Xi$ there is an approximate design $\tilde{\xi} = \sum_{j=1}^{k} w_j \delta_{\{x_j\}}$ with $k \leq \frac{1}{2}p(p+1)+1$ which satisfies $\mathbf{M}(\xi) = \mathbf{M}(\tilde{\xi})$. If $\mathbf{M}(\xi)$ is on the boundary of $\mathcal{M}$ there is a design $\tilde{\xi}$ with at most $\frac{1}{2}p(p+1)$ support points which satisfies $\mathbf{M}(\tilde{\xi}) = \mathbf{M}(\xi)$.

The proof is based on the fact, that $\mathcal{M}$ can be identified with a convex subset of $\mathbb{R}^{\frac{1}{2}p(p+1)}$ as all elements of $\mathcal{M}$ are symmetric $p \times p$-matrices. Furthermore, $\mathcal{M}$ is the convex hull of the information matrices of the one-point design measures. Caratheodory's theorem states that every element of the convex hull of a subset S of an n-dimensional Euclidean space can be expressed as a convex combination of at most $n+1$ elements of S and by at most n elements if it lies on the boundary of the convex hull.

We can therefore restrict our attention to designs with finite support. From now on for simplicity we will make no difference between the terms design and design measure.

3.3 Optimality criteria

In this section we consider the question on how to evaluate and compare the quality of designs.

In the simple case where $\boldsymbol{\beta}$ is only one-dimensional this question can easily be answered. It is reasonable to consider a design ξ^* to be optimal for the estimation of $\boldsymbol{\beta}$ if $\boldsymbol{\beta}$ is identifiable and if the variance $\sigma^2\mathbf{M}(\xi)^{-1}$ is minimized by ξ^*.

In the case where $\boldsymbol{\beta}$ has dimension greater than one, the situation is not so clear. A natural extension would be to call a design optimal if $\boldsymbol{\beta}$ is identifiable and all one-dimensional linear aspects can be estimated with the smallest variance, i. e. ξ^* would have to minimize $\mathbf{c}^\top \mathbf{M}(\xi)^{-1}\mathbf{c}$ for all p-dimensional vectors $\mathbf{c}$. This is equivalent to the condition that the optimal ξ^* would have to minimize $\mathbf{M}(\xi)^{-1}$ with respect to the Loewner partial ordering of non-negative definite matrices. Such a *uniformly optimal design*, however, can only be found in vary rare situations.

The way out of this situation is to evaluate the designs with respect to real-valued functionals Φ of the information matrix $\mathbf{M}(\xi)$ or its inverse $\mathbf{M}(\xi)^{-1}$. Φ should be chosen in a way that it fits reasonably well to the needs of the experimenter. Designs that minimize $\Phi(\mathbf{M}(\xi))$ will then be called Φ-optimal. Many different criteria have been proposed in the literature. In the following we will introduce only some of them.

We start with the popular D-criterion, where the determinant of the information matrix has to be maximized. This choice can be motivated by the fact, that under normality assumption (or asymptotically) a D-optimal design minimizes the volume of a confidence ellipsoid for the parameter vector.

Definition 3.6 *A design ξ^* is called D-optimal if $\mathbf{M}(\xi^*)$ minimizes*

$$\Phi_D(\mathbf{M}) = -\log\det(\mathbf{M})$$

on $\mathcal{M}$ or equivalently

$$\det(\mathbf{M}(\xi^*)) \geq \det(\mathbf{M}(\xi))$$

for all $\xi \in \Xi$.

One feature of the D-criterion is that the optimal designs are invariant with respect to a reparameterization of the form $\tilde{\boldsymbol{\beta}} = \mathbf{A}\boldsymbol{\beta}$, as for a regular matrix $\mathbf{A}$ the maximization of $\det(\mathbf{A}^\top \mathbf{M}(\xi)\mathbf{A})$ $= \det(\mathbf{A}^\top)\det(\mathbf{M}(\xi))\det(\mathbf{A})$ is independent of $\mathbf{A}$.

Tightly connected to the D-criterion – as we will see in the next section – is the G-criterion. Here, one is interested in minimizing the maximal variance of the point-wise prediction of the response within the design region $\mathcal{X}$.

Definition 3.7 *A design ξ^* is called G-optimal if*

$$\Phi_G(\mathbf{M}) = \max_{x\in\mathcal{X}} \mathbf{f}(x)^\top \mathbf{M}^{-1}\mathbf{f}(x)$$

is minimized by $\mathbf{M}(\xi^*)$, that is,

$$\max_{x\in\mathcal{X}} \mathbf{f}(x)^\top \mathbf{M}(\xi^*)^{-1}\mathbf{f}(x) \leq \max_{x\in\mathcal{X}} \mathbf{f}(x)^\top \mathbf{M}(\xi)^{-1}\mathbf{f}(x)$$

for all $\xi \in \Xi_\beta$.

An interesting class of optimality criteria are the *linear criteria*, where functionals that are linear in the inverse of the information matrix $\mathbf{M}(\xi)$ are minimized. Examples are the A-, the c-, or the IMSE-criterion.

With the A-criterion designs are chosen that minimize the expected mean squared deviation of the parameter estimates, which is given by the trace of the inverse of the information matrix:

Definition 3.8 *A design ξ^* is called A-optimal if*

$$\operatorname{tr}(\mathbf{M}(\xi^*)^{-1}) \leq \operatorname{tr}(\mathbf{M}(\xi)^{-1})$$

for all $\xi \in \Xi_\beta$, that is, if $\Phi_A(\mathbf{M}) = \operatorname{tr}(\mathbf{M}^{-1})$ is minimized by the information matrix of ξ^*.

The c-criterion is used if one wants to estimate a one-dimensional linear function $\varphi(\boldsymbol{\beta}) = \mathbf{c}^\top \boldsymbol{\beta}$ as good as possible.

Definition 3.9 *A design $\xi^* \in \Xi_\mathbf{c}$ is c-optimal if*

$$\mathbf{c}^\top \mathbf{M}(\xi^*)^- \mathbf{c} \leq \mathbf{c}^\top \mathbf{M}(\xi)^- \mathbf{c}$$

for all $\xi \in \Xi_\mathbf{c}$, that is, if $\Phi_c(\mathbf{M}) = \mathbf{c}^\top \mathbf{M}^- \mathbf{c}$ is minimized by ξ^*.

The c-criterion can also be used if the interest lies on the estimation of a non-linear function of $\boldsymbol{\beta}$. We will discuss this in Section 6.3.1.

Similar to the G-criterion the IMSE-criterion is also based on the variance function for the point-wise predictions on the design region. Now, however, we are interested in minimizing the expected integrated mean squared error of the prediction over the design region $\mathcal{X}$

$$\mathrm{E}\left(\int_\mathcal{X} (\mathbf{f}(x)^\top \boldsymbol{\beta} - \mathbf{f}(x)^\top \hat{\boldsymbol{\beta}})^2 \, \mu(dx) \right) \to \min,$$

where μ is a measure on the design region $\mathcal{X}$. This is equivalent to finding a design ξ^* such that $\mathbf{M}(\xi^*)$ minimizes

$$\Phi(\mathbf{M}) = \int_\mathcal{X} \mathbf{f}(x)^\top \mathbf{M}^{-1} \mathbf{f}(x) \, \mu(dx) = \operatorname{tr}\left(\left(\int_\mathcal{X} \mathbf{f}(x) \mathbf{f}(x)^\top \, \mu(dx) \right) \mathbf{M}^{-1} \right),$$

which is again a criterion that is linear in $\mathbf{M}^{-1}$. Note that all introduced linear criteria can be brought to the general form $\Phi(\mathbf{M}) = \operatorname{tr}(\mathbf{L}\mathbf{M}^{-1})$ with an appropriately chosen matrix $\mathbf{L}$.

3.4 Convex design theory

All the criteria described in Section 3.3 share the following two properties which allow the derivation of theoretical results and the construction of optimization algorithms:

Monotonicity: $\mathbf{M}_1 \geq \mathbf{M}_2 \implies \Phi(\mathbf{M}_1) \leq \Phi(\mathbf{M}_2)$,
that is, a design that is uniformly better than another one will also be judged as a better design by the optimality criterion, where '$\leq$' denotes the Loewner partial ordering of non-negative definite matrices.

Convexity: $\Phi(\alpha\mathbf{M}_1 + (1-\alpha)\mathbf{M}_2) \leq \alpha\Phi(\mathbf{M}_1) + (1-\alpha)\Phi(\mathbf{M}_2)$ for $\alpha \in [0,1]$
This condition assures that no local minima can occur. If Φ is strictly convex the optimal information matrix is unique. This, however, does not imply the uniqueness of the optimal design as two designs can share the same information matrix.

Note, however, that we do not give a strict formal definition of the term *optimality criterion* as in Gaffke and Heiligers (1996) but will use the term for any real-valued function on the set of non-negative definite matrices.

The requirements that the design region $\mathcal{X}$ should be compact and that the vector of regression functions $\mathbf{f}$ has to be continuous on $\mathcal{X}$ ensures that the set $\mathcal{M}$ of possible information matrices is a compact convex set (see Pukelsheim (1993), p. 29). This ensures the existence of a Φ-optimal design for the considered criteria.

If Φ is even strictly decreasing, the optimal information matrix has to be on the boundary of $\mathcal{M}$ and hence an optimal design with at most $\frac{1}{2}p(p+1)$ support points can be found (see Theorem 3.5).

To give some further results that can be found in Silvey (1980), we define the following directional derivative (Silvey (1980), p. 18):

Definition 3.10 *The Fréchet derivative of Φ at $\mathbf{M}_1$ in direction of $\mathbf{M}_2$ is given by*

$$F_\Phi(\mathbf{M}_1, \mathbf{M}_2) = \lim_{\alpha \downarrow 0} \frac{1}{\alpha} \left[\Phi((1-\alpha)\mathbf{M}_1 + \alpha\mathbf{M}_2) - \Phi(\mathbf{M}_1)\right].$$

The Fréchet derivative can also be defined by

$$F_\Phi(\mathbf{M}_1, \mathbf{M}_2) = \frac{d}{d\alpha}\Phi((1-\alpha)\mathbf{M}_1 + \alpha\mathbf{M}_2)\Big|_{\alpha=0^+}.$$

If Φ is differentiable in $\mathbf{M}_1$ in the usual sense (see Silvey (1980), Appendix 3, or Rockafellar (1972), p. 241), then $F_\Phi(\mathbf{M}_1, \sum a_i\mathbf{M}_{2i}) = \sum a_i F_\Phi(\mathbf{M}_1, \mathbf{M}_{2i})$ holds for $\alpha_i \in \mathbb{R}$ with $\sum_i \alpha_i = 1$.

We will now quote several theorems that can be found in Silvey (1980), pp. 19, in a similar form.

Theorem 3.11 (cf. Silvey (1980), Theorem 3.6)
Let Φ be convex on $\mathcal{M}$, then ξ^* Φ-optimal if and only if

$$F_\Phi(\mathbf{M}(\xi^*), \mathbf{M}(\xi)) \geq 0 \quad \textit{for all } \xi \in \Xi.$$

Figuratively, this means that a design is optimal if it does not improve if it is slightly changed in the direction of any other design. Under the condition of differentiability it is sufficient to check if the design is improved if it is changed in the direction of any one-point design measure.

Theorem 3.12 (cf. Silvey (1980), Theorem 3.7)
Let Φ be convex on $\mathcal{M}$ and differentiable at $\mathbf{M}(\xi^*)$ then ξ^* is Φ-optimal if and only if

$$F_\Phi\left(\mathbf{M}(\xi^*), \mathbf{M}(\delta_{\{x\}})\right) = F_\Phi\left(\mathbf{M}(\xi^*), \mathbf{f}(x)\mathbf{f}(x)^\top\right) \geq 0 \quad \textit{for all } x \in \mathcal{X}.$$

Theorem 3.13 (cf. Silvey (1980), Theorem 3.9)
Let Φ be convex on $\mathcal{M}$, differentiable at all points of $\mathcal{M}^+ = \{\mathbf{M} \in \mathcal{M} : \Phi(\mathbf{M}) < \infty\}$, and let a Φ-optimal design exist, then ξ^* is Φ-optimal if and only if

$$\min_{x \in \mathcal{X}} F_\Phi\left(\mathbf{M}(\xi^*), \mathbf{f}(x)\mathbf{f}(x)^\top\right) = \max_{\xi} \min_{x \in \mathcal{X}} F_\Phi\left(\mathbf{M}(\xi), \mathbf{f}(x)\mathbf{f}(x)^\top\right).$$

The famous theorem of the equivalence of D- and G-optimality (Kiefer and Wolfowitz (1960)) is a special case of the Theorems 3.12 and 3.13.

Theorem 3.14 (cf. Kiefer and Wolfowitz (1960))
The following three statements are equivalent

1. ξ^* *minimizes* $-\log(\det(\mathbf{M}(\xi)))$,
2. ξ^* *minimizes* $\max_{x \in \mathcal{X}} \mathbf{f}(x)^\top \mathbf{M}(\xi)^{-1} \mathbf{f}(x)$,
3. $\max_{x \in \mathcal{X}} \mathbf{f}(x)^\top \mathbf{M}(\xi^*)^{-1} \mathbf{f}(x) = p$.

This can be seen from the fact that for $\Phi(\mathbf{M}) = -\log(\det(\mathbf{M}))$

$$F_\Phi(\mathbf{M}_1, \mathbf{M}_2) = \operatorname{tr}(\mathbf{M}_1^{-1}(\mathbf{M}_1 - \mathbf{M}_2)) = \operatorname{tr}(\mathbf{I}_p) - \operatorname{tr}(\mathbf{M}_1^{-1}\mathbf{M}_2) = p - \operatorname{tr}(\mathbf{M}_1^{-1}\mathbf{M}_2).$$

Similar results can be obtained for other optimality criteria.

For linear criteria we get

Corollary 3.15 *Let $\xi^* \in \Xi_\beta$ and $\Phi(\mathbf{M}) = \operatorname{tr}(\mathbf{A}\mathbf{M}^{-1})$. Then ξ^* is Φ-optimal in Ξ_β if and only if*

$$\mathbf{f}(x)^\top \mathbf{M}(\xi^*)^{-1} \mathbf{A} \mathbf{M}(\xi^*)^{-1} \mathbf{f}(x) \leq \operatorname{tr}(\mathbf{A}\mathbf{M}(\xi^*)^{-1}) \quad \text{for all } x \in \mathcal{X}. \tag{3.2}$$

Note, that this corollary only holds for designs with regular information matrices. Considerations for optimal singular designs can be found, for example, in Silvey (1980), pp. 25.

These equivalence theorems can be used to check the optimality of a design at least graphically. For given ξ^* the left-hand-side of Theorem 3.14 *3.)* or Eq. (3.2) are functions in x which can be plotted to check if the maximum lies below the right-hand-side. In Figure 3.1 this is done for the D-criterion for a simple quadratic regression model, where the number of parameters $p = 3$.

The Fréchet derivatives also help us to construct optimal designs iteratively. The basic idea is as follows (see e. g. Silvey (1980), chapter 4). Suppose that we have a design ξ_n with corresponding information matrix $\mathbf{M}(\xi_n)$ and that Φ is differentiable in $\mathbf{M}(\xi_n)$. If ξ_n is not Φ-optimal, then according to Theorem 3.12 there exists a point $x_{n+1} \in \mathcal{X}$ such that $F_\Phi(\mathbf{M}(\xi_n), \mathbf{f}(x_{n+1})\mathbf{f}(x_{n+1})^\top) < 0$. We can hence get a better design by putting more weight on x_{n+1} then in ξ_n, that is, by using

$$\xi_{n+1} = (1 - \alpha_{n+1})\xi_n + \alpha_{n+1}\delta_{\{x_{n+1}\}}$$

with $\alpha_{n+1} \in [0, 1]$ chosen adequately.

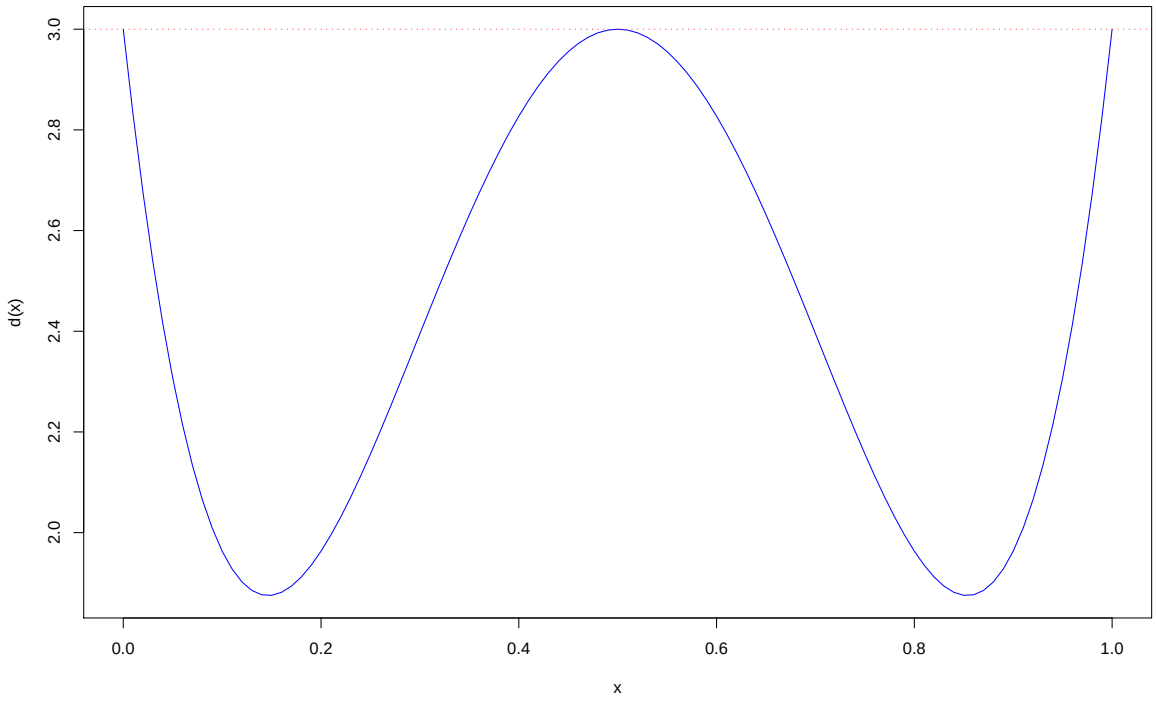


Figure 3.1: *Variance function for the D-optimal design in the quadratic regression model*

There are different methods for choosing α_{n+1}. One is to chose α_{n+1} such that Φ decreases maximally, that is, such that $F_\Phi(\mathbf{M}(\xi_{n+1}), \mathbf{f}(x_{n+1})\mathbf{f}(x_{n+1})^\top) = 0$. This is discussed in Fedorov (1972) for various Φ and referred to as V-algorithm (V for V. V. Fedorov). Another choice is to fix a sequence α_n with $\alpha_n \to 0$ and $\sum \alpha_n \to \infty$. This is referred to as W-algorithm (W for Wynn) (see also e. g. Fedorov (1972)). Fedorov (1972) also discusses the convergence of the algorithms for specific criteria.

4 Linear Mixed Models

Although the models appearing in the field of pharmacokinetics are usually of a nonlinear kind, in this chapter we first investigate linear mixed models. As we will see in Chapter 5, where the nonlinear mixed models are introduced, linear mixed models will also play an important role in the nonlinear situation. Especially for design problems usually a Taylor approximation of the nonlinear model function is used resulting in a linear mixed model, that is then used to derive results.

The general linear mixed model we use will have more or less the form given by Laird and Ware (1982). These models can be embedded in the more general class of GMANOVA models introduced by Potthoff and Roy (1964).

There is a vast amount of literature on the analysis of linear mixed models. In this chapter we will therefore only give a short summary of the model and its analysis. Most of the topics, especially the estimation methods, are covered in the books of Verbeke and Molenberghs (2000), Vonesh and Chinchilli (1997) or Davidian and Giltinan (1995) in much more detail.

In the first section we introduce the random coefficient regression model (RCR model). This is a special type of mixed model and it will be the model we mainly deal with in this text. This model is extended to a more general form as given by Laird and Ware (1982) in Section 4.2.

In Section 4.3 we give a short review of the estimation techniques for mixed models.

The most important instrument for design topics, the Fisher information matrix, is introduced in Section 4.4.

4.1 The random coefficient regression model

In this section we introduce one of the most often used type of mixed models, the random coefficient regression model (RCR model). As all mixed models it is a so-called two stage model (or hierarchical model) as it can be built up in two steps. In the first stage the variability occurring within the observations of one individual is modeled, referred to as *intra-individual* variability. Then, in a second step, the variability of the behavior of different individuals is assessed, referred to as *inter-individual* variability.

First stage (intra-individual model):

The jth observation of individual i is modeled by an ordinary linear model of the form

$$Y_{ij} = \mathbf{f}(x_{ij})^\top \boldsymbol{\beta}_i + \varepsilon_{ij}, \quad j = 1, \ldots, m_i, \tag{4.1}$$

where $\mathbf{f}(x_{ij})$ is a vector of known regression functions evaluated at the experimental setting x_{ij}, $\boldsymbol{\beta}_i$ is an unknown parameter vector and ε_{ij} is a random error term. The number of observations m_i can be different for the n individuals. In our context Y_{ij} is usually the concentration of a drug measured in individual i at time x_{ij} after the administration. The experimental settings do not have to be restricted to be one dimensional, like for example, blood sampling times. They can be elements of a very general *design region* $\mathcal{X}$ like, for example, the Cartesian product of possible sampling times and possible doses. As in the previous chapter we assume that $\mathbf{f}$ is continuous and that the design region $\mathcal{X}$ is compact.

The p-dimensional vector $\boldsymbol{\beta}_i$ contains the parameters related to the individual. Throughout this text the observational errors ε_{ij} are assumed to be independent identically normally distributed with zero mean and variance σ^2, unless otherwise stated. Some of the results however are not restricted to the normality of the errors.

The m_i observations taken from individual i can be summarized using vector/matrix notation by

$$\mathbf{Y}_i = \mathbf{F}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i.$$

Here, $\mathbf{F}_i := (\mathbf{f}(x_{i1}), \ldots, \mathbf{f}(x_{im_i}))^\top$ is the design matrix associated with the experimental settings of individual i and $\boldsymbol{\varepsilon}_i := (\varepsilon_{i1}, \ldots \varepsilon_{im_i})^\top$ is the vector of the corresponding observational errors. The vectors of observational errors for the different individuals are independently normally distributed with covariance matrix $\sigma^2\mathbf{I}_{m_i}$. Of course more complex forms for this covariance matrix could be used accounting for heteroscedasticity or dependence of the observational errors within one individual, however this will not be treated in this text.

Second stage (inter-individual model):

Explaining the name "random coefficient regression model", the vector of individual parameters (coefficients) $\boldsymbol{\beta}_i$ is random and modeled as the realization of a random vector described by a simple multivariate regression model of the form:

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{b}_i, \quad i = 1, \ldots, n \tag{4.2}$$

with $\mathbf{b}_i \sim i.i.d.\ \mathcal{N}_p(0, \sigma^2\mathbf{D})$. The rationale behind this is the assumption that the observed individuals are randomly chosen from the whole population. Each individual has then its own vector of parameters $\boldsymbol{\beta}_i$ which is a realization of of random vector. The vector $\boldsymbol{\beta}$ is called vector of *population parameters* as it describes the values of a typical individual from the population and the vector $\mathbf{b}_i$ is the vector of individual *random effects* and describes the deviation from the population parameters. Again, for some of the results the assumption of normality of the $\mathbf{b}_i$ is not necessary. Of course the model allows that some of the parameters can be fixed for the whole population if the corresponding rows and columns in the covariance matrix $\mathbf{D}$ are set to zero.

It is assumed that the observational errors and the random effects are independent.

We will now have a look at the correlation structure of the observations. Observations from the same individual are correlated with covariance structure $\mathrm{Cov}(Y_{ij}, Y_{ij'}) = \sigma^2\mathbf{f}(x_{ij})^\top\mathbf{D}\mathbf{f}(x_{ij'})$,

$j \neq j'$, while observations coming from different individuals are uncorrelated. The variance of a single observation Y_{ij} is $\mathrm{Var}(Y_{ij}) = \sigma^2[1 + \mathbf{f}(x_{ij})^\top \mathbf{D}\mathbf{f}(x_{ij})]$.

Then $\mathrm{E}(\mathbf{Y}_i) = \mathbf{F}_i\boldsymbol{\beta}$ and $\mathrm{Cov}(\mathbf{Y}_i) = \sigma^2\mathbf{V}_i$, with $\mathbf{V}_i = \mathbf{I}_{m_i} + \mathbf{F}_i\mathbf{D}\mathbf{F}_i^\top$.

We now summarize the observations of the whole sample of n individuals to $\mathbf{Y} = (\mathbf{Y}_1^\top, \ldots \mathbf{Y}_n^\top)^\top$. Then

$$\mathbf{Y} = \mathbf{F}\boldsymbol{\beta} + \mathbf{G}\mathbf{b} + \boldsymbol{\varepsilon},$$

where $\mathbf{F} = (\mathbf{F}_1^\top, \ldots, \mathbf{F}_n^\top)^\top$, $\mathbf{b} = (\mathbf{b}_1^\top, \ldots, \mathbf{b}_n^\top)^\top$, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1^\top, \ldots, \boldsymbol{\varepsilon}_n^\top)^\top$, and

$$\mathbf{G} = \begin{pmatrix} \mathbf{F}_1 & & 0 \\ & \ddots & \\ 0 & & \mathbf{F}_n \end{pmatrix}.$$

It follows that $\mathrm{E}(\mathbf{Y}) = \mathbf{F}\boldsymbol{\beta}$ and $\mathrm{Cov}(\mathbf{Y}) = \sigma^2\mathbf{V}$, where

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_1 & & 0 \\ & \ddots & \\ 0 & & \mathbf{V}_n \end{pmatrix}$$

is block diagonal.

Marginal model:

Under the assumed normality of the random components the RCR model induces the marginal model

$$\mathbf{Y}_i \sim \mathcal{N}_{m_i}\left(\mathbf{F}_i\boldsymbol{\beta}, \sigma^2(\mathbf{F}_i\mathbf{D}\mathbf{F}_i^\top + \mathbf{I}_{m_i})\right), \tag{4.3}$$

which is the basis for the estimation in this model.

Example 4.1 *One of the most simple examples of an RCR model is the linear regression model with random intercept. An observation of individual i at x_{ij} is modeled by*

$$Y_{ij} = \beta_0 + b_i + x_{ij}\beta_1 + \varepsilon_{ij},$$

where $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $b_i \sim \mathcal{N}(0, \sigma^2 d)$. In this situation the covariance of two observations Y_{ij} and $Y_{ij'}$ does not depend on the chosen settings, but is $\sigma^2 d$ (see Schwabe and Schmelter (2006) for a discussion of design issues for this model).

4.2 General model

The RCR model ((4.1) + (4.2)) can be extended to a more general mixed model as introduced by Laird and Ware (1982) and described in Verbeke and Molenberghs (2000) by modeling the second stage as a multivariate regression model of the form

$$\boldsymbol{\beta}_i = \mathbf{K}_i\boldsymbol{\beta} + \mathbf{b}_i, \quad i = 1, \ldots, n, \tag{4.4}$$

where $\mathbf{K}_i$ is a known $p \times q$ matrix, $\boldsymbol{\beta}$ is the (now) q-dimensional population parameter vector, and $\mathbf{b}_i$ is defined in the same way as for the RCR model. It is assumed that the matrix $\mathbf{K}_i$ does not depend on the experimental settings x_{ij}.

Note, that the RCR model is a special case of the general model where $\mathbf{K}_i = \mathbf{I}_p$ for all i.

Conditional on $\mathbf{b}_i$ the expectation of the observations of individual i is then

$$\mathrm{E}(\mathbf{Y}_i|\mathbf{b}_i) = \mathbf{F}_i(\mathbf{K}_i\boldsymbol{\beta} + \mathbf{b}_i)$$

and with that the covariance matrix of $\mathbf{Y}_i$ is

$$\mathrm{Cov}(\mathbf{Y}_i) = \mathrm{Cov}(\mathrm{E}(\mathbf{Y}_i|\mathbf{b}_i)) + \mathrm{E}(\mathrm{Cov}(\mathbf{Y}_i|\mathbf{b}_i)) = \sigma^2(\mathbf{F}_i\mathbf{D}\mathbf{F}_i^\top + \mathbf{I}_{m_i}) = \sigma^2\mathbf{V}_i,$$

which is the same as in the RCR model. As before, observations from different individuals are uncorrelated.

The marginal model can then be written as

$$\mathbf{Y}_i \sim \mathcal{N}_{m_i}(\mathbf{F}_i\mathbf{K}_i\boldsymbol{\beta}, \sigma^2(\mathbf{F}_i\mathbf{D}\mathbf{F}_i^\top + \mathbf{I}_{m_i})). \tag{4.5}$$

The model for the vector of all observations $\mathbf{Y}$ has the form

$$\mathbf{Y} = \mathbf{G}\mathbf{K}\boldsymbol{\beta} + \mathbf{G}\mathbf{b} + \boldsymbol{\varepsilon},$$

where $\mathbf{K} = (\mathbf{K}_1^\top, \ldots, \mathbf{K}_n^\top)^\top$, and $\mathbf{b}$ and $\mathbf{G}$ are defined as for the RCR model.

For identifiability reasons we require that $\mathbf{K}$ has full column rank; otherwise not the whole population parameter vector will be estimable.

We close this section with an example to see how this extension of the RCR model might be useful.

Example 4.2 *We consider a very simple study where a drug is administered in two different doses. One part of the sample population receives a high dose, the other one a low dose. We assume that the observed response variable is linear in time. The intercept (baseline measurement) is supposed to have the same mean in both groups while the slope can be different. This is modeled by a 3-dimensional population parameter vector* $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top$*, where* β_0 *describes the mean intercept in the population,* β_1 *the mean slope if the high dose is administered and* β_2 *the mean slope if the low dose is administered. The model for the jth observation taken from individual i at the experimental setting* x_{ij} *can then be formulated by*

$$Y_{ij} = (1\ x_{ij})\mathbf{K}_i \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + (1\ x_{ij}) \begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} + \varepsilon_{ij},$$

where $\mathbf{K}_i$ *equals either* $\mathbf{K}_H = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$ *if individual i was given the high dose or* $\mathbf{K}_i$ *equals* $\mathbf{K}_L = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ *if the individual was given the low dose. That is, in this example the matrix*

$\mathbf{K}_i$ *is used to select the parameter set of fixed effects depending on which dosage group individual* i *belongs to.*

4.3 Estimation

In this section we give a short review of the established estimation techniques for linear mixed models. In Section 4.3.1 we assume that the covariance structure $\mathbf{D}$ of the random effects is known. In this case the weighted least squares estimator can be used as best unbiased estimator for the population parameters. If the covariance structure is unknown, usually a maximum likelihood approach is used, which is briefly described in Section 4.3.2.

4.3.1 Estimation of fixed effects for known variance parameters

We will first restrict our attention to the RCR model defined in Section 4.1.

RCR model

If the matrix $\mathbf{D}$ is known, the fixed effects (population parameters) can be estimated using the well known theory for general linear models. Similarly to the discussion of the ordinary linear model (Chapter 3) we first consider the regular case that $\mathbf{F}$ has full column rank. Then, following from the Gauss-Markov theorem (see e. g. Christensen (1987), p. 34), the best linear unbiased estimator for $\boldsymbol{\beta}$ is the weighted least squares estimator

$$\hat{\boldsymbol{\beta}}_{\mathrm{WLS}} = \left(\mathbf{F}^\top \mathbf{V}^{-1} \mathbf{F}\right)^{-1} \mathbf{F}^\top \mathbf{V}^{-1} \mathbf{Y} \tag{4.6}$$

with

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{\mathrm{WLS}}) = \sigma^2 \left(\mathbf{F}^\top \mathbf{V}^{-1} \mathbf{F}\right)^{-1} = \sigma^2 \left(\sum_{i=1}^{n} \mathbf{F}_i^\top \mathbf{V}_i^{-1} \mathbf{F}_i\right)^{-1}. \tag{4.7}$$

Due to the block diagonal structure of the covariance matrix $\mathbf{V}$, the weighted least squares estimator can be rewritten as

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\mathrm{WLS}} &= \left(\sum_{i=1}^{n} \mathbf{F}_i^\top \mathbf{V}_i^{-1} \mathbf{F}_i\right)^{-1} \sum_{i=1}^{n} \mathbf{F}_i^\top \mathbf{V}_i^{-1} \mathbf{Y}_i \\ &= \left(\sum_{i=1}^{n} \mathbf{F}_i^\top \mathbf{V}_i^{-1} \mathbf{F}_i\right)^{-1} \sum_{i=1}^{n} \mathbf{F}_i^\top \mathbf{V}_i^{-1} \mathbf{F}_i \hat{\boldsymbol{\beta}}_i, \end{aligned}$$

where $\hat{\boldsymbol{\beta}}_i$ is an arbitrary weighted least squares solution for the individual first stage model, that is,

$$\hat{\boldsymbol{\beta}}_i = \left(\mathbf{F}_i^\top \mathbf{V}_i^{-1} \mathbf{F}_i\right)^{-} \mathbf{F}_i^\top \mathbf{V}_i^{-1} \mathbf{Y}_i.$$

This means that the weighted least squares estimator for $\boldsymbol{\beta}$ is a matrix-weighted average of estimates for the individual models.

Since $\mathbf{V}_i\mathbf{F}_i = (\mathbf{I}_{m_i} + \mathbf{F}_i\mathbf{D}\mathbf{F}_i^\top)\mathbf{F}_i = \mathbf{F}_i(\mathbf{I}_p + \mathbf{D}\mathbf{F}_i^\top\mathbf{F}_i) = \mathbf{F}_i\mathbf{U}_i$ for some $p \times p$-matrix $\mathbf{U}_i$, the individual weighted least squares solution coincides with an ordinary least squares solution,

$$\hat{\boldsymbol{\beta}}_i = \left(\mathbf{F}_i^\top\mathbf{V}_i^{-1}\mathbf{F}_i\right)^- \mathbf{F}_i^\top\mathbf{V}_i^{-1}\mathbf{Y}_i = (\mathbf{F}_i^\top\mathbf{F}_i)^-\mathbf{F}_i^\top\mathbf{Y}_i, \tag{4.8}$$

according to a result by Zyskind (1967).

If we look at the special case that the numbers of observations m_i and the chosen experimental settings are the same for all individuals and that $\mathbf{F}_i = \mathbf{F}_1, i = 1, \ldots, n$ has full column rank p, then the weighted least squares estimator simplifies to

$$\hat{\boldsymbol{\beta}}_{WLS} = (n\mathbf{F}_1^\top\mathbf{V}_1^{-1}\mathbf{F}_1)^{-1}\sum_{i=1}^n \mathbf{F}_1^\top\mathbf{V}_1^{-1}\mathbf{Y}_i = \frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\beta}}_i.$$

Hence, the calculation of the estimator does not require knowledge of the covariance matrix $\mathbf{D}$, and the weighted least squares estimator coincides with the ordinary least squares estimator (see e. g. Rao (1967) or Bischoff (1992) for related results, or the discussion in Entholzner et al. (2005)).

As in the ordinary linear model if $\mathbf{F}$ does not have full column rank p, it is still true that

$$\hat{\boldsymbol{\beta}}_{WLS} = (\mathbf{F}^\top\mathbf{V}^{-1}\mathbf{F})^-\mathbf{F}^\top\mathbf{V}^{-1}\mathbf{Y} \tag{4.9}$$

is a weighted least squares solution for any choice of the generalized inverse of $\mathbf{F}^\top\mathbf{V}^{-1}\mathbf{F}$. Moreover, if the linear aspect $\varphi(\boldsymbol{\beta}) = \mathbf{L}\boldsymbol{\beta}$ is identifiable, i. e. $\mathbf{L} = \mathbf{Q}\mathbf{F}$ for a suitable matrix $\mathbf{Q}$, then $\hat{\varphi} = \mathbf{L}\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator of φ with covariance matrix

$$\mathrm{Cov}(\hat{\varphi}) = \sigma^2\mathbf{L}\left(\sum_{i=1}^n \mathbf{F}_i^\top\mathbf{V}_i^{-1}\mathbf{F}_i\right)^- \mathbf{L}^\top. \tag{4.10}$$

With the arguments from above, if the same experimental settings are chosen for all individuals and we want to estimate $\varphi(\boldsymbol{\beta}) = \mathbf{L}\boldsymbol{\beta}$, then

$$\hat{\varphi} = \mathbf{L}\left(n\mathbf{F}_1^\top\mathbf{V}_1^{-1}\mathbf{F}_1\right)^- \sum_{i=1}^n \mathbf{F}_1^\top\mathbf{V}_1^{-1}\mathbf{Y}_i = \frac{1}{n}\sum_{i=1}^n \mathbf{L}\hat{\boldsymbol{\beta}}_i$$

holds and, hence, the estimates again do not depend on $\mathbf{D}$.

General model

In the general model the weighted least squares estimator can be used in the same manner as in the RCR model. In the formulas one only has to replace $\mathbf{F}$ by $\mathbf{GK}$ and $\mathbf{F}_i$ by $\mathbf{F}_i\mathbf{K}_i$ leading to

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{WLS} &= \left(\mathbf{K}^\top\mathbf{G}^\top\mathbf{V}^{-1}\mathbf{G}\mathbf{K}\right)^{-1}\mathbf{K}^\top\mathbf{G}^\top\mathbf{V}^{-1}\mathbf{Y} \\ &= \left(\sum_{i=1}^n \mathbf{K}_i^\top\mathbf{F}_i^\top\mathbf{V}_i^{-1}\mathbf{F}_i\mathbf{K}_i\right)^{-1}\sum_{i=1}^n \mathbf{K}_i^\top\mathbf{F}_i^\top\mathbf{V}_i^{-1}\mathbf{Y}_i\end{aligned}$$

with

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{WLS}) = \sigma^2 \left(\sum_{i=1}^{n} \mathbf{K}_i^\top \mathbf{F}_i^\top \mathbf{V}_i^{-1} \mathbf{F}_i \mathbf{K}_i \right)^{-1}.$$

Of course, the result that the weighted least squares estimator coincides with the ordinary least squares estimator in the case that the same settings are chosen for all individuals does not hold here in general because of the influence of the different $\mathbf{K}_i$.

4.3.2 Maximum likelihood estimation

Now, we are looking at the situation that the matrix $\mathbf{D}$ is unknown. We assume that $\mathbf{D}$ is parameterized by a ν-dimensional vector $\boldsymbol{\alpha}$, that is, $\mathbf{D} = \mathbf{D}(\boldsymbol{\alpha})$. To keep in mind that $\mathbf{V}_i$ then also depends on this vector $\boldsymbol{\alpha}$ we will use the notation $\mathbf{V}_i = \mathbf{V}_i(\boldsymbol{\alpha})$. One of the typical cases is that $\mathbf{D}$ is diagonal, where the natural parameterization $\mathbf{D} = \mathbf{D}(\boldsymbol{\alpha}) = \mathrm{diag}(\boldsymbol{\alpha})$ can be used.

One possibility to estimate the parameters (see e. g. Davidian and Giltinan (1995) is to perform the analysis in the same way as the model was constructed: in two steps. That is, in the first step the parameters $\boldsymbol{\beta}_i$ are estimated separately for each individual i by using the ordinary regression model (4.1). Then in the second step, the population parameters (and the corresponding variance parameters) are estimated from the regression model of the second stage (4.2) assuming the estimates for $\boldsymbol{\beta}_i$ were the true values. In practice, however, this can be problematic since it is often not possible to get reliable estimates of the individual parameters. Often it is not even possible to estimate the individual parameters at all. Especially in population pharmacokinetic studies it is very common to use a sparse sampling scheme where the number of measurements per individual is less than the number of population parameters p.

Therefore, the typical approach for the estimation in such models is to use maximum likelihood theory for the marginal model.

If both the observational errors and the random effects follow normal distributions, using the marginal models (4.3) for the RCR model or (4.5) for the general model, the likelihood function is given by

$$L_{\mathrm{ML}}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2; \mathbf{Y}) = \prod_{i=1}^{n} \left\{ (2\pi)^{-m_i/2} |\sigma^2 \mathbf{V}_i(\boldsymbol{\alpha})|^{-\frac{1}{2}} \right. \\ \left. \exp\left(-\frac{1}{2\sigma^2} (\mathbf{Y}_i - \mathbf{F}_i \boldsymbol{\beta})^\top \mathbf{V}_i(\boldsymbol{\alpha})^{-1} (\mathbf{Y}_i - \mathbf{F}_i \boldsymbol{\beta}) \right) \right\} \quad (4.11)$$

for the RCR model and by

$$L_{\mathrm{ML}}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2; \mathbf{Y}) = \prod_{i=1}^{n} \left\{ (2\pi)^{-m_i/2} |\sigma^2 \mathbf{V}_i(\boldsymbol{\alpha})|^{-\frac{1}{2}} \right. \\ \left. \exp\left(-\frac{1}{2\sigma^2} (\mathbf{Y}_i - \mathbf{F}_i \mathbf{K}_i \boldsymbol{\beta})^\top \mathbf{V}_i(\boldsymbol{\alpha})^{-1} (\mathbf{Y}_i - \mathbf{F}_i \mathbf{K}_i \boldsymbol{\beta}) \right) \right\} \quad (4.12)$$

for the general model.

Conditional on $\boldsymbol{\alpha}$ the maximum likelihood estimator for $\boldsymbol{\beta}$, obtained from maximizing Eq. (4.11)

and Eq. (4.12) resp., is given by

$$\hat{\boldsymbol{\beta}}(\alpha) = \left(\sum_{i=1}^{n} \mathbf{F}_i^\top \mathbf{V}_i(\boldsymbol{\alpha})^{-1} \mathbf{F}_i \right)^{-1} \sum_{i=1}^{n} \mathbf{F}_i^\top \mathbf{V}_i(\boldsymbol{\alpha})^{-1} \mathbf{Y}_i \tag{4.13}$$

and

$$\hat{\boldsymbol{\beta}}(\alpha) = \left(\sum_{i=1}^{n} \mathbf{K}_i^\top \mathbf{F}_i^\top \mathbf{V}_i(\boldsymbol{\alpha})^{-1} \mathbf{F}_i \mathbf{K}_i \right)^{-1} \sum_{i=1}^{n} \mathbf{K}_i^\top \mathbf{F}_i^\top \mathbf{V}_i(\boldsymbol{\alpha})^{-1} \mathbf{Y}_i \tag{4.14}$$

(see Laird and Ware (1982)) which is identical to the weighted least squares estimator for known $\mathbf{V}_i(\boldsymbol{\alpha})$ given in Eq. (4.6). If $\boldsymbol{\alpha}$ is not known but an estimate $\hat{\boldsymbol{\alpha}}$ is available, $\boldsymbol{\beta}$ can be estimated by plugging $\hat{\boldsymbol{\alpha}}$ into Eq. (4.13).

If the expression for $\hat{\boldsymbol{\beta}}$ from Eq. (4.13) is inserted into the likelihood function Eq. (4.11) the maximum likelihood estimation for $\boldsymbol{\alpha}$ can be calculated by maximizing that term.

Sometimes the so-called restricted maximum likelihood estimator (REML) is used to account for the fact that the maximum likelihood estimator for $\boldsymbol{\alpha}$ is biased. However, we will not discuss this here, as the asymptotic properties that we will use later are the same as for the maximum likelihood estimator. Details on the REML can be found in Harville (1974).

For the actual calculation of the estimates several methods are proposed in the literature (see Verbeke and Molenberghs (2000), Chapter 5, or Vonesh and Chinchilli (1997), Chapter 6, for an overview). One possibility is to use the EM algorithm proposed by Dempster et al. (1977), where the individual random effects are treated as missing data. As shown by Laird and Ware (1982) both the maximum likelihood and the restricted maximum likelihood estimates can be obtained with this procedure but the convergence is often slow. Today the ML or REML estimates for all parameters are usually calculated using Newton-Raphson-based procedures (see Lindstrom and Bates (1988)).

4.4 The Fisher information matrix

In the case that the variance parameters $\boldsymbol{\alpha}$ are known, it is easy to calculate the covariance matrix of the estimator for the population parameter vector $\boldsymbol{\beta}$ to assess the quality of the chosen experimental settings. When the variance parameters, however, are unknown the covariance matrix for the ML (or REML) estimators cannot be easily calculated or might not even exist. The usual way in such situations to assess the quality of the estimates is to use the Fisher information matrix, which is defined as follows:

Definition 4.3 *Given a statistical model* $\{f_{\mathbf{Y}}(y;\boldsymbol{\theta})\}$ *of a random vector* $\mathbf{Y}$*, the covariance of the score function*

$$U(y;\boldsymbol{\theta}) = \frac{\partial \ln f_{\mathbf{Y}}(y;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} = \left(\frac{\partial \ln f_{\mathbf{Y}}(y;\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial \ln f_{\mathbf{Y}}(y;\boldsymbol{\theta})}{\partial \theta_p} \right)$$

is called Fisher information matrix, denoted by $\mathfrak{M}(\boldsymbol{\theta}) = \mathrm{Cov}(U(\mathbf{Y};\boldsymbol{\theta}))$.

Under certain regularity conditions on the density (twice differentiability and interchangeability of integration and differentiation) it holds that

$$\mathfrak{M}(\boldsymbol{\theta}) = \mathrm{Cov}(U(\mathbf{Y};\boldsymbol{\theta})) = \mathrm{E}\left(U^\top(\mathbf{Y};\boldsymbol{\theta})U(\mathbf{Y};\boldsymbol{\theta})\right) = -\mathrm{E}\left(\frac{\partial U(\mathbf{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right).$$

The regularity conditions are satisfied for the density of the normal distribution.

If T is an unbiased estimator for $\boldsymbol{\theta}$, then

$$\mathrm{Cov}(T(\mathbf{Y})) \geq \mathfrak{M}(\boldsymbol{\theta})^{-1} \tag{4.15}$$

in the sense of the partial ordering (Loewner ordering) of nonnegative definite matrices (see e. g. Cox and Hinkley (2000)). That means that the inverse of the Fisher information matrix is a lower bound for the covariance matrix of any unbiased estimator for $\boldsymbol{\theta}$. The inequality is the so called Cramér-Rao bound.

Furthermore, it can be shown (see e. g. Cox and Hinkley (2000)) that under certain regularity conditions the maximum likelihood estimator is asymptotically normal and asymptotically reaches the Cramér-Rao bound. This motivates to use the inverse of the Fisher information matrix as an approximation of the covariance matrix of the estimator and use it as tool to evaluate the quality of the chosen experimental settings. Note that it can occur that the covariance matrix of the estimator does not exist.

To investigate the Fisher information matrix for the mixed model introduced in this chapter, we partition the matrix in the following way:

$$\mathfrak{M} = \begin{pmatrix} \mathfrak{M}^{\boldsymbol{\beta}} & \mathfrak{M}^{\boldsymbol{\beta\alpha}} & \mathfrak{M}^{\boldsymbol{\beta}\sigma^2} \\ \mathfrak{M}^{\boldsymbol{\beta\alpha}^\top} & \mathfrak{M}^{\boldsymbol{\alpha}} & \mathfrak{M}^{\boldsymbol{\alpha}\sigma^2} \\ \mathfrak{M}^{\boldsymbol{\beta}\sigma^{2\top}} & \mathfrak{M}^{\boldsymbol{\alpha}\sigma^{2\top}} & \mathfrak{M}^{\sigma^2} \end{pmatrix}. \tag{4.16}$$

Using rules of vector differential calculus (see the summary paper by Wand (2002) or the book by Magnus and Neudecker (1988)) it can be shown that for the RCR model

$$\begin{aligned} \mathfrak{M}^{\boldsymbol{\beta}} &= \tfrac{1}{\sigma^2}\mathbf{K}^\top\mathbf{F}^\top\mathbf{V}^{-1}(\boldsymbol{\alpha})\mathbf{F}\mathbf{K} \\ \mathfrak{M}^{\beta\alpha} &= 0 \\ \mathfrak{M}^{\beta\sigma^2} &= 0 \\ \mathfrak{M}^{\boldsymbol{\alpha}} &= \left(\tfrac{1}{2}\mathrm{tr}\left[\mathbf{V}(\boldsymbol{\alpha})^{-1}\tfrac{\partial \mathbf{V}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}_k}\mathbf{V}(\boldsymbol{\alpha})^{-1}\tfrac{\partial \mathbf{V}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}_l}\right]\right)_{k,l=1}^{v} \\ \mathfrak{M}^{\boldsymbol{\alpha}\sigma^2} &= \left(\tfrac{1}{2}\mathrm{tr}\left[\tfrac{1}{\sigma^2}\mathbf{V}(\boldsymbol{\alpha})^{-1}\tfrac{\partial \mathbf{V}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}_k}\right]\right)_{k=1}^{v} \\ \mathfrak{M}^{\sigma^2} &= \tfrac{m}{2\sigma^4} \end{aligned} \tag{4.17}$$

holds, where $m = \sum m_i$ is the total number of observations and v is the dimension of $\boldsymbol{\alpha}$.

The resulting matrix is hence block diagonal,

$$\mathfrak{M} = \begin{pmatrix} \mathfrak{M}^{\beta} & 0 & 0 \\ 0 & \mathfrak{M}^{\alpha} & \mathfrak{M}^{\alpha\sigma^2} \\ 0 & \mathfrak{M}^{\alpha\sigma^2\top} & \mathfrak{M}^{\sigma^2} \end{pmatrix}, \tag{4.18}$$

implicating that the ML estimators for the population parameter vector $\boldsymbol{\beta}$ and the variance parameters are asymptotically uncorrelated.

Note, that $\mathfrak{M}^{\beta^{-1}}$ is identical to the covariance matrix of the weighted least squares estimator for a known covariance structure of the random effects $\mathbf{D}(\boldsymbol{\alpha})$. So, in this case the Cramér-Rao bound is exactly met.

It is important to note that due to the block diagonal structure of $\mathbf{V}$ all the blocks of the information matrix can be additively split up into the contributions of the different individuals. That is, with

$$\begin{aligned} \mathfrak{M}_i^{\beta} &:= \tfrac{1}{\sigma^2}\mathbf{K}_i^{\top}\mathbf{F}_i^{\top}\mathbf{V}_i^{-1}(\boldsymbol{\alpha})\mathbf{F}_i\mathbf{K}_i \\ \mathfrak{M}_i^{\alpha} &:= \left(\tfrac{1}{2}\mathrm{tr}\left[\mathbf{V}_i(\boldsymbol{\alpha})^{-1}\tfrac{\partial \mathbf{V}_i(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}_k}\mathbf{V}_i(\boldsymbol{\alpha})^{-1}\tfrac{\partial \mathbf{V}_i(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}_l}\right]\right)_{k,l=1}^{\nu} \\ \mathfrak{M}_i^{\alpha\sigma^2} &:= \left(\tfrac{1}{2}\mathrm{tr}\left[\tfrac{1}{\sigma^2}\mathbf{V}_i(\boldsymbol{\alpha})^{-1}\tfrac{\partial \mathbf{V}_i(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}_k}\right]\right)_{k=1}^{\nu} \\ \mathfrak{M}_i^{\sigma^2} &:= \tfrac{m_i}{2\sigma^4} \end{aligned} \tag{4.19}$$

and

$$\mathfrak{M}_{\mathrm{ind},i} := \begin{pmatrix} \mathfrak{M}_i^{\beta} & 0 & 0 \\ 0 & \mathfrak{M}_i^{\alpha} & \mathfrak{M}_i^{\alpha\sigma^2} \\ 0 & \mathfrak{M}_i^{\alpha\sigma^2\top} & \mathfrak{M}_i^{\sigma^2} \end{pmatrix} \tag{4.20}$$

it holds that

$$\mathfrak{M} = \sum_{i=1}^{n} \mathfrak{M}_{\mathrm{ind},i}. \tag{4.21}$$

It is also important to mention that the information matrix depends on the unknown parameters $\boldsymbol{\alpha}$ and σ^2. This makes it impossible to judge about the quality of the chosen settings before the conduction of the experiment without some prior knowledge about these parameters. After the experiment has taken place the covariance matrix can be estimated by substituting the unknown parameters in the expression of the information matrix by their estimates. For the planning of experiments it is common to use best guesses for the unknown parameters to derive locally optimal settings, i. e. settings that will usually be optimal in a neighborhood of these best guesses.

One has to be careful if zero is chosen as best guess for the variance of one of the random effects. The matrix $\mathfrak{M}^{\alpha}$ will then usually become singular. This specific parameter should better be assumed to be known and be omitted when planning the experiment. As a variance of zero would lie on the boundary of the set of possible values for this component, the assumption of asymptotic normality of the estimates would be violated.

5 Nonlinear Mixed Models

In this chapter we give an introduction to nonlinear mixed models. They are much more important for the application in the area of pharmacokinetics than the linear ones, since most of the relevant kinetic models are derived from differential equations resulting in functions that are non-linear in the parameters (see Chapter 2). As for the linear case, a large amount of literature on the analysis of these models can be found, like for example, the books of Vonesh and Chinchilli (1997), Davidian and Giltinan (1995), or Pinheiro and Bates (2000). Therefore, we will not go further into detail in this chapter as most of the topics are covered in the cited references.

In Section 5.1 the nonlinear mixed model is introduced similar to the linear one in Chapter 4. The second section roughly covers maximum likelihood estimation methods. The most important part for design purposes is Section 5.3, where an approximation of the Fisher information matrix is given following Mentré et al. (1997) or Retout et al. (2001).

5.1 The model

Analogously to the linear mixed model introduced in Chapter 4, the nonlinear mixed model can be written as hierarchical two-stage model, where the first stage is used to model the intra-individual variability, whereas the second stage accounts for the inter-individual variability.

First stage (intra-individual model):

In the first stage on individual level the jth observation of individual i is modeled by a nonlinear regression model of the form

$$Y_{ij} = \eta(x_{ij}, \boldsymbol{\beta}_i) + \varepsilon_{ij}, \qquad j = 1, \ldots, m_i, \tag{5.1}$$

where, as before, x_{ij} is the experimental setting and $\boldsymbol{\beta}_i$ is the vector of individual parameters. Again the observational errors ε_{ij} are assumed to be independent and identically normally distributed with zero mean and variance σ^2. In contrast to before, the regression function η can now be nonlinear in the parameter vector $\boldsymbol{\beta}_i$. Note that the linear mixed model is a special case, where η has the form $\eta(x_{ij}, \boldsymbol{\beta}_i) = \mathbf{f}(x_{ij})^\top \boldsymbol{\beta}_i$. To avoid difficulties for later results we assume that η is continuous in x_{ij} and differentiable in $\boldsymbol{\beta}_i$.

If we define

$$\boldsymbol{\eta}_i(\boldsymbol{\beta}_i) = \begin{pmatrix} \eta(x_{i1}, \boldsymbol{\beta}_i) \\ \vdots \\ \eta(x_{im_i}, \boldsymbol{\beta}_i) \end{pmatrix},$$

all observations of individual i can be summarized using matrix/vector notation to

$$\mathbf{Y}_i = \boldsymbol{\eta}_i(\boldsymbol{\beta}_i) + \boldsymbol{\varepsilon}_i.$$

Second stage (inter-individual model):

In the second stage again the inter-individual variability is modeled by a linear multivariate regression model of the form

$$\boldsymbol{\beta}_i = \mathbf{K}_i\boldsymbol{\beta} + \mathbf{b}_i, \qquad i = 1, \ldots, n. \tag{5.2}$$

As before we assume that the individual random effects vectors $\mathbf{b}_i$, $i = 1, \ldots, n$, are independent and identically normally distributed, that is, $\mathbf{b}_i \sim \mathcal{N}_p(0, \sigma^2\mathbf{D})$ and that the $\mathbf{b}_i$ are independent of $\varepsilon_{i'j}$.

An important special case is the nonlinear random coefficient regression model, where, as for the linear RCR model, the matrices $\mathbf{K}_i$ are assumed to be identity matrices. This model will play the dominant role in this text. The model can be further extended by allowing the second stage to be also modeled by a nonlinear function.

Marginal model

The description of the marginal distribution of the individual observation vectors is not as simple as in the linear case. The density of the vector of observations can be expressed by the integral

$$f_{\mathbf{Y}_i}(\mathbf{y}_i; \boldsymbol{\beta}, \mathbf{D}, \sigma^2) = \int \frac{1}{(2\pi\sigma^2)^{m_i/2}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}_i - \boldsymbol{\eta}_i(\mathbf{K}_i\boldsymbol{\beta} + \mathbf{b}_i))^\top(\mathbf{y}_i - \boldsymbol{\eta}_i(\mathbf{K}_i\boldsymbol{\beta} + \mathbf{b}_i))\right] \frac{1}{(2\pi)^{p/2}|\sigma^2\mathbf{D}|^{1/2}} \exp\left[-\frac{1}{2\sigma^2}\mathbf{b}_i^\top\mathbf{D}^{-1}\mathbf{b}_i\right] d\mathbf{b}_i. \tag{5.3}$$

Due to the non-linearity of the regression function in the parameters, there is usually no closed-form solution of the integral. For a linear η Eq. (5.3) can be simplified and an expression of the form of Eq. (4.12) is obtained. Note that Eq. (5.3) has to be modified for singular $\mathbf{D}$.

5.2 Estimation

For the estimation we again assume that the matrix $\mathbf{D}$ can be parameterized by some vector $\boldsymbol{\alpha}$, that is, $\mathbf{D} = \mathbf{D}(\boldsymbol{\alpha})$. The parameters that have to be estimated are then $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and σ^2.

5.2.1 Two-stage procedure

As in the linear model, one possibility to estimate the parameters is to follow the two stages, in which the model was built up. That is, in the first step, individual estimates for $\boldsymbol{\beta}_i$ and an

estimate for σ^2 are obtained, for example, by using generalized least-squares. In the second step, estimates for the population parameter vector $\boldsymbol{\beta}$ and the covariance parameters $\boldsymbol{\alpha}$ are calculated. In the standard two-stage method (Steimer et al. (1984)) the estimates $\hat{\boldsymbol{\beta}}_i$ are treated as if they were the true $\boldsymbol{\beta}_i$ which leads to an upwardly biased estimator for $\mathbf{D}$ (see Davidian and Giltinan (1995), Chapter 5). Better estimates can be obtained by incorporating the uncertainty of the estimation for $\boldsymbol{\beta}_i$. This can be done by using the asymptotic covariance matrix for $\hat{\boldsymbol{\beta}}_i$ assuming that $\hat{\boldsymbol{\beta}}_i$ is approximately normally distributed (Davidian and Giltinan (1995), pp. 136).

5.2.2 Maximum likelihood

In many cases, the two-stage procedure described in the previous section cannot be used, since often reliable individual estimates cannot be obtained for all subjects. Especially in population pharmacokinetic studies, the number of blood samples per individual is often very small, sometimes only equal to or even smaller than the number of population parameters to be estimated.

Therefore, the usual estimation technique in this situation is maximum likelihood estimation based on the marginal model (5.3). As we saw in Eq. (5.3), the likelihood function of the vector of all observations $\mathbf{y}$ cannot be described in a closed form, but only as an integral of the form

$$L(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2; \mathbf{y}) = \prod_{i=1}^{n} \int \frac{1}{(2\pi\sigma^2)^{m_i/2}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{Y}_i - \boldsymbol{\eta}_i(\mathbf{K}_i\boldsymbol{\beta} + \mathbf{b}_i))^\top(\mathbf{Y}_i - \boldsymbol{\eta}_i(\mathbf{K}_i\boldsymbol{\beta} + \mathbf{b}_i))\right] \frac{1}{(2\pi)^{p/2}|\sigma^2\mathbf{D}(\boldsymbol{\alpha})|^{1/2}} \exp\left[-\frac{1}{2\sigma^2}\mathbf{b}_i^\top\mathbf{D}(\boldsymbol{\alpha})^{-1}\mathbf{b}_i\right] d\mathbf{b}_i. \quad (5.4)$$

To make the numerical optimization of this likelihood function a tractable problem, different approximations to Eq. (5.4) have been proposed in the literature (see, e. g., the reviews in Pinheiro and Bates (2000), Chapter 7, or in Davidian and Giltinan (1995)), Chapter 5 and 6,. Many of the procedures use a first-order Taylor expansion of the model function η around the expected value of the individual parameters $\boldsymbol{\beta}_i$ (see, e. g, Sheiner and Beal (1980)). This idea is, for example, implemented in the FO-method of NONMEM (Beal and Sheiner (1989)). The estimates are obtained in an iterative procedure, where in each step estimates for the population and covariance parameters are obtained using an extended least squares method and the linearization in the next step is calculated around the new estimates of the population parameters. An extension of this approach is to obtain empirical Bayes estimates for the individual parameters in each step of the iteration and to use a Taylor expansion around these individual parameters (see, e. g., Lindstrom and Bates (1990)). This is a better approximation to the true model, but is computationally more expensive. This type of procedure is implemented in the FOCE-method of NONMEM (Beal and Sheiner (1998)). A comparison and discussion of the different approximation methods can be found in Pinheiro and Bates (1995). There are also other approaches, like the SAEM proposed by Kuhn and Lavielle (2005), where instead of an approximation to the model function a stochastic approximation to the EM algorithm is used, which avoids the impact of a linearization. A comparison of several implementations of these

or other approaches with respect to their performance in a given pharmacokinetic model can be found in Girard and Mentré (2005).

As all the mentioned procedures deliver approximations to the maximum likelihood estimates, for designing the experiment it is not important, which method is actually used for the analysis.

5.3 Approximation of the Fisher information

The problem of the missing closed-form solution of the likelihood-function carries forward to the calculation of the Fisher information matrix, which is necessary for designing the experiment. The problem here gets even more complex, as the Fisher information is the expectation of the second derivatives of the log-likelihood function. Similar to some of the mentioned estimation procedures, the nonlinear model function is linearized around a best guess for the individual parameters $\boldsymbol{\beta}_i$ derived from some prior knowledge about the model (see, e. g., Mentré et al. (1997) or Retout et al. (2001) and Retout et al. (2002)). At least in the field of population pharmacokinetics this prior knowledge is often available from previously conducted ordinary pharmacokinetic studies.

Linearization in the non-linear RCR model

Again we first start with the simple RCR model. We assume that we have some idea about the values of the population parameter vector $\boldsymbol{\beta}$ and denote this best guess with $\boldsymbol{\beta}_0$. The best we can guess for $\mathbf{b}_i$ is its expectation $\mathrm{E}(\mathbf{b}_i)$, which is zero. Therefore, the best guess for the individual parameter vector $\boldsymbol{\beta}_i$ is also $\boldsymbol{\beta}_0$.

Similar as described in Section 5.2.2 for some of the estimation approaches, the non-linear model function η is linearized using a first-order Taylor approximation. We define

$$\mathbf{f}(x_{ij}, \boldsymbol{\beta}_0) := \left.\frac{\partial \eta(x_{ij}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}, \tag{5.5}$$

which is the gradient of η with respect to the second argument $\boldsymbol{\beta}$ evaluated at $\boldsymbol{\beta}_0$. The regression function η is then approximated by

$$\eta(x_{ij}, \boldsymbol{\beta}_i) \simeq \eta(x_{ij}, \boldsymbol{\beta}_0) + \mathbf{f}(x_{ij}, \boldsymbol{\beta}_0)^\top (\boldsymbol{\beta}_i - \boldsymbol{\beta}_0). \tag{5.6}$$

The first stage model (5.1) can then be approximated by

$$\begin{aligned} Y_{ij} &= \eta(x_{ij}, \boldsymbol{\beta}_i) + \varepsilon_{ij} \\ &\simeq \eta(x_{ij}, \boldsymbol{\beta}_0) + \mathbf{f}(x_{ij}, \boldsymbol{\beta}_0)^\top (\boldsymbol{\beta}_i - \boldsymbol{\beta}_0) + \varepsilon_{ij}. \end{aligned} \tag{5.7}$$

With $\tilde{Y}_{ij} := Y_{ij} - \eta(x_{ij}, \boldsymbol{\beta}_0)$ and $\tilde{\boldsymbol{\beta}}_i = \boldsymbol{\beta}_i - \boldsymbol{\beta}_0$ we can for – planning purposes – instead investigate the linear mixed model

$$\tilde{Y}_{ij} = \mathbf{f}(x_{ij}, \boldsymbol{\beta}_0)^\top \tilde{\boldsymbol{\beta}}_i + \varepsilon_{ij} \tag{5.8}$$

as an approximation to the original non-linear model Eq. (5.1). Model (5.8) has the form of a linear RCR model as defined in Eq. (4.1). To calculate the Fisher information matrix we can hence proceed as in Section 4.4. The resulting matrix, however, now does not only depend on the known or assumed variance parameters, but also on the chosen $\boldsymbol{\beta}_0$.

Note that in the case that η is already linear in the parameter vector, that is, if η has the form $\eta(x_{ij}, \boldsymbol{\beta}) = \mathbf{f}(x_{ij})^\top \boldsymbol{\beta}$, the gradient in Eq. (5.5) does not depend on $\boldsymbol{\beta}_0$ and the approximation in Eq. (5.7) is exact.

Linearization in the general model

In the general model the best guess for the individual parameter vector $\boldsymbol{\beta}_i$ is $\mathbf{K}_i\boldsymbol{\beta}_0$. Therefore, we approximate the non-linear mixed model by

$$Y_{ij} \simeq \eta(x_{ij}, \mathbf{K}_i\boldsymbol{\beta}_0) + \mathbf{f}(x_{ij}, \mathbf{K}_i\boldsymbol{\beta}_0)^\top \mathbf{K}_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \mathbf{f}(x_{ij}, \mathbf{K}_i\boldsymbol{\beta}_0)^\top \mathbf{b}_i + \varepsilon_{ij} \tag{5.9}$$

or equivalently by

$$\tilde{Y}_{ij} \simeq \mathbf{f}(x_{ij}, \mathbf{K}_i\boldsymbol{\beta}_0)^\top \mathbf{K}_i\tilde{\boldsymbol{\beta}} + \mathbf{f}(x_{ij}, \mathbf{K}_i\boldsymbol{\beta}_0)^\top \mathbf{b}_i + \varepsilon_{ij}. \tag{5.10}$$

This model now has the form of the general linear mixed model defined in Section 4.2 and the Fisher information matrix can be calculated as described there. Note, however, that the vector of regression functions does not only depend on $\boldsymbol{\beta}_0$, which is the same for all individuals, but also on $\mathbf{K}_i$, which can differ between the individuals. Thus, we have different regression functions for the different individuals depending on the shape of the respective $\mathbf{K}_i$.

The quality of this linearization approach depends on how close the chosen $\boldsymbol{\beta}_0$ is to the true $\boldsymbol{\beta}$ as well as on the nonlinearity of the model function in the parameter vector. The influence of the nonlinearity has been investigated by Merlé and Tod (2001) for two specific kinetic models. As a reference, they have numerically calculated the information matrices of several designs for the nonlinear model, which is computationally very exhaustive, and compared these with the information matrices derived from the linearized model together with curvature measures described in the literature for nonlinear regression models (see e. g., Cook and Goldberg (1986) and Bates et al. (1983)).

5.4 Proportional error models

In pharmacokinetic modeling the concentrations are often assumed to be log-normally distributed. Therefore, the observational errors (intra-individual variability) are often modeled as log-normal multiplicative errors, that is, the models for the individual observations have the form

$$Y_{ij} = \eta(x_{ij}, \boldsymbol{\beta}_i)\exp(\varepsilon_{ij}), \qquad j = 1, \ldots, m_i. \tag{5.11}$$

The design and the analysis can be performed in the same way as before by considering the log-transformed model

$$\log(Y_{ij}) = \log(\eta(x_{ij}, \boldsymbol{\beta}_i)) + \varepsilon_{ij} \tag{5.12}$$

instead of the original one, which leads to a model with additive normally distributed errors. Following from the chain rule for differentiation, the vector of regression functions in the corresponding linear approximation of the model has the form

$$\mathbf{f}(x_{ij}, \boldsymbol{\beta}_0) = \frac{1}{\eta(x_{ij}, \boldsymbol{\beta}_0)} \left.\frac{\partial \eta(x_{ij}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}.$$

Similarly, also the individual parameters (or some of them) are often assumed to be log-normally distributed within the population, that is,

$$\begin{aligned} & \boldsymbol{\beta}_i = \boldsymbol{\beta} \circ \exp(\mathbf{b}_i) \\ \Leftrightarrow \quad & \log(\boldsymbol{\beta}_i) = \log(\boldsymbol{\beta}) + \mathbf{b}_i \end{aligned}$$

for the RCR model. The exponential function and the logarithm in the equation are meant to be evaluated component-wise resulting in a vector and "$\circ$" denotes the component-wise product (Hadamard product) of two vectors or matrices of the same size. The model for the individual observations then has the form

$$Y_{ij} = \eta(x_{ij}, \boldsymbol{\beta} \circ \exp(\mathbf{b}_i)) + \varepsilon_{ij}. \tag{5.13}$$

This situation can also be led back to the originally introduced model with additive random effects by using a reparameterization. With $\tilde{\boldsymbol{\beta}}_i = \log(\boldsymbol{\beta}_i)$ and $\tilde{\boldsymbol{\beta}} = \log(\boldsymbol{\beta})$ the model can be written as

$$Y_{ij} = \eta(x_{ij}, \exp(\tilde{\boldsymbol{\beta}}_i)) + \varepsilon_{ij} = \eta(x_{ij}, \exp(\tilde{\boldsymbol{\beta}} + \mathbf{b}_i)) + \varepsilon_{ij},$$

where η is the same as in Eq. (5.13). The corresponding linear approximation then has the regression functions

$$\mathbf{f}(x_{ij}, \boldsymbol{\beta}_0) = \left.\frac{\partial \eta(x_{ij}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \circ \boldsymbol{\beta}_0.$$

We illustrate this with a small example.

Example 5.1 *We consider a simple one-compartment model without absorption (see Section 2.1), where the concentration of a drug observed at time x is modeled by*

$$C(x) = D\frac{1}{V_{ci}} e^{-k_{ei}x}.$$

Additive observational error / additive random effects: *We first investigate the standard case, that both the observational errors and the random effects are modeled as additive normal distributed random variables, that is, the concentration measured in individual i at time x_{ij} is modeled by*

$$Y_{ij} = D\frac{1}{V_{ci}} e^{-k_{ei}x_{ij}} + \varepsilon_{ij}$$

with $\varepsilon_{ij} \sim i.i.d.\, \mathcal{N}(0, \sigma^2)$ *and*

$$V_{ci} = V_c + b_{i1},$$
$$k_{ei} = k_e + b_{i2}$$

with $\mathbf{b}_i = (b_{i1}, b_{i2})^\top \sim \mathcal{N}_2(0, \sigma^2\mathbf{D})$. *If* V_{c0} *and* k_{e0} *are used as best guess for* V_c *and* k_e, *the linearized model has the regression functions*

$$f_1(x_{ij}, (V_{c0}, k_{e0})) = -D\frac{e^{-k_{e0}x_{ij}}}{V_{c0}^2},$$
$$f_2(x_{ij}, (V_{c0}, k_{e0})) = -D\frac{x_{ij}e^{-k_{e0}x_{ij}}}{V_{c0}}.$$

Proportional observational error / additive random effects: *Now we consider the case that the observational error is modeled as proportional error while the random effects are still additive random variables, that is, the concentration measured in individual* i *at time* x_{ij} *is modeled by*

$$Y_{ij} = D\frac{1}{V_{ci}}e^{-k_{ei}x_{ij}}\exp\varepsilon_{ij}.$$

The linearized log-transformed model has the regression functions

$$f_1(x_{ij}, (V_{c0}, k_{e0})) = -\frac{1}{V_{c0}},$$
$$f_2(x_{ij}, (V_{c0}, k_{e0})) = -x_{ij}.$$

Additive observational error / proportional random effects: *If the observational errors are assumed to be additive, while the random parameters are modeled by*

$$V_{ci} = V_c\exp(b_{i1}),$$
$$k_{ei} = k_e\exp(b_{i2})$$

the resulting regression functions of the linearized model are

$$f_1(x_{ij}, (V_{c0}, k_{e0})) = -D\frac{e^{-k_{e0}x_{ij}}}{V_{c0}},$$
$$f_2(x_{ij}, (V_{c0}, k_{e0})) = -D\frac{k_{e0}x_{ij}e^{-k_{e0}x_{ij}}}{V_{c0}}.$$

However, one has to keep in mind that now not k_e *and* V_d *but* $\log(k_e)$ *and* $\log(V_c)$ *are the parameters to be estimated. Note that this is a nonlinear reparameterization of the model.*

Proportional observational error / proportional random effects: *If both the observational error and the random effect are modeled as multiplicative log-normal, we obtain*

$$f_1(x_{ij}, (V_{c0}, k_{e0})) = -1,$$
$$f_2(x_{ij}, (V_{c0}, k_{e0})) = -k_{e0}x_{ij}$$

for the regression functions in the linearized model.

Note that in both cases, in which a proportional observational error was assumed, the resulting linearized model is a straight line regression.

6 Designs, Information Matrices and Criteria

In the first part of this chapter we adjust and extend the design definition for the ordinary linear model given in Section 3.2 to the needs in the mixed model setting. In the second section we derive representations for the blocks of the information matrices that correspond to these designs. We close this chapter with a short discussion about appropriate optimality criteria for designs in mixed models.

6.1 Designs

In the mixed model setup designing an experiment does not only mean to specify the sampling times (or more general the experimental settings), but also to determine how many individuals (or what proportions of the individuals) are to be observed at which of the specified times. Therefore, the term design will be used to specify

1. the number n of individuals that should be observed,
2. the number m_i of observations of individual i,
3. the settings x_{ij} (e. g., sampling times), under which the observations are taken.

Similar to the two stages occurring in the definition of the mixed models, designs for mixed models can be specified in two steps: first the definition of designs on individual level and then, on top of this, designs for the whole sample population.

6.1.1 Individual (elementary) designs

A design ξ_i on individual level is naturally defined as a set of experimental settings $x_{i1}, \ldots, x_{ik_i} \in \mathcal{X}$ together with the number of observations $m_{i1}, \ldots, m_{ik_i}$ to be taken at these settings, that is, with the number of replications (cf. Section 3.2). By this the observations that have to be taken for a single individual are specified.

Similar to the approximate designs defined for the ordinary linear model in Chapter 3 we will now allow that the replications m_{ij} may be non-integer, which will then be referred to as *approximate individual design*. We formalize this by means of the following definition.

Definition 6.1 *An (approximate) individual design ξ_i with m_i observations is a vector of experimental settings $(x_{i1}, \ldots, x_{ik_i}) \in \mathcal{X}^{k_i}$ together with the numbers of replications $(m_{i1}, \ldots, m_{ik_i}) \in \mathbb{R}^{k_i}$ to be taken at these settings, where $\sum m_{ij} = m_i$, $m_{ij} \geq 0$.*
Notation:

$$\xi_i = \begin{pmatrix} x_{i1} & \ldots & x_{ik_i} \\ m_{i1} & \ldots & m_{ik_i} \end{pmatrix}.$$

We will denote the set of all individual designs with m observations by $\Xi^{(m)}$.

The approximate individual designs can be identified with finite discrete measures on $\mathcal{X}$. As in the ordinary linear model, the set of possible approximate individual designs $\Xi^{(m)}$ with fixed sample size m is then convex and the optimization problems are easier to handle. Designs that are practically applicable can be found by rounding if necessary.

Individual designs, where all replications m_{ij} are integers, will be referred to as *exact individual designs*.

Note that in the mixed model setup it makes sense not to look at normalized design measures, as the corresponding individual information matrices are not proportional to the number of observations. We will see this in the following sections.

6.1.2 Population designs

In the second stage we now specify the experimental settings for the whole sample population with n individuals. We define a *population design* ζ as a set of individual designs $\xi_1, \ldots, \xi_l$ together with the proportions $g_1, \ldots, g_l$ of individuals of the sample population that should be observed using these designs. This means that ng_i individuals are to be observed with individual design ξ_i. This is formally summarized in the following definition.

Definition 6.2 *A population design ζ is a vector of individual designs $(\xi_1, \ldots, \xi_l)$ together with the vector of the proportions of individuals of the sample population $(g_1, \ldots, g_l) \in [0,1]^l$ with $\sum g_v = 1$ that should be observed under these designs.*
Notation:

$$\zeta = \begin{pmatrix} \xi_1 & \ldots & \xi_l \\ g_1 & \ldots & g_l \end{pmatrix}. \tag{6.1}$$

The set of all population designs will be denoted by Δ.

If both the ξ_i are exact designs and ng_v is integer for $v = 1, \ldots, l$, then the design ζ will be referred to as *exact population design*.

Note that in contrast to the elementary designs the standardization of population designs with respect to the number of individuals is sensible, as the population information matrices are proportional to the number of individuals.

6.2 Information matrices

In this section we have a closer look at the formulas for the information matrices described in Section 4.4. In particular, we derive a representation for the information matrix for approximate designs. As we have seen in Section 4.4, the information matrix of the whole sample population is the sum of the information matrices provided by the single individuals. It is hence natural to write

$$\mathfrak{M}_{\text{pop}}(\zeta) = \sum_{i=1}^{l} g_i \mathfrak{M}_{\text{ind}}(\xi_i),$$

where $\mathfrak{M}_{\text{ind}}(\xi_i)$ denotes the information provided by the individual design ξ_i. Thus $\mathfrak{M}_{\text{pop}}(\zeta)$ is the information matrix of the population design ζ normalized by the number of individuals.

We will now derive a formula for these individual information matrices that can be applied also to the approximate designs given in the previous section. For this, we treat the block for the population parameters and the block for the variance parameters one after another. We restrict our attention to the RCR model, that is, we assume that all the matrices $\mathbf{K}_i$ are identity matrices. It is, however, straight-forward to extend the result to the general case.

6.2.1 Population parameter block

In Eq. (4.19) in Section 4.4 we saw that the block in the individual information matrix accounting for $\boldsymbol{\beta}$ has the form

$$\frac{1}{\sigma^2}\mathbf{F}_i^\top \mathbf{V}_i^{-1}\mathbf{F}_i = \frac{1}{\sigma^2}\mathbf{F}_i^\top (\mathbf{F}_i \mathbf{D}\mathbf{F}_i^\top + \mathbf{I}_{m_i})^{-1}\mathbf{F}_i$$

with $\mathbf{F}_i = (\mathbf{f}(x_{i1}), \dots, \mathbf{f}(x_{im_i}))^\top$ being the design matrix of the observations for individual i.

In the case of an exact design $\xi_i = \begin{pmatrix} x_{i1} & \dots & x_{ik_i} \\ m_{i1} & \dots & m_{ik_i} \end{pmatrix}$, that is, in the case of a design with $m_{ij} \in \mathbb{N}$, $i = 1, \dots, k_i$, the individual information matrix can hence be calculated by

$$\mathfrak{M}_{\text{ind}}^{\boldsymbol{\beta}}(\xi_i) = \frac{1}{\sigma^2}\tilde{\mathbf{F}}_i^\top \left(\tilde{\mathbf{F}}_i \mathbf{D}\tilde{\mathbf{F}}_i^\top + \mathbf{I}_{m_i}\right)^{-1}\tilde{\mathbf{F}}_i,$$

where

$$\tilde{\mathbf{F}}_i = (\underbrace{\mathbf{f}(x_{i1}), \dots, \mathbf{f}(x_{i1})}_{m_{i1} \text{ times}}, \dots, \underbrace{\mathbf{f}(x_{ik_i}), \dots, \mathbf{f}(x_{ik_i})}_{m_{ik_i} \text{ times}})^\top.$$

This, however, does not show us, how the information matrix for an approximate individual design can be calculated. We use the following lemma to derive a different representation of the information matrix. This will then also be applicable for approximate designs.

Lemma 6.3 *Let* $\mathbf{F}$ *be a* $k \times p$*-matrix,* $\mathbf{D}$ *a* $p \times p$*-matrix, and* $\mathbf{S}$ *be an* $m \times k$*-matrix, for which* $\mathbf{S}^\top\mathbf{S}$ *is non-singular. Then*

$$\mathbf{F}^\top\mathbf{S}^\top \left(\mathbf{S}\mathbf{F}\mathbf{D}\mathbf{F}^\top\mathbf{S}^\top + \mathbf{I}_m\right)^{-1}\mathbf{S}\mathbf{F} = \mathbf{F}^\top \left(\mathbf{F}\mathbf{D}\mathbf{F}^\top + (\mathbf{S}^\top\mathbf{S})^{-1}\right)^{-1}\mathbf{F}.$$

Proof: It is sufficient to show that

$$\mathbf{S}^\top \left(\mathbf{SFDF}^\top\mathbf{S}^\top + \mathbf{I}_m\right)^{-1} \mathbf{S}\left(\mathbf{FDF}^\top + (\mathbf{S}^\top\mathbf{S})^{-1}\right) = \mathbf{I}_k.$$

$$\begin{aligned}
&\mathbf{S}^\top \left(\mathbf{SFDF}^\top\mathbf{S}^\top + \mathbf{I}_m\right)^{-1} \mathbf{S}\left(\mathbf{FDF}^\top + (\mathbf{S}^\top\mathbf{S})^{-1}\right) \\
&\qquad = \mathbf{S}^\top \left(\mathbf{SFDF}^\top\mathbf{S}^\top + \mathbf{I}_m\right)^{-1} \mathbf{S}\left(\mathbf{FDF}^\top + (\mathbf{S}^\top\mathbf{S})^{-1}\right)\mathbf{S}^\top\mathbf{S}(\mathbf{S}^\top\mathbf{S})^{-1} \\
&\qquad = \mathbf{S}^\top \left(\mathbf{SFDF}^\top\mathbf{S}^\top + \mathbf{I}_m\right)^{-1} \mathbf{S}\left(\mathbf{FDF}^\top\mathbf{S}^\top\mathbf{S} + \mathbf{I}_k\right)(\mathbf{S}^\top\mathbf{S})^{-1} \\
&\qquad = \mathbf{S}^\top \left(\mathbf{SFDF}^\top\mathbf{S}^\top + \mathbf{I}_m\right)^{-1} \left(\mathbf{SFDF}^\top\mathbf{S}^\top\mathbf{S} + \mathbf{S}\right)(\mathbf{S}^\top\mathbf{S})^{-1} \\
&\qquad\qquad = \mathbf{S}^\top \left(\mathbf{SFDF}^\top\mathbf{S}^\top + \mathbf{I}_m\right)^{-1} \left(\mathbf{SFDF}^\top\mathbf{S}^\top + \mathbf{I}_m\right)\mathbf{S}(\mathbf{S}^\top\mathbf{S})^{-1} = \mathbf{I}_k.
\end{aligned}$$

□

Let now $\mathbf{F}_i := (\mathbf{f}(x_{i1}), \ldots, \mathbf{f}(x_{ik_i}))^\top$ be the $k_i \times p$-dimensional design matrix, where each of the settings x_{ij} is contained only once, and $\mathbf{S}_i$ be a block matrix, where all the entries are either one or zero, defined by

$$\mathbf{S}_i := ((\mathbf{e}_1 \mathbb{1}_{m_{i1}}^\top), \ldots, (\mathbf{e}_{k_i} \mathbb{1}_{m_{ik_i}}^\top))^\top.$$

Here, $\mathbb{1}_{m_{ij}}$ denotes the m_{ij}-dimensional vector with all entries equal to one and $\mathbf{e}_i$ is the i-th unit vector of dimension k_i. Then $\tilde{\mathbf{F}}_i = \mathbf{S}_i\mathbf{F}_i$ and with $\mathbf{W}_i := \mathbf{S}_i^\top\mathbf{S}_i = \operatorname{diag}(m_{i1}, \ldots, m_{ik_i})$ and Lemma 6.3 it follows that

$$\mathfrak{M}_{\text{ind}}^{\beta}(\xi_i) = \frac{1}{\sigma^2}\tilde{\mathbf{F}}_i^\top \left(\tilde{\mathbf{F}}_i\mathbf{D}\tilde{\mathbf{F}}_i^\top + \mathbf{I}_{m_i}\right)^{-1} \tilde{\mathbf{F}}_i = \frac{1}{\sigma^2}\mathbf{F}_i^\top \left(\mathbf{F}_i\mathbf{D}\mathbf{F}_i{}^\top + \mathbf{W}_i^{-1}\right)^{-1} \mathbf{F}_i. \tag{6.2}$$

This representation can also be used for approximate designs. The part $\mathbf{F}_i\mathbf{D}\mathbf{F}_i^\top + \mathbf{W}_i^{-1}$ can be seen as a generalization of the covariance matrix $\mathbf{V}_i$. The more observations of one individual are taken at x_{ij}, the better the variance at x_{ij} can be assessed, which is expressed by the corresponding entry in $\mathbf{W}_i$.

With the following lemma we will see that the individual information matrix (6.2) in the RCR model can be expressed as a function of the information matrix of the same design in the corresponding ordinary linear model, that is, in the model without the random effects. This will later allow us to solve design problems by leading them back to the well known theory of optimal designs in the ordinary linear model setting.

Lemma 6.4 (cf. Schmelter (2006b)) *Let* $\mathbf{F}$ *be an arbitrary* $k \times p$*-matrix,* $\mathbf{W}$ *a symmetric positive definite* $k \times k$*-matrix, and* $\mathbf{D}$ *a symmetric non-negative definite* $p \times p$*-matrix. Then with* $\mathbf{M} := \mathbf{F}^\top\mathbf{W}\mathbf{F}$ *the following identity holds:*

$$\mathbf{F}^\top \left(\mathbf{FDF}^\top + \mathbf{W}^{-1}\right)^{-1} \mathbf{F} = \left(\mathbf{M}^+ + \mathbf{M}^+\mathbf{MDMM}^+\right)^+,$$

where $\mathbf{M}^+$ *denotes the Moore-Penrose inverse (see, e. g., Schott (1997), p. 171) of* $\mathbf{M}$*. If* $\mathbf{M}$ *is non-singular, the right-hand side simplifies to*

$$\left(\mathbf{M}^{-1} + \mathbf{D}\right)^{-1}.$$

For singular $\mathbf{M}$ *it holds that*

$$\lim_{\delta \to 0} \left((\mathbf{M} + \delta \mathbf{I}_p)^{-1} + \mathbf{D} \right)^{-1} = (\mathbf{M}^+ + \mathbf{M}^+ \mathbf{M} \mathbf{D} \mathbf{M} \mathbf{M}^+)^+ .$$

Proof: Without loss of generality we can assume $\mathbf{W} = \mathbf{I}_k$, since with $\tilde{\mathbf{F}} := \sqrt{\mathbf{W}}\mathbf{F}$, where $\sqrt{\mathbf{W}}$ denotes a non-singular matrix chosen such that $\sqrt{\mathbf{W}}\sqrt{\mathbf{W}}^\top = \mathbf{W}$, the left-hand side can be expressed as

$$\mathbf{F}^\top \left(\mathbf{F} \mathbf{D} \mathbf{F}^\top + \mathbf{W}^{-1} \right)^{-1} \mathbf{F} = \tilde{\mathbf{F}}^\top \left(\tilde{\mathbf{F}} \mathbf{D} \tilde{\mathbf{F}}^\top + \mathbf{I}_k \right)^{-1} \tilde{\mathbf{F}}.$$

Using a singular value decomposition (see, e. g., Searle (1982), p. 316, or Schott (1997), p. 131) $\mathbf{F}$ can be expressed as

$$\mathbf{F} = \mathbf{P}^\top \begin{pmatrix} \mathbf{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{Q}$$

with orthogonal matrices $\mathbf{P}$ and $\mathbf{Q}$ and a non-singular diagonal matrix $\mathbf{\Lambda}$ of size $r := \operatorname{rank}(\mathbf{F})$. Depending on the rank the $\mathbf{0}$-blocks in the matrix might vanish. Therefore,

$$\mathbf{F}^\top \left(\mathbf{F} \mathbf{D} \mathbf{F}^\top + \mathbf{I}_k \right)^{-1} \mathbf{F} = \mathbf{F}^\top \left(\mathbf{P}^\top \begin{pmatrix} \mathbf{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{Q} \mathbf{Q}^\top \begin{pmatrix} \mathbf{K} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{C} \end{pmatrix} \mathbf{Q} \mathbf{Q}^\top \begin{pmatrix} \mathbf{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{P} + \mathbf{I}_k \right)^{-1} \mathbf{F},$$

where $\mathbf{D}$ is expressed as $\mathbf{D} = \mathbf{Q}^\top \tilde{\mathbf{D}} \mathbf{Q}$

$$\text{with} \quad \tilde{\mathbf{D}} := \begin{pmatrix} \mathbf{K} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{C} \end{pmatrix} = \mathbf{Q} \mathbf{D} \mathbf{Q}^\top,$$

and the block $\mathbf{K}$ having the same size r as $\mathbf{\Lambda}$. Applying some matrix algebra leads to

$$\begin{aligned}
\mathbf{F}^\top \left(\mathbf{P}^\top \begin{pmatrix} \mathbf{\Lambda K \Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{P} + \mathbf{I}_k \right)^{-1} \mathbf{F} &= \mathbf{F}^\top \mathbf{P}^\top \begin{pmatrix} \mathbf{\Lambda K \Lambda} + \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{k-r} \end{pmatrix}^{-1} \mathbf{P} \mathbf{F} \\
&= \mathbf{Q}^\top \begin{pmatrix} \mathbf{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} (\mathbf{\Lambda K \Lambda} + \mathbf{I}_r)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{k-r} \end{pmatrix} \begin{pmatrix} \mathbf{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{Q} \\
&= \mathbf{Q}^\top \begin{pmatrix} (\mathbf{K} + \mathbf{\Lambda}^{-2})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{Q} \\
&= \left(\mathbf{Q}^\top \left(\begin{pmatrix} \mathbf{\Lambda}^{-2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + \begin{pmatrix} \mathbf{K} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \right) \mathbf{Q} \right)^+ .
\end{aligned}$$

Since

$$\mathbf{M} = \mathbf{Q}^\top \begin{pmatrix} \mathbf{\Lambda}^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{Q}, \qquad \mathbf{M}^+ = \mathbf{Q}^\top \begin{pmatrix} \mathbf{\Lambda}^{-2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{Q} \qquad \text{and} \qquad \mathbf{M}\mathbf{M}^+ = \mathbf{Q}^\top \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{Q},$$

it follows that

$$\begin{aligned}
\mathbf{F}^\top \left(\mathbf{F} \mathbf{D} \mathbf{F}^\top + \mathbf{I}_k \right)^{-1} \mathbf{F} &= \left(\mathbf{M}^+ + \mathbf{Q}^\top \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{K} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{C} \end{pmatrix} \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{Q} \right)^+ \\
&= (\mathbf{M}^+ + \mathbf{M}^+ \mathbf{M} \mathbf{D} \mathbf{M} \mathbf{M}^+)^+ .
\end{aligned}$$

For non-singular $\mathbf{M}$ all the Moore-Penrose inverses coincide with the ordinary matrix inverses.

To show the last part of the lemma, let $\mathbf{M}$ be singular. Then

$$\begin{aligned}\left((\mathbf{M}+\delta\mathbf{I}_p)^{-1}+\mathbf{D}\right)^{-1} &= \mathbf{Q}^\top\left(\begin{pmatrix}(\mathbf{\Lambda}^2+\delta\mathbf{I}_r)^{-1}+\mathbf{K} & \mathbf{B}\\ \mathbf{B}^\top & \frac{1}{\delta}\mathbf{I}_{p-r}+\mathbf{C}\end{pmatrix}^{-1}\right)\mathbf{Q}\\ &= \mathbf{Q}^\top\begin{pmatrix}\mathbf{T}_{11} & \mathbf{T}_{12}\\ \mathbf{T}_{21} & \mathbf{T}_{22}\end{pmatrix}\mathbf{Q}.\end{aligned}$$

With formulas for the inverse of partitioned matrices (Schott (1997), Theorem 7.1), the blocks $\mathbf{T}_{11}$, $\mathbf{T}_{22}$, $\mathbf{T}_{12}$, and $\mathbf{T}_{21}$ of appropriate size can be expressed as

$$\begin{aligned}\mathbf{T}_{11} &= \left[\left(\mathbf{\Lambda}^2+\delta\mathbf{I}_r\right)^{-1}+\mathbf{K}-\mathbf{B}\left(\tfrac{1}{\delta}\mathbf{I}_{p-r}+\mathbf{C}\right)^{-1}\mathbf{B}^\top\right]^{-1}\overset{\delta\to 0}{\longrightarrow}\left(\mathbf{\Lambda}^{-2}+\mathbf{K}\right)^{-1},\\ \mathbf{T}_{22} &= \left[\tfrac{1}{\delta}\mathbf{I}_{k-r}+\mathbf{C}-\mathbf{B}^\top\left(\left(\mathbf{\Lambda}^2+\delta\mathbf{I}_r\right)^{-1}+\mathbf{K}\right)^{-1}\mathbf{B}\right]^{-1}\overset{\delta\to 0}{\longrightarrow}0,\\ \mathbf{T}_{12} &= \mathbf{T}_{21}^\top=-\left((\mathbf{\Lambda}^2+\delta\mathbf{I}_r)^{-1}+\mathbf{K}\right)^{-1}\mathbf{B}\mathbf{T}_{22}\overset{\delta\to 0}{\longrightarrow}0\end{aligned}$$

leading to

$$\lim_{\delta\to 0}\mathbf{Q}^\top\begin{pmatrix}\mathbf{T}_{11} & \mathbf{T}_{12}\\ \mathbf{T}_{21} & \mathbf{T}_{22}\end{pmatrix}\mathbf{Q}=\mathbf{Q}^\top\begin{pmatrix}\left(\mathbf{\Lambda}^{-2}+\mathbf{K}\right)^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{Q}=\left(\mathbf{M}^+ + \mathbf{M}^+\mathbf{M}\mathbf{D}\mathbf{M}\mathbf{M}^+\right)^+.$$

□

In the context considered here the lemma implies that the $\boldsymbol{\beta}$-block of the information matrix of an individual design ξ_i can be represented by

$$\mathfrak{M}_{\text{ind}}^{\boldsymbol{\beta}}(\xi_i)=\frac{1}{\sigma^2}\mathbf{F}_i^\top\left(\mathbf{F}_i\mathbf{D}\mathbf{F}_i^\top+\mathbf{W}_i^{-1}\right)^{-1}\mathbf{F}_i=\frac{1}{\sigma^2}\left(\mathbf{M}_i^+ + \mathbf{M}_i^+\mathbf{M}_i\mathbf{D}\mathbf{M}_i\mathbf{M}_i^+\right)^+,$$

where $\mathbf{M}_i \coloneqq \mathbf{M}_i(\xi_i) \coloneqq \mathbf{F}_i^\top\mathbf{W}_i\mathbf{F}_i$ is the (non-standardized) information matrix of the corresponding fixed effect model (i. e., the ordinary linear model without the random effects). Note that different to the definition in Chapter 3, from now on $\mathbf{M}$ will be used to denote the non-standardized information matrix.

For regular designs the expression for the $\boldsymbol{\beta}$-block simplifies to $\frac{1}{\sigma^2}(\mathbf{M}_i^{-1}+\mathbf{D})^{-1}$, which can already be found in Liski et al. (2002) and Demidenko (2004).

For sake of clarity we will use $\mathfrak{M}_{\text{ind}}^{\boldsymbol{\beta}}(\mathbf{M}_i)=\mathfrak{M}_{\text{ind}}^{\boldsymbol{\beta}}(\mathbf{M}(\xi_i))$ synonymously for $\mathfrak{M}_{\text{ind}}^{\boldsymbol{\beta}}(\xi_i)$.

The extension to the general mixed model is straight forward:

$$\mathfrak{M}_{\text{ind}}^{\boldsymbol{\beta}}(\xi_i)=\frac{1}{\sigma^2}\mathbf{K}_i^\top\mathbf{F}_i^\top\left(\mathbf{F}_i\mathbf{D}\mathbf{F}_i^\top+\mathbf{W}_i^{-1}\right)^{-1}\mathbf{F}_i\mathbf{K}_i=\frac{1}{\sigma^2}\mathbf{K}_i^\top\left(\mathbf{M}_i^+ + \mathbf{M}_i^+\mathbf{M}_i\mathbf{D}\mathbf{M}_i\mathbf{M}_i^+\right)^+\mathbf{K}_i \tag{6.3}$$

and

$$\mathfrak{M}_{\text{ind}}^{\boldsymbol{\beta}}(\xi_i)=\frac{1}{\sigma^2}\mathbf{K}_i^\top(\mathbf{M}_i^{-1}+\mathbf{D})^{-1}\mathbf{K}_i \tag{6.4}$$

for the regular case.

With this we get the population parameter block of the information matrix for an elementary design in the mixed model as a function of the information matrix for the same design in the

corresponding ordinary linear model. As we will later see in Chapter 7, this function is monotone and convex with respect to the Loewner ordering of non-negative definite matrices. This will allow as to reduce some design problems in the mixed model case to problems already solved for ordinary linear models.

6.2.2 Variance parameter block (for diagonal D)

For a different representations of the blocks for the variance parameters in the information matrix, we will only consider the case that the covariance structure of the random effects $\mathbf{D}$ is a diagonal matrix, as this is the most relevant case in practice. We use the natural parameterization of $\mathbf{D}$, that is, $\mathbf{D} = \operatorname{diag}(\boldsymbol{\alpha})$. We can give representations for $\mathfrak{M}^{\boldsymbol{\alpha}}_{\text{ind}}$ and $\mathfrak{M}^{\boldsymbol{\alpha}\sigma^2}_{\text{ind}}$ that express these blocks as functions of $\mathfrak{M}^{\boldsymbol{\beta}}_{\text{ind}}$ and with this as functions of the information matrix $\mathbf{M}$ of the corresponding ordinary linear model. This then allows us to calculate the complete information matrix also for approximate designs.

Lemma 6.5 *In the RCR model the block in the information matrix representing the parameter vector $\boldsymbol{\alpha}$ can be written as*

$$\mathfrak{M}^{\boldsymbol{\alpha}}_{\text{ind}} = \frac{\sigma^4}{2}\mathfrak{M}^{\boldsymbol{\beta}}_{\text{ind}} \circ \mathfrak{M}^{\boldsymbol{\beta}}_{\text{ind}},$$

where $\circ$ denotes the Hadamard product, that is, the element-by-element product of two matrices.

Proof: As described in Section 4.4 Eq. (4.19), the entry in the kth row and lth column is

$$(\mathfrak{M}^{\boldsymbol{\alpha}}_{\text{ind}})_{kl} = \frac{1}{2}\operatorname{tr}\left[\mathbf{V}_i(\boldsymbol{\alpha})^{-1}\frac{\partial \mathbf{V}_i(\boldsymbol{\alpha})}{\partial \alpha_k}\mathbf{V}_i(\boldsymbol{\alpha})^{-1}\frac{\partial \mathbf{V}_i(\boldsymbol{\alpha})}{\partial \alpha_l}\right].$$

Since $\mathbf{V}_i = \mathbf{I}_{m_i} + \mathbf{F}_i\mathbf{D}\mathbf{F}_i^\top$ and $\mathbf{D}$ was assumed to be diagonal with the vector $\boldsymbol{\alpha}$ as entries on the diagonal,

$$\frac{\partial \mathbf{V}_i(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}_k} = \mathbf{F}_i\begin{pmatrix} 0 & & & \\ & \ddots & & \\ & & 1 & \\ & & & \ddots \\ & & & & 0 \end{pmatrix}\mathbf{F}_i^\top = \mathbf{F}_i\operatorname{diag}(\mathbf{e}_k)\mathbf{F}_i^\top. \tag{6.5}$$

Therefore,

$$\begin{aligned}
(\mathfrak{M}^{\alpha}_{\text{ind}})_{kl} &= \frac{1}{2}\operatorname{tr}\left[\mathbf{V}_i(\alpha)^{-1}\mathbf{F}_i\operatorname{diag}(\mathbf{e}_k)\mathbf{F}_i^\top\mathbf{V}_i(\alpha)^{-1}\mathbf{F}_i\operatorname{diag}(\mathbf{e}_l)\mathbf{F}_i^\top\right] \\
&= \frac{1}{2}\operatorname{tr}\left[\mathbf{F}_i^\top\mathbf{V}_i(\alpha)^{-1}\mathbf{F}_i\operatorname{diag}(\mathbf{e}_k)\mathbf{F}_i^\top\mathbf{V}_i(\alpha)^{-1}\mathbf{F}_i\operatorname{diag}(\mathbf{e}_l)\right] \\
&= \frac{1}{2}\operatorname{tr}\left[\sigma^2\mathfrak{M}^{\boldsymbol{\beta}}_{\text{ind}}\operatorname{diag}(\mathbf{e}_k)\sigma^2\mathfrak{M}^{\boldsymbol{\beta}}_{\text{ind}}\operatorname{diag}(\mathbf{e}_l)\right] = \frac{\sigma^4}{2}\left((\mathfrak{M}^{\boldsymbol{\beta}}_{\text{ind}})_{kl}\right)^2.
\end{aligned}$$

□

Lemma 6.6 *In the RCR model the block in the information matrix accounting for the interaction of the estimation of $\boldsymbol{\alpha}$ and σ^2 can be written as*

$$\mathfrak{M}^{\boldsymbol{\alpha}\sigma^2}_{\text{ind}} = \frac{1}{2}\left((\mathfrak{M}^{\boldsymbol{\beta}}_{\text{ind}})_{\nu\nu}\right)_{\nu=1}^{p}.$$

Proof: The $\mathfrak{M}_{\text{ind}}^{\alpha\sigma^2}$ block is just a $p \times 1$-matrix. The entry of the lth row is calculated as

$$(\mathfrak{M}_{\text{ind}}^{\alpha\sigma^2})_l = \frac{1}{2}\text{tr}\left[\frac{1}{\sigma^2}\mathbf{V}_i(\boldsymbol{\alpha})^{-1}\frac{\partial \mathbf{V}_i(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}_l}\right]$$

(see Section 4.4 Eq. (4.19)). With Eq. (6.5) this can be expressed as

$$\begin{aligned}(\mathfrak{M}_{\text{ind}}^{\alpha\sigma^2})_l &= \frac{1}{2}\text{tr}\left[\frac{1}{\sigma^2}\mathbf{V}_i(\boldsymbol{\alpha})^{-1}\mathbf{F}_i\text{diag}(\mathbf{e}_l)\mathbf{F}_i^\top\right] \\ &= \frac{1}{2}\text{tr}\left[\frac{1}{\sigma^2}\mathbf{F}_i^\top\mathbf{V}_i(\boldsymbol{\alpha})^{-1}\mathbf{F}_i\text{diag}(\mathbf{e}_l)\right] \\ &= \frac{1}{2}\text{tr}\left[\mathfrak{M}_{\text{ind}}^{\boldsymbol{\beta}}\text{diag}(\mathbf{e}_l)\right] = \frac{1}{2}(\mathfrak{M}_{\text{ind}}^{\boldsymbol{\beta}})_{ll}.\end{aligned}$$

□

For the general mixed model similar formulas for the blocks can be derived. As $\mathbf{V}_i$ does not depend on $\mathbf{K}_i$, we get

$$\mathfrak{M}_{\text{ind}}^{\boldsymbol{\alpha}}(\xi_i) = \frac{\sigma^4}{2}\tilde{\mathfrak{M}}_{\text{ind}}^{\boldsymbol{\beta}}(\xi_i) \circ \tilde{\mathfrak{M}}_{\text{ind}}^{\boldsymbol{\beta}}(\xi_i)$$

and

$$\mathfrak{M}_{\text{ind}}^{\alpha\sigma^2}(\xi_i) = \frac{1}{2}\left(\left(\tilde{\mathfrak{M}}_{\text{ind}}^{\boldsymbol{\beta}}(\xi_i)\right)_{vv}\right)_{v=1}^{p},$$

where $\tilde{\mathfrak{M}}_{\text{ind}}^{\boldsymbol{\beta}}$ denotes the $\boldsymbol{\beta}$-block of the information matrix of the corresponding RCR model, where the $\mathbf{K}_i$ are omitted.

Altogether the complete individual information matrix of design ξ_i (with non-singular $\mathbf{M}(\xi_i)$) in the general mixed model can be expressed by

$$\mathfrak{M}_{\text{ind}}(\xi_i) = \begin{pmatrix} \frac{1}{\sigma^2}\mathbf{K}_i^\top(\mathbf{M}_i^{-1}+\mathbf{D})^{-1}\mathbf{K}_i & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}(\mathbf{M}_i^{-1}+\mathbf{D})^{-1}\circ(\mathbf{M}_i^{-1}+\mathbf{D})^{-1} & \frac{1}{2\sigma^2}\text{diag}\left[(\mathbf{M}_i^{-1}+\mathbf{D})^{-1}\right] \\ \mathbf{0} & \frac{1}{2\sigma^2}\left[\text{diag}((\mathbf{M}_i^{-1}+\mathbf{D})^{-1})\right]^\top & \frac{m_i}{2\sigma^4} \end{pmatrix},$$

where m_i is the number of observations in design ξ_i and $\mathbf{M}_i = \mathbf{M}(\xi_i) = \mathbf{F}_i^\top\mathbf{W}_i\mathbf{F}_i$ is the information matrix of design ξ_i in the corresponding ordinary linear model.

As already mentioned, the information matrix of a population design $\zeta = \begin{pmatrix}\xi_1, & \ldots, & \xi_l \\ g_1, & \ldots, & g_l\end{pmatrix}$ is the weighted sum of the individual information matrices, that is,

$$\mathfrak{M}_{\text{pop}}(\zeta) = \sum_{i=1}^{l} g_i\mathfrak{M}_{\text{ind}}(\xi_i).$$

It is, however, important to note that none of the blocks of the individual information matrix can be expressed as the sum of the information matrices of the respective one-point designs,

that is, for $\xi_i = \begin{pmatrix} x_{i1} & \dots & x_{ik_i} \\ m_{i1} & \dots & m_{ik_i} \end{pmatrix}$

$$\mathfrak{M}_{\text{ind}}(\xi_i) \neq \sum_{v=1}^{k_i} m_{iv} \mathfrak{M}_{\text{ind}}(\delta_{\{x_{iv}\}}),$$

where $\delta_{\{x\}}$ denotes the one-point design measure in x. This is due to the dependence between the observations within one individual induced by $\mathbf{D}$.

6.3 Criteria

In principle the same criteria as for the ordinary linear model can also be used for the mixed models. In this section we will briefly discuss which criteria might be useful for which purpose. In Section 6.3.1 we concentrate on criteria based only on the population parameter block, whereas in Section 6.3.2 criteria are considered that are based on the complete information matrix.

6.3.1 Criteria based on the population parameter block

In this section criteria based only on the population parameter block of the information matrix are discussed. Due to the block diagonal structure of the information matrix, it makes no difference whether the remaining variance parameters are known or unknown, besides the general dependence of the population parameter block on the assumed or known matrix $\mathbf{D}$. In general, the same criteria as described in Section 3.3 for the ordinary linear model can also be applied for the assessment of the quality of a design for the estimation of the population parameters of the mixed model. However, one needs to be careful how the criteria have to be interpreted. The G-criterion

$$\Phi(\mathfrak{M}_{\text{pop}}^{\boldsymbol{\beta}}) = \max_{x \in \mathcal{X}} \mathbf{f}(x)^\top \mathfrak{M}_{\text{pop}}^{\boldsymbol{\beta}}{}^{-1} \mathbf{f}(x),$$

for example, is used to minimize the maximal variance of the prediction of the response of a "typical" individual, that is, an individual with $\boldsymbol{\beta}_i = \boldsymbol{\beta}$. A similar interpretation can be found for the IMSE-criterion.

In Section 3.3 we already mentioned, but did not discuss, criteria for the estimation of real-valued nonlinear functions $h(\boldsymbol{\beta})$ of the parameter vector. Similar as for the estimation of a linear function, the quality of a design can be assessed by a c-criterion, that is, by a criterion of the form

$$\Phi(\mathfrak{M}_{\text{pop}}^{\boldsymbol{\beta}}) = \mathbf{c}^\top \mathfrak{M}_{\text{pop}}^{\boldsymbol{\beta}}{}^{-1} \mathbf{c},$$

where $\mathbf{c}$ is here chosen as the column vector

$$\mathbf{c} = \left. \frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}.$$

This choice can be motivated by a linear approximation using the so-called δ-method (see e. g.,

Rao (1973), pp. 385), which is described in the following. Let $\hat{\boldsymbol{\beta}}^{(n)}$ denote the ML estimator of $\boldsymbol{\beta}$ if n individuals are observed using the population design ζ. Then, under the assumption of consistency of the estimator, we have

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{(n)} - \boldsymbol{\beta}\right) \xrightarrow{\mathcal{D}} \mathcal{N}_p(0, \mathfrak{M}^{\beta}(\zeta)^{-1}),$$

where $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution. If the function h is differentiable it can be shown by using a first-order Taylor approximation that

$$\sqrt{n}\left(h(\hat{\boldsymbol{\beta}}^{(n)}) - h(\boldsymbol{\beta})\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\top}} \mathfrak{M}^{\beta}(\zeta)^{-1} \frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right).$$

This means that if the variance of $h(\hat{\boldsymbol{\beta}}^{(n)})$ exists, it can be approximated by

$$\mathrm{Var}(h(\hat{\boldsymbol{\beta}}^{(n)})) \simeq \frac{1}{n} \frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\top}} \mathfrak{M}^{\beta}(\zeta)^{-1} \frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}, \tag{6.6}$$

which should be minimized to obtain a good design for the estimation of $h(\boldsymbol{\beta})$, otherwise the equation has to be interpreted as asymptotic variance. As the right-hand side of Eq. (6.6) still depends on the unknown vector $\boldsymbol{\beta}$, the usual approach for designing the experiment is to substitute $\boldsymbol{\beta}$ by a best guess $\boldsymbol{\beta}_0$ (as in Section 5.3 for the information matrix in the non-linear model) and to use

$$\Phi(\mathfrak{M}^{\beta}(\zeta)) = \left.\frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\top}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \mathfrak{M}^{\beta}(\zeta)^{-1} \left.\frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0},$$

which has the form of a c-criterion.

Example 6.7 *A typical example occurring in the field of pharmacokinetics is the estimation of the AUC (see Section 2.2). As the AUC is the integral of the concentration curve, it is a (usually nonlinear) function of the model parameters. In the case of a simple one-compartment model with absorption, where the concentration at time x is given by*

$$C(x) = D \frac{k_a}{V_c(k_a - \frac{Cl}{V_c})} \left(e^{-\frac{Cl}{V_c}x} - e^{-k_a x}\right)$$

the AUC is

$$AUC(k_a, Cl, V_c; D) = \int_0^{\infty} C(x)\, dx = \frac{D}{Cl}.$$

Hence, the quality of the estimation of the AUC only depends on the quality of the estimation of the clearance Cl.

One has to keep in mind that the theoretical c-optimal designs may be concentrated on fewer points than the number of population parameters and can hence be singular. In the case of a nonlinear model, it can then happen that the function of interest might not be estimable under the "optimal" design. An example for this can be found in Atkinson et al. (1993), where this problem is described for a one-compartment model without random effects.

6.3.2 Criteria based on the whole information matrix

Some of the criteria described in Section 3.3 for the ordinary linear model and in Section 6.3.1 for the estimation of the population parameters, like the D-criterion, can be applied analogously to the whole information matrix, although it should be discussed, whether the variance parameters should be treated in the same way as the population parameters. In the case that a very general covariance matrix $\mathbf{D}$ is used, there are $p(p+1)/2+1$ variance parameters to be estimated. If then the D-criterion is applied to the complete information matrix, a much higher importance is put onto the variance components than on the population parameters due to the big difference in the number of parameters.

For other criteria, like the G-criterion or the IMSE-criterion, there is no directly applicable extension to the whole information matrix.

In principle, c-criteria based on the complete information matrix are possible. For this, however, one needs reasonable functions that are based on both, the population and the variance parameters. In the following we give an idea for the construction of "reasonable" c-criteria based on the complete information matrix.

One might often not only be interested in the estimation of a (linear) function $h(\boldsymbol{\beta})$ of the population parameters, but also want to estimate certain quantiles of this function applied to the individual parameters $h(\boldsymbol{\beta}_i)$. For example, one might want to know the 95%-quantile of the individual AUC to be sure that the individual exposure to the drug is not to high. If $h(\boldsymbol{\beta}) = \mathbf{c}^\top\boldsymbol{\beta}$ and if we define $H_i = \mathbf{c}^\top\boldsymbol{\beta}_i$, then

$$H_i \sim \mathcal{N}(\mathbf{c}^\top\boldsymbol{\beta}, \sigma^2\mathbf{c}^\top\mathbf{D}(\boldsymbol{\alpha})\mathbf{c})$$

and hence the q-quantile of H_i is

$$H_i^q = \mathbf{c}^\top\boldsymbol{\beta} + z_q\sqrt{\sigma^2\mathbf{c}^\top\mathbf{D}(\boldsymbol{\alpha})\mathbf{c}},$$

where z_q denotes the q-quantile of the standard normal distribution. If we assume $\mathbf{D}$ to be diagonal with natural parameterization $\mathbf{D}(\boldsymbol{\alpha}) = \operatorname{diag}(\boldsymbol{\alpha})$, we get

$$H_i^q = \mathbf{c}^\top\boldsymbol{\beta} + z_q\sqrt{\sigma^2(\mathbf{c}\circ\mathbf{c})^\top\boldsymbol{\alpha}} =: g(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2),$$

which is a nonlinear function of all parameters. To find a good design for the estimation of this quantile, the variance $\operatorname{Var}(g(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\sigma}^2))$ (if it exists) should be minimized. The non-linearity can be handled as in the previous section by using the δ-method, that is, by using

$$\operatorname{Var}\left(g(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\sigma}^2)\right) \simeq \frac{1}{n}\frac{\partial g}{\partial(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2)}\mathfrak{M}_{\text{pop}}^{-1}\frac{\partial g}{\partial(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2)^\top}.$$

For the described situation we get

$$\begin{aligned}\frac{\partial g}{\partial \boldsymbol{\beta}} &= \mathbf{c},\\ \frac{\partial g}{\partial \boldsymbol{\alpha}} &= -z_q \frac{\sqrt{\sigma^2}}{2\sqrt{(\mathbf{c}\circ\mathbf{c})^\top \boldsymbol{\alpha}}}(\mathbf{c}\circ\mathbf{c}),\\ \frac{\partial g}{\partial \sigma^2} &= -z_q \frac{\sqrt{(\mathbf{c}\circ\mathbf{c})^\top \boldsymbol{\alpha}}}{2\sqrt{\sigma^2}},\end{aligned}$$

which are the coefficients for the c-criterion to be used. Non-linear functions of the population parameter vector can be treated by first linearizing the function around a best guess for the population parameters as described in Section 6.3.1 and then following the same procedure as just explained.

By the same principle a criterion similar to the IMSE-criterion can be constructed, where the integrated mean squared error of the prediction of the q-quantile of the individual responses over the design region $\mathcal{X}$ (cf. Section 3.3)

$$\mathrm{E}\left(\int_{\mathcal{X}} \left(\mathbf{f}(x)^\top \boldsymbol{\beta} + z_q\sqrt{\sigma^2 \mathbf{f}(x)^\top \mathbf{D}(\boldsymbol{\alpha})\mathbf{f}(x)} - \mathbf{f}(x)^\top \hat{\boldsymbol{\beta}} - z_q\sqrt{\hat{\sigma}^2 \mathbf{f}(x)^\top \mathbf{D}(\hat{\boldsymbol{\alpha}})\mathbf{f}(x)}\right)^2 \mu(dx)\right)$$

is to be minimized.

By linearizing this leads to a criterion of the form

$$\Phi(\mathfrak{M}_{\text{pop}}) = \mathrm{tr}(\mathbf{L}\mathfrak{M}_{\text{pop}}^{-1}),$$

where $\mathbf{L}$ has to be chosen as

$$\mathbf{L} = \int_{\mathcal{X}} \tilde{\mathbf{f}}(x)\tilde{\mathbf{f}}(x)^\top \mu(dx)$$

with

$$\begin{aligned}\tilde{\mathbf{f}}_1(x) &= -z_q \frac{\sqrt{\sigma^2}}{2\sqrt{(\mathbf{f}(x)\circ\mathbf{f}(x))^\top \boldsymbol{\alpha}}}(\mathbf{f}(x)\circ\mathbf{f}(x)),\\ \tilde{\mathbf{f}}_2(x) &= -z_q \frac{\sqrt{(\mathbf{f}(x)\circ\mathbf{f}(x))^\top \boldsymbol{\alpha}}}{2\sqrt{\sigma^2}},\\ \tilde{\mathbf{f}}(x) &= \left(\mathbf{f}(x)^\top, \tilde{\mathbf{f}}_1(x)^\top, \tilde{\mathbf{f}}_2(x)\right)^\top.\end{aligned}$$

Depending on the purpose of the experiment one can also apply different criteria to the two different blocks and combine them to a weighted criterion. One should, however, always try to keep the criterion interpretable.

One example for such a criterion could be a weighted standardized D-criterion like

$$\Phi(\mathfrak{M}_{\text{pop}}(\zeta)) = w_1 \frac{\det\left(\mathfrak{M}_{\text{pop}}^{\boldsymbol{\beta}}(\zeta)^{-1}\right)^{\frac{1}{p}}}{\min_{\zeta'\in\Delta}\det\left(\mathfrak{M}_{\text{pop}}^{\boldsymbol{\beta}}(\zeta')^{-1}\right)^{\frac{1}{p}}} + w_2 \frac{\det\left(\mathfrak{M}_{\text{pop}}^{(\alpha,\sigma^2)}(\zeta)^{-1}\right)^{\frac{1}{r}}}{\min_{\zeta'\in\Delta}\det\left(\mathfrak{M}_{\text{pop}}^{(\alpha,\sigma^2)}(\zeta')^{-1}\right)^{\frac{1}{r}}},$$

where $\mathfrak{M}_{\text{pop}}^{(\alpha,\sigma^2)}$ is used to denote the block in the information matrix accounting for all variance parameters, that is, for $\boldsymbol{\alpha}$ and σ^2, and r denotes the number of variance parameters. The weights w_1 and w_2 can be chosen depending on how much importance one wants to give each of the blocks. The idea of standardized optimality criteria in general is described in Dette (1997).

7 Single-Group and Group-wise Identical Approximate Designs for Population Parameter Estimation

In this chapter we will give special results for designs for the estimation of the population parameter vector. We will consider criteria that are only based on the $\boldsymbol{\beta}$-block of the information matrix as described in Section 6.3.1. Theoretical results in this context can be found in the book by Fedorov and Hackl (1997) where an equivalence theorem for the D-optimality in RCR models can be found, to which we will come back in Chapter 8. In the book by Liski et al. (2002) and the dissertation of Luoma (2000) special results for polynomial regression models are derived.

As we have seen in Chapter 6, designs for mixed models can be defined in a very general way leading to a very complex set of possible population designs, which makes the search for optimal designs quite difficult. In this chapter we will show how the set of designs can be restricted without reducing the quality of the optimal designs. Section 7.1 is concerned with single-group designs for the RCR model. The derived result is extended to the general mixed model in the Sections 7.2 and 7.3.

Throughout the chapter we can without loss of generality assume that $\sigma^2 = 1$. This is due to the fact that $\frac{1}{\sigma^2}$ appears only as a constant factor in front of the information matrix and has, hence, no influence on the design optimization if a sensible criterion is used.

7.1 Single-group designs in RCR models

As we have already seen in Section 4.3.1, the estimation of the population parameters in the linear RCR model simplifies dramatically if all individuals are observed under the same conditions, that is, under the same elementary design, since then the weighted least squares estimator is identical to the ordinary one. Hence, from the point of view of analyzing the models it is favorable to use designs, where this is the case. We will call these designs *single-group designs*, as the individuals are not split up in different groups having different designs.

Besides the easier analysis, the single-group designs have another advantage. Following from Lemma 6.4 the population parameter block of the information of a regular single-group design $\zeta = \begin{pmatrix} \xi \\ 1 \end{pmatrix}$ can be written as

$$\mathfrak{M}^{\boldsymbol{\beta}}_{\text{pop}}(\zeta) = \mathfrak{M}^{\boldsymbol{\beta}}_{\text{ind}}(\xi) = (\mathbf{M}(\xi)^{-1} + \mathbf{D})^{-1}.$$

If now a linear criterion of the form $\Phi(\mathbf{M}) = \mathrm{tr}(\mathbf{L}\mathbf{M}^{-1})$ is used, we get

$$\Phi(\mathfrak{M}^{\boldsymbol{\beta}}_{\mathrm{pop}}(\zeta)) = \mathrm{tr}(\mathbf{L}(\mathbf{M}^{-1} + \mathbf{D})) = \mathrm{tr}(\mathbf{L}\mathbf{M}^{-1}) + \mathrm{tr}(\mathbf{L}\mathbf{D}).$$

The second summand has no influence on the design optimization with respect to Φ, as it is constant (see, e. g., Liski et al. (2002) or Entholzner et al. (2005)). Therefore, the optimal single-group design is identical to the one, which is obtained in the corresponding ordinary linear model. This fact also holds for singular designs if the c-criterion is used. Let $h(\boldsymbol{\beta}) = \mathbf{c}^\top\boldsymbol{\beta}$ be a linear function of the parameter vector. If $h(\boldsymbol{\beta})$ is identifiable under the single-group design $\zeta = \begin{pmatrix}\xi\\ 1\end{pmatrix}$, there exists a matrix $\mathbf{Q}$ such that $\mathbf{c}^\top = \mathbf{Q}\mathbf{M}$, where $\mathbf{M} := \mathbf{M}(\xi)$ is the information matrix of the corresponding ordinary linear model. We then get with Lemma 6.4

$$\begin{aligned}
\Phi(\mathfrak{M}^{\boldsymbol{\beta}}_{\mathrm{pop}}(\zeta)) &= \mathbf{c}^\top\mathfrak{M}^{\boldsymbol{\beta}}_{\mathrm{pop}}(\zeta)^+\mathbf{c} = \mathbf{c}^\top((\mathbf{M}^+ + \mathbf{M}^+\mathbf{M}\mathbf{D}\mathbf{M}\mathbf{M}^+)^+)^+\mathbf{c}\\
&= \mathbf{c}^\top\mathbf{M}^+\mathbf{c} + \mathbf{c}^\top\mathbf{M}^+\mathbf{M}\mathbf{D}\mathbf{M}\mathbf{M}^+\mathbf{c}\\
&= \mathbf{c}^\top\mathbf{M}^+\mathbf{c} + \mathbf{Q}\mathbf{M}\mathbf{M}^+\mathbf{M}\mathbf{D}\mathbf{M}\mathbf{M}^+\mathbf{M}\mathbf{Q}^\top\\
&= \mathbf{c}^\top\mathbf{M}^+\mathbf{c} + \mathbf{Q}\mathbf{M}\mathbf{D}\mathbf{M}\mathbf{Q}^\top = \mathbf{c}^\top\mathbf{M}^+\mathbf{c} + \mathbf{c}^\top\mathbf{D}\mathbf{c},
\end{aligned}$$

that is, again the second summand is independent of the chosen design. Thus the c-optimal single-group design is the same one as in the ordinary linear model. Note that this is not true for non-linear criteria like the D-criterion.

Now, the question arises if, despite this welcome properties of single-group designs, it might be better to leave the single-group designs and to assign the individuals to different designs to get a lower value of the used criterion, that is, to get a more efficient design.

To investigate this question, we formally define the following two classes of approximate designs:

(i) The class of *single-group designs* of the form $\zeta = \begin{pmatrix}\xi\\ 1\end{pmatrix}$, where all individuals are observed under the same approximate individual design ξ, and the number of observations per individual is m. We denote this class with Δ^m_A.

(ii) The class of *more-group designs* of the form $\zeta = \begin{pmatrix}\xi_1 & \dots & \xi_l\\ g_1 & \dots & g_l\end{pmatrix}$, where the individuals can be observed under different approximate individual designs, and the mean number of observations per individual $\sum g_i m_i$ is fixed to m. We denote this class with Δ^m_B.

Note that Δ^m_A is a subset of Δ^m_B.

We make the following minor assumptions for the criterion Φ under consideration. Let in the following $nnd(\mathbb{R}^{p\times p})$ denote the set of symmetric non-negative definite matrices in $\mathbb{R}^{p\times p}$.

Assumption 7.1 *1. Φ is a real-valued function defined on $nnd(\mathbb{R}^{p\times p})$, that is, $\Phi : nnd(\mathbb{R}^{p\times p}) \to (-\infty, \infty]$.*

2. Φ *is monotone (with respect to the Loewner partial ordering) on* $nnd(\mathbb{R}^{p\times p})$ *in the sense that* $\mathbf{M}_1 \geq \mathbf{M}_2 \Rightarrow \Phi(\mathbf{M}_1) \leq \Phi(\mathbf{M}_2)$ *for* $\mathbf{M}_1, \mathbf{M}_2 \in nnd(\mathbb{R}^{p\times p})$.

Note that the assumptions are satisfied for most of the common criteria, including the D-criterion and all linear criteria.

We will show that for a criterion fulfilling Assumption 7.1 a design that is Φ-optimal in the class of single-group designs Δ_A^m, is also Φ-optimal in the much larger class of more-group designs Δ_B^m.

For the proof of this result we need the following lemma.

Lemma 7.2 (cf. Schmelter (2006b)) *Let* $\mathbf{M}_1$, $\mathbf{M}_2$ *be symmetric non-negative definite matrices. Then*

$$\mathfrak{M}_{\text{ind}}^{\boldsymbol{\beta}}(\alpha\mathbf{M}_1 + (1-\alpha)\mathbf{M}_2) \geq \alpha\mathfrak{M}_{\text{ind}}^{\boldsymbol{\beta}}(\mathbf{M}_1) + (1-\alpha)\mathfrak{M}_{\text{ind}}^{\boldsymbol{\beta}}(\mathbf{M}_2)$$

holds with respect to the Loewner partial ordering of symmetric non-negative definite matrices.

Proof: The following holds for symmetric positive definite matrices $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{D}$:

$$\begin{aligned}\left[(\alpha\mathbf{M}_1 + (1-\alpha)\mathbf{M}_2)^{-1} + \mathbf{D}\right]^{-1} &= \mathbf{D}^{-1} - \mathbf{D}^{-1}\left[\alpha\mathbf{M}_1 + (1-\alpha)\mathbf{M}_2 + \mathbf{D}^{-1}\right]^{-1}\mathbf{D}^{-1} \\ &\geq \mathbf{D}^{-1} - \mathbf{D}^{-1}\left[\alpha(\mathbf{M}_1 + \mathbf{D}^{-1})^{-1} + (1-\alpha)(\mathbf{M}_2 + \mathbf{D}^{-1})^{-1}\right]\mathbf{D}^{-1} \\ &= \alpha(\mathbf{M}_1^{-1} + \mathbf{D})^{-1} + (1-\alpha)(\mathbf{M}_2^{-1} + \mathbf{D})^{-1},\end{aligned}$$

where the equalities hold according to a general matrix equality (Schott (1997), Corollary 1.7.1) and the inequality because of $(\alpha\mathbf{A} + (1-\alpha)\mathbf{B})^{-1} \leq \alpha\mathbf{A}^{-1} + (1-\alpha)\mathbf{B}^{-1}$ for positive definite matrices $\mathbf{A}$ and $\mathbf{B}$ (see Fedorov and Hackl (1997), p. 107). The other situations (singular matrices) can be shown with the last part of Lemma 6.4 and a continuity argument. □

Using Lemma 7.2 we can now proof our statement:

Theorem 7.3 (cf. Schmelter (2006b))

Let $\zeta^* = \begin{pmatrix}\xi^* \\ 1\end{pmatrix} \in \Delta_A^m$ *be a design that is* Φ*-optimal in the class of single-group designs* Δ_A^m*, where* Φ *is a criterion function satisfying Assumption 7.1. Then* ζ^* *is also* Φ*-optimal in the larger class of more-group designs* Δ_B^m.

Proof: Let $\zeta^* = \begin{pmatrix}\xi^* \\ 1\end{pmatrix}$ be a Φ-optimal single-group design (with m observations per individual), then

$$\Phi(\mathfrak{M}_{\text{pop}}^{\boldsymbol{\beta}}(\zeta^*)) = \Phi(\mathfrak{M}_{\text{ind}}^{\boldsymbol{\beta}}(\xi^*)) \leq \Phi(\mathfrak{M}_{\text{ind}}^{\boldsymbol{\beta}}(\xi))$$

holds for all individual designs ξ with m observations.

Let $\tilde{\zeta} = \begin{pmatrix}\tilde{\xi}_1 & \cdots & \tilde{\xi}_l \\ g_1 & \cdots & g_l\end{pmatrix}$ be Φ-optimal in Δ_B^m, then because of Lemma 7.2

$$\mathfrak{M}_{\text{pop}}^{\boldsymbol{\beta}}(\tilde{\zeta}) = \sum_{i=1}^{l} g_i\mathfrak{M}_{\text{ind}}^{\boldsymbol{\beta}}(\tilde{\xi}_i) \leq \mathfrak{M}_{\text{ind}}^{\boldsymbol{\beta}}\left(\sum_{i=1}^{l} g_i\mathbf{M}(\tilde{\xi}_i)\right) = \mathfrak{M}_{\text{ind}}^{\boldsymbol{\beta}}(\tilde{\xi}),$$

where $\bar{\xi}$ is the individual design constructed as the convex combination (with weights $g_1, \ldots, g_l$) of the individual designs $\tilde{\xi}_1, \ldots, \tilde{\xi}_l$ and is hence also an individual design with m observations.

Therefore,

$$\Phi(\mathfrak{M}^{\beta}_{\text{pop}}(\tilde{\zeta})) \geq \Phi(\mathfrak{M}^{\beta}_{\text{ind}}(\bar{\xi})) \geq \Phi(\mathfrak{M}^{\beta}_{\text{ind}}(\xi^*)) = \Phi(\mathfrak{M}^{\beta}_{\text{pop}}(\zeta^*).$$

□

Note, that the result does not hold in general if only exact individual designs are allowed, as one can see in the following example.

Example 7.4 *We consider a simple linear regression model of the form*

$$Y_{ij} := \beta_0 + b_i + \beta_1 x_{ij} + \varepsilon_{ij}, \quad x_{ij} \in \mathcal{X} = [0,1]$$

where only the intercept is random. We set the mean number of observations per individual to $m = 3$. *Let* d_0 *denote the variance of* b_i. *It can be shown that then* $\zeta^* = \begin{pmatrix} \xi^* \\ 1 \end{pmatrix}$ *with* $\xi^* = \begin{pmatrix} 0 & 1 \\ 1.5 & 1.5 \end{pmatrix}$ *is an approximate D-optimal single-group design independent of* d_0.
If only exact single-group designs are allowed, however, either all individuals should be observed under the design $\xi_1 = \begin{pmatrix} 0 & 1 \\ 2 & 1 \end{pmatrix}$ *or under the design* $\xi_2 = \begin{pmatrix} 0 & 1 \\ 1 & 2 \end{pmatrix}$, *that is,* $\zeta_1 = \begin{pmatrix} \xi_1 \\ 1 \end{pmatrix}$ *and* $\zeta_2 = \begin{pmatrix} \xi_2 \\ 1 \end{pmatrix}$ *are both D-optimal in the set of exact single-group designs. The more-group design* $\zeta_3 = \begin{pmatrix} \xi_1 & \xi_2 \\ 0.5 & 0.5 \end{pmatrix}$, *however, has a higher efficiency.*

7.2 Group-wise identical designs in the general mixed model

The result of the previous section can be extended to the general mixed model from Section 4.2, where the response of individual i observed at x_{ij} is modeled by

$$Y_{ij} = \mathbf{f}(x_{ij})^\top \mathbf{K}_i \boldsymbol{\beta} + \mathbf{f}(x_{ij})^\top \mathbf{b}_i + \varepsilon_{ij}.$$

We assume that the factor defining the shapes of the matrices $\mathbf{K}_i$ is known and not under the control of the investigator. By this, the individuals under observation can be subdivided into different groups according to the shape of the respective $\mathbf{K}_i$. We assume that the number or the proportion of individuals in each group is known and fixed. As the individual information matrices depend on the matrices $\mathbf{K}_i$, which can differ from individual to individual, we cannot show that all individuals should be observed under the same design, although this is often true in the linear mixed model due to symmetry arguments. However, we can show that in each of the groups defined by the different shapes of the $\mathbf{K}_i$ only one elementary design is necessary, that is, in each of the groups a single-group design can be used.

For ease of notation, we assume that we have only two different shapes of $\mathbf{K}_i$, which we will denote by $\mathbf{K}_a$ and $\mathbf{K}_b$. The individuals in the sample population can hence be split up into

two different groups. This could be, for example, two groups of the sample population getting different doses of a drug as in Example 4.2 or receiving different treatments.

For each of the two groups a and b we can prescribe a population design $\zeta_a = \begin{pmatrix} \xi_1^a & \dots & \xi_{l_a}^a \\ g_1^a & \dots & g_{l_a}^a \end{pmatrix}$ and $\zeta_b = \begin{pmatrix} \xi_1^b & \dots & \xi_{l_b}^b \\ g_1^b & \dots & g_{l_b}^b \end{pmatrix}$, under which the individuals of the respective group should be observed.

As in the previous section, we formally define two classes of population designs:

1. $\tilde{\Delta}_A^{m_a,m_b}$, the class of designs, where all individuals of one group are observed under the same approximate individual design *(group-wise identical designs)*. Both ζ_a and ζ_b are single-group designs with m_a and m_b observations per individual, and
2. $\tilde{\Delta}_B^{m_a,m_b}$, the class of designs, where the experimental settings can be different for each of the individuals and only the mean number of observations per individual is prescribed for each group ($\sum g_v^a m_v^a = m_a$ and $\sum g_v^b m_v^b = m_b$).

We show that for an optimality criterion Φ that satisfies Assumption 7.1, a design that is Φ-optimal in class $\tilde{\Delta}_A^{m_a,m_b}$ of group-wise identical designs is also Φ-optimal in the larger second class $\tilde{\Delta}_B^{m_a,m_b}$ of general more-group designs.

To be able to use results from the previous section, we express the information matrices of the underlying model by means of the information matrices of the corresponding random coefficient regression model, that is, of the model where the matrices $\mathbf{K}_i$ are omitted. We hence consider the RCR model

$$\mathbf{Y}_i = \mathbf{F}_i\tilde{\boldsymbol{\beta}} + \mathbf{F}_i\tilde{\mathbf{b}}_i + \boldsymbol{\varepsilon}_i, \tag{7.1}$$

where $\tilde{\mathbf{b}}_i$, the vector of random effects of individual i, has the same covariance matrix $\mathbf{D}$ as in the originally considered mixed model. Let $\tilde{\mathfrak{M}}_{\text{ind}}^{\boldsymbol{\beta}}(\xi_i) = \mathbf{F}_i^\top(\mathbf{F}_i\mathbf{D}\mathbf{F}_i^\top + \mathbf{W}_i^{-1})^{-1}\mathbf{F}_i$ denote the individual information matrix of design ξ_i in model (7.1) and let $\tilde{\mathfrak{M}}_{\text{pop}}^{\boldsymbol{\beta}}(\zeta_a)$ ($\tilde{\mathfrak{M}}_{\text{pop}}^{\boldsymbol{\beta}}(\zeta_b)$ resp.) be the population information matrix of design ζ_a (ζ_b resp.) in that model. Then, the information matrices of the general model can be expressed as (cf. Eq. (6.4))

$$\begin{aligned} \mathfrak{M}_{\text{ind}}^{\boldsymbol{\beta}}(\xi_i) &= \mathbf{K}_i^\top\tilde{\mathfrak{M}}_{\text{ind}}^{\boldsymbol{\beta}}(\xi_i)\mathbf{K}_i, \\ \mathfrak{M}_{\text{pop}}^{\boldsymbol{\beta}}(\zeta) &= \frac{n_a}{n_a+n_b}\mathbf{K}_a^\top\tilde{\mathfrak{M}}_{\text{pop}}^{\boldsymbol{\beta}}(\zeta_a)\mathbf{K}_a + \frac{n_b}{n_a+n_b}\mathbf{K}_b^\top\tilde{\mathfrak{M}}_{\text{pop}}^{\boldsymbol{\beta}}(\zeta_b)\mathbf{K}_b, \end{aligned}$$

where n_a and n_b are the numbers of individuals in the two groups.

Note that in general the individual information matrices $\mathfrak{M}_{\text{ind}}^{\boldsymbol{\beta}}(\xi_i)$ are singular.

We now keep the design ζ_b for group b fixed. The contribution of group b to the information matrix is then

$$\mathbf{C} = \frac{n_b}{n_a+n_b}\mathbf{K}_b^\top\tilde{\mathfrak{M}}_{\text{pop}}^{\boldsymbol{\beta}}(\zeta_b)\mathbf{K}_b.$$

As we can see with the following lemma, for any criterion Φ satisfying Assumption 7.1 and any symmetric non-negative definite matrix $\mathbf{C}$ of appropriate dimension, the criterion Ψ, defined

as

$$\Psi(\mathbf{M}) = \Phi\left(\frac{n_a}{n_a + n_b}\mathbf{K}_a^\top \mathbf{M}\mathbf{K}_a + \mathbf{C}\right),$$

also satisfies Assumption 7.1, that is, Ψ is also a monotone criterion.

Lemma 7.5 (cf. Schmelter (2006a)) *Let $\Phi : nnd(\mathbb{R}^{p\times p}) \to \mathbb{R}$ be a monotone decreasing optimality criterion, $\mathbf{C}$ a non-negative definite $\mathbb{R}^{p\times p}$-matrix, $\mathbf{K}$ an arbitrary $\mathbb{R}^{p\times q}$-matrix, and $\lambda > 0 \in \mathbb{R}$. Then*

$$\Psi : nnd(\mathbb{R}^{q\times q}) \to \mathbb{R}, \quad \Psi(\mathbf{M}) = \Phi(\lambda \mathbf{K}^\top \mathbf{M}\mathbf{K} + \mathbf{C})$$

is also a monotone decreasing optimality criterion.

Proof: Let $\mathbf{M}_1$ and $\mathbf{M}_2$ be non-negative definite $p \times p$-matrices. Then

$$\begin{aligned} \mathbf{M}_1 \geq \mathbf{M}_2 \quad &\Rightarrow \quad \mathbf{K}^\top \mathbf{M}_1 \mathbf{K} \geq \mathbf{K}^\top \mathbf{M}_2 \mathbf{K} \\ &\Rightarrow \quad \lambda \mathbf{K}^\top \mathbf{M}_1 \mathbf{K} + \mathbf{C} \geq \lambda \mathbf{K}^\top \mathbf{M}_2 \mathbf{K} + \mathbf{C} \\ &\Rightarrow \quad \Phi(\lambda \mathbf{K}^\top \mathbf{M}_1 \mathbf{K} + \mathbf{C}) \leq \Phi(\lambda \mathbf{K}^\top \mathbf{M}_2 \mathbf{K} + \mathbf{C}). \end{aligned}$$

□

Hence, if we keep the design ζ_b for group b fixed, the search for a design ζ_a^* that minimizes $\Phi(\mathfrak{M}_{\text{pop}}^{\boldsymbol{\beta}}(\zeta_a, \zeta_b))$ in the mixed model is equivalent to the search of a design that minimizes $\Psi(\tilde{\mathfrak{M}}_{\text{pop}}^{\boldsymbol{\beta}}(\zeta_a))$ in the random coefficient regression model (6.5) if we set $\mathbf{C} = {}^{n_b}\!/\!_{(n_a+n_b)}\mathfrak{M}_{\text{pop}}^{\boldsymbol{\beta}}(\zeta_b)$. For random coefficient regression models we have just shown in the previous section that the optimal approximate balanced design (single-group design) with respect to a criterion that satisfies Assumption 7.1 is also optimal in the larger class of group designs, where different individual designs are allowed for the different individuals. Therefore, no matter which design is used in one of the groups, we can always find an optimal single-group design for the other group. Thus, for each of the two groups, single-group designs $\zeta_a^* = (\xi^{a*}; 1)$ and $\zeta_b^* = (\xi^{b*}; 1)$ can be found such that $\zeta^* = (\zeta_a^*, \zeta_b^*)$ is optimal in the class of designs, where different individual designs are allowed for the different individuals. The generalization to more than two groups is straight-forward.

We summarize this result to the following theorem.

Theorem 7.6 (cf. Schmelter (2006a))
Consider a linear mixed model as described in Section 4.2 and an optimality criterion Φ satisfying Assumption 7.1. Then a design that is Φ-optimal in class $\tilde{\Delta}_A^{m_a,m_b}$ of group-wise identical designs is also Φ-optimal in the larger class $\tilde{\Delta}_B^{m_a,m_b}$, where the individual designs can vary within the groups.

For illustration we now have another look at Example 4.2.

Example 7.7 *We consider the same situation as described in Example 4.2 and now additionally assume that the covariance structure of the random effects $\mathbf{D}$ is diagonal, that is, $\mathbf{D} = \text{diag}(d_1, d_2)$. For simplicity we assume that the design region $\mathcal{X}$ of possible experimental settings x_{ij} is $\mathcal{X} = [0, 1]$.*

Due to a symmetry argument the search for optimal designs can be restricted to designs that coincide in both groups. The same argument provides that the number of individuals should be the

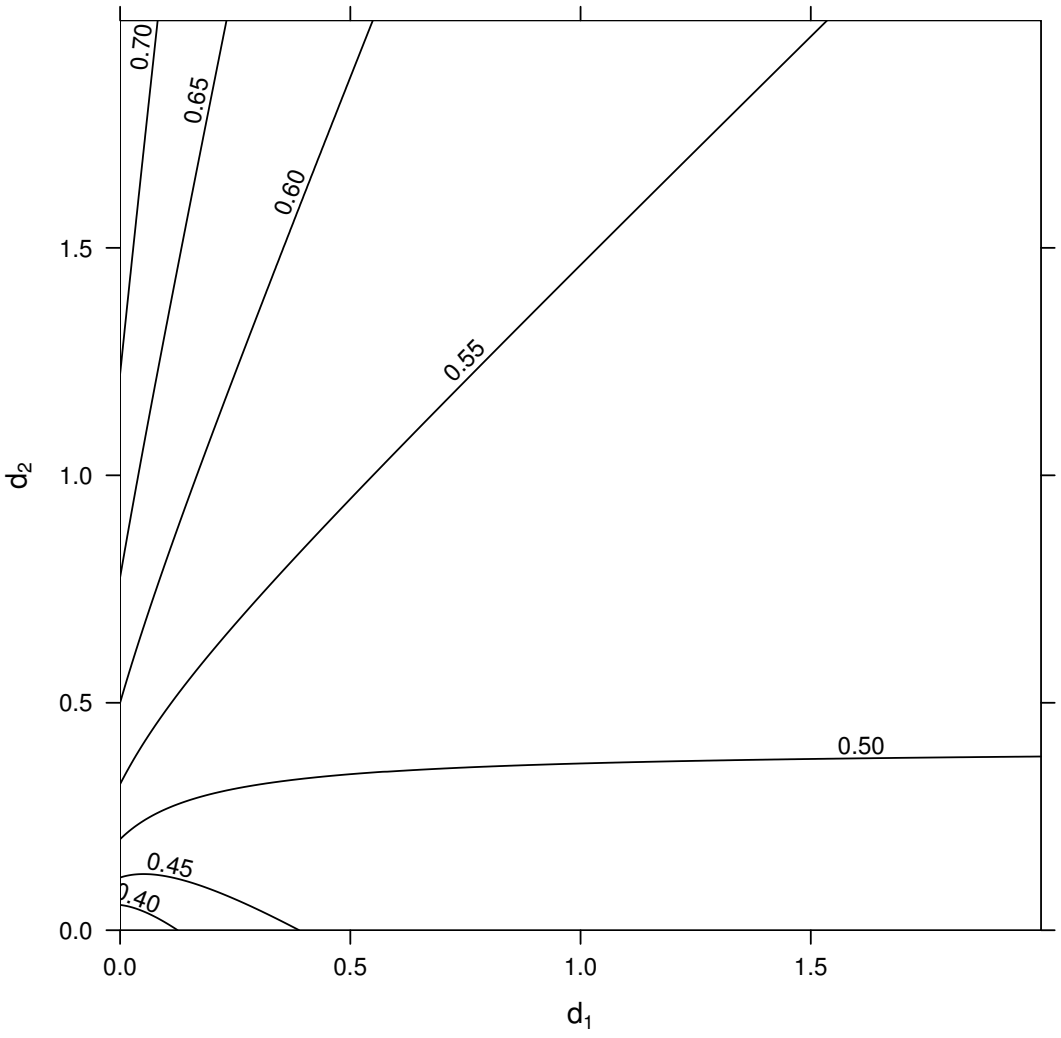


Figure 7.1: *Contour plot showing the weight on the observation at zero in dependence on d_1 and d_2 for the D-optimal design in a two-group linear random coefficient regression model with common intercept in the two groups.*

same in both groups. Furthermore, we can restrict the search on designs that are concentrated on the end points of the design region 0 *and* 1 *(see, e. g., Liski et al. (2002)). Hence, the optimal individual design in both groups has the form* $\xi^* = \begin{pmatrix} 0 & 1 \\ mw & m(1-w) \end{pmatrix}$ *and the only values to determine in this case are the weights* w *and* $(1-w)$ *in dependence on* d_1, d_2, *and* m.

The contour plot in Figure 7.1 shows a graphical representation of the D-optimal designs for the case that the number of observations per individual is $m_i = m = 10$. *The curves in the figure show the weight* w *put on the observation at* $x_1 = 0$ *in dependence on* d_1 *and* d_2.

Since the response at $x_1 = 0$ *is the same in both groups, one should expect a lower weight on this*

point. This is true in the case that $d_1 = d_2 = 0$, where we are in the situation of a ordinary linear regression model without random effects. Here, the known result applies that in both groups one third of the observations should be taken at $x_1 = 0$ and two thirds at $x_2 = 1$ (see Schwabe (1996)).

However, a counter-intuitive phenomenon can be observed if the variance d_2 of the slope parameters is increased: The weight w on the observation at $x_{ij} = 0$ then also increases. That means that the D-optimal design suggests to take more observations where it is easy (where the variance is low) and reduce the observations where the variance is high (see Schmelter et al. (2006) for a general discussion of such phenomena in random slope models).

7.3 Further extension

The result can be extended in two ways. First, the covariance structure $\mathbf{D}$ for the random effects does not have to be the same in both groups. In the derivation of the result, both groups were looked at independently, and it was only assumed that the other group provided some fixed information matrix $\mathbf{C}$.

Secondly, for the same reason, also the regression functions may differ between the groups, that is, the regression functions may depend on some additional known parameter vector $\boldsymbol{\gamma}_i$. If we look at such a model

$$Y_{ij} = \mathbf{f}(x_{ij}, \boldsymbol{\gamma}_i)^\top \mathbf{K}_i \boldsymbol{\beta} + \mathbf{f}(x_{ij}, \boldsymbol{\gamma}_i)^\top \mathbf{b}_i + \varepsilon_{ij}, \tag{7.2}$$

it can be easily seen that the dependence on $\boldsymbol{\gamma}_i$ has no influence on the derivation of the result in the previous section as long as $\boldsymbol{\gamma}_i$ does not vary within the group.

At first glance this extension looks rather artificial. However, it can be useful for the application of the result in non-linear mixed models. As we saw in Section 5.3, the linearization in the general non-linear mixed model just leads to this situation. In Eq. (5.10) we saw that we can use the general linear mixed model

$$\tilde{Y_{ij}} \simeq \mathbf{f}(x_{ij}, \mathbf{K}_i \boldsymbol{\beta}_0)^\top \mathbf{K}_i \tilde{\boldsymbol{\beta}} + \mathbf{f}(x_{ij}, \mathbf{K}_i \boldsymbol{\beta}_0)^\top \mathbf{b}_i + \varepsilon_{ij}$$

for planning purposes, which has the form of the model in Eq. (7.2). The following example gives an idea how this can be applied in praxis.

Example 7.8 *As an application of the result, one could imagine a population pharmacokinetic study with the aim to investigate the food effect on the absorption of a drug. The absorption of a drug is often slower if the drug is taken together with food. The study population is therefore split up into two groups, one group is fasting, the other one getting food together with the drug. We assume that the concentration of the drug in the body can be described by the following non-linear function:*

$$\eta(t, D; (V_c, Cl, k_a)^\top) = D \frac{k_a}{V_c(k_a - \frac{Cl}{V_c})} \left(e^{-\frac{Cl}{V_c} t} - e^{-k_a t} \right),$$

where D is the given (fixed) dose, V_c is the volume of distribution (the volume of the blood in the body), Cl is the clearance and k_a is the absorption rate constant. The absorption rate constant is assumed to be greater for the fasting patients. Using the notation of the previous sections the

concentration of individual i at time t_{ij} is now modeled by

$$Y_{ij} = \eta\left(t_{ij}, D; \mathbf{K}_i\boldsymbol{\beta} + \mathbf{b}_i\right) + \varepsilon_{ij},$$

where $\boldsymbol{\beta} = (V_c, Cl, k_a^{(L)}, k_a^{(H)})^\top$ is the vector of population parameters and

$$\mathbf{K}_i = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad \textit{or} \quad \mathbf{K}_i = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

depending on the diet of the respective patient i. If one assumes different prior guesses for $k_a^{(L)}$ and $k_a^{(H)}$, which would be natural, since a food effect is expected, the regression functions in the linearized model used for the planning of the experiment are different for the two groups, as they depend on the respective parameter. This will then usually lead to different designs in the two groups as the symmetry argument used in Example 7.7 does not hold. Due to the shown result of Theorem 7.6, however, it will be sufficient to optimize one elementary design for each group. A numerical example will be given in Example 8.9.

8 Equivalence Theorems

In this chapter we will transfer the theory of convex design theory that we have briefly described in Section 3.4 for the ordinary linear model to the mixed model setting. In Section 3.4 we cited three theorems, which can be found in the book of Silvey (1980), that characterized Φ-optimal designs in the approximate setting with the help of directional derivatives of the criterion function.

In Section 8.1 we will state Theorems 3.11, 3.12, and 3.13 in a more general way than in Silvey (1980) going away from the context of design optimization. This makes it easier to formulate equivalence theorems for various situations in the mixed model setup, which will be done in Sections 8.2– 8.4.

8.1 General formulation

Theorems 3.11, 3.12, and 3.13 can be stated in a more general way without the direct relation to the design problem. We consider the following situation.

Assumption 8.1 1. *$\bar{\mathcal{M}}$ is a compact subset of an Euclidean space $\mathbb{R}^k$, $k \in \mathbb{N}$.*

2. *$\mathcal{M} = \operatorname{conv}(\bar{\mathcal{M}})$ is the convex hull of $\bar{\mathcal{M}}$.*

3. *$\Phi : \mathcal{M} \to (-\infty, \infty]$ is a convex function,*

In the following sections, the roles of $\bar{\mathcal{M}}$ and $\mathcal{M}$ will be taken over by different sets of information matrices. This general framework allows an easier derivation of equivalence theorems for different classes of designs.

Again we will denote by

$$F_\Phi(\mathbf{M}_1, \mathbf{M}_2) = \lim_{\alpha \downarrow 0} \frac{1}{\alpha} \left[\Phi((1-\alpha)\mathbf{M}_1 + \alpha\mathbf{M}_2) - \Phi(\mathbf{M}_1)\right] = \frac{d}{d\alpha}\Phi((1-\alpha)\mathbf{M}_1 + \alpha\mathbf{M}_2)\Big|_{\alpha=0^+}$$

the Fréchet directional derivative of Φ at $\mathbf{M}_1$ in direction of $\mathbf{M}_2$. Note that as Φ is convex, $F_\Phi(\mathbf{M}_1, \mathbf{M}_2)$ exists for all $\mathbf{M}_2$ if Φ is finite in $\mathbf{M}_1$, no matter whether Φ is differentiable or not (see Rockafellar (1972), p. 213). Further note that $F_\Phi(\mathbf{M}, \mathbf{M}) = 0$ by definition. As already mentioned in Section 3.4, if Φ is differentiable in $\mathbf{M}^*$ in the usual sense (see Rockafellar (1972), p. 241), $F_\Phi(\mathbf{M}^*, \sum_i a_i \mathbf{M}_i) = \sum_i a_i F_\Phi(\mathbf{M}^*, \mathbf{M}_i)$ holds for $a_i \in \mathbb{R}$ with $\sum_i a_i = 1$.

We will now restate the theorems of Section 3.4. The proofs can be directly transferred from the respective theorems in Silvey (1980).

Theorem 8.2 (cf. Silvey (1980), Theorem 3.6)
Let Assumption 8.1 be satisfied. Then $\mathbf{M}^* \in \mathcal{M}$ *minimizes* Φ *on* $\mathcal{M}$ *if and only if*

$$F_\Phi(\mathbf{M}^*, \mathbf{M}) \geq 0 \quad \textit{for all } \mathbf{M} \in \mathcal{M}.$$

Theorem 8.3 (cf. Silvey (1980), Theorem 3.7)
Let Assumption 8.1 be satisfied and Φ *be differentiable at* $\mathbf{M}^*$*. Then* $\mathbf{M}^* \in \mathcal{M}$ *minimizes* Φ *on* $\mathcal{M}$ *if and only if*

$$F_\Phi(\mathbf{M}^*, \mathbf{M}) \geq 0 \quad \textit{for all } \mathbf{M} \in \bar{\mathcal{M}}.$$

Theorem 8.4 (cf. Silvey (1980), Theorem 3.9)
Let Assumption 8.1 be satisfied, Φ *be differentiable at all points of* $\mathcal{M}^+ = \{\mathbf{M} \in \mathcal{M} : \Phi(\mathbf{M}) < \infty\}$*, and let the minimum of* Φ *on* $\mathcal{M}$ *exist. Then* $\mathbf{M}^*$ *minimizes* Φ *on* $\mathcal{M}$ *if and only if*

$$\min_{\mathbf{M} \in \bar{\mathcal{M}}} F_\Phi(\mathbf{M}^*, \mathbf{M}) = \max_{\mathbf{N} \in \mathcal{M}} \min_{\bar{\mathbf{N}} \in \bar{\mathcal{M}}} F_\Phi(\mathbf{N}, \bar{\mathbf{N}}).$$

This general formulation of the theorems can now be applied to various situations occurring in the context of mixed models.

Often the condition $F_\Phi(\mathbf{M}^*, \mathbf{M}) \geq 0$ can be brought into the form $\phi(\mathbf{M}, \mathbf{M}^*) \leq C(\mathbf{M}^*)$. The function ϕ on the left hand side is usually called *sensitivity function*, as it tells how sensitively the quality of a design reacts if it is slightly changed into the direction of $\mathbf{M}$. Often, the design or the design points that $\mathbf{M}$ is derived from are used instead of $\mathbf{M}$ as the first argument of ϕ. For the D-optimality in the ordinary linear model, the sensitivity function is given by the variance function $d(x, \mathbf{M}^*) = \mathbf{f}(x)^\top \mathbf{M}^{*-1} \mathbf{f}(x)$ and C was the number of parameters and therefore also independent of $\mathbf{M}^*$.

One should note that if $\mathbf{M}^* = \sum_i \alpha_i \mathbf{M}_i$, with $\mathbf{M}^*, \mathbf{M}_i \in \mathcal{M}$ and $\sum_i \alpha_i = 1, \alpha_i > 0$, minimizes Φ on $\mathcal{M}$ and Φ is differentiable in $\mathbf{M}^*$, then

$$F_\Phi(\mathbf{M}^*, \mathbf{M}_i) = 0 \quad \text{for all } i.$$

This is due to the fact that

$$0 = F_\Phi(\mathbf{M}^*, \mathbf{M}^*) = F_\Phi(\mathbf{M}^*, \sum \alpha_i \mathbf{M}_i) = \sum \alpha_i \underbrace{F_\Phi(\mathbf{M}^*, \mathbf{M}_i)}_{\geq 0},$$

which can only be true if all summands $F_\Phi(\mathbf{M}^*, \mathbf{M}_i)$ are equal to zero.

8.2 Equivalence theorems for single-group designs for the estimation of the population parameters in the RCR model

In the Chapter 7 we have shown how the set of designs, in which the optimization has to be carried out, can be restricted. This, however, does not immediately tell us how optimal designs can be found or how the optimality of a given design can be checked. Unfortunately, the set of

information matrices of single-group designs cannot easily shown to be convex. The welcome feature in the ordinary linear model that the information matrix of the convex combination of two designs is the convex combination of the information matrices of the two designs does not hold any more in the mixed model setting, that is, in general we have

$$\mathfrak{M}^{\boldsymbol{\beta}}_{\text{ind}}(\alpha\xi_1 + (1-\alpha)\xi_2) \neq \alpha\mathfrak{M}^{\boldsymbol{\beta}}_{\text{ind}}(\xi_1) + (1-\alpha)\mathfrak{M}^{\boldsymbol{\beta}}_{\text{ind}}(\xi_2).$$

However, we can again make use of the representation of the $\boldsymbol{\beta}$-block of the information matrix derived in Section 6.2.1. By defining

$$\Psi(\mathbf{M}) := \Phi\left(\mathfrak{M}^{\boldsymbol{\beta}}_{\text{ind}}(\mathbf{M})\right) = \Phi\left((\mathbf{M}^+ + \mathbf{M}^+\mathbf{M}\mathbf{D}\mathbf{M}\mathbf{M}^+)^+\right),$$

any criterion Φ on the set of information matrices of single group designs in the RCR model can be transfered to a criterion Ψ on the set of information matrices of the corresponding ordinary linear model.

Furthermore, if Φ is monotone and convex on the set of information matrices of the mixed model, Ψ is monotone and convex on the set of information matrices of the corresponding ordinary linear model, as we show in the following lemma.

Lemma 8.5 *Let* $\Phi : nnd(\mathbb{R}^{p\times p}) \to (-\infty, \infty]$ *be a monotone, convex and continuous optimality criterion. Then* $\Psi : nnd(\mathbb{R}^{q\times q}) \to (-\infty, \infty]$, $\Psi(\mathbf{M}) = \Phi(\mathfrak{M}^{\boldsymbol{\beta}}_{\text{ind}}(\mathbf{M})) = \Phi((\mathbf{M}^+ + \mathbf{M}^+\mathbf{M}\mathbf{D}\mathbf{M}\mathbf{D}^+)^+)$ *is also monotone and convex.*

Proof: For positive definite matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ the following holds:

MONOTONICITY:

$$\begin{aligned}
\mathbf{M}_1 \geq \mathbf{M}_2 &\Leftrightarrow \mathbf{M}_1^{-1} \leq \mathbf{M}_2^{-1} \\
&\Leftrightarrow (\mathbf{M}_1^{-1} + \mathbf{D}) \leq (\mathbf{M}_2^{-1} + \mathbf{D}) \\
&\Leftrightarrow (\mathbf{M}_1^{-1} + \mathbf{D})^{-1} \geq (\mathbf{M}_2^{-1} + \mathbf{D})^{-1} \\
&\Leftrightarrow \Phi((\mathbf{M}_1^{-1} + \mathbf{D})^{-1}) \leq \Phi((\mathbf{M}_2^{-1} + \mathbf{D})^{-1}) \Leftrightarrow \Psi(\mathbf{M}_1) \leq \Psi(\mathbf{M}_2).
\end{aligned}$$

CONVEXITY: Lemma 7.2 states:

$$\mathfrak{M}^{\boldsymbol{\beta}}_{\text{ind}}((1-\alpha)\mathbf{M}_1 + \alpha\mathbf{M}_2) \geq (1-\alpha)\mathfrak{M}^{\boldsymbol{\beta}}_{\text{ind}}(\mathbf{M}_1) + \alpha\mathfrak{M}^{\boldsymbol{\beta}}_{\text{ind}}(\mathbf{M}_2)$$

and hence

$$\begin{aligned}
\Psi((1-\alpha)\mathbf{M}_1 + \alpha\mathbf{M}_2) &= \Phi(\mathfrak{M}^{\boldsymbol{\beta}}_{\text{ind}}((1-\alpha)\mathbf{M}_1 + \alpha\mathbf{M}_2)) \\
&\leq \Phi((1-\alpha)\mathfrak{M}^{\boldsymbol{\beta}}_{\text{ind}}(\mathbf{M}_1) + \alpha\mathfrak{M}^{\boldsymbol{\beta}}_{\text{ind}}(\mathbf{M}_2)) \\
&\leq (1-\alpha)\Phi(\mathfrak{M}^{\boldsymbol{\beta}}_{\text{ind}}(\mathbf{M}_1)) + \alpha\Phi(\mathfrak{M}^{\boldsymbol{\beta}}_{\text{ind}}(\mathbf{M}_2)) = (1-\alpha)\Psi(\mathbf{M}_1) + \alpha\Psi(\mathbf{M}_2).
\end{aligned}$$

The results for singular $\mathbf{M}_1$ and/or $\mathbf{M}_2$ follow with the second part of Lemma 6.4 and a continuity argument. □

This allows us to directly apply convex design theory as described for the ordinary linear model in Section 3.4. We have reduced the problem of finding a Φ-optimal design for the RCR model

to finding a Ψ-optimal design in the ordinary linear model. We can now formulate equivalence theorems similar to the one for D-optimality for the ordinary linear model in Section 3.4. In the following we use the same sets of designs Δ_A^m and Δ_B^m as in Section 7.1.

We start with an equivalence theorem for the D-criterion for the RCR-model. A similar version can be found in Fedorov and Hackl (1997).

Theorem 8.6
A single-group design $\zeta^* = \begin{pmatrix} \xi^* \\ 1 \end{pmatrix} \in \Delta_A^m$ *for an RCR-model is D-optimal on* Δ_A^m *if and only if*

$$m\mathbf{f}(x)^\top \mathbf{M}(\xi^*)^{-1} \mathfrak{M}_{\text{ind}}^{\beta}(\xi^*) \mathbf{M}(\xi^*)^{-1} \mathbf{f}(x) \leq \text{tr}[\mathfrak{M}_{\text{ind}}^{\beta}(\xi^*) \mathbf{M}(\xi^*)^{-1}] \quad \textit{for all } x \in \mathcal{X}.$$

Proof: The criterion to investigate is $\Psi(\mathbf{M}) = -\log(\det((\mathbf{M}^{-1}+\mathbf{D})^{-1})) = \log(\det(\mathbf{M}^{-1}+\mathbf{D}))$. We have to calculate $F_\Psi(\mathbf{M}_1, \mathbf{M}_2) = \frac{d}{d\alpha} \log(\det(((1-\alpha)\mathbf{M}_1 + \alpha\mathbf{M}_2)^{-1} + \mathbf{D}))$. By applying rules for matrix/vector differential calculus (see, e. g., the book by Magnus and Neudecker (1988) or Wand (2002)) we get

$$F_\Psi(\mathbf{M}_1, \mathbf{M}_2) = \text{tr}\left[\mathfrak{M}_{\text{ind}}^{\beta}(\mathbf{M}_1)\mathbf{M}_1^{-1}(\mathbf{M}_2 - \mathbf{M}_1)\mathbf{M}_1^{-1}\right].$$

With Theorem 3.12 (or Theorem 8.3) the assertion of the theorem follows. □

According to Theorem 7.3 ζ^* will also be optimal in the set Δ_B^m. The sensitivity function for the D-criterion in this situation has the form

$$\phi(x, \xi) = m\mathbf{f}(x)^\top \mathbf{M}(\xi)^{-1} \mathfrak{M}_{\text{ind}}^{\beta}(\xi) \mathbf{M}(\xi)^{-1} \mathbf{f}(x).$$

Different to the ordinary linear model the constant $C(\xi) = \text{tr}[\mathfrak{M}_{\text{ind}}^{\beta}(\xi)\mathbf{M}(\xi)^{-1}]$ now really depends on ξ. With this, it can be immediately seen that in the RCR model the D-criterion is not equivalent to the G-criterion, where

$$\Phi_G(\mathfrak{M}_{\text{ind}}(\xi)) = \max_{x \in \mathcal{X}} \mathbf{f}(x)^\top \mathfrak{M}_{\text{ind}}^{\beta}(\xi)^{-1} \mathbf{f}(x)$$

has to be minimized. In fact, D- and G-criterion can lead to designs that substantially differ. This can already be seen in a simple straight line regression with random slope (see Schmelter et al. (2006)).

Another special property of D-optimal designs in the ordinary linear model does not hold for the mixed model case. In the ordinary linear model, D-optimal designs with minimal support (i.e., with the number of support points equal to the number of parameters to be estimated) have equal weights on each support point. This is not the case in the mixed model setting, which can be seen in Example 8.7.

As we already mentioned, if a linear criterion is used, the optimal design does not depend on the covariance structure of the random effects and hence the optimal design is the same as in the corresponding ordinary linear model. Therefore, of course also the same equivalence theorems as in the ordinary linear model can be applied, that is, Corollary 3.15 can also be applied in the RCR model.

We illustrate the results with the following example.

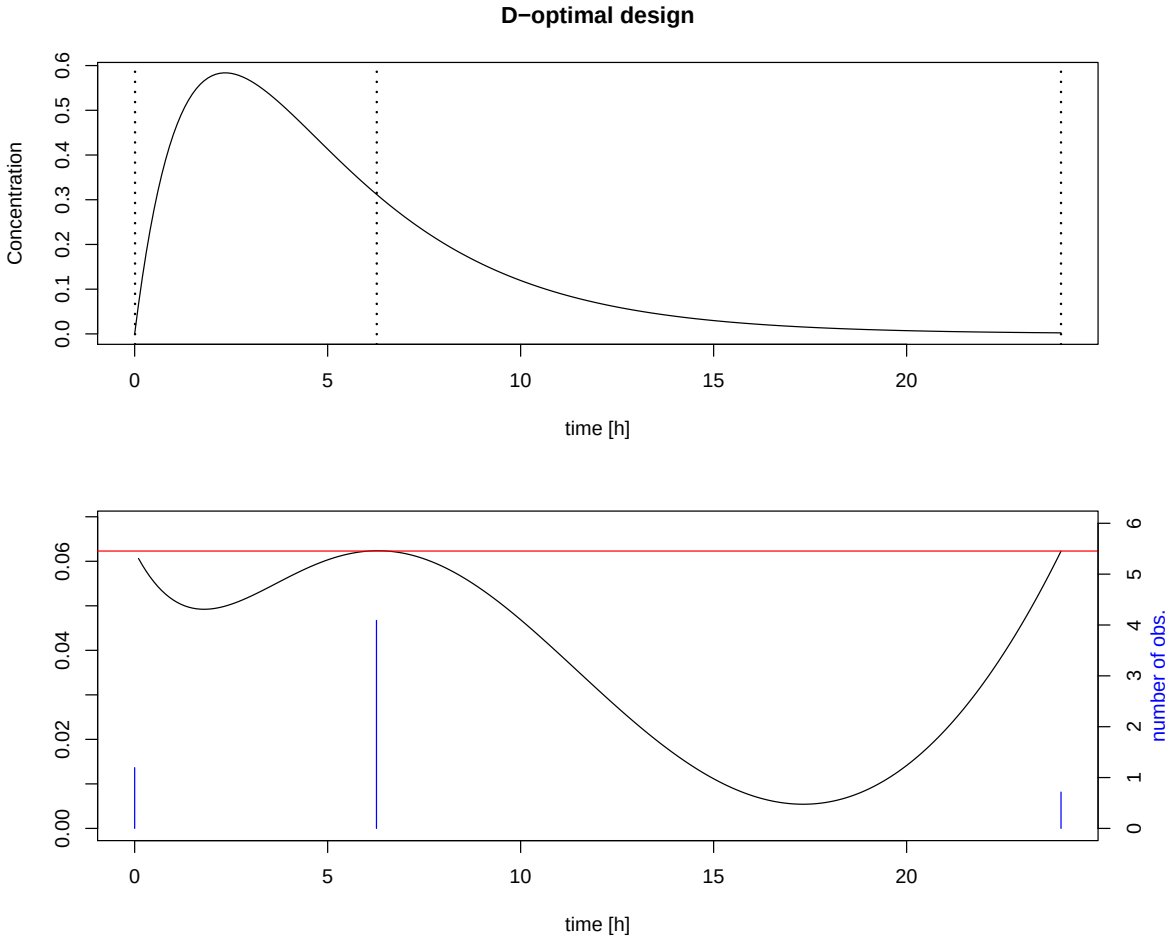


Figure 8.1: *Illustration of the locally D-optimal design for a one-compartment model with absorption, where both the observational errors and the random effects are modeled as multiplicative log-normal random variables. Top: Response curve of a typical individual. Bottom: Sensitivity function and optimal sampling times with number of replications.*

Example 8.7 *We consider a one-compartment model with absorption. The observational error and the individual random effects are modeled by multiplicative log-normal random variables. Thus, the observed concentration of individual i at time x_{ij} is given by*

$$Y_{ij} = D \frac{k_{ai}}{V_{ci}(k_{ai} - \frac{Cl_i}{V_{ci}})} \left(e^{-\frac{Cl_i}{V_{ci}} x_{ij}} - e^{-k_{ai} x_{ij}} \right) \exp(\varepsilon_{ij})$$

with

$$k_{ai} = k_a \exp(b_{1i}), \quad Cl_i = Cl \exp(b_{2i}), \quad V_{ci} = V_c \exp(b_{3i})$$

and

$$\varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \quad (b_{1i}, b_{2i}, b_{3i})^\top \overset{iid}{\sim} \mathcal{N}_3(\mathbf{0}, \sigma^2 \mathbf{D}).$$

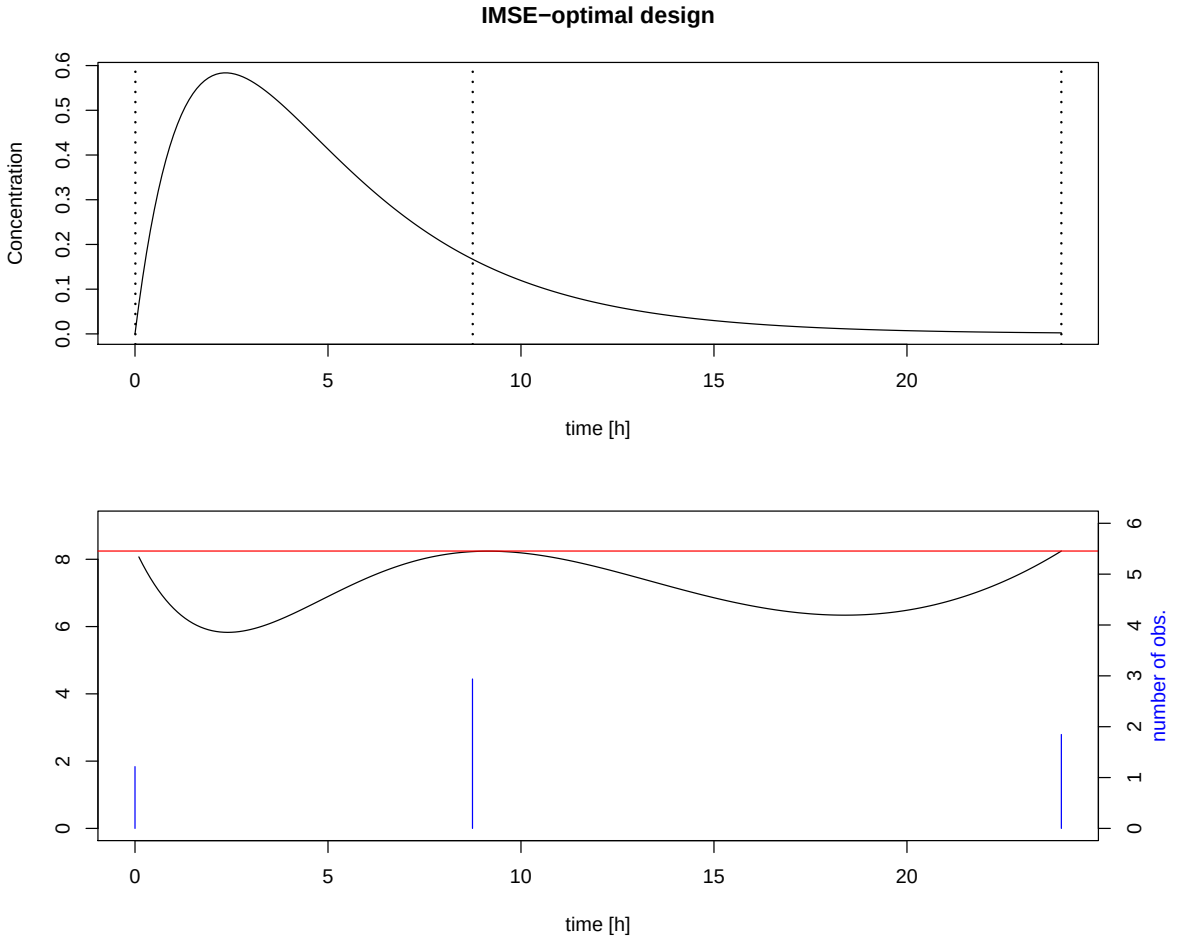


Figure 8.2: *Illustration of the locally IMSE-optimal design for a one-compartment model with absorption, where both the observational errors and the random effects are modeled as multiplicative log-normal random variables.. Top: Response curve of a typical individual. Bottom: Sensitivity function and optimal sampling times with number of replications.*

For planning purposes we use the following information from an earlier study:

$$k_{a0} = 0.61, \quad Cl_0 = 25, \quad V_{c0} = 88,$$

$$\sigma^2 = 0.01, \quad \textit{and} \quad \mathbf{D} = \begin{pmatrix} 89.3 & 0 & 0 \\ 0 & 12.5 & 0 \\ 0 & 0 & 9.0 \end{pmatrix}.$$

We consider the D- and the IMSE-criterion and calculate the optimal approximate one-group designs for six observations per individual. We restrict the design region $\mathcal{X}$ to the interval $[0.01, 24.00]$. We furthermore set the number of observations per individual to $m_i = 6$. The numerical optimization results in

locally D-optimal design

Time	*0.01*	*6.27*	*24.00*
Replication	*1.18*	*4.10*	*0.72*

locally IMSE-optimal design

Time	*0.01*	*8.95*	*24.00*
Replication	*1.23*	*2.93*	*1.84*

Note that the locally IMSE-optimal design only depends on the assumed value of $\boldsymbol{\beta}$ and not on $\mathbf{D}$, as the IMSE-criterion is a linear criterion (see Section 7.1). The results are illustrated in Figure 8.1 for the D-criterion and in Figure 8.2 for the IMSE-criterion. In the top pictures the concentration profile of a typical individual can be seen. The dotted vertical lines indicate the positions of the optimal design points. In the bottom figures the sensitivity functions of the respective criteria are plotted for the optimized designs. The horizontal lines give the constants C derived in the equivalence theorems. As can be seen, the sensitivity functions do not cross these lines. The vertical bars indicate the numbers of replication at the optimal design points. Note that we did not give a theoretical proof of the optimality of the two designs, but only applied the equivalence theorems graphically.

8.3 Equivalence theorems in the general mixed model

Similar results as in the previous section can also be obtained for the general mixed model. However, as we will later see, they are not as easily to apply. For simplicity we again assume without loss of generality that there are only two groups, a and b, such that a population design ζ can be represented by $\zeta = (\zeta_a, \zeta_b)$. Furthermore, we take the number (or proportion) of subjects in each of the groups as given and fixed. We define

$$\Psi(\mathbf{M}_a, \mathbf{M}_b) = \Phi\left(\frac{n_a}{n_a + n_b}\mathbf{K}_a^\top \tilde{\mathfrak{M}}_{\text{ind}}^{\boldsymbol{\beta}}(\mathbf{M}_a)\mathbf{K}_a + \frac{n_b}{n_a + n_b}\mathbf{K}_b^\top \tilde{\mathfrak{M}}_{\text{ind}}^{\boldsymbol{\beta}}(\mathbf{M}_b)\mathbf{K}_b\right),$$

where $\tilde{\mathfrak{M}}_{\text{ind}}$ is used as before to denote the information matrix of the RCR model without $\mathbf{K}_a$ and $\mathbf{K}_b$. We can now interpret a criterion Φ defined on the set of information matrices of the general mixed model as a criterion defined on the Cartesian product of the sets of information matrices of the two corresponding ordinary linear models, that is, $\Psi : \mathcal{M}_a \times \mathcal{M}_b \to (-\infty, \infty]$, where $\mathcal{M}_a$ and $\mathcal{M}_b$ are the sets of individual information matrices for the two groups. If Φ is convex and monotone decreasing, Ψ is also convex and monotone decreasing in both components.

As $\mathcal{M}_a$ and $\mathcal{M}_b$ are the convex hulls of the information matrices of the respective one-point design measures, $\mathcal{M}_a \times \mathcal{M}_b$ is the convex hull of the Cartesian product of the information matrices of the one-point designs. This means that the Theorems 8.2, 8.3, and 8.4 can be applied using $\mathcal{M}_a \times \mathcal{M}_b$ instead of $\mathcal{M}$. Note that for differentiable criteria we have

$$\begin{aligned} F_\Psi((\mathbf{M}_1^*, \mathbf{M}_2^*), (\mathbf{M}_1, \mathbf{M}_2)) &= F_\Psi((\mathbf{M}_1^*, \mathbf{M}_2^*), (\mathbf{M}_1, \mathbf{M}_2^*) + (\mathbf{M}_1^*, \mathbf{M}_2) - (\mathbf{M}_1^*, \mathbf{M}_2^*)) \\ &= F_\Psi((\mathbf{M}_1^*, \mathbf{M}_2^*), (\mathbf{M}_1, \mathbf{M}_2^*)) + F_\Psi((\mathbf{M}_1^*, \mathbf{M}_2^*), (\mathbf{M}_1^*, \mathbf{M}_2)) - F_\Psi((\mathbf{M}_1^*, \mathbf{M}_2^*), (\mathbf{M}_1^*, \mathbf{M}_2^*)) \\ &= F_\Psi((\mathbf{M}_1^*, \mathbf{M}_2^*), (\mathbf{M}_1, \mathbf{M}_2^*)) + F_\Psi((\mathbf{M}_1^*, \mathbf{M}_2^*), (\mathbf{M}_1^*, \mathbf{M}_2)), \quad (8.1) \end{aligned}$$

which follows from the fact that in general $F_\Phi(\mathbf{M}^*, \sum_i a_i \mathbf{M}_i) = \sum_i a_i F_\Phi(\mathbf{M}^*, \mathbf{M}_i)$ holds for $a_i \in \mathbb{R}$ with $\sum_i a_i = 1$ for Φ differentiable in $\mathbf{M}^*$ (see Rockafellar (1972), p. 244).

We now give an equivalence theorem for the D-criterion. We use the same definition of $\tilde{\Delta}_A^{m_a,m_b}$ and $\tilde{\Delta}_B^{m_a,m_b}$ as in Section 7.2 and restrict our attention to non-singular elementary designs, that is, to designs, for which the corresponding information matrix in the ordinary linear model is non-singular.

Theorem 8.8
Let $\zeta^* = (\zeta_a^*, \zeta_b^*) = \left(\begin{pmatrix} \xi_a^* \\ 1 \end{pmatrix}, \begin{pmatrix} \xi_b^* \\ 1 \end{pmatrix}\right) \in \tilde{\Delta}_A^{m_a,m_b}$ *be a group-wise identical design for a general mixed model and* ξ_a^* *and* ξ_b^* *be non-singular. The* ζ^* *is D-optimal on* $\tilde{\Delta}_A^{m_a,m_b}$ *if and only if*

$$m_a \mathbf{f}(x)^\top \mathbf{M}(\xi_a^*)^{-1} \tilde{\mathfrak{M}}_{\text{ind}}^{\boldsymbol{\beta}}(\xi_a^*) \mathbf{K}_a \mathfrak{M}_{\text{pop}}^{\boldsymbol{\beta}}(\zeta^*)^{-1} \mathbf{K}_a^\top \tilde{\mathfrak{M}}_{\text{ind}}^{\boldsymbol{\beta}}(\xi_a^*) \mathbf{M}(\xi_a^*)^{-1} \mathbf{f}(x)$$
$$\leq \operatorname{tr}\left[\mathfrak{M}_{\text{pop}}(\zeta^*)^{-1} \mathbf{K}_a^\top \tilde{\mathfrak{M}}_{\text{ind}}^{\boldsymbol{\beta}}(\xi_a^*) \mathbf{M}(\xi_a^*)^{-1} \tilde{\mathfrak{M}}_{\text{ind}}^{\boldsymbol{\beta}}(\xi_a^*) \mathbf{K}_a \right] \quad \textit{for all } x \in \mathcal{X}$$

and

$$m_b \mathbf{f}(x)^\top \mathbf{M}(\xi_b^*)^{-1} \tilde{\mathfrak{M}}_{\text{ind}}^{\boldsymbol{\beta}}(\xi_b^*) \mathbf{K}_b \mathfrak{M}_{\text{pop}}^{\boldsymbol{\beta}}(\zeta^*)^{-1} \mathbf{K}_b^\top \tilde{\mathfrak{M}}_{\text{ind}}^{\boldsymbol{\beta}}(\xi_b^*) \mathbf{M}(\xi_b^*)^{-1} \mathbf{f}(x)$$
$$\leq \operatorname{tr}\left[\mathfrak{M}_{\text{pop}}(\zeta^*)^{-1} \mathbf{K}_b^\top \tilde{\mathfrak{M}}_{\text{ind}}^{\boldsymbol{\beta}}(\xi_b^*) \mathbf{M}(\xi_b^*)^{-1} \tilde{\mathfrak{M}}_{\text{ind}}^{\boldsymbol{\beta}}(\xi_b^*) \mathbf{K}_b \right] \quad \textit{for all } x \in \mathcal{X}$$

Proof: Similar to the proof of Theorem 8.6 we now have to investigate the criterion

$$\Psi(\mathbf{M}^a, \mathbf{M}^b) = -\log\det\left(\mathbf{K}_a^\top \tilde{\mathfrak{M}}_{\text{ind}}^{\boldsymbol{\beta}}(\mathbf{M}^a) \mathbf{K}_a + \mathbf{K}_b^\top \tilde{\mathfrak{M}}_{\text{ind}}^{\boldsymbol{\beta}}(\mathbf{M}^b) \mathbf{K}_b \right).$$

For the design (ζ_a^*, ζ_b^*) to be optimal it is necessary that both

$$F_\Psi((\mathbf{M}(\xi_a^*), \mathbf{M}(\xi_b^*)), (m_a \mathbf{f}(x)\mathbf{f}(x)^\top, \mathbf{M}(\xi_b^*)) \geq 0 \quad \text{for all } x \in \mathcal{X}$$

and

$$F_\Psi((\mathbf{M}(\xi_a^*), \mathbf{M}(\xi_b^*)), (\mathbf{M}(\xi_a^*), m_b \mathbf{f}(x)\mathbf{f}(x)^\top) \geq 0 \quad \text{for all } x \in \mathcal{X}$$

holds, as this figuratively means that the design cannot be improved by slightly changing the design of one of the two groups, which are special cases of the general necessary condition. On the other hand the two conditions are also sufficient for the optimality, as (ζ_a^*, ζ_b^*) is Ψ-optimal if

$$F_\Psi((\mathbf{M}(\xi_a^*), \mathbf{M}(\xi_b^*)), (m_a \mathbf{f}(x)\mathbf{f}(x)^\top, m_b \mathbf{f}(y)\mathbf{f}(y)^\top)$$
$$= F_\Psi((\mathbf{M}(\xi_a^*), \mathbf{M}(\xi_b^*)), (m_a \mathbf{f}(x)\mathbf{f}(x)^\top, \mathbf{M}(\xi_b^*)) + F_\Psi((\mathbf{M}(\xi_a^*), \mathbf{M}(\xi_b^*)), (\mathbf{M}(\xi_a^*), m_b \mathbf{f}(y)\mathbf{f}(y)^\top) \geq 0$$

holds for all $x, y \in \mathcal{X}$ according to Theorem 8.3 and this condition is satisfied if both of the summands are greater or equal zero. □

Note that according to Theorem 7.6, ζ^* is then also optimal in $\Delta_B^{m_a,m_b}$.

Example 8.9 *We consider the situation of Example 7.8, where the sample population is split up into two groups, in which the absorption rate constant is assumed to be different. As a numerical example we assume for one of the groups the same prior information as we used in Example 8.7. For the other group we assume that all parameters except the absorption rate* k_a *are the same. For* k_a *in this group we assume a value that is twice as high as in the first group. Again both*

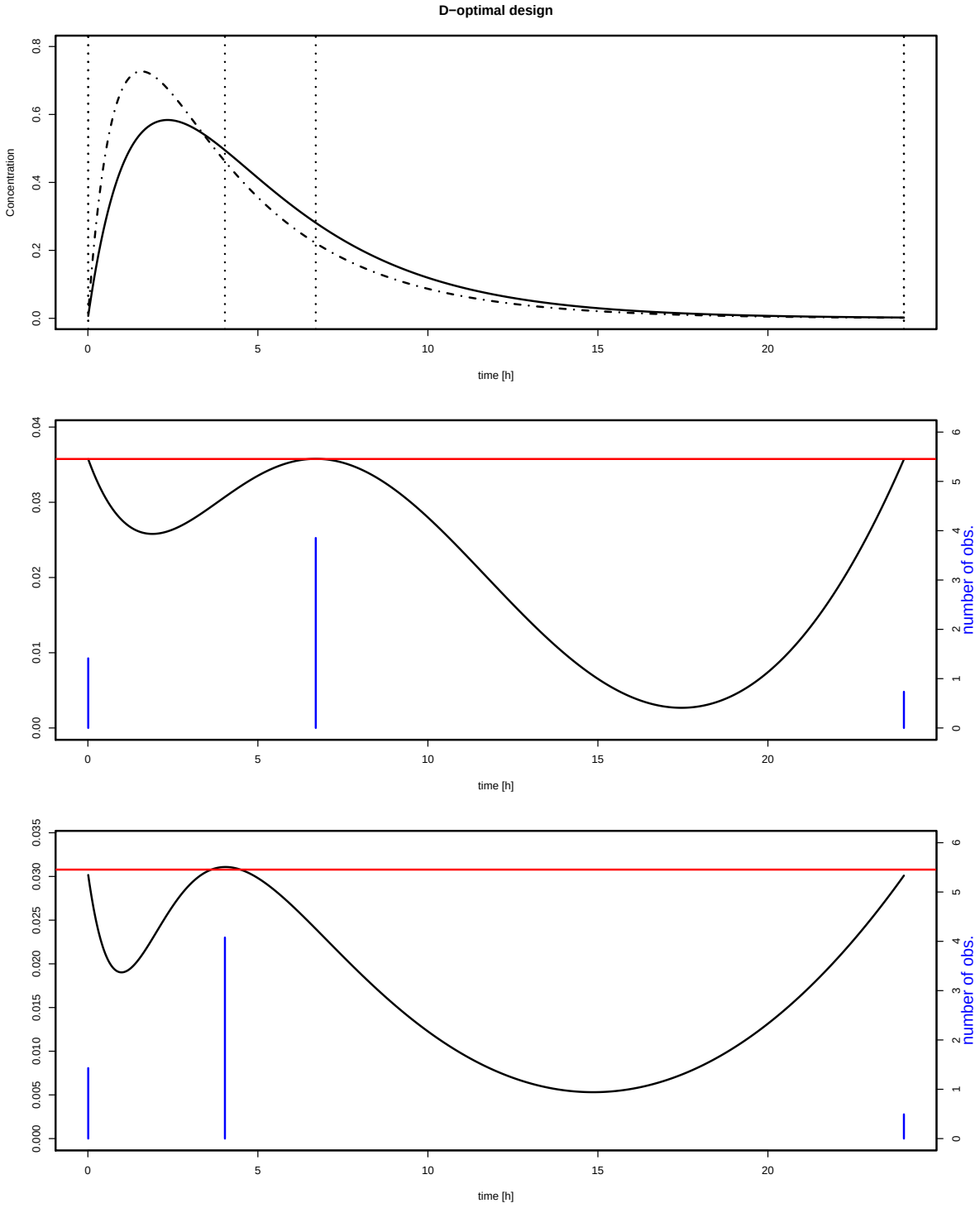


Figure 8.3: *Illustration of the locally D-optimal design for two groups in a one-compartment model with absorption, where both the observational error and the random effects are modeled as multiplicative log-normal random variables. Top: Response curves for typical individuals of the two groups. Middle: Sensitivity function and optimal sampling times for group a. Bottom: Sensitivity function and optimal sampling times for group b.*

the observational errors as well as the random effects are modeled by multiplicative log-normal random variables. The locally optimized approximate designs for the D-criterion for the two groups are given in the following tables.

Group a			
Time	*0.01*	*6.70*	*24.00*
Replications	*1.41*	*3.85*	*0.74*

Group b			
Time	*0.01*	*4.03*	*24.00*
Replications	*1.43*	*4.07*	*0.50*

Figure 8.3 shows in the top picture the concentration profiles of the typical individuals for each group (group a = solid line, group b = dashed line). Below this the two sensitivity functions of the locally D-optimal design for the two groups are shown. The vertical bars denote the optimal sampling times, where the height of the bars is proportional to the number of replications. Note that the designs are not theoretically proven to be optimal but only graphically checked.

In the general mixed model the fact that for linear criteria the same optimal designs are obtained as for the ordinary linear model cannot be shown. Thus, it is reasonable to obtain equivalence theorems also for this situation.

Theorem 8.10
Let Φ be a linear criterion of the form $\Phi(\mathbf{M}) = \mathrm{tr}[\mathbf{L}\mathbf{M}^{-1}]$ and $\zeta^* = \left(\begin{pmatrix}\xi_a^*\\ 1\end{pmatrix}, \begin{pmatrix}\xi_b^*\\ 1\end{pmatrix}\right) \in \tilde{\Delta}_A^{m_a,m_b}$ be a group-wise identical design for a general mixed model such that ξ_a^* and ξ_b^* are non-singular matrices. Then ζ^* is Φ-optimal on $\tilde{\Delta}_A^{m_a,m_b}$ if and only if

$$m_a \mathbf{f}(x)^\top \mathbf{M}(\xi_a^*)^{-1} \tilde{\mathfrak{M}}_{\mathrm{ind}}^{\boldsymbol{\beta}}(\xi_a^*) \mathbf{K}_a \mathfrak{M}_{\mathrm{pop}}^{\boldsymbol{\beta}}(\zeta^*)^{-1} \mathbf{L} \mathfrak{M}_{\mathrm{pop}}^{\boldsymbol{\beta}}(\zeta^*)^{-1} \mathbf{K}_a^\top \tilde{\mathfrak{M}}_{\mathrm{ind}}^{\boldsymbol{\beta}}(\xi_a^*) \mathbf{M}(\xi_a^*)^{-1} \mathbf{f}(x)$$
$$\leq \mathrm{tr}\left[\mathfrak{M}_{\mathrm{pop}}^{\boldsymbol{\beta}}(\zeta^*)^{-1} \mathbf{L} \mathfrak{M}_{\mathrm{pop}}^{\boldsymbol{\beta}}(\zeta^*)^{-1} \mathbf{K}_a^\top \tilde{\mathfrak{M}}_{\mathrm{ind}}^{\boldsymbol{\beta}}(\xi_a^*) \mathbf{M}(\xi_a^*)^{-1} \tilde{\mathfrak{M}}_{\mathrm{ind}}^{\boldsymbol{\beta}}(\xi_a^*) \mathbf{K}_a\right] \quad \textit{for all } x \in \mathcal{X}$$

and

$$m_b \mathbf{f}(x)^\top \mathbf{M}(\xi_b^*)^{-1} \tilde{\mathfrak{M}}_{\mathrm{ind}}^{\boldsymbol{\beta}}(\xi_b^*) \mathbf{K}_b \mathfrak{M}_{\mathrm{pop}}^{\boldsymbol{\beta}}(\zeta^*)^{-1} \mathbf{L} \mathfrak{M}_{\mathrm{pop}}^{\boldsymbol{\beta}}(\zeta^*)^{-1} \mathbf{K}_b^\top \tilde{\mathfrak{M}}_{\mathrm{ind}}^{\boldsymbol{\beta}}(\xi_b^*) \mathbf{M}(\xi_b^*)^{-1} \mathbf{f}(x)$$
$$\leq \mathrm{tr}\left[\mathfrak{M}_{\mathrm{pop}}^{\boldsymbol{\beta}}(\zeta^*)^{-1} \mathbf{L} \mathfrak{M}_{\mathrm{pop}}^{\boldsymbol{\beta}}(\zeta^*)^{-1} \mathbf{K}_b^\top \tilde{\mathfrak{M}}_{\mathrm{ind}}^{\boldsymbol{\beta}}(\xi_b^*) \mathbf{M}(\xi_b^*)^{-1} \tilde{\mathfrak{M}}_{\mathrm{ind}}^{\boldsymbol{\beta}}(\xi_b^*) \mathbf{K}_b\right] \quad \textit{for all } x \in \mathcal{X}.$$

Proof: Analogous to proof of Theorem 8.8. □

The extension of Theorem 8.8 and Theorem 8.10 to more than two groups is straight-forward.

8.4 Equivalence theorems based on population designs

There are reasons to consider equivalence theorems not only on the level of the elementary designs (as we have done in the previous sections), but also one level above this on the level of population designs.

One reason is that the results described in the previous sections do not help us if we consider criteria that are based on the complete information matrix also including the variance parameters. Although we gave representations for the different blocks of the information matrix, which expressed them as functions of $\mathbf{M}$, the resulting criteria $\Psi(\mathbf{M}) = \Phi(\mathfrak{M}_{\text{ind}}(\mathbf{M}))$ are not convex in $\mathbf{M}$, which makes it impossible to transfer the results.

Another reason is that the results of the previous section do not help us if we restrict the set of allowed individual designs to exact elementary designs, as then the set, on which the optimization is performed, is not convex anymore.

To derive results anyway we consider the following situation. Let Ξ' be an arbitrary subset of the set Ξ of all individual designs. This could, for example, be the set of all exact elementary designs with a certain given number of observations. We define

$$\mathcal{M}_{\text{ind}}(\Xi') = \{\mathfrak{M}_{\text{ind}}(\xi) | \xi \in \Xi'\}$$

and

$$\Delta(\Xi') = \left\{ \zeta \mid \zeta = \begin{pmatrix} \xi_1, & \dots, & \xi_l \\ g_1, & \dots, & g_l \end{pmatrix} \text{ with } \xi_1, \dots, \xi_l \in \Xi' \right\}.$$

The set of the information matrices of all population designs based on Ξ'

$$\mathcal{M}_{\text{pop}}(\Xi') = \left\{ \mathfrak{M}_{\text{pop}}(\zeta) \mid \zeta \in \Delta(\Xi') \right\} \tag{8.2}$$

is a convex set, as it is the convex hull of the set of information matrices of the elementary designs contained in Ξ', that is, $\mathcal{M}_{\text{pop}}(\Xi') = \text{conv}(\mathcal{M}_{\text{ind}}(\Xi'))$.

Hence the requirements for the application of the theorems stated in the beginning of this chapter are met. The chosen set Ξ' takes over the function that the design space $\mathcal{X}$ had in the previous sections.

We can state the following theorem.

Theorem 8.11
Let $\Xi' \subset \Xi$ be a subset of the set of all elementary designs, $\mathcal{M}_{pop}(\Xi')$ be defined as in Eq. (8.2), and Φ be convex optimality criterion that is differentiable in $\mathfrak{M}_{\text{pop}}(\zeta^*)$, $\zeta^* \in \Delta(\Xi')$. Then ζ^* is Φ-optimal on $\Delta(\Xi')$ if and only if

$$F_\Phi(\mathfrak{M}_{\text{pop}}(\zeta^*), \mathfrak{M}_{\text{ind}}(\xi)) \geq 0 \quad \textit{for all } \xi \in \Xi'.$$

Proof: Follows directly from Theorem 8.3. □

For the D-criterion the following equivalence theorem can be shown, which is similar to the multivariate version of equivalence theorem stated in Fedorov (1972), p. 212.

Theorem 8.12
A population design $\zeta^* \in \Delta(\Xi')$ is D-optimal on $\Delta(\Xi')$ if and only if

$$\text{tr}(\mathfrak{M}_{\text{pop}}(\zeta^*)^{-1}\mathfrak{M}_{\text{ind}}(\xi)) \leq r \quad \textit{for all } \xi \in \Xi',$$

where r is the dimension of the information matrix, that is, the number of parameters to be estimated.

Proof: $\Phi(\mathbf{M}) = -\log(\det(\mathbf{M}))$ is convex on the convex set $\mathcal{M}_{\text{pop}}(\Xi') \subset nnd(\mathbb{R}^{r\times r})$ and differentiable on the subset of positive definite matrices. The assertion then follows with Theorem 8.3. □

For linear criteria, the form of the equivalence theorem is also similar to the one for the ordinary linear model.

Theorem 8.13
Let Φ be a linear criterion of the form $\Phi(\mathbf{M}) = \text{tr}(\mathbf{A}\mathbf{M}^{-1})$. A non-singular population design $\zeta^* \in \Delta(\Xi')$ is Φ-optimal on $\Delta(\Xi')$ if and only if

$$\text{tr}(\mathfrak{M}_{\text{pop}}(\zeta^*)^{-1}\mathbf{A}\mathfrak{M}_{\text{pop}}(\zeta^*)^{-1}\mathfrak{M}_{\text{ind}}(\xi)) \leq \text{tr}(\mathfrak{M}_{\text{pop}}(\zeta^*)^{-1}\mathbf{A}) \quad \textit{for all } \xi \in \Xi'.$$

The results can be restricted to subsets of the parameters in the usual way like, for example, to only the population parameters.

Different to the equivalence theorems in the previous section, the application of this theorem is only useful for low dimensions, that is, for a low number of observations per individual like two or three.

It should be also mentioned that the results in this chapter not only allow the checking of the optimality of given designs but also allow the construction of iterative optimization algorithms. The general idea briefly described at the end of Section 3.4 can be directly applied here.

We close this section with an example for the application of two equivalence theorems just introduced.

Example 8.14 *We consider the same situation as in Example 8.7 and use the same prior information as there. However, we now restrict the set of allowed elementary designs to exact designs with two observations per individual. The following tables show the optimized designs for the D- and the IMSE-criterion (based only on the population parameter block).*

locally D-optimal design

Elementary design	*(0.01, 4.35)*	*(0.01, 24.00)*	*(5.71, 24.00)*
Weight	*0.28*	*0.29*	*0.43*

locally IMSE-optimal design

Elementary design	*(0.01, 18.31)*	*(3.85, 17.65)*	*(9.49, 24.00)*
Weight	*0.28*	*0.32*	*0.40*

Figure 8.4 and Figure 8.5 show the contour plots of the sensitivity functions of optimal designs for the D- and the IMSE-criterion. The grey circles indicate the elementary two-point designs. The radius of the circles is proportional to the weight on the respective design.

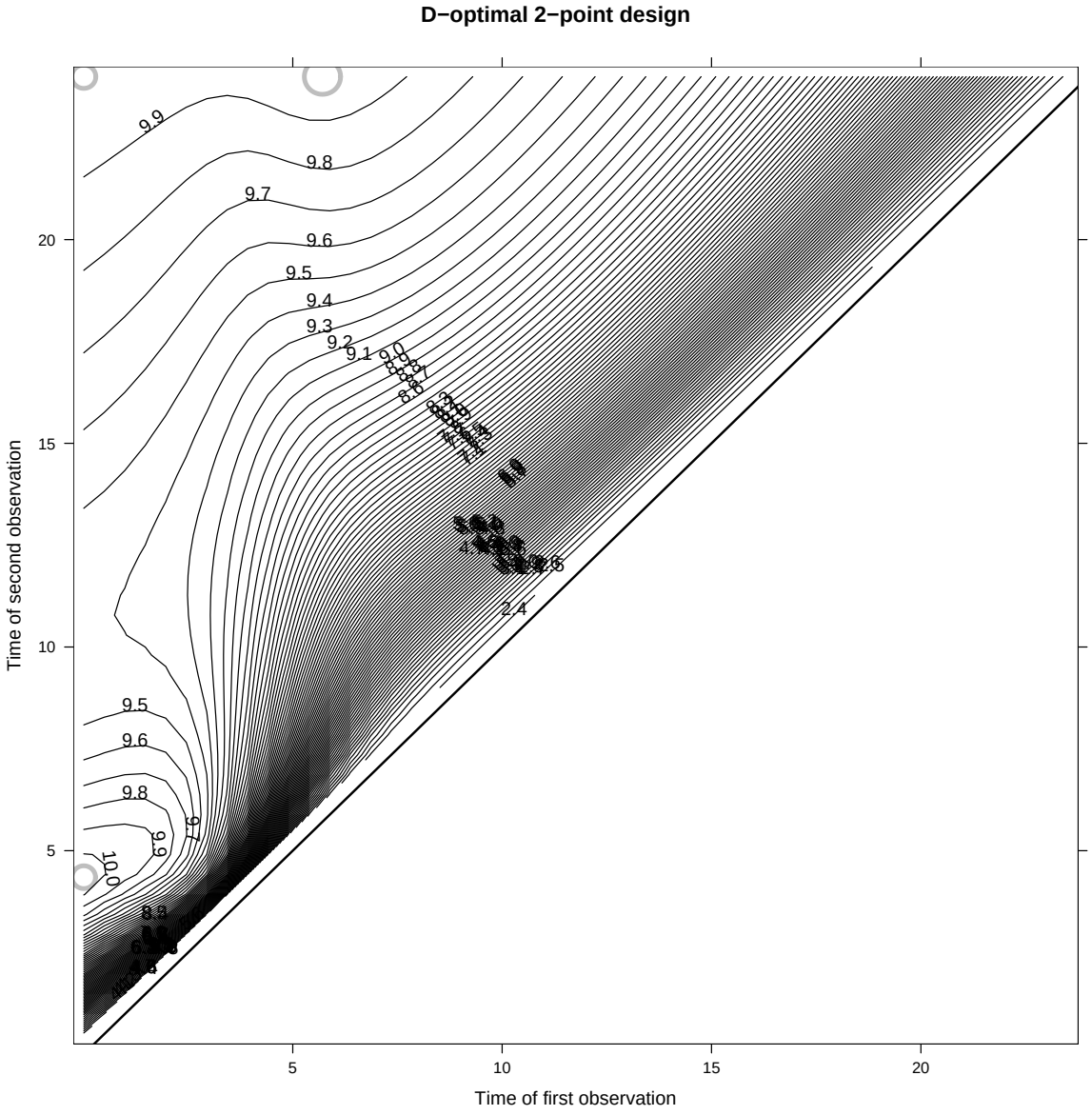


Figure 8.4: *Contour plot of the sensitivity function of the D-criterion for the D-optimal 2-point population design for a one-compartment model with absorption. Both the observational errors and the random effects are modeled as multiplicative random variables. The grey circles in the figure mark the optimal elementary two-point designs. The radii of the circles are proportional to the weight.*

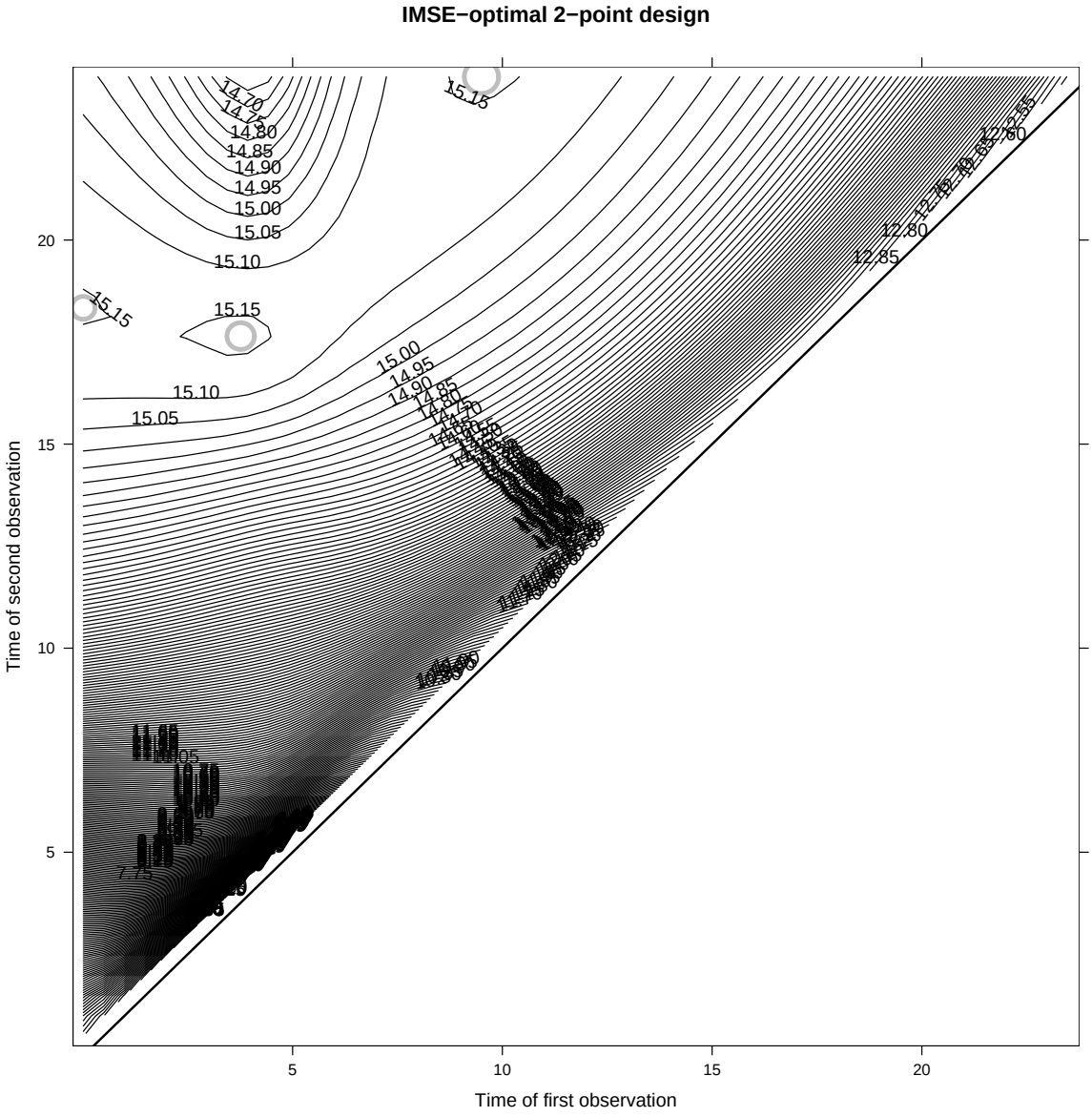


Figure 8.5: *Contour plot of the sensitivity function of the IMSE-criterion for the IMSE-optimal 2-point population design for a one-compartment model with absorption. Both the observational errors and the random effects are modeled as multiplicative random variables. The grey circles in the figure mark the optimal elementary two-point designs. The radii of the circles are proportional to the weight*

9 Practical Considerations and Discussion

There are some limitations theoretically optimized designs have when they are applied in practice. We will use this chapter to discuss some of these issues, like possibly non-integer numbers of observations or the dependence of the designs on prior information. Against this background we will also discuss the results of this thesis.

9.1 Non-integer replications

Usually, when single-group designs or group-wise identical designs for the estimation of population parameters are optimized, the numbers of replications for the different sampling times will not be integer. As for the ordinary linear model (see, e. g., Pukelsheim (1993)) efficient designs can usually be found by rounding. However, it may occur that the number of support points of a design is larger than the number of measurements allowed in the design. Then efficient rounding of the replications is usually not possible and it is necessary to step back from the single group designs and allow more than one group or to allow more observations per individual. For this, the number of support points in the optimal approximate single-group design is a good indicator for how many distinct time points to include into the applicable population design. In any case, the optimized one-group designs can be used as benchmark designs. The theoretical framework makes them easy to calculate and, as the exact designs are a subset of the approximate designs, no exact design with the same constraints on the number of observations can have a higher efficiency, that is, a lower value of the criterion.

Similar consideration have to be made for the weights on the elementary designs in optimized population designs. It might occur that the number of individuals ng_i to be observed with design ξ_i is not integer. Again rounding of these numbers usually leads to efficient designs.

The importance of the benchmark property should not be underestimated. Optimal approximate designs can, on the one hand, usually be relatively easily numerically calculated by using, for example, the first-order algorithm and, on the other hand, be shown to be optimal with the help of one of the equivalence theorems discussed in Chapter 8. If a candidate for an efficient applicable design is found, the efficiency can easily be checked by comparison with the optimal approximate design. Depending on the efficiency one can then decide whether it is worth the effort to further optimize the design.

9.2 Repeated measurements

In the previous section we just discussed the problem that we might have non-integer numbers of replications. In the practice of pharmacokinetic studies, however, it is usually even impossible to get more than one blood sample at a given time point. That is, numbers of replications that are not equal to one are usually impossible to implement in practice. One might get the idea to obtain a larger blood sample at a given time and split the amount of blood into several samples to circumvent this problem. By this, however, the assumed conditional independence of the blood samples of one individual is usually violated.

There are studies, however, where replicated measurements are possible in a certain sense. This is the case, for example, in multiple dose studies. There, it is possible to optimize an elementary design for one dosing interval in steady-state. The replications can than be distributed to the other dosing intervals. We illustrate this with the following example.

Example 9.1 *We consider the same one-compartment model as in Example 8.7. We assume that we have the same prior information for the planning of the experiment. Both the observational errors and the random effects are again assumed to be multiplicative log-normal random variables.*

This time we assume that the drug is not only administered once, but repeatedly every six hours. As described in Section 2.1.3, after several administrations the concentration profile approaches a steady-state and the profiles look the same for each subsequent dosing interval. An efficient design for such a multiple-dose study can now be found by first optimizing the design for the steady-state model, rounding the replications, and then allocating the replications to different dosing intervals. By this, the problem with impossible replications of measurements at the same time point within one subject vanishes. Figure 9.1 illustrates this idea. The top figure shows the concentration profile at steady state. The bars indicate the D-optimal sampling time points. The height of the bars gives the number of replications at the respective sampling time. The bottom picture shows the concentration profile for several repeated administrations. After the fourth drug intake the steady-state is reached. The D-optimal sampling times are allocated to the fourth, fifth, and sixth dosing interval. In this example, the D-efficiency of the rounded design, which is defined by

$$\mathrm{Eff}_D(\zeta) = \frac{\det(\mathfrak{M}(\zeta))^{\frac{1}{p}}}{\det(\mathfrak{M}(\zeta^*))^{\frac{1}{p}}},$$

is still 99%.

Note that for multiple dose experiments often additional variance components are introduced to model the inter-occasional variability, that is, the variability of the kinetic parameters between the dosing intervals within one subject. For the planning stage, however, this component is often neglected.

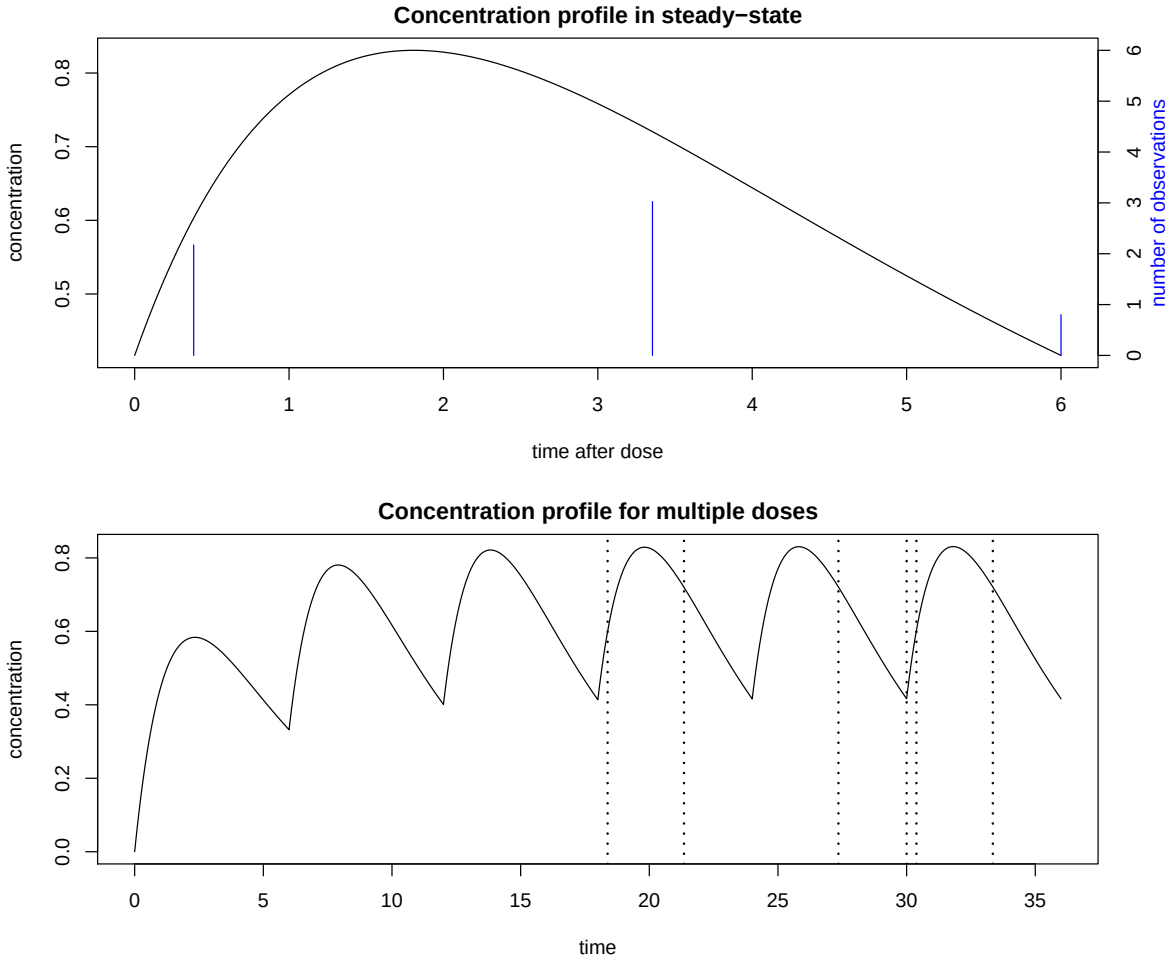


Figure 9.1: *D-efficient design for a multiple-dose experiment. Top figure: concentration profile of a one-compartment model in steady state with optimal design points. Bottom figure: concentration profile for multiple doses with D-efficient sample time points.*

9.3 Dependence of optimal designs on unknown parameters

The probably most severe problem with theoretical design optimization in mixed models is the dependence of the optimal designs on the prior information one has about the unknown parameters. In the linear mixed model these are only the variance parameters, in the non-linear model these are also the population parameters. The dependence on the variance parameters is not as severe as the one on the population parameters. If the focus lies on the estimation of the population parameters and a linear criterion is used, like the IMSE-criterion, then the optimal design does not depend on the variance parameters at all (see Section 7.1).

Proposals to reduce the dependence of the optimal designs on the unknown parameters are to use an Bayesian or pseudo-Bayesian approach or to use minimax criteria.

In the "real" Bayesian approach both the planning and the analysis are performed within a Bayesian framework, that is, an a-priori distribution for the unknown parameters is used for the planning and the Bayesian analysis of the experiment. A general description of Bayesian analysis of population pharmacokinetic models is given in Lunn et al. (2002). Examples for Bayesian designs for mixed models are given in Han and Chaloner (2004) and Stroud et al. (2001) besides others. We will not further discuss Bayesian approaches, as all the rest of this thesis was written from a frequentist's perspective.

With pseudo-Bayesian, or also called averaged, criteria we mean approaches, where a prior distribution is used for the experimental design but the analysis is non-Bayesian (see Fedorov and Hackl (1997), pp. 101). There is a statement by Atkinson and Cox (1974) "The classical Fisherian approach is that prior information (not usually quantitatively expressed) is highly relevant in selecting a design but that the analysis should not depend on this information".

The usual pseudo-Bayesian approach is outlined in the following. To emphasize the dependence of the information matrix on a vector of unknown parameters $\boldsymbol{\theta}$, we will in this section denote the information matrix by $\mathfrak{M}_{\text{pop}}(\zeta;\boldsymbol{\theta})$. The vector $\boldsymbol{\theta}$ can contain only the variance parameters or both the variance and the population parameters depending on the situation. We express our assumptions on the unknown parameters by a prior distribution μ. The pseudo-Bayesian or averaged criterion defined on a given set of designs now has the form

$$\Psi(\zeta) = \int \Phi(\mathfrak{M}_{\text{pop}}(\zeta;\boldsymbol{\theta}))\mu(d\boldsymbol{\theta})$$

(see, e. g., Fedorov and Hackl (1997), pp. 101). Depending on the chosen prior distribution, the calculation of optimal designs is much more complicated due to the integrals that have to be solved. Note that Ψ cannot be expressed as a function of the information matrix of a design. However, for many criteria Φ one can also find equivalence theorems for the corresponding averaged criterion. The results of Chapter 7 still hold for averaged criteria due to the linearity of the integral. This means that the optimization can still be restricted to single-group designs and group-wise identical designs if a criterion based on the population parameters is used. One should also note that the upper bound for the number of support points for an optimal design given by Caratheodory's theorem does not hold in this situation, as the criterion is not a function of a single information matrix anymore. The increase of support points can be seen in

Atkinson et al. (1993), where optimal pseudo-Bayesian designs for a one-compartment model without random effects are calculated.

The optimization of locally optimal designs can be seen as an extreme case of the pseudo-Bayesian approach, as the used best guess can be seen as a degenerate prior distribution.

Alternatively to the formulation given above, one can also base averaged criteria on the averaged information matrix or the averaged inverse of the information matrix instead of the averaged criterion value, that is, using

$$\Psi(\zeta) = \Phi\left(\int \mathfrak{M}_{\text{pop}}(\zeta;\boldsymbol{\theta})\mu(\boldsymbol{\theta})\right) \quad \text{or} \quad \Psi(\zeta) = \Phi\left(\left(\int \mathfrak{M}_{\text{pop}}(\zeta;\boldsymbol{\theta})^{-1}\mu(\boldsymbol{\theta})\right)^{-1}\right).$$

Tod and Rocchisani (1997) have compared three different variants of averaged criteria for the D-criterion in pharmacokinetic models.

Dodds et al. (2005) investigate the ED criterion, where the averaged determinant of the information matrix is maximized, for a population pharmacokinetic model. Tod et al. (1998) do the same for the EID criterion, where the averaged determinant of the inverse of the information matrix is minimized.

A similar approach are so-called minimax criteria. Here, not a specific prior distribution is specified for the unknown parameters, but only a set Θ, in which they are assumed to lie. To protect oneself against the worst case scenario one now minimizes the maximum value of the criterion over the set Θ, that is, one minimizes

$$\Phi(\zeta) = \max_{\theta\in\Theta} \Phi(\mathfrak{M}_{\text{pop}}(\zeta;\theta)).$$

Again, depending on the chosen set Θ the solution of this problem is not trivial. The results of Theorem 7.3 and 7.6 stay valid also for this kind of criterion. Theoretical results for a similar criterion, where a standardized minimax D-criterion is used, for a compartment model without random effects have been obtained by Biedermann et al. (2004).

9.4 Further practical implications

The result of a theoretical design optimization is usually a set of few exact sampling time points. If the design is seriously followed, the blood samples have to be taken at exactly these time points to assure the calculated efficiency. In practice this is in most cases not feasible. On the other hand, it is often also not acceptable. If the samples are obtained at only very few distinct time points, there is no robustness against model misspecification. The model used for the planning is usually developed using data from an ordinary kinetics study conducted in few healthy volunteers. The population, for which the study is planned, often has a higher variability. The sources of variability are often not known in advance and are investigated after the study has been conducted by stepwise including or excluding additional covariates, like weight, age, etc. For sake of both, robustness against deviation from the model and practicability of the study, often not single time points are prescribed for the blood samples but time windows, in which the samples should be taken. These time windows are usually oriented at the optimal sampling

times. Chenel et al. (2005) describe an approach to obtain efficient time windows against the background of a phase II study. Bogacka et al. (2006) present an idea based on the equivalence theorem to find D-efficient time windows for models without random effects.

For practical application also another kind of robustness of the design is needed: robustness against convergence problems of the fitting algorithm. Even in the case that a theoretical optimal design is used, it can happen that fitting of the model with, for example, NONMEM is not successful due to numerical instabilities. Simulation runs are in any case advised before implementing a design for a clinical study. By this it can be checked if the optimized design performs reasonably well if it is applied in a practical situation and numerical problems can be recognized in advance.

9.5 Conclusion

Despite all the described difficulties that might appear in the practical application of the theoretically optimized designs, the results presented in this thesis form a good framework for design optimization.

By allowing approximate designs on both levels (individual and population level) the problem becomes theoretically better tractable. We showed that within this setting, under certain conditions, a design that is optimal in the class of single-group designs is also optimal in the class of general population designs. By this the set in which an optimal design is to be searched for can be drastically reduced without losing efficiency.

Furthermore, the derived equivalence theorems not only allow for the checking of optimality of a given design, but also offer a starting point for the construction of optimization algorithms.

Nomenclature

$()^+$	Moore-Penrose inverse of a matrix	44
$()^-$	generalized inverse of a matrix	28
$\circ$	component-wise product (Hadamard product) of two matrices	38
$\mathbb{1}_m$	m-dimensional vector containing only ones	44
$\boldsymbol{\gamma}_i$	additional parameter in the regression function	62
AUC	area under the curve	9
$\mathbf{b}_i$	vector of random effects of individual i	24
$C(t)$	concentration at time t	5
Cov	covariance of a random variable	13
$c_{\max}$	maximum concentration	9
D	dose	4
$\mathbf{D}$	covariance structure for the random effects	24
d	design	14
E()	expectation of a random variable	13
$\mathbf{e}_i$	i-th unit vector	44
F	bioavailability of a drug	5
F_Φ	Fréchet derivative of Φ	19
$\mathbf{F}$	design matrix	13
$\mathbf{F}_i$	design matrix of individual i	24
f_1	regression function	13
$\mathbf{f}$	vector of regression functions	13

$\delta_{\{x\}}$	one-point measure in x	15
ε_i	observational error	13
ε_{ij}	observational error of jth observation of individual i	24
$\boldsymbol{\varepsilon}$	vector of observational errors	13
$\boldsymbol{\varepsilon}_i$	vector of observational errors of individual i	24
ζ	population design	42
η	nonlinear regression function	33
μ	measure	18
Ξ	set of (elementary) designs	16
Ξ_β	set of regular (elementary) designs	16
Ξ_φ	set of (elementary) designs that allow estimation of φ	16
$\Xi^{(m)}$	set of elementary designs with m observations	42
ξ	(elementary) design	15
σ^2	variance of observational error	13
τ	time between two administrations of a drug	8
Φ	optimality criterion	17
$\phi()$	sensitivity function	66
Ψ	optimality criterion	60

Bibliography

Al-Banna, M. K., Kelman, A. W., and Whiting, B. (1990). Experimental design and efficient parameter estimation in population pharmacokinetics. *Journal of Pharmacokinetics and Biopharmaceutics*, 18(4):347–360.

Atkinson, A. C. and Bailey, R. A. (2001). One hundred years of the design of experiments on and off the pages of Biometrika. *Biometrika*, 88(1):53–97.

Atkinson, A. C., Chaloner, K., Herzberg, A. M., and Juritz, J. (1993). Optimum experimental designs for properties of a compartmental model. *Biometrics*, 49:325–337.

Atkinson, A. C. and Cox, D. R. (1974). Planning experiments for discriminating between models (with discussion). *Journal of the Royal Statistical Society, Series B*, 36:321–348.

Atkinson, A. C. and Donev, A. N. (1996). *Optimum Experimental Designs*. Oxford Statistical Science Series. Oxford University Press, Oxford.

Bates, D. M., Hamilton, D. C., and Watts, D. G. (1983). Calculation of intrinsic and parameter effects curvature for nonlinear regression models. *Communications in Statistics: Simulation and Computation*, 12:469–477.

Beal, S. L. and Sheiner, L. B. (1989). *NONMEM Users Guide – Part I: Users Basic Guide*. NONMEM Project Group, University of California, San Francisco.

Beal, S. L. and Sheiner, L. B. (1998). *NONMEM Users Guide – Part VII: Conditional Estimation Methods*. NONMEM Project Group, University of California, San Francisco.

Biedermann, S., Dette, H., and Pepelyshev, A. (2004). Maximin optimal designs for the compartmental model. In Bucchianico, A. D., Läuter, H., and Wynn, H. P., editors, *mODa 7 – Advances in Model-Oriented Design and Analysis*, Contributions to Statistics, pages 41–48, Heidelberg. Physica-Verlag.

Bischoff, W. (1992). On exact D-optimal designs for regression models with correlated observations. *Annals of the Institute of Statistical Mathematics*, 44(2):229–238.

Bogacka, B., Johnson, P., Jones, B., and Volkov, O. (2006). D-efficient window experimental designs. *Journal of Statistical Planning and Inference (to appear)*.

Chenel, M., Ogungbenro, K., Duval, V., Laveille, C., Jochemsen, R., and Aarons, L. (2005). Optimal blood sampling time windows for parameter estimation using a population approach: design of a phase ii trial. *Journal of Pharmacokinetics and Pharmacodynamics*, 32(5–6):737–756.

Christensen, R. (1987). *Plane Answers to Complex Questions: The Theory of Linear Models*. Springer, New York.

Cook, R. D. and Goldberg, M. L. (1986). Curvatures for parameter subsets in nonlinear regression. *Annals of Statistics*, 14(4):1399–1418.

Cox, D. R. and Hinkley, D. V. (2000). *Theoretical Statistics*. Chapman & Hall/CRC, Boca Raton.

D'Argenio, D. Z. and Schumitzky, A. (1997). *Adapt II User's guide: Pharmacokinetic/Pharmacodynamic Systems Analysis Software*. Biomedical Simulations Resource, Los Angeles.

Davidian, M. and Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*. Chapman and Hall, London.

Demidenko, E. (2004). *Mixed Models – Theory and Applications*. Wiley Series in Probability and Statistics. Wiley, New Jersey.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Derendorf, H., Gramatté, T., and Schäfer, H. G. (2002). *Pharmakokinetik – Einführung in die Theorie und Relevanz für die Arzneimitteltherapie*. Wissenschaftliche Verlagsgesellschaft mbH, Stuttgart. (in German).

Dette, H. (1997). Designing experiments with respect ot 'standardized' optimality criteria. *Journal of the Royal Statistical Society Series B*, 59:97–110.

Dodds, M. G., Hooker, A. C., and Vicini, P. (2005). Robust population pharmacokinetic experiment design. *Journal of Pharmacokinetics and Pharmacodynamics*, 32(1):33–63.

Duffull, S. B., Retout, S., and Mentré, F. (2002). The use of simulated annealing for finding optimal population designs. *Computer Methods and Programs in Biomedicine*, 69:25–35.

EMEA-CHMP (2006a). *Guideline on reporting the results of population pharmacokinetic analysis (Draft)*. EMEA, London.

EMEA-CHMP (2006b). *Guideline on the role of pharmacokinetics in the development of medicinal products in the paediatric population*. EMEA, London.

Entholzner, M., Benda, N., Schmelter, T., and Schwabe, R. (2005). A note on designs for estimating population parameters. *Biometrical Letters*, 42:25–41.

FDA (1999). *Guidance for Industry - Population Pharmacokinetics*. U.S. Department of Health and Human Services, Food and Drug Administration, Rockville.

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

Fedorov, V. V. and Hackl, P. (1997). *Model-Oriented Design of Experiments*, volume 125 of *Lecture Notes in Statistics*. Springer, New York.

Gaffke, N. and Heiligers, B. (1996). Approximate designs for polynomial regression: Invariance, admissibility, and optimality. In Ghosh, S. and Rao, C. R., editors, *Handbook of Statistics*, volume 13, pages 1149–1199. Elsevier, Amsterdam.

Girard, P. and Mentré, F. (2005). A comparison of estimation methods in nonlinear mixed effects models using a blind analysis. PAGE 14 Abstr. 834.

Gladitz, J. and Pilz, J. (1982). Construction of optimal designs in random coefficient regression models. *Mathematische Operationsforschung und Statistik*, 13:371–385.

Green, B. and Duffull, S. B. (2003). Prospective evaluation of a D-optimal designed population pharmacokinetic study. *Journal of Pharmacokinetics and Pharmacodynamics*, 30(2):145–161.

Han, C. and Chaloner, K. (2004). Bayesian experimental design for nonlinear mixed-effects models with application to HIV dynamics. *Biometrics*, 60:25–33.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, 61:383–385.

Hennig, S., Waterhouse, T. H., Bell, S. C., France, M., Wainwright, C. E., Miller, H., Charles, B. G., and Duffull, S. B. (2006). A D-optimal designed population pharmacokinetic study of oral itraconazole in adult cystic fibrosis patients. *British Journal of Clinical Pharmacology*, page available online.

Kiefer, J. and Wolfowitz, J. (1959). Optimum designs in regression problems. *The Annals of Mathematical Statistics*, 30:271–294.

Kiefer, J. and Wolfowitz, J. (1960). The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366.

Kuhn, E. and Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics and Data Analysis*, 49:1020–1038.

Laird, N. M. and Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics*, 38:963–974.

Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the Amarican Statistical Association*, 83:1014–1022.

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed-effects models for repeated measures data. *Biometrics*, 46(673–687).

Liski, E. P., Mandal, N. K., Shah, K. R., and Sinha, B. K. (2002). *Topics in Optimal Design*, volume 163 of *Lecture Notes in Statistics*. Springer, New York.

Lunn, D. J., Best, N., Thomas, A., Wakefield, J., and Spiegelhalter, D. (2002). Bayesian analysis of population PK/PD model: General concepts and software. *Journal of Pharmacokinetics and Pharmacodynamics*, 29(3):271–307.

Luoma, A. (2000). *Optimal Designs in Linear Regression Models*. Number 747 in Acta Universitatis Tamperensis. Academic Dissertation, University of Tampere, Department of Mathematics, Statistics and Philosophy.

Magnus, J. R. and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Economics*. Wiley Series in Probability and Mathematical Statistics. Wiley, Chichester.

Mentré, F., Dubruc, C., and Thènout, J.-P. (2001). Population pharmacokinetic analysis and optimization of the experimental design for mizolastine solution in children. *Journal of Pharmacokinetics and Pharmacodynamics*, 28(3):299–319.

Mentré, F., Mallet, A., and Baccar, D. (1997). Optimal design in random-effects regression models. *Biometrika*, 84(2):429–442.

Merlé, Y. and Tod, M. (2001). Impact of pharmacokinetic-pharmacodynamic model linearization on the accuracy of population information matrix and optimal design. *Journal of Pharmacokinetics and Pharmacodynamics*, 28(4):363–387.

Pázman, A. (1986). *Foundations of Optimum Experimental Design*. Mathematics and its Applications. D. Reidel Publishing Company.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4(1):12–35.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing. Springer, New York.

Potthoff, R. F. and Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51:313–326.

Pukelsheim, F. (1993). *Optimal Design of Experiments*. Wiley series in Probability and Mathematical Statistics. John Wiley & Sons, New York.

Rao, C. R. (1967). Least squares theory using an estimated dispersion matrix and its applications to measurement of signals. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 355–372. University of California Press.

Rao, C. R. (1973). *Linear Statistical Inference an its Applications*. Wiley, New York, 2. edition.

Retout, S., Dufful, S., and Mentré, F. (2001). Development and implementation of the population Fisher information matrix for the evaluation of population pharmacokinetic designs. *Computer Methods and Programs in Biomedicine*, 65:141–151.

Retout, S. and Mentré, F. (2003). Optimization of individual and population designs using Splus. *Journal of Pharmacokinetics and Pharmacodynamics*, 30(6):417–443.

Retout, S., Mentré, F., and Bruno, R. (2002). Fisher information matrix for non-linear mixed-effects models: evaluation and application for optimal design of enoxaparin population pharmacokinetics. *Statistics in Medicine*, 21:2623–2639.

Rockafellar, R. T. (1972). *Convex Analysis*. Princeton University Press, Princeton.

Schmelter, T. (2006a). Considerations on group-wise identical designs for linear mixed models. *Journal of Statistical Planning and Inference (to appear)*.

Schmelter, T. (2006b). The optimality of single-group designs for certain mixed models. *Metrika (to appear; available online)*.

Schmelter, T., Benda, N., and Schwabe, R. (2006). Some curiosities in optimal designs for random slopes. *O.-v.-G.-University Magdeburg, Faculty for Mathematics, Preprint*, (06-39). (to appear in mODa 8).

Schott, J. R. (1997). *Matrix Analysis for Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, New York.

Schwabe, R. (1996). *Optimum Designs for Multi-Factor Models*, volume 113 of *Lecture notes in statistics*. Springer, New York.

Schwabe, R. and Schmelter, T. (2006). On optimal designs in random intercept models. *O.-v.-G.-University Magdeburg, Faculty for Mathematics, Preprint*, (06-42). (to appear in Tatra Mountains Mathematical Publications).

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York.

Sheiner, L. and Wakefield, J. (1999). Population modelling in drug development. *Statistical Methods in Medical Research*, 8:183–193.

Sheiner, L. B. and Beal, S. L. (1980). Evaluation of methods for estimating population pharmacokinetic parameters. I. Michaelis-Menten model: Routine clinical pharmacokinetic date. *Journal of Pharmacokinetics and Biopharmaceutics*, 8(6):553–571.

Sheiner, L. B. and Beal, S. L. (1981). Evaluation of methods for estimating population pharmacokinetic parameters II. biexponential model and experimental pharmacokinetic data. *Journal of Pharmacokinetics and Biopharmaceutics*, 9:635–651.

Sheiner, L. B., Rosenberg, B., and Marathe, V. (1977). Estimation of population characteristics of pharmacokinetic parameters from routine clinical data. *Journal of Pharmacokinetics and Biopharmaceutics*, 5:445–479.

Sheiner, L. B., Rosenberg, B., and Melmon, K. L. (1972). Modelling of individual pharmacokinetics for computer-aided drug dosage. *Computers and Biomedical Research*, 5:441–459.

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

Smith, K. (1918). On the standad 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.

Steimer, J.-L., Mallet, A., Golmard, J. L., and Boisvieux, J. F. (1984). Alternative approaches to estimation of population pharmacokinetic parameters: Comparison with the nonlinear mixed effect model. *Drug Metabolism Reviews*, 15:265–292.

Stroud, J. R., Müller, P., and Rosner, G. L. (2001). Optimal sampling times in population pharmacokinetic studies. *Applied Statistics*, 50:345–359.

Tod, M., Mentré, F., Merlé, Y., and Mallet, A. (1998). Robust optimal design for the estimation of hyperparameters in population pharmacokinetics. *Journal of Pharmacokinetics and Pharmacodynamics*, 26(6):689–716.

Tod, M. and Rocchisani, J.-M. (1997). Comparison of ED, EID, and API criterio for the robust optimization of sampling times in pharmacokinetics. *Journal of Pharmacokinetics and Biopharmaceutics*, 25(4):515–537.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics. Springer, New York.

Vonesh, E. F. and Chinchilli, V. M. (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. Marcel Dekker, New York.

Wand, M. P. (2002). Vector differential calculus in statistics. *The American Statistician*, 56(1):55–62.

Wang, J. and Endrenyi, L. (1992). A computationally efficient approach for the design of population pharmacokinetic studies. *Journal of Pharmacokinetics and Biopharmaceutics*, 20(3):279–294.

Wikipedia (2006). Pharmacokinetics. Wikipedia, The Free Encyclopedia. [Online; accessed December 20, 2006].

Zyskind, G. (1967). On canonical forms, non-negative covariance matrices and best and simple least squares linear estimators in linear models. *Annals of Mathematical Statistics*, 38(1092–1109).