Approximations of the Fisher Information for the Construction of Efficient Experimental Designs in Nonlinear Mixed Effects Models

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat)

von **Dipl.-Math. Tobias Mielke**
geb. am **13.11.1982** in **Essen**

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

Gutachter: Prof. Dr. Rainer Schwabe
Prof. Dr. Luc Pronzato

eingereicht am: 02.11.2011
Verteidigung am: 17.02.2012

Acknowledgment

Many people supported me during the process of this work, but I would specially like to thank my advisor Prof. Dr. Rainer Schwabe for his supporting comments and interesting discussions on and off the topic of experimental design. It was very helpful to have an advisor, who always encouraged me to present my findings at conferences, what led to useful deadlines for writing down my results.

I am very grateful to the institute of mathematical stochastics at the Otto-von-Guericke University of Magdeburg for a wonderful working atmosphere and to Prof. Dr. Herbert Henning. For every problem, there were open doors where helpful answers or coffee were available. I would like to thank Kerstin Altenkirch for many cups of coffee and training lessons in basketball. Thanks to Prof. Dr. Norbert Gaffke for an interesting course on measure theory and stochastics, for many enlightening discussions and for the daily newspaper of Magdeburg. I was lucky to work together with Fritjof Freise, Dr. Katrin Roth and Dr. Christoph Riethmüller, who were of big help in the case of technical problems and for settling in Magdeburg.

This work is a result of the BMBF project SKAVOE in collaboration with three universities and two pharmaceutic companies. I would like to thank Dr. Thomas Schmelter for his suggestions on interesting topics of science within this project and specially for his work on mixed effects models, which mainly built the foundations of my knowledge on this topic.

My family and friends in Dresden, Magdeburg and Velbert were a great support for starting and finishing this thesis. In particular I am grateful to the Magdeburg/Dresden part of my family for their work on convincing me to start these doctoral studies and my parents for asking me frequently how my studies were going.

I am very glad for the supporting love and patience of Olga Mitrofanova and very thankful for her encouragement, motivation and inspiration.

This work was supported by the BMBF grant SKAVOE 03SCPAB3.

Summary

Mixed effects models are applied for the analysis of grouped data in various areas of science. The fundamental idea of mixed modeling is the assumption that observations within groups of individuals follow a common response structure, whereas observations between groups differ by group wise varying parameters, which are in population studies assumed to be realizations of random variables.

The statistical properties of estimators in nonlinear mixed effects models with bounded individual sample sizes are not sufficiently known, such that experimental designs are based in the examined models on the Fisher information matrix. Approximations of the Fisher information are usually applied for the construction of optimal designs, as the Fisher information can not be represented in a closed form in nonlinear mixed effects models. Aim of this thesis is the derivation of reliable approximations of the Fisher information and the study of the influence of approximations on the designs of the planned studies.

Linear and nonlinear regression models and the basic terms of the experimental design theory are presented after an introductory chapter in the second chapter. A first definition of mixed effects models and estimation methods, designs and information matrices in the linear case are given in the third chapter. Two approximations of the Fisher information are introduced in the following chapter on nonlinear mixed effects models after a short summary of estimation methods. A small example motivates further studies of the Fisher information.

Novel approximations of the Fisher information are developed in the fifth chapter, which are based on a representation of the Fisher information as the variance of the conditional mean of the individual parameters for given observations. This leads to an alternative motivation for the application of the Fisher information resulting from linear mixed effects models in the problematic nonlinear case. An example illustrates the accuracy and the dependence of different approximations on the variance of the individual parameters.

The generalization of the optimal design theory to mixed effects models is the topic of the sixth chapter. The theory on approximations and designs is applied in an example of two pharmacokinetic models in the seventh chapter, which leads to the concluding findings that the approximations proposed in the literature provide a good foundation for the construction of efficient designs. Designs of higher efficiency can however be derived by taking the here introduced approximations of the Fisher information into account.

An outlook on further considerations and a summarizing discussion of the approximations are given in the concluding chapter eight.

Zusammenfassung

Gemischte Modelle werden zur Analyse gruppierter Daten in verschiedensten Forschungsgebieten angewendet. Die grundlegende Idee ist hierbei, dass Beobachtungen innerhalb einer Gruppe einer gemeinsamen Grundstruktur folgen, die jeweiligen Gruppen sich jedoch durch individuelle Parameter unterscheiden, welche in Populationsstudien als Realisierungen von Zufallsvariablen modelliert werden.

Da in nichtlinearen gemischten Modellen die statistischen Eigenschaften der Schätzer im Falle begrenzter individueller Stichprobengrößen nicht ausreichend bekannt sind, wird als Grundlage für die Versuchsplanung die Fisher-Information genutzt, die sich jedoch in den betrachteten Modellen nicht in einer geschlossenen Form darstellen lässt. Ziel der vorliegenden Arbeit ist nun die Herleitung zuverlässiger Approximationen der Fisher-Information und die Überprüfung des Einflusses der Approximationen auf die Effizienz der Versuchspläne.

Nach einem in das Thema einleitenden Kapitel werden lineare und nichtlineare Regressionsmodelle und die Theorie der optimalen Versuchsplanung in diesen Modellen im zweiten Kapitel eingeführt. Modelle mit gemischten Effekten werden erstmals im dritten Kapitel vorgestellt und Schätzer, Versuchspläne und Informationsmatrizen im linearen Fall behandelt. Im folgenden Kapitel über nichtlineare gemischte Modelle werden nach der Zusammenfassung einiger Schätzmethoden zwei in der Literatur verwendete Approximationen der Fisher-Information eingeführt. Ein kleines Beispiel motiviert das weitere Studium der Fisher-Information.

Im fünften Kapitel der Arbeit werden neue Approximationen der Fisher-Information hergeleitet. Die Basis der Approximationen bildet eine Darstellung der Fisher-Information als Kovarianzmatrix des bedingten Erwartungswertes der individuellen Parameter für gegebene Beobachtungen. Dies liefert eine weitere Motivation für die Anwendung der Fisher-Information linearer gemischter Modelle im nichtlinearen Fall. Ein Beispiel illustriert die Abhängigkeit verschiedener Approximationen von der Varianz der individuellen Parameter.

Die Verallgemeinerung der Versuchsplanungstheorie auf Modelle mit gemischten Effekten ist das Thema des sechsten Kapitels, bevor im siebten Kapitel die in der Arbeit vorgestellte Theorie an zwei Beispielen der Populationspharmakokinetik angewandt wird. Zusammenfassend zeigt sich, dass die bisher in der Literatur genutzten Verfahren eine gute Basis zur Konstruktion effizienter Versuchspläne bilden. Eine höhere Effizienz kann jedoch mit Hilfe der neu eingeführten Approximationen erreicht werden.

Im abschließenden achten Kapitel werden die Approximationen zusammenfassend diskutiert und Ausblicke für weiterführende Überlegungen gegeben.

Contents

1 Introduction

Experiments are conducted in sciences in order to obtain ideas on relations, to verify or to discard theories. Bandemer and Bellmann (1994) interpret experiments as questions to the nature, which should be adequately posed to obtain reasonable and useful answers. In these terms the first ideas on optimally questioning were published in an Biometrika article of Kirstine Smith (1918) about a century ago. Although Fisher published a book with the title "The design of experiments" already in 1935, the main development in experimental design theory took place in the second half of the last century with the works of Elfving (1952) and Kiefer and Wolfowitz (1959) on optimally planning experiments. The resulting convex design theory, optimality criteria and equivalence theorems nowadays still form the foundations for the construction of optimal experimental designs.

Atkinson and Bailey (2001) briefly summarize the development of experimental designs in the last century and name some recent areas of research. Specially mentioned are designs for studies with complicated variance structures, designs for computer experiments, adaptive designs and designs for training neural networks. Complicated variance structures might be caused by individual wise varying parameter vectors, which are met in population studies and described by mixed effects models. The development of statistical tools for the analysis of mixed effects models and the new computational possibilities allow the application of population modeling in wide areas in the life sciences and explain hence the growing demand for optimal designs in mixed effects models, which are the topic of this thesis.

The theory on least squares estimators for location parameters in linear mixed effects models was already developed in 1965 by Rao. For generalized linear and nonlinear mixed effects models, reliable estimators are harder to obtain. Sheiner *et al.* (1972) describe some estimation methods in nonlinear mixed effects models, which were further refined and discussed in the following works on evaluations of methods for estimating population parameters in clinical studies by Sheiner and Beal (1980, 1981, 1983). Relatively new approaches aim in minimizing iteratively reweighted sums of squares (Lindstrom and Bates (1990)) or in maximizing the likelihood function by stochastic algorithms (Kuhn and Lavielle (2001)). Quasi-likelihood estimators might be used in generalized linear mixed models for gaining reliable insight in the population behavior (Niaparast (2010)). A generalization of this method for the estimation in nonlinear mixed effects models is in general not possible, as closed form representations of mean and variance of the observations as functions of the unknown parameters are possible only in few cases.

Mixed effects models are frequently used in the analysis of grouped data. In clinical studies the observed individuals usually share common response structures, such that information from individual responses can be merged to obtain efficient estimates. The particular example of pharmacokinetics in paediatrics often serves as a motivation for population modeling. Pharmacokinetic studies are applied in order to examine the absorption, distribution, metabolism and excretion of drugs in the body. The time-concentration relationship of drugs in the body is usually described by compartmental models, which are extensively presented in the monograph on nonlinear regression by Seber and Wild (2003, ch. 8) and in the dissertation of Schmelter (2007a). Experimental designs with many measurements in few individual tend to be inefficient for the estimation of the population mean parameter vector compared to designs with few measurements on many individuals. Moreover, blood samples with dense individual sampling schemes are due to ethical and logistical reasons usually not possible in the target population of ill patients. Nonlinear mixed effects models are used in the analysis of pharmacokinetic studies and encouraged specially for sparse sampling schemes by the EMEA (2006):

> "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."

The main problem in designing experiments for nonlinear mixed effects models is the missing knowledge of the probabilistic behavior of the parameter estimators in the case of limited numbers of individual measurements. The usual approach for circumventing this problem is in the literature the use of the inverse of the Fisher information matrix as a lower bound of the variance of any unbiased estimator. Two different approximations of the Fisher information matrix for nonlinear mixed effects models compete in the literature.

Mentré *et al.* (1997) and Schmelter (2007a) approximate the nonlinear mixed effects model by a linear mixed effects model and optimize the design with respect to the Fisher information matrix of the resulting model. Algorithms and equivalence theorems for optimal designs in linear mixed effects models can be readily generalized from the book of Fedorov (1972). Optimal designs for individual predictions in mixed effects models were assumed to follow from results of Gladitz and Pilz (1982), but recently Prus and Schwabe (2011) reviewed this assumption and presented updated equivalence theorems for this situation.

The second information approximation is based on the Fisher information matrix in heteroscedastic normal models and was developed by Retout and Mentré (2003). Some results on designs in heteroscedastic models were published by Atkinson and Cook (1995) and in the dissertation of Holland-Letz (2009). Although both mentioned approaches for approximating the Fisher information are based on similar linearizations, the resulting information matrices and designs might severely differ, as was discussed in Mielke and Schwabe (2010).

In mixed effects models optimal design problems can be solved analytically only in few cases, but various numerical techniques for addressing design problems might be applied in these models. However, because of the unknown structure of the Fisher information matrix in nonlinear mixed effects models, one has to return to the statement of Bandemer and Bellmann (1994): the questions have to be posed in the right way in order to obtain reasonable and useful answers to the optimal design problem in nonlinear mixed effects models.

Based on linear and nonlinear regression models, we will introduce the fundamental terms in optimal designs of experiments in the second chapter. Different estimation methods are presented in the second chapter, which will be useful for the estimation of the population mean parameter vector in the third chapter on linear mixed effects models. The linear mixed effects models build the foundations for nonlinear mixed effects models, which are introduced in the fourth chapter. After the presentation of different estimation methods in nonlinear mixed effects models, the last subsection of the fourth chapter describes the two mentioned approximations of the Fisher information matrix by the information matrix of a linear mixed effects model and a nonlinear heteroscedastic normal model. This motivates further studies on approximations of the Fisher information, which are presented in the fifth chapter. After these considerations on information matrices in nonlinear mixed effects models, we return in the sixth chapter to the design problem in mixed effects models before presenting the impact of different information matrices on the design of pharmacokinetic studies in two models in the seventh chapter. The concluding eighth chapter summarizes the findings of this thesis and addresses further considerations on experimental designs in nonlinear mixed effects models.

2 Design in Linear and Nonlinear Regression Models

Many general definitions and results on experimental design can be well explained within the topic of ordinary linear regression models and then be generalized to other model classes, as nonlinear and mixed effects models. This chapter provides the foundations for the theory and results of the following chapters.
Experimental designs in linear models are well discussed in many publications. The monographs by Fedorov (1972), Silvey (1980) and Bandemer and Bellmann (1994) build a good introduction to optimal designs of experiments. The book of Pukelsheim (1993) on experimental design describes optimal design problems on an abstract mathematical level. Specially the dissertation of Schmelter (2007a) yields an insight in the topic of optimal designs for mixed effects models. This introductory chapter on optimal designs in linear regression models is based on the mentioned publications.

The linear regression model and some connections to nonlinear models are presented in section 2.1. The second part of this chapter describes the fundamental terms in the theory of optimal designs of experiments. Analytical and numerical approaches to the design problem are given in section 2.3.

2.1 Classical Regression Models

Let the j-th observation under experimental setting $x_j \in \mathcal{X}$ be described by

$$Y_j = \eta(\beta, x_j) + \epsilon_j,\ j = 1, ..., m \tag{2.1}$$

with a real valued response function η, a p dimensional unknown parameter vector β and an observation error $\epsilon_j \in \mathbb{R}$ with zero mean and variance $\sigma^2 > 0$. In the classical regression framework the errors ϵ_j are assumed to be uncorrelated and homoscedastic. The real valued response function η is assumed to be differentiable in β and continuous on the *design region* $\mathcal{X}$, which for technical reasons is assumed to be a compact set. Primary aim is the estimation of the underlying response function $\eta(\beta, x)$, $x \in \mathcal{X}$.

2.1.1 Linear Models

In linear regression models the function η is linear in the parameter vector β, such that

$$Y_j = f(x_j)^T\beta + \epsilon_j,\ \ j = 1, ..., m \tag{2.2}$$

with a vector of known real valued regression functions $f(x) = \big(f_1(x), ..., f_p(x)\big)^T$. Note that the *design locus* $f(\mathcal{X}) := \{y|y = f(x), x \in \mathcal{X}\}$ is a compact set, as $f(x)$ is continuous on the compact design region $\mathcal{X}$. The vector $Y = (Y_1, ..., Y_m)^T$ summarizes the m observations in one model with the *design matrix* $F = \big(f(x_1), ..., f(x_m)\big)^T$ and an error vector $\epsilon = (\epsilon_1, ..., \epsilon_m)^T$ as

$$Y = F\beta + \epsilon. \tag{2.3}$$

The above assumption on the correlation of the observation errors leads to a model of Y with expectation and variance as

$$E(Y) = F\beta \ \text{ and } \ Cov(Y) = \sigma^2 I_m.$$

The best linear unbiased estimator is obtained by minimizing the squared distance of the proposed model from the observed values:

$$L_{OLS}(\beta; y) := (y - F\beta)^T(y - F\beta) \rightarrow \min_{\beta\in\mathbb{R}^p}.$$

and results for design matrices F with full column rank in

$$\widehat{\beta}_{OLS} := (F^TF)^{-1}F^TY.$$

Under the classical assumptions of uncorrelated and homoscedastic observation errors with variance σ^2, the covariance of the ordinary least squares estimator is obtained as

$$Cov(\widehat{\beta}_{OLS}) = \sigma^2(F^TF)^{-1}.$$

Note however, that the classical assumptions are generally not fulfilled. The observation errors are not necessarily uncorrelated and the variance may depend on the experimental settings. Then the variance matrix of the observation vector is of the form

$$Cov(Y) = \sigma^2 V$$

with a matrix V, which is here assumed to be positive definite and independent of the vector β. In this new model, the best linear unbiased estimator takes the variance of the measurements into account. The solution of

$$L_{WLS}(\beta; y) := (y - F\beta)^T V^{-1}(y - F\beta) \rightarrow \min_{\beta\in\mathbb{R}^p}$$

is called the weighted least squares estimator and is obtained for matrices F of full column rank as

$$\widehat{\beta}_{WLS} := (F^TV^{-1}F)^{-1}F^TV^{-1}Y.$$

The covariance of the estimator results in

$$Cov(\widehat{\beta}_{WLS}) = \sigma^2 (F^T V^{-1} F)^{-1}.$$

For unknown variance matrices V, generalized least squares procedures are often applied, which iteratively modify a weight matrix for optimally estimating the vector β. Estimates of the true variance matrix V are usually used as weight matrices. Further details on the generalized least squares estimation will be presented in section 3.2.2.

Notice that for rank deficient design matrices F no linear unbiased estimator of β exists. If just the estimation of a linear function $\psi(\beta) = L_\psi \beta$ is of interest, the existence of a matrix Q with $L_\psi = QF$ is sufficient for the identifiability of $\psi(\beta)$:

$$\widehat{\psi} = L_\psi (F^T V^{-1} F)^- F^T V^{-1} Y$$

with a generalized inverse $(F^T V^{-1} F)^-$ of $F^T V^{-1} F$. The covariance of the estimator results in

$$Cov(\widehat{\psi}) = \sigma^2 L_\psi (F^T V^{-1} F)^- L_\psi^T.$$

2.1.2 Nonlinear Models

The estimation for nonlinear functions η can be conducted in a similar way. Let the vector valued response function be therefore described as $\eta(\beta) = \big(\eta(\beta, x_1), ..., \eta(\beta, x_m)\big)^T$.
In the nonlinear model with uncorrelated and homoscedastic observation errors ϵ_i with variance $\sigma^2 > 0$, the squared distance

$$L_{OLS}(\beta; y) := \big(y - \eta(\beta)\big)^T \big(y - \eta(\beta)\big)$$

is to be minimized. The derivatives of the response function with respect to the parameter take here the role of the design matrix, such that we define:

$$F_\beta := \frac{\partial \eta(\beta)}{\partial \beta^T},$$

what depends on the unknown parameter vector β, which is to be estimated. The ordinary least-squares estimator is obtained as a root of the estimating equation

$$l'_{OLS}(\beta; y) := F_\beta^T \big(y - \eta(\beta)\big) = \frac{\partial \eta(\beta)^T}{\partial \beta} \big(y - \eta(\beta)\big) \stackrel{!}{=} 0.$$

With n replications of the measurements under the experimental settings $(x_1, ..., x_m)$ and under appropriate regularity conditions, the ordinary least squares estimator $\widehat{\beta}_{OLS}$ is asymptotically normally distributed (e.g. Jennrich (1969)):

$$\sqrt{n}(\widehat{\beta}_{OLS} - \beta) \xrightarrow{\mathfrak{L}} \mathcal{N}\big(0, \sigma^2 (F_\beta^T F_\beta)^{-1}\big) \text{ as } n \to \infty.$$

In the more general model of possibly correlated and heteroscedastic observations errors, the variance of the observation vector results in

$$Cov(Y) = \sigma^2 V$$

with a matrix V as in 2.1.1. The uncertainty of the observations should then be taken into account for the estimation. The weighted squared distance

$$L_{WLS}(\beta; y) := \big(y - \eta(\beta)\big)^T V^{-1} \big(y - \eta(\beta)\big)$$

is minimized by the weighted least squares estimator $\widehat{\beta}_{WLS}$, which under some regularity conditions is asymptotically normally distributed as

$$\sqrt{n}(\widehat{\beta}_{WLS} - \beta) \xrightarrow{\mathfrak{L}} \mathcal{N}\big(0, \sigma^2 (F_\beta^T V^{-1} F_\beta)^{-1}\big) \text{ as } n \to \infty.$$

For the numerical derivation of nonlinear least squares estimates, the response function might be linearized around a guess β_0 of the true parameter vector β. For β_0 close enough to β, this leads under the assumption of negligible linearization errors to a linear model as in 2.1.1. Note that the root of the estimating equation needs not be unique.

2.1.3 Maximum Likelihood Estimation

Least squares estimation yields in linear models unbiased estimates of the parameter β specifying the location, even without knowledge of the underlying distribution of the observation error. If the error distribution is assumed to be an element of a parametric family of probability distributions, *Maximum Likelihood Estimation* can be applied for fitting the model to the data. Let $f_Y(y, \theta)$ describe the probability density of Y with a parameter θ to be estimated. For given realizations y of Y, the maximization of the likelihood function $L(\theta; y)$ or equivalently the log-likelihood function $l(\theta; y)$:

$$L(\theta; y) := f_Y(y, \theta) \quad \text{and} \quad l(\theta; y) := \log[L(\theta; y)]$$

with respect to the parameter θ yields the maximum likelihood estimate $\widehat{\theta}_{ML}$. The maximum likelihood estimate defines the probability distribution, which optimally describes the observed data within the proposed class of probability distributions $f_Y(y, \theta)$ and is obtained as a root of the score function:

$$l'(\theta; y) := \frac{\partial l(\theta; y)}{\partial \theta} \stackrel{!}{=} 0.$$

Maximum likelihood estimators are under some regularity conditions asymptotically optimal within the class of asymptotically unbiased estimators. The variance of the maximum likelihood estimator then asymptotically reaches the Cramer-Rao bound, a lower bound of the variance of any unbiased estimator, which corresponds to the inverse of the Fisher information:

Definition 2.1. *For a given parametric model $f_Y(y, \theta)$ of a random vector Y with*

$$E\left(\frac{\partial l(\theta; Y)}{\partial \theta}\right) = 0$$

and under appropriate regularity conditions, the covariance of the score function is called the Fisher information:

$$\mathfrak{M} := E\left(\frac{\partial l(\theta; Y)}{\partial \theta} \frac{\partial l(\theta; Y)}{\partial \theta^T}\right).$$

With n replications of the experimental settings $(x_1, ..., x_m)$, the maximum likelihood estimator is under some more regularity conditions asymptotically normally distributed:

$$\sqrt{n}(\widehat{\theta}_{ML} - \theta) \xrightarrow{\mathfrak{L}} \mathcal{N}(0, \mathfrak{M}^{-1}) \text{ as } n \to \infty.$$

Maximum likelihood and least squares estimators generally yield different results. For the classical assumption of uncorrelated and homoscedastic normally distributed observation errors ϵ_i with a constant variance $\sigma^2 > 0$, maximum likelihood and ordinary least squares estimator in the presented model (2.1) coincide. Specially for nonlinear response functions η the least squares estimators share the asymptotic results of maximum likelihood estimators on unbiasedness and asymptotic efficiency. However, maximum likelihood estimators might lead for misspecified families of distributions of Y to biased inefficient estimates, whereas the asymptotic results on least squares estimators are based on the zero expectation and the variance matrix $\sigma^2 V$ of the observation errors only. Hence least squares estimators are more robust with respect to model missspecifications compared to maximum likelihood estimators. For the introduction on design of experiments we will thus concentrate on designs minimizing the variance of least squares estimators.

2.2 Experimental Design

In the following sections we will focus on linear models with uncorrelated and homoscedastic observation errors ϵ, such that $Cov(\epsilon) = \sigma^2 I_m$. The presented results can be readily generalized for nonlinear response functions, by taking the dependence of the asymptotic covariance matrix of the estimator on the unknown parameter vector β into account. The situation of more general variance structures $\sigma^2 V$ will be considered in the following chapters.

2.2.1 Information and Design

Given the linear model (2.3), the variance of the best linear unbiased estimator of the location parameter vector β depends on the sample settings x_j, $j = 1, ..., m$ and the variance of the observations only.

Definition 2.2. *An exact design $\mathbf{x}$ of size m is a vector $(x_1, ..., x_m)$ of experimental settings x_i, $i = 1, ..., m$ chosen from a design region $\mathcal{X}$. The settings x_i, $i = 1, ..., m$ are called support points of the design $\mathbf{x}$. The set of all exact designs of size m is given by the m-dimensional design region $\mathcal{X}^m$.*

For optimally estimating the vector β, the variance of the estimator has to be minimized with respect to the design settings

$$Cov(\widehat{\beta}_{OLS}) = \sigma^2 \big(F(\mathbf{x})^T F(\mathbf{x})\big)^{-1} \to \min_{\mathbf{x} \in \mathcal{X}^m},$$

with design matrices $F(\mathbf{x})$. The influence of the design $\mathbf{x}$ on the information is contained in the matrix $F(\mathbf{x})^T F(\mathbf{x})$, which is therefore generally called *information matrix*. Up to the constant $\frac{1}{\sigma^2}$, the information matrix coincides with the Fisher information for normally distributed homoscedastic observation errors.

Definition 2.3. *For an exact design* $\mathbf{x}$ *of size* m *the matrix*

$$\mathbf{M}(\mathbf{x}) := \frac{1}{m} F(\mathbf{x})^T F(\mathbf{x})$$

is called the normalized information matrix of the design $\mathbf{x}$.

Exact designs with k distinct experimental settings x_i and m_i measurement replications under the experimental settings x_i, $i = 1, ..., k$, are in the literature often described with the notation

$$\mathbf{x} \sim \begin{pmatrix} x_1 & ... & x_k \\ m_1 & ... & m_k \end{pmatrix}.$$

With a matrix of weights $W(\mathbf{x}) = diag(\omega_1, ..., \omega_k)$, where $\omega_i := \frac{m_i}{m}$ and now a $k \times p$ design matrix $F(\mathbf{x}) := \big(f(x_1), ..., f(x_k)\big)^T$ the normalized information can be described as

$$\mathbf{M}(\mathbf{x}) = \frac{1}{m} \sum_{j=1}^{k} m_j f(x_j) f(x_j)^T = \sum_{j=1}^{k} \omega_j f(x_j) f(x_j)^T = F(\mathbf{x})^T W(\mathbf{x}) F(\mathbf{x}),$$

such that the resulting matrix does not depend on the total number of measurements m any more, but only on the proportions $\omega_1, ..., \omega_k$ with

$$\sum_{i=1}^{k} \omega_i = 1$$

of the experimental settings x_i, $i = 1, ..., k$ in the design $\mathbf{x}$. Hence designs can be described by measures $\xi_{\mathbf{x}}$ as in Silvey (1980):

$$\xi_{\mathbf{x}} = \sum_{i=1}^{k} \omega_i \delta_{x_i}.$$

For constructing optimal designs thus probability distributions on the design region $\mathcal{X}$ are sought, which minimize the variance of the estimator $\widehat{\beta}_{OLS}$ by maximizing the information

$$\mathbf{M}(\xi_{\mathbf{x}}) := E_{\xi_{\mathbf{x}}}\big(f(x) f(x)^T\big) = \int_{\mathcal{X}} f(x) f(x)^T \xi_{\mathbf{x}}(dx).$$

Definition 2.4. *The measure* $\xi = \sum_{i=1}^{k} \omega_i \delta_{x_i}$ *with support points* $x_i \in \mathcal{X}$ *and weights* $\omega_i \geq 0$, $i = 1, ..., k$; $\sum_{i=1}^{k} \omega_i = 1$, *for some* $k \in \mathbb{N}$, *is called approximate design on* $\mathcal{X}$.

$$\xi := \begin{pmatrix} x_1 & ... & x_k \\ \omega_1 & ... & \omega_k \end{pmatrix}.$$

Contrary to exact designs, the approximate designs build a convex set and its closure leads to the set of all probability measures on $\mathcal{X}$.

Definition 2.5. *Let*

(i) Ξ *denote the set of all approximate designs on* $\mathcal{X}$*,*

(ii) Ξ_{L_ψ} *denote the set of all approximate designs, under which the linear aspect* $\psi(\beta) = L_\psi \beta$ *is identifiable,*

(iii) $\Xi_\beta := \Xi_{I_p}$.

Designs $\xi \in \Xi \setminus \Xi_\beta$ *are called singular designs.*

The set Ξ induces the set of information matrices

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

The linearity of the integral yields with the compactness of the design locus the convexity and compactness of the set $\mathcal{M}$. Even more, with the proposed assumptions on uncorrelated observations, every information matrix $\mathbf{M} \in \mathcal{M}$ can be constructed as a convex combination of information matrices of design measures with mass on one support point only.

Since the information matrices are symmetric, they can be represented as vector valued functions with image in $\mathbb{R}^{\frac{1}{2}p(p+1)}$. With an application of Caratheodory's theorem, the number of support points of optimal designs can be limited, since every point in the convex hull of the set $\mathcal{M}$ can be represented as a convex combination of a finite number of points of $\mathcal{M}$.

Theorem 2.6. *For every design* $\xi \in \Xi$ *an approximate design* $\bar{\xi}$ *exists with* $k \leq \frac{1}{2}p(p+1)+1$ *support points, satisfying* $\mathbf{M}(\xi) = \mathbf{M}(\bar{\xi})$*. For boundary elements* $\mathbf{M}(\xi)$ *of* $\mathcal{M}$ *a design* $\bar{\xi}$ *exists with at most* $\frac{1}{2}p(p+1)$ *support points and* $\mathbf{M}(\xi) = \mathbf{M}(\bar{\xi})$*.*

As a consequence it is sufficient to optimize designs in the class of approximate designs with at most $\frac{1}{2}p(p+1)+1$ support points only.

2.2.2 Design Criteria

Note that the covariance of the estimator is for parameter vectors of dimension $p \geq 2$ a matrix. Hence covariances induced by different experimental designs cannot be compared straightforwardly. The construction of optimal designs with respect to the Loewner partial ordering on the set of non-negative definite matrices would be desirable:

Definition 2.7. *The Loewner partial ordering for non-negative definite* (*n.n.d.*) *matrices* $\mathbf{M}_1$ *and* $\mathbf{M}_2$ *is defined as*

$$\mathbf{M}_1 \leq \mathbf{M}_2 :\Leftrightarrow (\mathbf{M}_2 - \mathbf{M}_1) \ \textit{n.n.d.}$$

Unfortunately this optimization task is not practicable, as generally no Loewner optimal design exists. To circumvent this problem, *design criteria* Φ are used to determine efficient designs with respect to the aim of the planned studies. Design criteria are real valued functions Φ of the information matrices or alternatively of the designs defining the information matrices.

With the proposed asymptotic normality of the least squares estimator for β, the content of the confidence ellipsoid is inversely proportional to the square root of the determinant of the information matrix, yielding the D-optimality criterion. The determinant is in this thesis described by $\|\cdot\|$. The D-optimality criterion then takes the form:

Definition 2.8. *A design ξ^* is called D-optimal if*

$$\| \mathbf{M}(\xi^*) \| \;\; \geq \;\; \| \mathbf{M}(\xi) \|, \; \forall \; \xi \in \Xi.$$

The design ξ^* minimizes $\Phi_D\big(\mathbf{M}(\xi)\big) := -\log\big(\| \mathbf{M}(\xi) \| \big)$ on Ξ.

The D-optimality criterion is one of the most popular design criteria in the literature. Welcome properties of D-optimal designs are the invariance with respect to linear reparameterizations of the parameter vector and the equivalence of D- and G-optimal designs in the case of uncorrelated homoscedastic measurement errors. The G-optimality criterion is used for minimizing the maximal variance of the estimate of a response over the design region $\mathcal{X}$:

Definition 2.9. *A design ξ^* is called G-optimal if*

$$\max_{x \in \mathcal{X}} \; f(x)^T \mathbf{M}(\xi^*)^{-1} f(x) \;\; \leq \;\; \max_{x \in \mathcal{X}} \; f(x)^T \mathbf{M}(\xi)^{-1} f(x), \; \forall \; \xi \in \Xi_\beta.$$

The design ξ^* minimizes $\Phi_G\big(\mathbf{M}(\xi)\big) := \max_{x \in \mathcal{X}} \; f(x)^T \mathbf{M}(\xi)^{-1} f(x)$ on Ξ_β.

Alternatively *linear criteria* are induced by non-negative definite $p \times p$ matrices L. Designs can be optimized with respect to different aspects in dependence of the choice of the matrix L. Designs with singular information matrices might be optimal for the L-optimality criteria, such that a generalized inverse $\mathbf{M}^-$ of the information matrix is used to define these criteria.

Definition 2.10. *A design ξ^* is called L-optimal if*

$$\text{tr} \, \big(L\mathbf{M}(\xi^*)^-\big) \;\; \leq \;\; \text{tr} \, \big(L\mathbf{M}(\xi)^-\big), \; \forall \; \xi \in \Xi_L.$$

The design ξ^* minimizes $\Phi_L\big(\mathbf{M}(\xi)\big) := \text{tr} \, \big(L\mathbf{M}(\xi)^-\big)$ on Ξ_L.

The $IMSE$-optimality is used if interest lies in minimizing the mean variance of the estimated response over the design region. Depending on a probability measure μ on the design region $\mathcal{X}$, the **I**ntegrated **M**ean **S**quared **E**rror is minimized, what results for ordinary linear models and the least squares estimator in:

$$E\Big(\int_{\mathcal{X}} [f(x)^T\beta - f(x)^T\widehat{\beta}]^2 \mu(dx) \Big) = \text{tr} \, \Big(\int_{\mathcal{X}} f(x) f(x)^T \mu(dx) \; \mathbf{M}(\xi)^{-1} \Big) \to \min_{\xi \in \Xi_\beta},$$

such that with

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

the $IMSE$-criterion is a particular case of definition 2.10.

Other particular linear criteria are the A-optimality criterion and the c-optimality. A-optimal designs minimize the average variance of the parameter estimator, such that for A-optimality the matrix L is the p-dimensional identity. Note that the value of the optimality criterion is influenced by the scale of the components of the parameter vector for the A-criterion.

Aim of the c-optimality is the minimization of the variance of an estimator of the linear aspect $c^T\beta$. Only the linear aspect $c^T\beta$ has here to be identifiable, such that the design optimization is based on the set Ξ_c instead of Ξ_β:

Definition 2.11. *A design ξ^* is called c-optimal if*

$$c^T\mathbf{M}(\xi^*)^-c \quad \leq \quad c^T\mathbf{M}(\xi)^-c, \ \forall \ \xi \in \Xi_c.$$

The design ξ^* minimizes $\Phi_c\big(\mathbf{M}(\xi)\big) := c^T\mathbf{M}(\xi)^-c$ on Ξ_c.

The presented criteria can be similarly applied for the optimization of designs for the estimation of arbitrary linear aspects $\psi(\beta)$. Design optimization is then constrained on the set Ξ_{L_ψ}.

This list of optimality criteria is in no way comprehensive. Possible nonstandard criteria are E- and MV-optimality. E-optimal designs minimize the maximal eigenvalue of the variance matrix, whereas MV-optimal designs minimize the maximal diagonal element of the variance matrix, and hence an upper bound of the variance of the estimator.

In nonlinear models the information matrix generally depends on the prior unknown parameter vector β. Bayesian optimality criteria circumvent this problem by including the uncertainty on the parameter in the criterion function via a probability distribution on the unknown parameter vector β. Another approach in order to circumvent the problem of parameter misspecifications is the use of adaptive designs (e.g. Pronzato (2010)).

Desired properties of design criteria are the monotonicity with respect to the Loewner partial ordering and the convexity:

Definition 2.12. *An optimality criterion is called*

(i) monotone, if for information matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ holds

$$\mathbf{M}_1 \geq \mathbf{M}_2 \Rightarrow \Phi(\mathbf{M}_1) \leq \Phi(\mathbf{M}_2).$$

(ii) convex, if for information matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ holds

$$\Phi\big(\alpha\mathbf{M}_1 + (1-\alpha)\mathbf{M}_2\big) \leq \alpha\Phi(\mathbf{M}_1) + (1-\alpha)\Phi(\mathbf{M}_2), \ \forall \ \alpha \in [0,1].$$

For strictly monotone criteria, the optimal information matrix $\mathbf{M}(\xi^*)$ is a boundary point of the set $\mathcal{M}$. Hence each optimal design ξ^* can be represented as a convex combination of at most $\frac{1}{2}p(p+1)$ one-point designs. Notice that strictly convex design criteria yield an unique optimal information matrix, whereas the optimal design is not necessarily unique.

2.3 Design Optimization

2.3.1 Equivalence Theorems

Equivalence theorems are used in the convex design theory to prove the optimality of designs and to construct optimal designs. With the Fréchet derivative of Φ at $\mathbf{M}_1$ in the direction of $\mathbf{M}_2$:

$$F_\Phi(\mathbf{M}_1, \mathbf{M}_2) := \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}\big(\Phi[(1-\varepsilon)\mathbf{M}_1 + \varepsilon\mathbf{M}_2] - \Phi[\mathbf{M}_1]\big),$$

a design $\xi^* \in \Xi$ is Φ-optimal if and only if

$$F_\Phi\big(\mathbf{M}(\xi^*), \mathbf{M}(\xi)\big) \geq 0 \ \forall \ \xi \in \Xi \text{ (e.g. Silvey (1980))}.$$

The linearity of $F_\Phi\big(\mathbf{M}_1, \mathbf{M}_2\big)$ in the second argument can be shown for differentiable criteria Φ, such that for approximate designs ξ with support points x_i and weights ω_i, $i = 1, ..., k$, holds:

$$F_\Phi\big(\mathbf{M}(\xi^*), \mathbf{M}(\xi)\big) = \sum_{i=1}^{k} \omega_i F_\Phi\big(\mathbf{M}(\xi^*), f(x_i)f(x_i)^T\big),$$

what directly yields the equivalence theorem:

Theorem 2.13. *(Silvey (1980)) If Φ is differentiable on $\mathcal{M}^+ := \{\mathbf{M} \in \mathcal{M},\ \Phi(\mathbf{M}) < \infty\}$ and a Φ-optimal design exists, then ξ^* is Φ optimal if and only if*

$$\min_{x\in\mathcal{X}} F_\Phi\big(\mathbf{M}(\xi^*), f(x)f(x)^T\big) = \max_{\xi} \min_{x\in\mathcal{X}} F_\Phi\big(\mathbf{M}(\xi^*), f(x)f(x)^T\big),$$

where the maximum with respect to ξ is the maximum over $\{\xi \in \Xi,\ \mathbf{M}(\xi) \in \mathcal{M}^+\}$.

Note that support points of the optimal design ξ^* are minima of the Fréchet derivative in ξ^*. The Fréchet derivative shows the effect on the criterion function, when marginally moving from a matrix $\mathbf{M}_1$ to a matrix $\mathbf{M}_2$. Alternatively it can be derived as

$$F_\Phi(\mathbf{M}_1, \mathbf{M}_2) := \frac{\partial}{\partial\varepsilon}\Phi[(1-\varepsilon)\mathbf{M}_1 + \varepsilon\mathbf{M}_2]|_{\varepsilon=0}.$$

For the proposed optimality criteria of D- and L-optimality the equivalence theorems can be readily calculated with some matrix differential calculus (e.g. Fedorov (1972, ch. 2)):

Theorem 2.14. *The design ξ^* is D- and G- optimal if and only if*

$$\max_{x\in\mathcal{X}} \text{tr}\ [\mathbf{M}(\xi^*)^{-1} f(x)f(x)^T] \leq p.$$

Notice that this theorem holds generally only under the assumption of uncorrelated observation errors. The result for linear optimality criteria and regular information matrices is similar:

Theorem 2.15. *The design ξ^* is L-optimal if and only if*

$$\text{tr}\ [\mathbf{M}(\xi^*)^{-1}L\mathbf{M}(\xi^*)^{-1}\big(f(x)f(x)^T - \mathbf{M}(\xi^*)\big)] \leq 0\ \forall\ x \in \mathcal{X}.$$

Silvey (1978) discusses the problem of optimal design measures with singular information matrices. Note that the above equivalence theorem for L-optimality is then not applicable and alternative considerations to prove optimality have to be undertaken.
The optimality of certain approximate designs $\xi \in \Xi_\beta$ can be relatively easy verified using the functions

$$\begin{aligned} g_{D,\xi}(x) &:= \text{tr}\ [\mathbf{M}(\xi)^{-1}f(x)f(x)^T] - p \text{ and} \\ g_{L,\xi}(x) &:= \text{tr}\ \big[\mathbf{M}(\xi)^{-1}L\mathbf{M}(\xi)^{-1}\big(f(x)f(x)^T - \mathbf{M}(\xi)\big)\big]. \end{aligned}$$

which are some kind of sensitivity functions for D- and L-optimality. Designs with non-positive functions $g_{D,\xi}$ or $g_{L,\xi}$ on $\mathcal{X}$ are optimal with respect to the corresponding optimality criterion.

2.3.2 Construction of Optimal Designs

The analytical construction of optimal designs is just in special cases possible. The information provided by the Fréchet derivative can be used in hill climbing optimization algorithms as the V- or the W-algorithm. A design ξ_n with $x_n \in \mathcal{X}$ fulfilling $g_{\cdot,\xi_n}(x_n) > 0$ can be improved by adding weight to the point x_n, such that a new design ξ_{n+1} results in

$$\xi_{n+1} := (1-\alpha_n)\xi_n + \alpha_n\delta_{x_n}$$

The convergence of the algorithm depends on the choice of the sequence of step lengths $\alpha_n \in (0,1)$ and the added design point x_n. In the V-algorithm the step length is defined to maximize the decrease of the criterion function in α_n for given x_n $\big($Fedorov (1972)$\big)$. The steepest decent is attained when adding optimal weight to the point x_n maximizing the sensitivity function. Alternatively the step length can be defined by sequences fulfilling

$$\alpha_n \to 0 \text{ and } \sum \alpha_n \to \infty,$$

what then leads to the W-algorithm $\big($Wynn (1970)$\big)$. Typically step lengths as $\alpha_n = n^{-1}$ are applied. Specially for design regions $\mathcal{X}$ with higher dimensions, the location of the maximum of the functions $g_{\cdot,\xi_n}$ complicates the problem.

With the proposed result on the representation of the optimal design by a design with at most $k = \frac{1}{2}p(p+1)+1$ support points, the design problem can be formulated as

$$\Phi\big(\mathbf{M}(\xi)\big) \to \min_{\xi\in\Xi_k} \text{ with } \Xi_k := \left\{\mathcal{X}^k \times [0,1]^{k-1},\ \sum_{i=1}^{k-1}\omega_i \leq 1\right\}.$$

such that standard numerical algorithms can be applied for solving this optimization problem.

Figure 2.1 shows the functions g_{D,ξ_n} for D-optimality in a nonlinear model with the exponential decay $\eta(\beta,x) = \beta_1\exp(x\beta_2)$, resulting from iterations of the V- and a $BFGS$-algorithm.

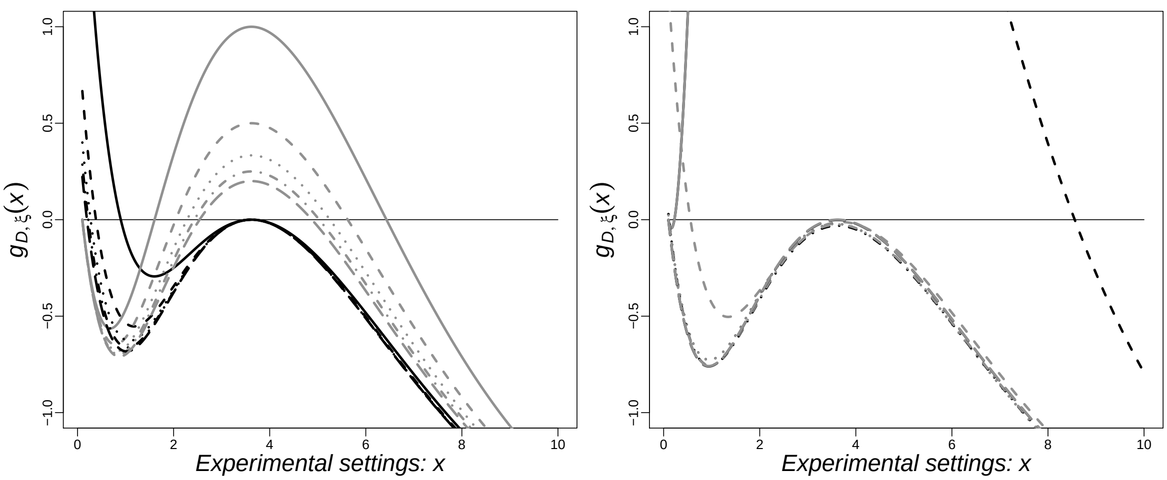

Figure 2.1: 10 iterations of the optimization algorithms: Left: V-algorithm; Right: $BFGS$-Algorithm

Notice that optimal designs in the present nonlinear example depend on the value of the parameter $\beta = (\beta_1, \beta_2)$ and are hence only locally optimal. For both algorithms the same initial design was used and the functions g_{D,ξ_n} were plotted for the first ten iterations. An often observed behavior of standard design algorithms as the V-algorithm is the cycling on a set of points as described by Silvey (1980). Every second iteration in the present example yields a function g_{D,ξ_n} with approximately the same maxima as two iterations before. The design points in these iterations just slightly change, only the weights get updated until convergence.

In the proposed example, optimal designs can be analytically derived without big problems. The sign of the derivative of the function g_{D,ξ_n} here depends on a quadratic polynomial of x, such that at most one maximum in the interior of $\mathcal{X}$ exists. Hence the D-optimal design is supported on at most two points with equal weights on each and results for real design regions $\mathcal{X} = [x_l, x_u]$ and a parameter $\beta = (\beta_1, \beta_2)^T$ in

$$\xi^* = \begin{pmatrix} x_l & \min(x_l + \frac{1}{|\beta_2|}, x_u) \\ 0.5 & 0.5 \end{pmatrix} \text{ for } \beta_2 < 0 \text{ and } \xi^* = \begin{pmatrix} \max(x_u - \frac{1}{\beta_2}, x_l) & x_u \\ 0.5 & 0.5 \end{pmatrix} \text{ for } \beta_2 > 0.$$

3 Linear Mixed Effects Models

The presented results in the second chapter were mainly based on the uncorrelated homoscedastic observation errors. If measurements from a population of individuals are drawn, the observations within one individual are possibly correlated, such that the model assumptions of chapter 2 are then generally not fulfilled. In the analysis of grouped data often mixed effects models are applied for modeling the correlated observations by individual and observation-wise varying random effects. Before discussing nonlinear mixed effects models, this chapter presents some definitions and results on linear mixed effects models.
The books by Davidian and Giltinan (1995) and Demidenko (2005) provide an insight on estimators and their distributions in mixed effects models. Numerical approaches for solving occurring nonlinear optimization problems are extensively discussed in Demidenko's book. Pinheiro and Bates (2000) discuss computational methods and tools for analyzing mixed effects models with normally distributed random effects in the statistical software **S**.

The linear mixed effects model will be introduced in the first section with special emphasis on the random coefficient regression model. Estimation methods for the linear mixed effects model will be described in the second section. The experimental designs in mixed effects models usually consist of two levels, which will be described together with the information matrices for mixed effects models in the third section.

3.1 Model Formulation

Mixed effects models are generally defined in two stages. In the random coefficient regression model, the observations of each individual are assumed to follow the same statistical model (*intra-individual*) with individual-wise varying parameters (*inter-individual*). A generalized influence of individual effects will be given at the end of this section.

3.1.1 Intra-Individual Model

The j-th observation of the i-th individual, $i = 1, ..., N$ under experimental settings $x_{ij} \in \mathcal{X}$ is with a vector of known regression functions $f(x_{ij})$ described by

$$Y_{ij} = f(x_{ij})^T \beta_i + \epsilon_{ij}, \quad j = 1, ..., m_i,$$

with uncorrelated and homoscedastic observation errors ϵ_{ij}, $j = 1, ..., m_i$ of zero mean and variance $\sigma^2 > 0$. The vector of regression functions $f(x)$ is assumed to be continuous on the compact design region $\mathcal{X}$. Hence the *intra-individual* model is an ordinary linear regression model with an unknown parameter vector $\beta_i \in \mathbb{R}^p$. The vector of the m_i observations under experimental settings $(x_{i1}, ..., x_{im_i})$ of the i-th individual can be summarized as

$$Y_i = F_i\beta_i + \epsilon_i,$$

with a design matrix $F_i := \big(f(x_{i1}), ..., f(x_{im_i})\big)^T$ and an error vector $\epsilon_i = (\epsilon_{i1}, ..., \epsilon_{im_i})^T$. Note that the number of measurements m_i is not necessarily constant among the individuals and that more complicated covariance structures of the observation errors than $\sigma^2 I_{m_i}$ might be considered as well. General covariance structures will be briefly discussed in the last subsection of this chapter.

3.1.2 Inter-Individual Model

For modeling the observations of different individuals together in one model, the individual parameter vectors in the random coefficient regression model are assumed to be realizations of uncorrelated and identically distributed random variables with the so called *population parameters*

$$E(\beta_i) = \beta, \text{ and } Cov(\beta_i) = \sigma^2 D.$$

The individual parameter vector can alternatively be interpreted as the sum of a fixed population parameter β and an individual effect b_i with zero mean and variance $\sigma^2 D$. If some of the components of β_i are assumed not to vary between the individuals, the correspondent rows and columns of the covariance matrix $\sigma^2 D$ are zero.

Example 3.1. *In the case of quadratic regression the difference of the individual response functions*

$$\eta(\beta_i, x) := \beta_{i;1} + \beta_{i;2}x + \beta_{i;3}x^2.$$

for matrices $D = diag(d_1, 0, 0)$ with $d_1 > 0$ is only induced by the individual-wise varying intercepts, i.e. $\beta_i = (\beta_{i;1}, \beta_2, \beta_3)$.

Observation errors and individual parameter vectors are generally assumed to be uncorrelated. For the observations within one individual this yields

$$Y_i = F_i\beta + F_i b_i + \epsilon_i,$$
$$E(Y_i) = F_i\beta, \; Cov(Y_i) = \sigma^2 V_i, \text{ with } V_i := I_{m_i} + F_i D F_i^T,$$

and for $F := (F_1^T, ..., F_N^T)^T$, $b = (b_1^T, ..., b_N^T)^T$, $\epsilon = (\epsilon_1^T, ..., \epsilon_N^T)^T$ and $G := diag(F_1, ..., F_N)$, the model of all observations can be summarized as

$$Y = F\beta + Gb + \epsilon.$$

Hence the linear mixed effects model is a linear model as in chapter 2, with mean and variance as

$$E(Y) = F\beta \text{ and } Cov(Y) = \sigma^2 V := \sigma^2 diag(V_1, ..., V_N),$$

where the matrices F and V depend on the individual experimental settings.

Of special interest are models with normally distributed random effects b_i and ϵ_{ij}. For the marginal distribution of Y_i then straightforwardly follows:

$$Y_i \sim \mathcal{N}_{m_i}\big(F_i\beta, \sigma^2(I_{m_i} + F_iDF_i^T)\big).$$

A generalized model description is frequently used in the literature, considering a population location parameter $\beta \in \mathbb{R}^{p_1}$ and individual random effects $b_i \in \mathbb{R}^p$. The observation vector of the i-th individual is with design matrices $F_{i,1}$ and $F_{i,2}$ of appropriate dimensions then modeled by

$$Y_i = F_{i,1}\beta + F_{i,2}b_i + \epsilon_i.$$

One possible application of this model is the inclusion of fixed group effects as described by Schmelter (2007a) with a matrix K_i. The individual parameter vectors are then modeled as in Laird and Ware (1982) or Verberke and Molenberghs (2001) by

$$\beta_i = K_i\beta + b_i$$

and the design matrix $F_{i,1}$ is obtained as $F_{i,1} = F_{i,2}K_i$. Mean and covariance result for the marginal model of the individual observation vectors in

$$E(Y_i) = F_{i,1}\beta \text{ and } Cov(Y_i) = \sigma^2(I_{m_i} + F_{i,2}DF_{i,2}^T),$$

such that the proposed model can still be interpreted as a linear regression model with a block diagonal variance matrix.

3.2 Estimation

The linear mixed effects model as described in the previous section is a linear regression model with correlated observations. The estimation in linear regression models with general covariance matrices $\sigma^2 V$ was in section 2.1 briefly discussed. In dependence on the knowledge of the variance matrix, different estimation procedures were proposed.

3.2.1 Weighted Least Squares Estimation

A known variance σ^2 of the observation errors and variance matrix $\sigma^2 D$ of the individual parameter vectors yield the complete knowledge of the variance matrix of the observation vector Y. The weighted least squares estimator results with the block diagonal structure of V in:

$$\begin{aligned}\widehat{\beta}_{WLS} &= (F^TV^{-1}F)^-F^TV^{-1}Y = (\sum_{i=1}^N F_i^TV_i^{-1}F_i)^- \sum_{i=1}^N F_i^TV_i^{-1}Y_i \\ &= (\sum_{i=1}^N F_i^TV_i^{-1}F_i)^- \sum_{i=1}^N F_i^TV_i^{-1}F_i\widehat{\beta}_{WLS,i},\end{aligned}$$

where

$$\widehat{\beta}_{WLS,i} := (F_i^T V_i^{-1} F_i)^- F_i^T V_i^{-1} Y_i, \; i = 1, ..., N$$

describe the individual weighted least squares estimators. With a matrix inversion formula (e.g. Schott (1997, p. 9)) follows

$$F_i^T V_i^{-1} = F_i^T - F_i^T F_i D (I_p + F_i^T F_i D)^{-1} F_i^T = (I_p + F_i^T F_i D)^{-1} F_i^T,$$

such that the estimators $\widehat{\beta}_{WLS,i}$ do not depend on the matrix D, as the estimating equations for the individual weighted and ordinary least squares estimators coincide:

$$F_i^T V_i^{-1}(Y_i - F_i \beta) = 0 \Leftrightarrow F_i^T (Y_i - F_i \beta) = 0 \quad \Rightarrow \quad \widehat{\beta}_{WLS,i} = \widehat{\beta}_{OLS,i}.$$

If all individuals are observed under identical individual sampling schemes, $m_i = m_1$ and $F_i = F_1$, $i = 1, ..., N$, the weighted least squares estimator results in the arithmetic mean of the individual ordinary least squares estimates:

$$\widehat{\beta}_{WLS} = \frac{1}{N} (F_1^T V_1^{-1} F_1)^- \sum_{i=1}^{N} F_1^T V_1^{-1} Y_i = \frac{1}{N} \sum_{i=1}^{N} \widehat{\beta}_{OLS,i}.$$

Hence the inter-individual variance $\sigma^2 D$ has for these special designs no influence on the estimation (Entholzner *et al.* (2005)).

Note however, that the weighted least squares estimator for β generally depends on the matrix D if not all individual sampling schemes are identical.

For a matrix F of full column rank and a known variance matrix $\sigma^2 V$, the proposed estimator $\widehat{\beta}_{WLS}$ is the best linear unbiased estimator. The covariance is obtained as in chapter 2 by

$$Cov(\widehat{\beta}_{WLS}) = \sigma^2 (F^T V^{-1} F)^{-1} = \sigma^2 (\sum_{i=1}^{N} F_i^T V_i^{-1} F_i)^{-1}.$$

The results of section 2.1.1 on the estimation of identifiable linear aspects can be readily applied. The linear aspect $\psi(\beta) = L_\psi \beta$ is identifiable, if a matrix Q with $L_\psi = QF$ exists. Then

$$\widehat{\psi} = L_\psi (F^T V^{-1} F)^- F^T V^{-1} Y \text{ with } Cov(\widehat{\psi}) = \sigma^2 L_\psi (F^T V^{-1} F)^- L_\psi^T$$

is the best linear unbiased estimator for $\psi(\beta)$.

3.2.2 Generalized Least Squares Estimation

Weighted and ordinary least squares estimators generally not coincide if not all individual sampling schemes are identical. If the variance parameters σ^2 and D are unknown, an unbiased estimator for β is given by the ordinary least squares estimator. Theoretically, the estimation can be improved by taking the variances into account. Generalized least squares procedures are described in the literature for iteratively estimating the parameter vector β and the variance parameters based on the estimated parameter $\widehat{\beta}$. Davidian and Giltinan (1995, p. 35) describe for general nonlinear regression models the following procedure:

(1) Estimate β, e.g. using the ordinary least squares estimator $\widehat{\beta}_{OLS}$

(2) Form estimated weights based on the actual estimate of $\beta \rightarrow$ estimate V by some $\widehat{V}$

(3) Reestimate β by weighted least squares, using $\widehat{V}$ as weight matrix, return to step (2).

The final estimate is denoted as generalized least squares estimate $\widehat{\beta}_{GLS}$ and is sometimes also referred to as iteratively reweighted least squares estimator. Different approaches might be applied for the second step of this algorithm. Notice that the true variance of the observations $\sigma^2 V$ depends on the unknown parameters σ^2 and D, where the matrix D can be represented in vector notation by a ν-dimensional vector α as in Schmelter (2007a), with $\nu \leq \frac{1}{2}p(p+1)$. Davidian and Carroll (1987) discuss for heteroscedastic regression models different methods for estimating variance functions. Methods based on estimators maximizing normal likelihoods were proposed by the authors for estimating the unknown parameters. Demidenko (2005, ch. 3) develops distribution-free unbiased estimators for the variance parameters in linear mixed effects models. The described estimators are non-iterative and Demidenko states their consistency under certain conditions, such that the asymptotic distributions for numbers of individuals $N \rightarrow \infty$ of the weighted and generalized least squares estimators for β in linear mixed effects models coincide (Davidian and Giltinan (1995), Demidenko (2005)).

3.2.3 Maximum Likelihood Estimation

Linear Mixed Effects Models are in the literature usually discussed under the assumption of normally distributed random effects:

$$\beta_i \sim \mathcal{N}(\beta, \sigma^2 D) \quad \text{and} \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2 I_{m_i}).$$

The resulting individual observations vectors Y_i are then normally distributed, as described in section 3.1.2:

$$Y_i \sim \mathcal{N}\big(F_i\beta, \sigma^2(I_{m_i} + F_i D F_i^T)\big),$$

such that the log-likelihood function for N individuals with m_i measurements and an observation vector $y = (y_1, ..., y_N)^T$ results with the independence of the observations of different individuals for the parameter vector $\theta = (\beta^T, \sigma^2, \alpha^T)^T$ and a parameterization α of the matrix $D = D(\alpha)$, in

$$\begin{aligned} l(\theta; y) &= \sum_{i=1}^{N} l(\theta; y_i) \\ &= -\frac{1}{2}\sum_{i=1}^{N} m_i \log(2\pi\sigma^2) + \log \parallel V_i \parallel + \frac{1}{\sigma^2}(y_i - F_i\beta)^T V_i^{-1}(y_i - F_i\beta), \end{aligned}$$

where the matrix V_i depends on α and $\parallel \cdot \parallel$ describes the determinant. Maximization of the likelihood with respect to σ^2 yields with $m_T = \sum_{i=1}^{N} m_i$:

$$\widehat{\sigma}^2_{ML} = \frac{1}{m_T}\sum_{i=1}^{N}(y_i - F_i\beta)^T V_i^{-1}(y_i - F_i\beta).$$

The profile likelihood function is obtained by including the estimate $\widehat{\sigma}^2_{ML}$ in the original likelihood (Demidenko (2005)):

$$l_p(\beta, \alpha; y) := l\big((\beta^T, \widehat{\sigma}^2_{ML}, \alpha^T); y\big) \;=\; -\frac{1}{2}\big(m_T \log[\sum_{i=1}^N (y_i - F_i\beta)^T V_i^{-1}(y_i - F_i\beta)] + m_T[\log(m_T) - 1 + 2\pi] + \sum_{i=1}^N \log \parallel V_i \parallel \big),$$

what is maximized in β with

$$\widehat{\beta}_{ML} = (\sum_{i=1}^N F_i^T V_i^{-1} F_i)^- \sum_{i=1}^N F_i^T V_i^{-1} Y_i.$$

Maximum likelihood estimator and weighted least squares estimator obviously coincide for normally distributed random effects with known variances σ^2 and $\sigma^2 D$. When plugging the estimator $\widehat{\beta}_{ML}$ in the profile likelihood, the maximum likelihood estimation results in a nonlinear optimization problem on the set of non-negative definite matrices for D, or the vector α parameterizing D respectively.

Note that the maximum likelihood estimators for the variance parameters are biased. A slightly modified likelihood function

$$l_{REML}(\theta; y) \;:=\; -\frac{1}{2}\big((m_T - p)\log(2\pi\sigma^2) + \log \parallel \sum_{i=1}^N (F_i^T V_i^{-1} F_i) \parallel + \sum_{i=1}^N [\log \parallel V_i \parallel + \frac{1}{\sigma^2}(y_i - F_i\beta)^T V_i^{-1}(y_i - F_i\beta)]\big)$$

is maximized by the restricted maximum likelihood estimator. The function can be profiled on the resulting estimator for σ^2:

$$\widehat{\sigma}^2_{REML} = \frac{1}{m_T - p} \sum_{i=1}^N (y_i - F_i\beta)^T V_i^{-1}(y_i - F_i\beta).$$

The restricted maximum likelihood estimator for β is of the same form as the maximum likelihood estimator, however now depending on the matrix D maximizing the profile restricted likelihood function. The asymptotic properties of maximum likelihood and restricted maximum likelihood estimators coincide (Demidenko (2005)).

3.3 Information and Design

Experimental designs in linear regression models were described by probability measures on the design region $\mathcal{X}$ and optimized by minimizing a real valued function of the information matrix depending on the design. As mentioned in the preceding subsections, linear mixed effects models are linear models with a special covariance structure of the observation vector.

This special covariance is induced by the correlated observations within the individuals and uncorrelated observations of different individuals, which imply an impact of the different individual experimental settings on the covariance of the final estimator.

The first part of this section hence defines experimental designs in mixed effects models. With the introduced designs, we describe in the second part of this section the information provided by the weighted least squares and maximum likelihood estimator. Considerations on more general covariance structures will be presented in the last subsection.

3.3.1 Experimental Designs in Mixed Effects Models

The two stages in modeling the observations of different individuals carry forward to the design of population studies. In the second subsection the weighted least squares estimator could be shown to coincide for identical individual designs with the arithmetic mean of the ordinary individual least squares estimators. Individual designs ξ define the experimental settings for a group of individuals. Two different kinds of individual designs in mixed effects models are often considered. The exact individual designs are based on a bounded natural number of measurements in the individuals:

Definition 3.2. *The exact individual design ξ of size m describes the m experimental settings x_j from the design region $\mathcal{X}$:*

$$\xi := (x_1, ..., x_m), \; m \in \mathbb{N}.$$

The set of all exact individual designs of size m hence coincides with the m-dimensional design region $\mathcal{X}^m$. For exact designs ξ, the design matrix is defined as in section 2.2.1 by

$$F(\xi) := \big(f(x_1), ..., f(x_m)\big)^T \quad \text{and} \quad V(\xi) := I_m + F(\xi) D F(\xi)^T$$

results up to the constant σ^2 as the variance of the individual observations for a given exact design ξ.

Similar to exact designs in ordinary linear regression models, the class of exact individual designs can be extended to the class of approximate designs, by taking non-integer numbers of measurement replications into account:

Definition 3.3. *The approximate individual design ξ of size m describes the experimental settings $(x_1, ..., x_l) \in \mathcal{X}^l$, for some $l \in \mathbb{N}$, with the according numbers of measurement replications $(m_1, ..., m_l) \in \mathbb{R}^l_+$ with $\sum m_j = m$:*

$$\xi := \begin{pmatrix} x_1 & ... & x_l \\ m_1 & ... & m_l \end{pmatrix}.$$

The set of approximate individual designs of size m will be denoted as Ξ^m.

The design and variance matrix can for the so defined approximate individual design be represented with a weight matrix $W(\xi) = (m_1, ..., m_l)$ by

$$F_a(\xi) := \big(f(x_1), ..., f(x_l)\big)^T \quad \text{and} \quad V_a(\xi) := W(\xi)^{-1} + F_a(\xi) D F_a(\xi)^T.$$

Note that exact individual designs of size m are elements of Ξ^m. Hence optimal approximate individual designs of size m are at least as good as exact individual designs. However, exact individual designs are in most population studies more realistic, as the number of measurements on each individual is usually small, such that approximations of the exact designs by approximate designs are generally not satisfying. Schmelter (2007b, ch. 7) showed some optimality results on approximate individual designs, which might help designing experiments for exact designs.

The second stage of the mixed effects model is described by the inter-individual variation and motivates the population designs. Generally not all individuals have to be observed under the same experimental settings. The population design describes the proportions of individual sampling schemes in the whole population:

Definition 3.4. *The population design ζ is a vector of individual designs $(\xi_1, ..., \xi_k)$, for some $k \in \mathbb{N}$, with a vector of according proportions $(\omega_1, ..., \omega_k)$, where $\omega_i \geq 0$ and $\sum \omega_i = 1$:*

$$\zeta := \begin{pmatrix} \xi_1 & \dots & \xi_k \\ \omega_1 & \dots & \omega_k \end{pmatrix}.$$

This definition does not specify the number of measurements in the individual designs. For designing experiments we will generally consider individual designs consisting of the same number of observations. The above definition corresponds in the special case of population designs with exact individual designs of equal size $m_i = m$, $i = 1, ..., k$, to the definition of approximate designs in linear regression models on a design region $\mathcal{X}^m$ given in chapter 2.

3.3.2 Information in Mixed Models

For known variance parameters σ^2 and D, the covariance matrix of the best linear unbiased estimator $\widehat{\beta}_{WLS}$ under a regular population design ζ with N individuals and k distinct exact individual designs $\xi_i \in \mathcal{X}^{m_i}$ is easily obtained as:

$$Cov(\widehat{\beta}_{WLS}) = \frac{\sigma^2}{N} \left(\sum_{i=1}^{k} \omega_i F(\xi_i)^T V(\xi_i)^{-1} F(\xi_i) \right)^{-1}.$$

The information matrix of an estimator is here defined as the inverse of the covariance matrix of the estimator. In mixed effects models two stages of information matrices are of interest. The population information matrix of the weighted least squares estimators is for a population design ζ defined by

$$\mathbf{M}_{\widehat{\beta}_{WLS};pop}(\zeta) := \frac{1}{\sigma^2} \sum_{i=1}^{k} \omega_i F(\xi_i)^T V(\xi_i)^{-1} F(\xi_i).$$

The information matrix of the population design ζ is a weighted sum of matrices depending on the individual designs $\xi_i \in \mathcal{X}^{m_i}$ supporting the population design and the matrices

$$\mathbf{M}_{\widehat{\beta}_{WLS};ind}(\xi_i) := \frac{1}{\sigma^2} F(\xi_i)^T V(\xi_i)^{-1} F(\xi_i)$$

are called individual information matrices. The population information matrix contains the complete influence of the experimental settings on the accuracy of the parameter estimates. Note that the information matrix for the whole parameter vector $\theta = (\beta^T, \sigma^2, \alpha^T)^T$ generally depends on the estimators, which are applied for estimating the variance parameters σ^2 and α.

The Fisher information matrix provides under certain regularity conditions an upper bound for the information of any unbiased estimator with respect to the Loewner partial ordering. Given a parametric model $f_{Y_i}(y_i, \theta)$ and the independence of observations Y_i of the N different individuals, the joint density of the whole observation vector Y is easily derived as the product of the individual densities, such that likelihood and log-likelihood function result in

$$L(\theta; y) = \prod_{i=1}^{N} f_{Y_i}(y_i, \theta) \text{ and } l(\theta; y) = \sum_{i=1}^{N} \log\left(f_{Y_i}(y_i, \theta)\right).$$

The independence of the individual observations carries forward to the Fisher information. The normalized Fisher information of a population design ζ is defined as

$$\mathfrak{M}_{pop}(\zeta) := \frac{1}{N} E\left(\frac{\partial l(\theta; Y)}{\partial \theta} \frac{\partial l(\theta; Y)}{\partial \theta^T}\right),$$

and results with the independence of observations of different individuals in the weighted sum of the individual Fisher information matrices $\mathfrak{M}_{ind}(\xi_i)$:

$$\mathfrak{M}_{pop}(\zeta) := \frac{1}{N} E\left(\frac{\partial l(\theta; Y)}{\partial \theta} \frac{\partial l(\theta; Y)}{\partial \theta^T}\right) = \sum_{i=1}^{k} \omega_i E\left(\frac{\partial l(\theta; Y_i)}{\partial \theta} \frac{\partial l(\theta; Y_i)}{\partial \theta^T}\right) = \sum_{i=1}^{k} \omega_i \mathfrak{M}_{ind}(\xi_i),$$

where

$$\mathfrak{M}_{ind}(\xi_i) := E\left(\frac{\partial l(\theta; Y_i)}{\partial \theta} \frac{\partial l(\theta; Y_i)}{\partial \theta^T}\right).$$

In the case of mixed effects models with an unknown parameter vector $\theta = (\beta^T, \sigma^2, \alpha^T)^T$, where the ν-dimensional vector α parameterizes the matrix D, the individual and population Fisher information matrices are of the form

$$\mathfrak{M}_{\cdot}(\cdot) = \begin{pmatrix} \mathfrak{M}_{\cdot}^{\beta} & \mathfrak{M}_{\cdot}^{\beta,\sigma^2} & \mathfrak{M}_{\cdot}^{\beta,\alpha} \\ \mathfrak{M}_{\cdot}^{\beta,\sigma^2\ T} & \mathfrak{M}_{\cdot}^{\sigma^2} & \mathfrak{M}_{\cdot}^{\sigma^2,\alpha} \\ \mathfrak{M}_{\cdot}^{\beta,\alpha^T} & \mathfrak{M}_{\cdot}^{\sigma^2,\alpha^T} & \mathfrak{M}_{\cdot}^{\alpha} \end{pmatrix}.$$

For homoscedastic normally distributed random effects β_i and ϵ_i, the entries of the Fisher information matrix can be readily calculated and result on the individual level for an exact individual design $\xi \in \mathcal{X}^m$ as presented by Mentré *et al.* (1997) in

$$\begin{aligned}
\mathfrak{M}^{\beta}_{ind}(\xi) &= \tfrac{1}{\sigma^2} F(\xi)^T V(\xi)^{-1} F(\xi) \\
\left(\mathfrak{M}^{\beta,\sigma^2}_{ind}(\xi)\right)_j &= 0,\ j=1,...,p \\
\left(\mathfrak{M}^{\beta,\alpha}_{ind}(\xi)\right)_{j,k} &= 0,\ j=1,...,p,\ k=1,...\nu \\
\mathfrak{M}^{\sigma^2}_{ind}(\xi) &= \tfrac{m}{2\sigma^4} \\
\left(\mathfrak{M}^{\sigma^2,\alpha}_{ind}(\xi)\right)_j &= \tfrac{1}{2\sigma^2}\mathrm{tr}\left[V(\xi)^{-1}\frac{\partial V(\xi)}{\partial \alpha_j}\right], j=1,...,\nu \\
\left(\mathfrak{M}^{\alpha}_{ind}(\xi)\right)_{j,k} &= \tfrac{1}{2}\mathrm{tr}\left[V(\xi)^{-1}\frac{\partial V(\xi)}{\partial \alpha_j}V(\xi)^{-1}\frac{\partial V(\xi)}{\partial \alpha_k}\right],\ j,k=1,...,\nu.
\end{aligned}$$

The components of the Fisher information matrix depend on the usually unknown variance parameters σ^2 and α, such that optimal designs are generally just locally optimal. The Fisher information in linear mixed effects models with normally distributed and homoscedastic random effects is block-diagonal. Hence efficient estimates of the variance and location parameters are uncorrelated. Weighted least squares information and Fisher information for β coincide under the assumed normal distribution of the random effects.

Schmelter (2007a) generalized the presented results to information matrices of approximate individual designs $\xi \in \Xi^m$. For approximate individual designs with l distinct support points $x_j \in \mathcal{X}$, corresponding real valued measurement replications $m_j > 0$ and with the matrices $F_a(\xi)$, $V_a(\xi)$ and $W(\xi)$ as defined in the preceding subsection, the individual weighted least squares information for β results in

$$\mathbf{M}_{\widehat{\beta}_{WLS};ind,a}(\xi) = \frac{1}{\sigma^2}F_a(\xi)^T V_a(\xi)^{-1} F_a(\xi) = \frac{1}{\sigma^2}\lim_{\delta\to 0}\left([F_a(\xi)^T W(\xi) F_a(\xi) + \delta I_p]^{-1} + D\right)^{-1}.$$

This yields for regular matrices $F_a(\xi)^T W(\xi) F_a(\xi)$ the representation of the information as in Liski *et al.* (2002):

$$\mathbf{M}_{\widehat{\beta}_{WLS};ind,a}(\xi) = \frac{1}{\sigma^2}\left([F_a(\xi)^T W(\xi) F_a(\xi)]^{-1} + D\right)^{-1}.$$

Under the normality assumptions on the random effects and the assumption of a diagonal covariance matrix $D = diag(\alpha)$ of the individual random effects, the components of the variance parameter blocks of the Fisher information were specified for approximate and exact individual designs by Schmelter (2007a, p. 47)

$$\begin{aligned}
\mathfrak{M}^{\alpha}_{ind}(\xi) &= \tfrac{\sigma^4}{2}\mathfrak{M}^{\beta}_{ind}(\xi)\circ\mathfrak{M}^{\beta}_{ind}(\xi) \text{ and} \\
\left(\mathfrak{M}^{\sigma^2,\alpha}_{ind}(\xi)\right)_j &= \tfrac{1}{2}\mathfrak{M}^{\beta}_{ind}(\xi)_{j,j}\ ,\ j=1,...,\nu,
\end{aligned}$$

with $\circ$ describing the Hadamard product of matrices.

The results in the more general model

$$Y_i = F_{i,1}\beta + F_{i,2}b_i + \epsilon_i$$

follow analogously with the appropriate variance model

$$Cov(Y_i) = \sigma^2(I_{m_i} + F_{i,2}DF_{i,2}^T)$$

of the vector Y_i.

3.3.3 General Covariance Structures

Proportional error models with normally distributed random effects β_i and ϵ_{ij}, as for example

$$Y_{ij} = f(x_{ij})^T \beta_i \exp(\epsilon_{ij}) \quad \text{or} \quad Y_{ij} = f(x_{ij})^T \beta_i (1 + \epsilon_{ij}),$$

are often considered in the pharmacokinetic literature. For both models generally no closed form representation of the likelihood function exists. The first model might be transformed using the logarithm (Schmelter (2007a)), what yields a nonlinear mixed effects model with additive normal errors:

$$\log(Y_{ij}) = \log\left(f(x_{ij})^T \beta_i\right) + \epsilon_{ij}.$$

Note however, that the transformation cannot be done straightforwardly for linear functions, as negative values of the response function for normally distributed individual parameter vectors β_i are possible. Nonlinear mixed effects models will be discussed in the following chapter.

Retout and Mentré (2003) approximate the second model by a model with an error variance, depending on the population parameter vector β instead of the individual parameter vector and by assuming the observation errors to be uncorrelated. These assumptions lead to the model

$$Y_{ij} = f(x_{ij})^T \beta_i + \widetilde{\epsilon}_{ij}, \text{ where } \widetilde{\epsilon}_{ij} \sim \mathcal{N}\big(0, \sigma^2 \cdot (f(x_{ij})^T \beta)^2\big).$$

The individual observation vector Y_i hence follows with a design $\xi \in \mathcal{X}^m$ under the assumption of normally distributed individual parameters β_i a heteroscedastic normal model, similar to the models presented in Atkinson and Cook (1995):

$$Y_i \sim \mathcal{N}\big(F(\xi)\beta, \sigma^2 V(\xi)\big) \text{ with } V(\xi) := diag\left((f(x_{i1})^T \beta)^2, ..., (f(x_{im})^T \beta)^2\right) + F(\xi) D F(\xi)^T.$$

The components of the Fisher information matrix result with the dependence of the matrix $V(\xi)$ on the parameters β and α in

$$\begin{aligned}
\mathfrak{M}^{\beta}_{ind}(\xi) &= \tfrac{1}{\sigma^2} F(\xi)^T V(\xi)^{-1} F(\xi) + \tfrac{1}{2} S \text{ with} \\
(S)_{j,k} &:= \operatorname{tr}\left[V(\xi)^{-1} \frac{\partial V(\xi)}{\partial \beta_j} V(\xi)^{-1} \frac{\partial V(\xi)}{\partial \beta_k}\right], \; j,k = 1, ..., p \\
\left(\mathfrak{M}^{\beta,\sigma^2}_{ind}(\xi)\right)_j &= \tfrac{1}{2\sigma^2} \operatorname{tr}\left[V(\xi)^{-1} \frac{\partial V(\xi)}{\partial \beta_j}\right], \; j = 1, ..., p \\
\left(\mathfrak{M}^{\beta,\alpha}_{ind}(\xi)\right)_{j,k} &= \tfrac{1}{2} \operatorname{tr}\left[V(\xi)^{-1} \frac{\partial V(\xi)}{\partial \beta_j} V(\xi)^{-1} \frac{\partial V(\xi)}{\partial \alpha_k}\right], \; j = 1, ..., p, \; k = 1, ...\nu \\
\mathfrak{M}^{\sigma^2}_{ind}(\xi) &= \tfrac{m}{2\sigma^4} \\
\left(\mathfrak{M}^{\sigma^2,\alpha}_{ind}(\xi)\right)_j &= \tfrac{1}{2\sigma^2} \operatorname{tr}\left[V(\xi)^{-1} \frac{\partial V(\xi)}{\partial \alpha_j}\right], \; j = 1, ..., \nu \\
\left(\mathfrak{M}^{\alpha}_{ind}(\xi)\right)_{i,k} &= \tfrac{1}{2} \operatorname{tr}\left[V(\xi)^{-1} \frac{\partial V(\xi)}{\partial \alpha_j} V(\xi)^{-1} \frac{\partial V(\xi)}{\partial \alpha_k}\right], \; j,k = 1, ..., \nu.
\end{aligned}$$

Notice specially that the information for the location parameter β does not correspond to the weighted least squares information any more. Additional information in the estimation of β might be gained by the dependence of the observation variance on the vector β, such that the parameter β might in these models be not efficiently estimated with the weighted least squares estimator.

4 Nonlinear Mixed Effects Models

Nonlinear mixed effects models generalize the presented linear mixed effects model to response functions nonlinearly depending on the individual parameter vector β_i. The nonlinear influence of the parameter vector β_i on the response function η complicates estimation in mixed effects models extremely. Specially in population pharmacokinetic models the response functions are generally nonlinear in the vector β_i and insight in the distribution of estimators in the proposed models is sought for optimally planning pharmacokinetic studies. Nonlinear mixed effects models are discussed in the literature by various authors. As in linear mixed effects models, the books by Davidian and Giltinan (1995) and Demidenko (2005) extensively describe the analysis in nonlinear mixed effects models.

The first part of this chapter briefly describes the model formulation, before estimation procedures are presented in the second section. The missing closed form representation of the likelihood function carries forward to the construction of the Fisher information. Topic of the third section are information matrices based on in the literature proposed linearizations of the response function.

4.1 Model Formulation

The nonlinear mixed effects model can be motivated by a two-stage model, similar to the linear mixed effects model. In this section the intra- and inter-individual models are described under the assumption of normally distributed random effects.

4.1.1 Intra-Individual Model

The j-th observation of the i-th individual under experimental setting $x_{ij} \in \mathcal{X}$ is modeled by

$$Y_{ij} = \eta(\beta_i, x_{ij}) + \epsilon_{ij},$$

with a real valued response function η, a p dimensional individual parameter vector β_i and a real valued observation error ϵ_{ij}. To avoid difficulties, the response function η is assumed to be continuous on $\mathcal{X}$ and differentiable in β_i.

The exact individual experimental design $\xi_i = (x_{i1}, ..., x_{im_i}) \in \mathcal{X}^{m_i}$ describes the experimental settings of the i-th individual. The response function for the whole m_i-dimensional individual observation vector Y_i is then vector valued and denoted as

$$\eta(\beta_i, \xi_i) := \big(\eta(\beta_i, x_{i1}), ..., \eta(\beta_i, x_{im_i})\big)^T.$$

The vector of the m_i observations within the i-th individual is for a given parameter vector β_i completely described up to the unknown normally distributed observation error vector $\epsilon_i = (\epsilon_{i1}, ..., \epsilon_{im_i})^T$ by the response function η and the individual experimental settings ξ_i. The intra-individual statistical model results in

$$Y_i|_{\beta_i=\mathrm{\beta}_i} \sim \mathcal{N}\big(\eta(\mathrm{\beta}_i, \xi_i), \sigma^2 I_{m_i}\big). \tag{4.1}$$

4.1.2 Inter-Individual Model

The individual parameter vectors β_i are assumed to be independent and identically distributed as

$$\beta_i \sim \mathcal{N}(\beta, \sigma^2 D), \tag{4.2}$$

inducing the inter-individual variation. The observation error vectors ϵ_i and individual parameter vectors β_i are considered to be stochastically independent and the observations of different individuals are stochastically independent as well. The vector $\theta = (\beta^T, \sigma^2, \alpha^T)^T$ summarizes the population parameters, with an appropriate vector-valued parameterization α of the matrix D.
For linear response functions η, the normality of the random effects ϵ_i and β_i yields the normality of the random variable Y_i, as described in chapter 3. For nonlinear response functions, $\sigma^2 > 0$ and a positive definite matrix D, the probability density $f_{Y_i}(y_i)$ of Y_i cannot be represented in a closed form. The likelihood of observations y_i results in integral form in

$$L(\theta; y_i) := f_{Y_i}(y_i) \quad = \quad \int_{\mathbb{R}^p} \phi_{Y_i|\mathrm{\beta}_i}(y_i)\phi_{\beta_i}(\mathrm{\beta}_i)d\mathrm{\beta}_i$$

The population parameter θ influences the likelihood by the normal densities $\phi_{Y_i|\mathrm{\beta}_i}$ and ϕ_{β_i} with mean and variance as in the models 4.1 and 4.2:

$$\begin{aligned} \phi_{Y_i|\mathrm{\beta}_i}(y_i) &= \sqrt{2\pi\sigma^2}^{\,-m_i} \exp\left[-\frac{1}{2\sigma^2}\big(y_i - \eta(\mathrm{\beta}_i, \xi_i)\big)^T\big(y_i - \eta(\mathrm{\beta}_i, \xi_i)\big)\right] \\ \phi_{\beta_i}(\mathrm{\beta}_i) &= \sqrt{2\pi\sigma^2}^{\,-p}\sqrt{\| D \|}^{\,-1} \exp\left[-\frac{1}{2\sigma^2}(\mathrm{\beta}_i - \beta)^T D^{-1}(\mathrm{\beta}_i - \beta)\right], \end{aligned}$$

where $\| \cdot \|$ describes the determinant. With

$$\tilde{l}(\mathrm{\beta}_i, \theta; y_i) := \big(y_i - \eta(\mathrm{\beta}_i, \xi_i)\big)^T\big(y_i - \eta(\mathrm{\beta}_i, \xi_i)\big) + (\mathrm{\beta}_i - \beta)^T D^{-1}(\mathrm{\beta}_i - \beta)$$

the integrand then results in

$$\phi_{Y_i|\mathrm{\beta}_i}(y_i)\phi_{\beta_i}(\mathrm{\beta}_i) = \frac{1}{c}\exp\left[-\frac{1}{2\sigma^2}\tilde{l}(\mathrm{\beta}_i, \theta; y_i)\right]$$

with the constant $c = \sqrt{2\pi\sigma^2}^{(m_i+p)}\sqrt{\| D \|}$.
As in the previous chapter, the model of the individual parameter vector β_i can be generalized. The p-dimensional individual parameter vector β_i can be modeled by

$$\beta_i = \gamma_i(\beta) + b_i$$

with individual random effects $b_i \in \mathbb{R}^p$ and differentiable functions $\gamma_i : \mathbb{R}^{p_1} \to \mathbb{R}^p$ of the population parameters $\beta \in \mathbb{R}^{p_1}$. Fixed group effects might be considered, as in the preceding chapter, by a function $\gamma_i(\beta) = K_i\beta$ with an appropriate matrix K_i. In this chapter we concentrate on the simpler case of $p_1 = p$ with $\gamma_i(\beta) = I_p\beta$ and cover the general case in some remarks in the next chapter only.

The matrix

$$F_{\beta_0} := \frac{\partial\eta(\beta_i, \xi_i)}{\partial\beta_i^T}|_{\beta_i=\beta_0},\ \beta_0 \in \mathbb{R}^p$$

describes in the following the design matrix in nonlinear mixed effect models.

4.2 Estimation

The ordinary nonlinear regression model can be represented as the limiting case of a nonlinear mixed effects models with inter-individual variance $\sigma^2 D \to 0$. Nonlinear least squares estimators can then be applied for estimating the location parameter vector β. Another special case is obtained for an intra-individual variance $\sigma^2 \to 0$, while the inter-individual variance $\sigma^2 D = const$ remains constant. If the individual models are identifiable, nonlinear least squares estimates $\widehat{\beta}_i$ of the individual parameter vector β_i then coincide with the true individual parameter vector, such that the mean of the individual estimates yields the maximum likelihood estimator of the population location vector β.

In more realistic scenarios, different estimation procedures can be applied, which generally yield different estimates of the parameter vector β. The proposed estimators are generally based either on linearizations of the model, on the two-stages of the model or on approximations of the likelihood function. In this section we briefly summarize results on often applied estimators.

4.2.1 Maximum Likelihood Estimation

Maximum likelihood estimators are desirable for estimating the unknown parameter θ as they yield under appropriate regularity conditions asymptotically efficient normally distributed estimates of the parameter θ. Note that the maximum likelihood estimator is generally not unbiased, but in dependence on the numbers of individuals N asymptotically unbiased.

The score function for β results with the proposed model 4.2 of the individual vectors β_i with the log-likelihood function $l(\theta; y_i)$ and the assumption on interchangeability of differentiation and integration in

$$\begin{aligned}
\frac{\partial l(\theta; y_i)}{\partial \beta} &= \frac{1}{f_{Y_i}(y_i)} \int_{\mathbb{R}^p} \phi_{Y_i|\beta_i}(y_i) \frac{\partial \phi_{\beta_i}(\beta_i)}{\partial \beta} d\beta_i \\
&= \frac{1}{f_{Y_i}(y_i)} \int_{\mathbb{R}^p} \phi_{Y_i|\beta_i}(y_i) \phi_{\beta_i}(\beta_i) \frac{1}{\sigma^2} D^{-1}(\beta_i - \beta) d\beta_i \\
&= \frac{1}{\sigma^2} D^{-1}\big(E(\beta_i|Y_i = y_i) - \beta\big).
\end{aligned}$$

Analogously the score functions for σ^2 and α are with $D = D(\alpha)$ obtained as

$$\begin{aligned}
\frac{\partial l(\theta; y_i)}{\partial \sigma^2} &= E\left[\frac{1}{2\sigma^4}\tilde{l}(\beta_i, \theta; y_i)|Y_i = y_i\right] - \frac{m_i + p}{2\sigma^2} \text{ and} \\
\frac{\partial l(\theta; y_i)}{\partial \alpha_j} &= E\left[\frac{1}{2\sigma^2}(\beta_i - \beta)^T D^{-1} \frac{\partial D}{\partial \alpha_j} D^{-1}(\beta_i - \beta)|Y_i = y_i\right] - \frac{1}{2}\text{tr } D^{-1}\frac{\partial D}{\partial \alpha_j}, \ j = 1, ..., \nu.
\end{aligned}$$

The missing closed form representation of the likelihood function carries forward to the score functions, such that maximum likelihood estimates cannot be obtained straightforwardly. Kuhn and Lavielle (2001) propose a stochastic version of the expectation maximization algorithm for the estimation and state the convergence of the proposed algorithm under some conditions. Alternatively, Pinheiro and Bates (1995) propose approximations of the log-likelihood function by importance sampling, Gaussian quadrature rules and Laplace approximations for estimating the parameters of interest. Unfortunately Gaussian quadrature rules tend to be inefficient if accurate results for the likelihood-approximation are sought.

Approximations of the likelihood based on linearizations of the penalized least squares term $\tilde{l}$ in empirical Bayes estimates for β_i were proposed by Beal and Sheiner (1998) and yield similar results to an algorithm presented by Lindstrom and Bates (1990), which will be discussed in subsection 4.2.3. Note that the properties of estimators based on approximated likelihood functions, generally not coincide with the asymptotic properties of the maximum likelihood estimator. The convergence rates of maximum likelihood estimators and approximated maximum likelihood estimators in mixed effects models were described by Nie (2007). Demidenko (2005, ch. 8.9) presents an example of a one-parametric exponential decay model and shows the relative asymptotic bias of the Lindstrom and Bates estimator for small individual sample sizes.

4.2.2 Two-Stage Estimation

As the model is build in two stages, the estimation of the population parameters can be conducted in two stages as well. Let $\widehat{\beta}_i$ here describe the individual ordinary least squares estimates:

$$\widehat{\beta}_i := \underset{\beta_i \in \mathbb{R}^p}{\operatorname{argmin}} \big(y_i - \eta(\beta_i, \xi_i)\big)^T \big(y_i - \eta(\beta_i, \xi_i)\big).$$

Then a naive first estimator of β is obtained by the mean of the individual estimates

$$\widehat{\beta} = \frac{1}{N}\sum_{i=1}^{N} \widehat{\beta}_i,$$

what is similar to the weighted least squares estimators in the case of identical individual designs in section 3.2.1.

The inter-individual variance can be analogously estimated by

$$\widehat{\sigma^2 D} = \frac{1}{N-1}\sum_{i=1}^{N}(\widehat{\beta}_i - \widehat{\beta})(\widehat{\beta}_i - \widehat{\beta})^T,$$

what yields in the special case of mixed effects models without observation errors efficient estimates for the inter-individual variance. Davidian and Giltinan (1995) state that these estimators do not require normally distributed random effects and that the estimator for the matrix D is generally upwardly biased. Alternatively, the uncertainty in the individual estimates can be included in the estimation of the population parameters, what leads to the *global two-stage method*. Therefore the asymptotic theory on nonlinear least squares estimators (Jennrich (1969)) provides an approximation of the marginal distribution of the individual estimates. Since $\widehat{\beta}_i$ is for an observation vector y_i with the true, but unknown individual parameter vector β_i, a root of the score function

$$\begin{aligned} 0 &= F_{\widehat{\beta}_i}^T\big(y_i - \eta(\widehat{\beta}_i,\xi_i)\big) \approx F_{\widehat{\beta}_i}^T\big(y_i - \eta(\beta_i,\xi_i) + F_{\widehat{\beta}_i}(\beta_i - \widehat{\beta}_i)\big) \\ \Rightarrow \quad \widehat{\beta}_i &\approx \beta_i + (F_{\widehat{\beta}_i}^T F_{\widehat{\beta}_i})^{-1} F_{\widehat{\beta}_i}^T \epsilon_i, \end{aligned}$$

the individual estimators $\widehat{\beta}_i$ are considered to be normally distributed (Demidenko (2005)):

$$\widehat{\beta}_i \overset{app.}{\sim} \mathcal{N}\left(\beta, Cov(\widehat{\beta}_i)\right),\; i = 1, ..., N,$$

where the covariance of the estimator $\widehat{\beta}_i$ is approximated by the matrix

$$Cov(\widehat{\beta}_i) \approx Cov\left(\beta_i + (F_{\widehat{\beta}_i}^T F_{\widehat{\beta}_i})^{-1} F_{\widehat{\beta}_i}^T \epsilon_i\right) = \sigma^2[(F_{\widehat{\beta}_i}^T F_{\widehat{\beta}_i})^{-1} + D].$$

The observed individual estimates $\widehat{\beta}_i$ are for the estimation of the population parameters assumed to be realizations of normally distributed random variables. The intra-individual variance parameter σ^2 is estimated by

$$\widehat{\sigma}^2 = \frac{1}{\sum m_i - p}\sum_{i=1}^{N}\left(y_i - \eta(\widehat{\beta}_i,\xi_i)\right)^T\left(y_i - \eta(\widehat{\beta}_i,\xi_i)\right)$$

and is in the following assumed to be given.

The parameters β and α can then be estimated by maximum likelihood estimation under the distribution assumption on the individual estimates $\widehat{\beta}_i$. The two-stage maximum likelihood estimator of the parameter vector β results in

$$\widehat{\beta}_{TS} = \big(\sum_{i=1}^{N}[(F_{\widehat{\beta}_i}^T F_{\widehat{\beta}_i})^{-1} + D]^{-1}\big)^{-1}\sum_{i=1}^{N}[(F_{\widehat{\beta}_i}^T F_{\widehat{\beta}_i})^{-1} + D]^{-1}\widehat{\beta}_i$$

and the two-stage likelihood function is then profiled on $\widehat{\beta}_{TS}$ and maximized with respect to the vector α parameterizing D. Alternatively a restricted likelihood version or methods of moments can be applied for the estimation of the inter-individual variance D. The Fisher information for the assumed distribution of the individual parameter estimates is obtained, yielding the asymptotic independence of the estimates $\widehat{\beta}_{TS}$ and $\widehat{D}$, since for α parameterizing D holds:

$$E\left(\frac{\partial^2 l(\theta;\widehat{\beta}_i)}{\partial\alpha_j\partial\beta}\right) = -\frac{1}{\sigma^2}[(F_{\widehat{\beta}_i}^T F_{\widehat{\beta}_i})^{-1} + D(\alpha)]^{-1}\frac{\partial D(\alpha)}{\partial\alpha_j}[(F_{\widehat{\beta}_i}^T F_{\widehat{\beta}_i})^{-1} + D(\alpha)]^{-1}E(\widehat{\beta}_i - \beta) = 0.$$

Hence the variance matrix for the estimation of β is approximated by

$$\widehat{Cov(\widehat{\beta}_{TS})} = \widehat{\sigma}^2\big(\sum_{i=1}^{N}[(F_{\widehat{\beta}_i}^T F_{\widehat{\beta}_i})^{-1} + \widehat{D}(\alpha)]^{-1}\big)^{-1},$$

with estimates $\widehat{\sigma}^2$ and $\widehat{D}(\alpha)$ as described.

The quality of two stage estimators heavily depends on the individual sampling schemes, as the presented two-stage likelihood approach is based on the normality assumption of the individual estimates $\widehat{\beta}_i$, which for small individual sample sizes is generally not given. Two stage estimators are asymptotically efficient, when the number of individual observations m_i and the number of individuals N tend to infinity.

4.2.3 Generalized Least Squares Estimation

If the individual nonlinear least squares problems cannot be uniquely solved, generalized least squares estimation might be applied for estimating the population parameter β. For a given vector of observations $y = (y_1^T, ..., y_N^T)^T$ of N individuals and a given inter-individual variance matrix D, the penalized nonlinear least squares objective function (Pinheiro and Bates (2000))

$$\begin{aligned} L_{PNLS}(\widetilde{\mathrm{\beta}}_i, \beta, D; y) &= \sum_{i=1}^{N} \tilde{l}(\mathrm{\beta}_i, \theta; y_i) \\ &= \sum_{i=1}^{N} \big(y_i - \eta(\mathrm{\beta}_i, \xi_i)\big)^T \big(y_i - \eta(\mathrm{\beta}_i, \xi_i)\big) + (\mathrm{\beta}_i - \beta)^T D^{-1}(\mathrm{\beta}_i - \beta) \end{aligned}$$

has to be minimized with respect to the individual parameter vectors $\widetilde{\mathrm{\beta}}_i = (\mathrm{\beta}_1^T, ..., \mathrm{\beta}_N^T)^T$ and the population parameter β. Note that the minimization of the objective function L_{PNLS} is equivalent to the maximization of the integrand in the likelihood function for the observation vector y.

For unknown variance parameters σ^2 and D, the penalized least squares estimates $\mathrm{\beta}_i^*$ and $\widehat{\beta}$ yield the first step of the Lindstrom and Bates Algorithm (Lindstrom and Bates (1990)). The algorithm consists of two alternating steps:

1.) *Penalized least squares*: Minimization of $L_{PNLS}(\widetilde{\mathrm{\beta}}_i, \beta, D; y)$ with respect to $\widetilde{\mathrm{\beta}}_i$ and β for fixed D.

2.) *Linear mixed effects approximation:* The pseudo observation w_i is assumed to follow a normal model:

$$w_i = y_i - \eta(\beta_i^*, \xi_i) + F_{\beta_i^*}\beta_i^* \overset{app.}{\sim} \mathcal{N}\big(F_{\beta_i^*}\beta, \sigma^2(I_{m_i} + F_{\beta_i^*}DF_{\beta_i^*}^T)\big).$$

Application of maximum (restricted) likelihood estimation for the parameters σ^2, D and β in the proposed model.

The likelihood function in the second step of the Lindstrom and Bates algorithm can be profiled to be a function of the D parameterizing vector α only. Note that the final estimates $\widehat{\beta}_{LB}$ of β at convergence coincide in the two different steps of the algorithm (Demidenko (2005)). The original algorithm was generalized by Davidian and Giltinan (1995) to models with individual parameter vectors as nonlinear functions $d(a_i, \beta, b_i)$ of the population parameter β and the individual random effects b_i. Also the intra-individual model was allowed to include covariance matrices depending on the unknown individual parameters. The proposed algorithm can be interpreted as a generalized least squares algorithm, yielding for large N under the assumption of negligible linearization errors to an asymptotically normally distributed estimator for β (Davidian and Giltinan (1995)):

$$\sqrt{N}(\widehat{\beta}_{LB} - \beta) \xrightarrow{\mathfrak{L}} \mathcal{N}\big(0, \sigma^2 \lim_{N\to\infty}(\frac{1}{N}\sum_{i=1}^{N} F_{\beta_i^*}^T[I_{m_i} + F_{\beta_i^*}DF_{\beta_i^*}^T]^{-1}F_{\beta_i^*})^{-1}\big).$$

Note however, that an example on the exponential decay by Demidenko (2005, ch. 8.9) shows that the linearization error is generally not negligible, as the estimator is for small individual sample sizes biased.

Wolfinger (1993) and Vonesh (1996) state that the Lindstrom and Bates algorithm can be alternatively derived by an application of Laplacian approximations, what fits the observation of Beal and Sheiner (1998) of the similar results, which are produced by the FOCE method in the program NONMEM and the Lindstrom and Bates algorithm. Other generalized least squares algorithms for nonlinear mixed effects models are described by Vonesh and Carter (1992) and Davidian and Giltinan (1995). The main difference of these algorithms consists in the inclusion of individual estimates.

4.3 Information under Linearization

In the previous chapters the information matrices for linear and linear mixed effects models were presented. Linearizations of the response function η are applied in the literature in order to circumvent the problem of the missing closed form representation of the Fisher information in nonlinear mixed effects models. Different models, which lead to different information matrices, are obtained in dependence on the support point of the Taylor approach. In the first part of this section we describe the information resulting from a linearization in a guess β_0 of the true parameter vector β, as it was presented in Retout *et al.* (2001) and Schmelter (2007a). Topic of the second subsection is the information obtained from a linearization in the true parameter vector β. This approach yields results as in Retout and Mentré (2003).

The difference of the resulting information matrices is shown on a small example in the third subsection.

4.3.1 Linear Mixed Effects Approximation

A common approach for approximating the Fisher information in nonlinear mixed effects models is based on linear mixed effects models. With a first order Taylor approach in a guess β_0 of the true population location parameter β follows under the assumption of negligible linearization errors

$$\begin{aligned} Y_i &= \eta(\beta_i, \xi_i) + \epsilon_i \\ &\approx \eta(\beta_0, \xi_i) + F_{\beta_0}(\beta - \beta_0) + F_{\beta_0}(\beta_i - \beta) + \epsilon_i, \end{aligned}$$

as proposed in Schmelter (2007a). The distribution assumptions on β_i and ϵ_i then yield a linear mixed effects model:

$$Y_i \overset{app.}{\sim} \mathcal{N}\big(\eta(\beta_0, \xi_i) + F_{\beta_0}(\beta - \beta_0), \sigma^2(I_{m_i} + F_{\beta_0} D F_{\beta_0}^T)\big).$$

The theory developed in the chapter on linear mixed effects models straightforwardly leads to the information matrix in the approximated model, such that the Fisher information is with variance parameters σ^2 and α and a matrix

$$V_{\beta_0} := I_{m_i} + F_{\beta_0} D F_{\beta_0}^T$$

approximated by

$$\mathbf{M}_1 := \begin{pmatrix} \mathbf{M}_{1,\beta} & \mathbf{M}_{1,\beta,\sigma^2} & \mathbf{M}_{1,\beta,\alpha} \\ \mathbf{M}_{1,\beta,\sigma^2}^T & \mathbf{M}_{1,\sigma^2} & \mathbf{M}_{1,\sigma^2,\alpha} \\ \mathbf{M}_{1,\beta,\alpha}^T & \mathbf{M}_{1,\sigma^2,\alpha}^T & \mathbf{M}_{1,\alpha} \end{pmatrix}$$

with entries

$$\begin{aligned} \mathbf{M}_{1,\beta} &:= \frac{1}{\sigma^2} F_{\beta_0}^T V_{\beta_0}^{-1} F_{\beta_0} \\ \big(\mathbf{M}_{1,\beta,\sigma^2}\big)_j &:= 0, \ j = 1, ..., p \\ (\mathbf{M}_{1,\beta,\alpha})_{j,k} &:= 0, \ j = 1, ..., p, \ k = 1, ...\nu \\ \mathbf{M}_{1,\sigma^2} &:= \frac{m_i}{2\sigma^4} \\ \big(\mathbf{M}_{1,\sigma^2,\alpha}\big)_j &:= \frac{1}{2\sigma^2} \mathrm{tr} \left[F_{\beta_0}^T V_{\beta_0}^{-1} F_{\beta_0} \frac{\partial D}{\partial \alpha_j} \right], \ j = 1, ..., \nu \\ (\mathbf{M}_{1,\alpha})_{j,k} &:= \frac{1}{2} \mathrm{tr} \left[F_{\beta_0}^T V_{\beta_0}^{-1} F_{\beta_0} \frac{\partial D}{\partial \alpha_j} F_{\beta_0}^T V_{\beta_0}^{-1} F_{\beta_0} \frac{\partial D}{\partial \alpha_k} \right], \ j, k = 1, ..., \nu. \end{aligned}$$

Specially the Fisher information for the location parameter vector β under assumed knowledge of the variance parameters σ^2 and α is approximated by

$$\mathbf{M}_{1,\beta} := \frac{1}{\sigma^2} F_{\beta_0}^T V_{\beta_0}^{-1} F_{\beta_0},$$

with $\beta_0 = \beta$ for planning purposes. A second motivation for this approximation will be presented in the next chapter.

4.3.2 Nonlinear Heteroscedastic Model Approximation

Instead of linearizing the response function in a guess β_0, a Taylor approach in the true location parameter vector β as described by Davidian and Giltinan (1995) yields

$$\begin{aligned} Y_i &= \eta(\beta_i, \xi_i) + \epsilon_i \\ &\approx \eta(\beta, \xi_i) + F_\beta(\beta_i - \beta) + \epsilon_i \end{aligned}$$

and with the assumption of negligible linearization errors, the nonlinear mixed effects model is approximated by a nonlinear heteroscedastic normal model:

$$Y_i \overset{app.}{\sim} \mathcal{N}\big(\eta(\beta, \xi_i), \sigma^2(I_{m_i} + F_\beta D F_\beta^T)\big).$$

The Fisher information is in this nonlinear model of the same form as in section 3.3.3, such that for the information of the location parameter vector under assumed knowledge of the variance parameters follows

$$\begin{aligned} \mathbf{M}_{2,\beta} &:= \frac{1}{\sigma^2} F_\beta^T V_\beta^{-1} F_\beta + \frac{1}{2} S \text{ where} \\ S_{j,k} &= \text{tr}\left[V_\beta^{-1} \frac{\partial V_\beta}{\partial \beta_j} V_\beta^{-1} \frac{\partial V_\beta}{\partial \beta_k}\right],\ j,k = 1, ..., p. \end{aligned}$$

Notice that location and variance parameter estimates are not uncorrelated under the nonlinear heteroscedastic model approximation. The complete information matrix is of the form

$$\mathbf{M}_2 := \begin{pmatrix} \mathbf{M}_{2,\beta} & \mathbf{M}_{2,\beta,\sigma^2} & \mathbf{M}_{2,\beta,\alpha} \\ \mathbf{M}_{2,\beta,\sigma^2}^T & \mathbf{M}_{2,\sigma^2} & \mathbf{M}_{2,\sigma^2,\alpha} \\ \mathbf{M}_{2,\beta,\alpha}^T & \mathbf{M}_{2,\sigma^2,\alpha}^T & \mathbf{M}_{2,\alpha} \end{pmatrix},$$

where only the entries $\mathbf{M}_{2,\beta}$ and

$$\begin{aligned} \left(\mathbf{M}_{2,\beta,\sigma^2}\right)_j &:= \frac{1}{2\sigma^2} \text{tr}\left[V_\beta^{-1} \frac{\partial V_\beta}{\partial \beta_j}\right],\ j = 1, ..., p, \\ \left(\mathbf{M}_{2,\beta,\alpha}\right)_{j,k} &:= \frac{1}{2} \text{tr}\left[V_\beta^{-1} \frac{\partial V_\beta}{\partial \beta_j} V_\beta^{-1} F_\beta \frac{\partial D}{\partial \alpha_k} F_\beta^T\right],\ j = 1, ..., p,\ k = 1, ..., \nu \end{aligned}$$

differ from the information matrix of the linear mixed effects approximation. Specially the additional non-negative definite matrix term $\frac{1}{2}S$ is in the approximation of the Fisher information by the nonlinear heteroscedastic model approximation of big interest. The effect of this additional term will be studied on a small example in the next subsection.

Note that both here presented information matrices are based on the statistical models, which are obtained after similar linearizations of the model equation. The difference in the information matrices is generally not negligible.

4.3.3 Influence of the Linearization

Mielke and Schwabe (2010) discussed the difference of the proposed approximations and showed in an example problems of the nonlinear heteroscedastic model approximation for a nonlinear mixed effects model without observation errors:

$$\epsilon_{ij} \quad \sim \mathcal{N}(0, \sigma_\epsilon^2),\ \sigma_\epsilon^2 \to 0,\ \text{while } \beta_i \sim \mathcal{N}(\beta, \sigma_{\beta_i}^2 D),\ \sigma_{\beta_i}^2 D = const.$$

In the proposed example of a log-normally distributed vector of observations, this led to the conclusion that the nonlinear heteroscedastic model approximation systematically generates information by missspecifying the observations by a normal model. Note however, that the presented example illustrated just a boundary case. We here assume the variance parameters to be known and show the difference of the approximations $\mathbf{M}_{1,\beta}$ and $\mathbf{M}_{2,\beta}$ in dependence on the inter-individual variance in a simple example:

Example 4.1. *Let the individual observations be described by*

$$Y_i = \exp(\beta_i) + \epsilon_i$$

with scalar valued random effects:

$$\beta_i \sim \mathcal{N}(\beta, d) \quad \textit{and } \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Then the approximated informations for β result in

$$\begin{aligned} \mathbf{M}_{1,\beta} &= \frac{\exp(2\beta)}{\sigma^2 + d\exp(2\beta)} \quad \textit{and} \\ \mathbf{M}_{2,\beta} &= \frac{\exp(2\beta)}{\sigma^2 + d\exp(2\beta)} + \frac{2d^2\exp(4\beta)}{\sigma^4 + 2d\sigma^2\exp(2\beta) + d^2\exp(4\beta)}. \end{aligned}$$

In figure 4.1 the ratios of the deduced informations are plotted for 5 different parameters β in dependence on the intra-individual variance σ^2 and the inter-individual variance d, which are here parameterized by the ratios

$$\begin{aligned} \sigma^2 = \frac{\rho_\sigma}{1-\rho_\sigma} \quad &\Rightarrow \quad \rho_\sigma = \frac{\sigma^2}{1+\sigma^2},\ \rho_\sigma \in [0,1) \quad \text{and} \\ d = \frac{\rho_d}{1-\rho_d} \quad &\Rightarrow \quad \rho_d = \frac{d}{1+d},\ \rho_d \in [0,1). \end{aligned}$$

On the left hand side of figure 4.1, the observation errors are assumed to follow a standard normal distribution and the variance of the individual parameter β_i is varied. The right hand side illustrates the ratios of the information in dependence on the variance of the observation errors for the case of normally distributed individual parameters β_i with variance $d = 1$.

The ratio of the information converges for all $\beta \in \mathbb{R}$ and inter-individual variances $d \geq 0$:

$$\frac{\mathbf{M}_{1,\beta}}{\mathbf{M}_{2,\beta}} \longrightarrow \frac{1}{1+2d} \quad \text{as} \quad \sigma^2 \longrightarrow 0.$$

This describes the earlier mentioned gain in information, which is also obtained for $\rho_d \to 1$. It will be seen in the next chapter that the information $\mathbf{M}_{1,\beta}$ coincides in these boundary cases with the true Fisher information. Although both approximations are similarly motivated, the observed differences show the need of further investigations of information approximations.

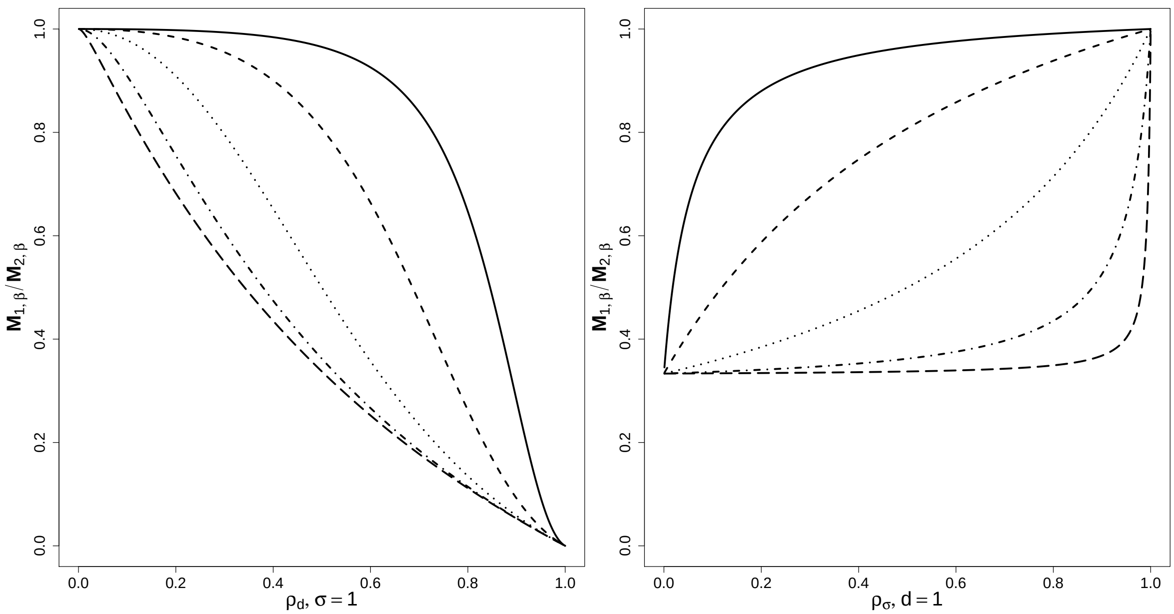

Figure 4.1: Ratios of $\mathbf{M}_{1,\beta}$ and $\mathbf{M}_{2,\beta}$ in dependence on β;
solid: $\beta = -2$, dashed: $\beta = -1$, dotted: $\beta = 0$, dot-dash: $\beta = 1$, longdash: $\beta = 2$

5 Approximation of the Fisher Information

The numerical and analytical calculation of the Fisher information is in nonlinear mixed effects models generally only possible for given experimental settings and is computationally very intensive, such that approximations are applied for optimizing designs in the literature. The linear mixed effects approximation to nonlinear mixed effects models, as described in the preceding section, was developed in Retout *et al.* (2001). This linearization of the model equation was applied already earlier in Mentré *et al.* (1995, 1997) and Tod *et al.* (1998). Merlé and Tod (2001) studied the impact of the linearization on the accuracy of the information matrix and on optimal designs in a pharmacodynamic and a pharmacokinetic model. In the appendix to this article a method for the stochastic computation of the Fisher information based on conditional moments is mentioned.

Approximations of the Fisher information for the parameter vector β, which are based on conditional moments are deduced in this chapter. In the first section the needed representation of the Fisher information is therefor derived. The Laplace approximation and an approach invented by Tierney and Kadane (1986) for approximating posterior densities and conditional moments are presented in the second section. A similar approach is applied for obtaining approximations of the Fisher information in two subsections of the second section. All proposed approximations of the Fisher information are compared on an example in the third section. The final section presents further considerations on the Fisher information for singular inter-individual variance matrices.

5.1 Fisher Information for β

Throughout this chapter we assume the model of the observations of the i-th individual to follow the structure as in the preceding chapter:

$$Y_i = \eta(\beta_i, \xi_i) + \epsilon_i$$

with an exact individual design $\xi_i = (x_{i1}, ..., x_{im_i}) \in \mathcal{X}^{m_i}$ and normally distributed random effects

$$\beta_i \sim \mathcal{N}(\beta, \sigma^2 D),\ \epsilon_i \sim \mathcal{N}(0, \sigma^2 I_{m_i}),$$

with a positive definite matrix D and $\sigma^2 > 0$. The vector valued response function η is considered to be differentiable in β_i and continuous on the m_i-dimensional design region

$\mathcal{X}^{m_i}$. The design matrix is defined in dependence on a vector $\beta_0 \in \mathbb{R}^p$ by the matrix of derivatives:

$$F_{\beta_0} := \frac{\partial \eta(\beta_i, \xi_i)}{\partial \beta_i^T}|_{\beta_i = \beta_0}.$$

Under certain regularity conditions, the Fisher information was defined in the second chapter as the covariance of the score function:

$$\mathfrak{M} := E\left(\frac{\partial l(\theta; Y)}{\partial \theta} \frac{\partial l(\theta; Y)}{\partial \theta^T}\right),$$

with a parameter θ, which in nonlinear mixed effects models comprises the location and variance parameters $\theta = (\beta^T, \sigma^2, \alpha^T)^T$, for a vector α parameterizing the matrix D. The Fisher information was for linear mixed effects models introduced in two stages in the third chapter. Note that the representation depends on the intra and inter-individual statistical model and can be analogously deduced for nonlinear mixed effects models such that the normalized population information

$$\mathfrak{M}_{pop}(\zeta) := \frac{1}{N} E\left(\frac{\partial l(\theta; Y)}{\partial \theta} \frac{\partial l(\theta; Y)}{\partial \theta^T}\right)$$

describes the normalized Fisher information in dependence on the population design ζ, whereas the individual Fisher information is defined as

$$\mathfrak{M}_{ind}(\xi_i) := E\left(\frac{\partial l(\theta; Y_i)}{\partial \theta} \frac{\partial l(\theta; Y_i)}{\partial \theta^T}\right),$$

for exact individual designs $\xi_i \in \mathcal{X}^{m_i}$. As the population Fisher information matrix consists of the weighted sum of the individual Fisher information matrices, knowledge of the structure of the individual Fisher information matrices is sufficient for designing experiments. We thus concentrate in this chapter on the approximation of the individual Fisher information matrix for an exact design ξ_i of sample size m_i, as the results for the population information matrix readily follow by the summation of individual information matrices.

The likelihood function is with known variance parameters σ^2 and α a function of the population location parameter β only and the individual Fisher information for an exact individual design ξ_i then results in

$$\mathfrak{M}^{\beta}_{ind}(\xi_i) = E\left(\frac{\partial l(\theta; Y_i)}{\partial \beta} \frac{\partial l(\theta; Y_i)}{\partial \beta^T}\right).$$

The score function is obtained with the assumption on interchangeability of differentiation and integration as in chapter 4 by

$$\frac{\partial l(\theta; y_i)}{\partial \beta} = \frac{1}{\sigma^2} D^{-1} \big(E(\beta_i | Y_i = y_i) - \beta\big).$$

Hence for the individual Fisher information of the parameter vector β follows:

$$\begin{aligned}
\mathfrak{M}^{\beta}_{ind}(\xi_i) &= E\left(\frac{\partial l(\theta;Y_i)}{\partial\beta}\frac{\partial l(\theta;Y_i)}{\partial\beta^T}\right) \\
&= \frac{1}{\sigma^4}D^{-1}E\left[\big(E(\beta_i|Y_i)-\beta\big)\big(E(\beta_i|Y_i)-\beta\big)^T\right]D^{-1} \\
&= \frac{1}{\sigma^4}D^{-1}Cov\big(E(\beta_i|Y_i)\big)D^{-1} \\
&= \frac{1}{\sigma^2}D^{-1}-\frac{1}{\sigma^4}D^{-1}E\big(Cov(\beta_i|Y_i)\big)D^{-1},
\end{aligned}$$

where the last equality is a direct consequence of the distribution assumptions on the individual parameter vector β_i, since:

$$Cov(\beta_i)=E\big(Cov(\beta_i|Y_i)\big)+Cov\big(E(\beta_i|Y_i)\big)=\sigma^2 D,$$

such that knowledge of the conditional moments of the individual parameter vector is needed for gaining an insight in the Fisher information matrix.

An alternative representation of the Fisher information can be obtained by taking another representation of the score function into account. The score function results with the probability density for the individual parameter vector β_i

$$\phi_{\beta_i}(\upbeta_i)=\sqrt{2\pi\sigma^2}^{\,-p}\sqrt{\parallel D\parallel}^{\,-1}\exp\left[-\frac{1}{2\sigma^2}(\upbeta_i-\beta)^T D^{-1}(\upbeta_i-\beta)\right]$$

and the application of partial integration in

$$\begin{aligned}
\frac{\partial l(\theta;y_i)}{\partial\beta} &= \frac{1}{f_{Y_i}(y_i)}\int_{\mathbb{R}^p}\phi_{Y_i|\upbeta_i}(y_i)\frac{\partial\phi_{\beta_i}(\upbeta_i)}{\partial\beta}d\upbeta_i \\
&= \frac{1}{f_{Y_i}(y_i)}\int_{\mathbb{R}^p}\phi_{Y_i|\upbeta_i}(y_i)\phi_{\beta_i}(\upbeta_i)\frac{1}{\sigma^2}D^{-1}(\upbeta_i-\beta)d\upbeta_i \\
&= -\frac{1}{f_{Y_i}(y_i)}\int_{\mathbb{R}^p}\phi_{Y_i|\upbeta_i}(y_i)\frac{\partial\phi_{\beta_i}(\upbeta_i)}{\partial\upbeta_i}d\upbeta_i \\
&= \frac{1}{f_{Y_i}(y_i)}\int_{\mathbb{R}^p}\frac{\partial\phi_{Y_i|\upbeta_i}(y_i)}{\partial\upbeta_i}\phi_{\beta_i}(\upbeta_i)d\upbeta_i \\
&= \frac{1}{f_{Y_i}(y_i)}\int_{\mathbb{R}^p}\frac{1}{\sigma^2}F^T_{\upbeta_i}[y_i-\eta(\upbeta_i,\xi_i)]\phi_{Y_i|\upbeta_i}(y_i)\phi_{\beta_i}(\upbeta_i)d\upbeta_i \\
&= \frac{1}{\sigma^2}E\big(F^T_{\beta_i}[y_i-\eta(\beta_i,\xi_i)]|Y_i=y_i\big).
\end{aligned}$$

This form of the score function alternatively results when modeling the individual parameter vector as $\beta_i=\beta+b_i$, where the individual random effects b_i have mean zero and variance

$\sigma^2 D$. For the Fisher information follows with this structure of the score function:

$$\begin{aligned}\mathfrak{M}_{ind}^{\beta}(\xi_i) &= E\left(\frac{\partial l(\theta;Y_i)}{\partial \beta}\frac{\partial l(\theta;Y_i)}{\partial \beta^T}\right)\\ &= \frac{1}{\sigma^4}Cov\big[E\big(F_{\beta_i}^T[Y_i-\eta(\beta_i,\xi_i)]|Y_i\big)\big]\\ &= \frac{1}{\sigma^2}E(F_{\beta_i}^T F_{\beta_i}) - \frac{1}{\sigma^4}E\big[Cov\big(F_{\beta_i}^T[Y_i-\eta(\beta_i,\xi_i)]|Y_i\big)\big].\end{aligned}$$

Note that upper bounds of the Fisher information can be constructed with the illustrated representations of the Fisher information:

Lemma 5.1. *Let $Y_i = \eta(\beta_i,\xi_i)+\epsilon_i$, with $\beta_i \sim \mathcal{N}(\beta,\sigma^2 D)$ and $\epsilon_i \sim \mathcal{N}(0,\sigma^2 I_{m_i})$ stochastically independent. Then*

$$\mathfrak{M}_{ind}^{\beta}(\xi_i) \le \min\{\frac{1}{\sigma^2}D^{-1}, \frac{1}{\sigma^2}E(F_{\beta_i}^T F_{\beta_i})\}$$

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

Proof: Since

$$\begin{aligned}\frac{1}{\sigma^2}D^{-1} &= \mathfrak{M}_{ind}^{\beta}(\xi_i) + \frac{1}{\sigma^4}D^{-1}E\big(Cov(\beta_i|Y_i)\big)D^{-1} \text{ and}\\ \frac{1}{\sigma^2}E(F_{\beta_i}^T F_{\beta_i}) &= \mathfrak{M}_{ind}^{\beta}(\xi_i) + \frac{1}{\sigma^4}E\big[Cov\big(F_{\beta_i}^T[Y_i-\eta(\beta_i,\xi_i)]|Y_i\big)\big]\end{aligned}$$

follows with

$$E\big(Cov(\beta_i|Y_i)\big) \ge 0 \text{ and } E\big[Cov\big(F_{\beta_i}^T[Y_i-\eta(\beta_i,\xi_i)]|Y_i\big)\big] \ge 0$$

the inequality $\mathfrak{M}_{ind}^{\beta}(\xi_i) \le \min\{\frac{1}{\sigma^2}D^{-1}, \frac{1}{\sigma^2}E(F_{\beta_i}^T F_{\beta_i})\}$. □

Specially the effect of very small and big inter-individual variances can be well illustrated with the use of the above presented representations of the Fisher information. Let therefor the inter-individual variance matrix be additionally scaled by a scalar τ: $Cov(\beta_i) = \tau\sigma^2 D$. In dependence on τ then easily follows

$$\mathfrak{M}_{ind}^{\beta}(\xi_i) \to \frac{1}{\sigma^2}F_{\beta}^T F_{\beta} \text{ for } \tau \to 0 \text{ and } \mathfrak{M}_{ind}^{\beta}(\xi_i) \to 0 \text{ for } \tau \to \infty,$$

where the first result is a consequence of the resulting distribution of the observation vector Y_i for $\tau\sigma^2 D = 0$. The regression problem then collapses to an ordinary nonlinear regression problem. The second behavior is provided with the upper bound of the Fisher information in the Lemma:

$$\mathfrak{M}_{ind}^{\beta}(\xi_i) \le \frac{1}{\tau\sigma^2}D^{-1} \to 0,\ \tau \to \infty.$$

The assumption on the normality of the individual parameter vectors β_i is often not realistic. In pharmacokinetics the individual parameters describe volumes, elimination rates and absorption rates, which generally cannot take negative values, such that they are alternatively modeled by multiplicative log-normally distributed random individual effects. Schmelter (2007a) describes how the response function and parameters have to be transformed to return to the nonlinear mixed effects model as defined in the preceding chapter. The mean of the resulting individual parameter vectors β_i might then depend on some differentiable function γ_i of the population location parameters β:

$$\beta_i = \gamma_i(\beta) + b_i,\ b_i \sim \mathcal{N}(0, \sigma^2 D)$$

with a differentiable function γ_i. The presented results can be straightforwardly generalized for this case as in Mielke (2011a):

Corollary 5.2. *Let for $j = 1, ..., m_i$*

$$Y_{ij} \quad = \quad \eta(\beta_i, x_{ij}) + \epsilon_{ij},\ \text{with } \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \text{ and } \beta_i \sim \mathcal{N}_p\big(\gamma_i(\beta), \sigma^2 D\big),$$

where γ_i is a differentiable function $\gamma_i : \mathbb{R}^{p_1} \to \mathbb{R}^p$ with $(p \times p_1)$-Jacobi-Matrix $G_i(\beta)$. The Fisher information for the location parameter β then results in

$$\begin{aligned}\mathfrak{M}^{\beta}_{ind}(\xi_i) &= \frac{1}{\sigma^4} G_i(\beta)^T D^{-1} Cov\big(E(\beta_i|Y_i)\big) D^{-1} G_i(\beta) \\ &= \frac{1}{\sigma^2} G_i(\beta)^T \Big(D^{-1} - \frac{1}{\sigma^2} D^{-1} E\big(Cov(\beta_i|Y_i)\big) D^{-1}\Big) G_i(\beta).\end{aligned}$$

Proof: Note that now

$$\phi_{\beta_i}(\upbeta_i) = \sqrt{2\pi\sigma^2}^{\,-p} \sqrt{\| D \|}^{\,-1} \exp[-\frac{1}{2\sigma^2}\big(\upbeta_i - \gamma_i(\beta)\big)^T D^{-1}\big(\upbeta_i - \gamma_i(\beta)\big)],$$

what yields

$$\frac{\partial \phi_{\beta_i}(\upbeta_i)}{\partial \beta} = \frac{1}{\sigma^2} G_i(\beta)^T D^{-1}\big(\upbeta_i - \gamma_i(\beta)\big)\phi_{\beta_i}(\upbeta_i).$$

For the score function readily follows

$$\frac{\partial l(\theta; y_i)}{\partial \beta} \quad = \quad \frac{1}{\sigma^2} G_i(\beta)^T D^{-1}\big(E(\beta_i|Y_i = y_i) - \gamma_i(\beta)\big),$$

such that

$$\mathfrak{M}^{\beta}_{ind}(\xi_i) = \frac{1}{\sigma^4} G_i(\beta)^T D^{-1} E[\big(E(\beta_i|Y_i) - \gamma_i(\beta)\big)\big(E(\beta_i|Y_i) - \gamma_i(\beta)\big)^T] D^{-1} G_i(\beta). \qquad \square$$

Reliable approximations of the conditional moments might lead with the above description of the Fisher information to reliable approximations of the information. For estimating conditional moments often Monte-Carlo methods and quadrature rules are applied, which unfortunately already for small sample sizes and small dimensions of the parameter vector are computationally intensive and hence practicable only in few cases, such that analytic approximations of the Fisher information are of big interest. Tierney and Kadane (1986) propose an approximation of conditional moments by applications of the Laplace approximation. This approximation is computationally less intensive and yields under a slight modification approximations of the Fisher information in a closed form.

5.2 Laplace Approximation

The Laplace approximation is generally used for approximating integrals with an exponential integrand by applying a second order Taylor approach in the mode of the integrand. In nonlinear mixed effects models with normally distributed random effects, the probability density of the observations Y_i can be expressed as an integral depending on the parameter vector $\theta = (\beta^T, \sigma^2, \alpha^T)^T$ with an appropriate parameterization α of the inter-individual variance matrix D:

$$f_{Y_i}(y_i) \quad = \quad \int_{\mathbb{R}^p} \phi_{Y_i|\mathrm{b}_i}(y_i)\phi_{\beta_i}(\mathrm{b}_i)d\mathrm{b}_i,$$

as in the preceding chapter. The exponent of the integrand in $f_{Y_i}(y_i)$ is for nonlinear mixed effects models with normally distributed random effects proportional to the penalized sum of squares:

$$\tilde{l}(\mathrm{b}_i, \theta; y_i) := \big(y_i - \eta(\mathrm{b}_i, \xi_i)\big)^T \big(y_i - \eta(\mathrm{b}_i, \xi_i)\big) + (\mathrm{b}_i - \beta)^T D^{-1}(\mathrm{b}_i - \beta),$$

such that the support point of the Taylor approach in the Laplace approximation is chosen as the individual penalized least squares estimate b_i^*:

$$\mathrm{b}_i^* := \underset{\mathrm{b}_i \in \mathbb{R}^p}{\operatorname{argmin}}\, \tilde{l}(\mathrm{b}_i, \theta; y_i).$$

Since b_i^* fulfills

$$\frac{\partial \tilde{l}(\mathrm{b}_i, \theta; y_i)}{\partial \mathrm{b}_i}|_{\mathrm{b}_i = \mathrm{b}_i^*} = 0,$$

the linear term in the Taylor approach vanishes, yielding

$$\tilde{l}(\mathrm{b}_i, \theta; y_i) \approx \tilde{l}(\mathrm{b}_i^*, \theta; y_i) + \frac{1}{2}(\mathrm{b}_i - \mathrm{b}_i^*)\frac{\partial^2 \tilde{l}(\mathrm{b}_i, \theta; y_i)}{\partial \mathrm{b}_i \partial \mathrm{b}_i^T}|_{\mathrm{b}_i = \mathrm{b}_i^*}(\mathrm{b}_i - \mathrm{b}_i^*).$$

This quadratic form of b_i leads to the approximation of the probability density by

$$\int_{\mathbb{R}^p} \phi_{Y_i|\mathrm{b}_i}(y_i)\phi_{\beta_i}(\mathrm{b}_i)d\mathrm{b}_i \approx \tfrac{1}{c} \cdot \exp[-\tfrac{1}{2\sigma^2}\tilde{l}(\mathrm{b}_i^*, \theta; y_i)], \quad \text{where}$$

$$c = \sqrt{2\pi\sigma^2}^{\,m_i} \sqrt{\| D \| \| \tfrac{1}{2}\tfrac{\partial^2 \tilde{l}(\mathrm{b}_i, \theta; y_i)}{\partial \mathrm{b}_i \partial \mathrm{b}_i^T}|_{\mathrm{b}_i = \mathrm{b}_i^*} \|}.$$

The Hesse matrix of $\tilde{l}$ is in the literature (e.g. Wolfinger and Lin (1997)) often approximated by ignoring the term induced by the second derivatives of the response function η:

$$\begin{aligned} \frac{\partial^2 \tilde{l}(\mathrm{b}_i, \theta; y_i)}{\partial \mathrm{b}_i \partial \mathrm{b}_i^T} &= -2\frac{\partial^2 \eta(\mathrm{b}_i, \xi_i)}{\partial \mathrm{b}_i \partial \mathrm{b}_i^T}\big(y_i - \eta(\mathrm{b}_i, \xi_i)\big) + 2\frac{\partial \eta(\mathrm{b}_i, \xi_i)^T}{\partial \mathrm{b}_i}\frac{\partial \eta(\mathrm{b}_i, \xi_i)}{\partial \mathrm{b}_i^T} + 2D^{-1} \\ &\approx 2\frac{\partial \eta(\mathrm{b}_i, \xi_i)^T}{\partial \mathrm{b}_i}\frac{\partial \eta(\mathrm{b}_i, \xi_i)}{\partial \mathrm{b}_i^T} + 2D^{-1}. \end{aligned}$$

5.2.1 Second Order Approximation

For approximating conditional moments of positive functions $h(\beta_i)$:

$$E\big(h(\beta_i)|Y_i = y_i\big) = \frac{\int h(\beta_i)\exp[-\frac{1}{2\sigma^2}\tilde{l}(\beta_i,\theta;y_i)]d\beta_i}{\int \exp[-\frac{1}{2\sigma^2}\tilde{l}(\beta_i,\theta;y_i)]d\beta_i}, \tag{5.1}$$

Tierney and Kadane (1986) propose the application of Laplace approximations to both, the numerator and the denominator integrals. If the function $h(\beta_i)$ is not positive on the whole space $\mathbb{R}^p$, Tierney *et al.* (1989) suggest the addition of a big constant to $h(\beta_i)$, in order to make the resulting function positive. Notice that the derivation of the mode β_i^* for numerator and denominator is generally not possible in a closed form, such that this heuristic cannot be readily applied for approximating the Fisher information.

Alternatively a similar approach can be applied to the conditional density of β_i for given observations y_i:

$$f_{\beta_i|Y_i=y_i}(\beta_i) := \frac{\phi_{Y_i|\beta_i}(y_i)\phi_{\beta_i}(\beta_i)}{f_{Y_i}(y_i)}. \tag{5.2}$$

With Taylor approaches in different support points for the denominator and numerator, approximations of the conditional density can be obtained. However, it is not guaranteed that the resulting expressions are probability densities. For applications of the Taylor approach in the same support points $\widehat{\beta}_i$ for the numerator and denominator, approximations of the conditional density of β_i for given observations y_i by normal densities are obtained in Mielke (2011a):

Theorem 5.3. *Let $Y_i = \eta(\beta_i,\xi_i) + \epsilon_i$, with $\beta_i \sim \mathcal{N}(\beta,\sigma^2 D)$ and $\epsilon_i \sim \mathcal{N}(0,\sigma^2 I_{m_i})$ stochastically independent and let for $y_i \in \mathbb{R}^{m_i}$ and $\beta_i \in \mathbb{R}^p$*

$$\begin{aligned}
\tilde{l}(\beta_i,\theta;y_i) &:= \big(y_i - \eta(\beta_i,\xi_i)\big)^T\big(y_i - \eta(\beta_i,\xi_i)\big) + (\beta_i - \beta)^T D^{-1}(\beta_i - \beta),\\
\widetilde{F}_{\widehat{\beta}_i} &:= \frac{1}{2}\frac{\partial\tilde{l}(\beta_i,\theta;y_i)}{\partial\beta_i}\Big|_{\beta_i=\widehat{\beta}_i}, \quad \textit{and}\\
\widetilde{M}_{\widehat{\beta}_i} &:= \frac{1}{2}\frac{\partial^2\tilde{l}(\beta_i,\theta;y_i)}{\partial\beta_i\partial\beta_i^T}\Big|_{\beta_i=\widehat{\beta}_i}.
\end{aligned}$$

The approximation of $\tilde{l}$ by a second order Taylor expansion in an estimate $\widehat{\beta}_i$ of β_i yields as an approximation for the conditional distribution of β_i given y_i

$$\beta_i|_{Y_i=y_i} \overset{app.}{\sim} \mathcal{N}\left(\widehat{\beta}_i - \widetilde{M}_{\widehat{\beta}_i}^{-1}\widetilde{F}_{\widehat{\beta}_i}, \sigma^2\widetilde{M}_{\widehat{\beta}_i}^{-1}\right).$$

Proof: The second order Taylor approximation of $\tilde{l}$ in $\widehat{\beta}_i$ yields

$$\begin{aligned}
\tilde{l}(\beta_i,\theta;y_i) &\approx \tilde{l}(\widehat{\beta}_i,\theta;y_i) + 2\widetilde{F}_{\widehat{\beta}_i}(\beta_i - \widehat{\beta}_i) + (\beta_i - \widehat{\beta}_i)^T\widetilde{M}_{\widehat{\beta}_i}(\beta_i - \widehat{\beta}_i)\\
&= \tilde{l}(\widehat{\beta}_i,\theta;y_i) - \widetilde{F}_{\widehat{\beta}_i}^T\widetilde{M}_{\widehat{\beta}_i}^{-1}\widetilde{F}_{\widehat{\beta}_i}\\
&\quad + \big(\beta_i - \widetilde{M}_{\widehat{\beta}_i}^{-1}(\widetilde{M}_{\widehat{\beta}_i}\widehat{\beta}_i - \widetilde{F}_{\widehat{\beta}_i})\big)^T\widetilde{M}_{\widehat{\beta}_i}\big(\beta_i - \widetilde{M}_{\widehat{\beta}_i}^{-1}(\widetilde{M}_{\widehat{\beta}_i}\widehat{\beta}_i - \widetilde{F}_{\widehat{\beta}_i})\big).
\end{aligned}$$

This approximation in the integrand of the probability density of Y_i implies with a constant $c := \sqrt{2\pi\sigma^2}^{\,m_i+p}\sqrt{\| D \|}$:

$$\begin{aligned}
&\int_{\mathbb{R}^p} \phi_{Y_i|\beta_i}(y_i)\phi_{\beta_i}(\beta_i)d\beta_i \\
\approx\;& \exp\big(-\frac{1}{2\sigma^2}[\tilde{l}(\widehat{\beta}_i,\theta;y_i) - \widetilde{F}^T_{\widehat{\beta}_i}\widetilde{M}^{-1}_{\widehat{\beta}_i}\widetilde{F}_{\widehat{\beta}_i}]\big) \\
&\times \int_{\mathbb{R}^p} \frac{1}{c}\cdot \exp\big(-\frac{1}{2\sigma^2}[\beta_i - \widetilde{M}^{-1}_{\widehat{\beta}_i}(\widetilde{M}_{\widehat{\beta}_i}\widehat{\beta}_i - \widetilde{F}_{\widehat{\beta}_i})]^T\widetilde{M}_{\widehat{\beta}_i}[\beta_i - \widetilde{M}^{-1}_{\widehat{\beta}_i}(\widetilde{M}_{\widehat{\beta}_i}\widehat{\beta}_i - \widetilde{F}_{\widehat{\beta}_i})]\big)d\beta_i \\
=\;& \frac{1}{\sqrt{2\pi\sigma^2}^{\,m_i}\sqrt{\| D \|\| \widetilde{M}_{\widehat{\beta}_i} \|}}\exp\big(-\frac{1}{2\sigma^2}[\tilde{l}(\widehat{\beta}_i,\theta;y_i) - \widetilde{F}^T_{\widehat{\beta}_i}\widetilde{M}^{-1}_{\widehat{\beta}_i}\widetilde{F}_{\widehat{\beta}_i}]\big).
\end{aligned}$$

An analogue approximation to the numerator of the conditional density (5.2) yields

$$\begin{aligned}
f_{\beta_i|Y_i=y_i}(\beta_i) \;\approx\;& \sqrt{2\pi\sigma^2}^{\,-p}\sqrt{\| \widetilde{M}_{\widehat{\beta}_i} \|}\times \\
& \exp\big(-\frac{1}{2\sigma^2}[\beta_i - \widetilde{M}^{-1}_{\widehat{\beta}_i}(\widetilde{M}_{\widehat{\beta}_i}\widehat{\beta}_i - \widetilde{F}_{\widehat{\beta}_i})]^T\widetilde{M}_{\widehat{\beta}_i}[\beta_i - \widetilde{M}^{-1}_{\widehat{\beta}_i}(\widetilde{M}_{\widehat{\beta}_i}\widehat{\beta}_i - \widetilde{F}_{\widehat{\beta}_i})]\big),
\end{aligned}$$

such that $\beta_i|_{Y_i=y_i} \overset{app.}{\sim} \mathcal{N}(\widehat{\beta}_i - \widetilde{M}^{-1}_{\widehat{\beta}_i}\widetilde{F}_{\widehat{\beta}_i}, \sigma^2\widetilde{M}^{-1}_{\widehat{\beta}_i})$. □

The individual parameter vectors under given observations are generally not normally distributed. The presented theorem yields just an approximation of the true conditional density for given observations y_i by a normal density.

As the point β_i^* maximizing the function $\tilde{l}$ depends on the observations y_i and the parameter θ, it usually cannot be presented in a closed form. Another problem is met for sparse individual sampling schemes as the function $\tilde{l}$ might then be multimodal. However, the following approximation is given when applying the Taylor approach in a mode β_i^* of $\tilde{l}$:

Remark 5.4. *The approximated conditional distribution of β_i given y_i is for the Laplacian approximation of the form*

$$\beta_i|_{Y_i=y_i} \overset{app.}{\sim} \mathcal{N}\left(\beta_i^*, \sigma^2\widetilde{M}^{-1}_{\beta_i^*}\right).$$

Proof: The result readily follows, since

$$\widetilde{F}_{\beta_i^*} \;:=\; \frac{1}{2}\frac{\partial\tilde{l}(\beta_i,\theta;y_i)}{\partial\beta_i}|_{\beta_i=\beta_i^*} = 0.$$ □

The conditional mean of the individual parameter vector β_i is for the maximum likelihood estimation and for the approximation of the Fisher information of interest. Remark 5.4 presents an approximation of the conditional mean by the conditional mode. The nonlinearity of the response function η implies that the conditional density of β_i for given observations y_i is not symmetrical around β_i^*. Hence mode and mean need not coincide, such that maximum likelihood estimation based on this approximation may become biased.

The conditional expectation and variance might be used for the calculation of the Fisher

information. Note however, that a second level of approximations is needed for deriving the expectation or variance of the conditional moments, as the mode β_i^* generally nonlinearly depends on y_i.
The method presented by Tierney and Kadane (1986) for accurate approximations of posterior moments is based on Laplace approximations with different support points of the Taylor approaches in the numerator and denominator of (5.1). The benefit of using similar approximations to numerator and denominator is that the leading terms of the errors implied by the Taylor expansion cancel when the ratio is taken.

5.2.2 First Order Approximation

First-Order Taylor expansions might be alternatively used for approximating the function $\tilde{l}$. Instead of the application of a second order Taylor expansion of $\tilde{l}$, only the function η is linearly approximated around an estimate $\widehat{\beta}_i$ of the individual parameter vector β_i:

Theorem 5.5. *Let $Y_i = \eta(\beta_i, \xi_i) + \epsilon_i$, with $\beta_i \sim \mathcal{N}(\beta, \sigma^2 D)$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2 I_{m_i})$ stochastically independent, and let for $y_i \in \mathbb{R}^{m_i}$ and $\beta_i \in \mathbb{R}^p$*

$$\begin{aligned}
\tilde{l}(\beta_i, \theta; y_i) &:= \big(y_i - \eta(\beta_i, \xi_i)\big)^T \big(y_i - \eta(\beta_i, \xi_i)\big) + (\beta_i - \beta)^T D^{-1} (\beta_i - \beta), \\
F_{\widehat{\beta}_i} &:= \frac{\partial \eta(\beta_i, \xi_i)}{\partial \beta_i^T}\Big|_{\beta_i = \widehat{\beta}_i}, \text{ and} \\
M_{\widehat{\beta}_i} &:= F_{\widehat{\beta}_i}^T F_{\widehat{\beta}_i} + D^{-1}.
\end{aligned}$$

The approximation of $\tilde{l}$ by a first order Taylor expansion of $\eta(\beta_i, \xi_i)$ in an estimate $\widehat{\beta}_i$ of β_i yields

$$\begin{aligned}
\beta_i|_{Y_i = y_i} &\overset{app.}{\sim} \mathcal{N}\left(\mu(y_i, \widehat{\beta}_i, \beta), \sigma^2 M_{\widehat{\beta}_i}^{-1}\right), \text{ with} \\
\mu(y_i, \widehat{\beta}_i, \theta) &:= M_{\widehat{\beta}_i}^{-1} \Big(F_{\widehat{\beta}_i}^T \big(y_i - \eta(\widehat{\beta}_i, \xi_i) + F_{\widehat{\beta}_i} \widehat{\beta}_i\big) + D^{-1} \beta\Big).
\end{aligned}$$

Proof: With the first order Taylor expansion of the response function η around the estimate $\widehat{\beta}_i$ one obtains

$$\eta(\beta_i, \xi_i) \approx \eta(\widehat{\beta}_i, \xi_i) + F_{\widehat{\beta}_i}(\beta_i - \widehat{\beta}_i).$$

Let $\widetilde{y}_i := y_i - \eta(\widehat{\beta}_i, \xi_i) + F_{\widehat{\beta}_i} \widehat{\beta}_i$. Then

$$\begin{aligned}
\tilde{l}(\beta_i, \theta; y_i) &\approx (\widetilde{y}_i - F_{\widehat{\beta}_i} \beta_i)^T (\widetilde{y}_i - F_{\widehat{\beta}_i} \beta_i) + (\beta_i - \beta)^T D^{-1} (\beta_i - \beta) \\
&= \widetilde{y}_i^T \widetilde{y}_i + \beta^T D^{-1} \beta - (F_{\widehat{\beta}_i}^T \widetilde{y}_i + D^{-1}\beta)^T M_{\widehat{\beta}_i}^{-1} (F_{\widehat{\beta}_i}^T \widetilde{y}_i + D^{-1}\beta) \\
&\quad + \big(\beta_i - M_{\widehat{\beta}_i}^{-1}(F_{\widehat{\beta}_i}^T \widetilde{y}_i + D^{-1}\beta)\big)^T M_{\widehat{\beta}_i} \big(\beta_i - M_{\widehat{\beta}_i}^{-1}(F_{\widehat{\beta}_i}^T \widetilde{y}_i + D^{-1}\beta)\big).
\end{aligned}$$

As in the proof of theorem 5.3 one obtains for the approximation of the integral with a constant $c := \sqrt{2\pi\sigma^2}^{\,m_i+p}\sqrt{\| D \|}$:

$$\begin{aligned}
&\int_{\mathbb{R}^p} \phi_{Y_i|\beta_i}(y_i)\phi_{\beta_i}(\beta_i)d\beta_i \\
\approx\ & \exp\Big(-\frac{1}{2\sigma^2}[\widetilde{y}_i^T\widetilde{y}_i - \mu(y_i,\widehat{\beta}_i,\beta)^T M_{\widehat{\beta}_i}\mu(y_i,\widehat{\beta}_i,\beta) + \beta^T D^{-1}\beta]\Big) \\
& \times \int_{\mathbb{R}^p} \frac{1}{c}\cdot \exp\Big(-\frac{1}{2\sigma^2}[\beta_i - \mu(y_i,\widehat{\beta}_i,\beta)]^T M_{\widehat{\beta}_i}[\beta_i - \mu(y_i,\widehat{\beta}_i,\beta)]\Big)d\beta_i \\
=\ & \frac{\sqrt{2\pi\sigma^2}^{\,-m_i}}{\sqrt{\| D \|\| M_{\widehat{\beta}_i} \|}} \exp\Big(-\frac{1}{2\sigma^2}[\widetilde{y}_i^T\widetilde{y}_i - \mu(y_i,\widehat{\beta}_i,\beta)^T M_{\widehat{\beta}_i}\mu(y_i,\widehat{\beta}_i,\beta) + \beta^T D^{-1}\beta]\Big)
\end{aligned}$$

Applying the same approximation to the numerator of (5.2) yields for the conditional density

$$f_{\beta_i|Y_i=y_i}(\beta_i) \approx \frac{\sqrt{\| M_{\widehat{\beta}_i} \|}}{\sqrt{2\pi\sigma^2}^{\,p}} \exp\Big(-\frac{1}{2\sigma^2}[\beta_i - \mu(y_i,\widehat{\beta}_i,\beta)]^T M_{\widehat{\beta}_i}[\beta_i - \mu(y_i,\widehat{\beta}_i,\beta)]\Big)$$

such that $\beta_i|_{Y_i=y_i} \overset{app.}{\sim} \mathcal{N}\big(\mu(y_i,\widehat{\beta}_i,\theta), \sigma^2 M_{\widehat{\beta}_i}^{-1}\big)$. □

A specific result for an approximation of the conditional distribution is obtained by taking a look at the penalized least squares estimate β_i^* of β_i:

Remark 5.6. *The approximated conditional distribution of β_i given y_i resulting from a First-Order-Linearization in β_i^* is of the form*

$$\beta_i|_{Y_i=y_i} \overset{app.}{\sim} \mathcal{N}\left(\beta_i^*, \sigma^2 M_{\beta_i^*}^{-1}\right).$$

Proof: With β_i^* minimizing $\tilde{l}(\beta_i,\theta;y_i)$ follows

$$\begin{aligned}
& -F_{\beta_i^*}^T\big(y_i - \eta(\beta_i^*,\xi_i)\big) + D^{-1}(\beta_i^* - \beta) = 0 \\
\Leftrightarrow\ & D^{-1}\beta_i^* = F_{\beta_i^*}^T\big(y_i - \eta(\beta_i^*,\xi_i)\big) + D^{-1}\beta,
\end{aligned}$$

such that

$$\begin{aligned}
& M_{\beta_i^*}^{-1}\Big(F_{\beta_i^*}^T\big(y_i - \eta(\beta_i^*,\xi_i) + F_{\beta_i^*}\beta_i^*\big) + D^{-1}\beta\Big) \\
=\ & M_{\beta_i^*}^{-1}\Big(F_{\beta_i^*}^T\big(y_i - \eta(\beta_i^*,\xi_i)\big) + D^{-1}\beta + F_{\beta_i^*}^T F_{\beta_i^*}\beta_i^*\Big) \\
=\ & M_{\beta_i^*}^{-1}(D^{-1}\beta_i^* + F_{\beta_i^*}^T F_{\beta_i^*}\beta_i^*) = \beta_i^*.
\end{aligned}$$

□

The only difference between remark 5.4 and remark 5.6 is the resulting approximation of the conditional variance. This result motivates the earlier presented simplified approximation of the Hesse matrix of $\tilde{l}$. The nonlinear dependence of β_i^* on y_i carries forward to a nonlinear dependence of the conditional moments on y_i, such that estimates of the information still cannot be obtained straightforwardly without yet another approximation.

Besides the big advantage that just first derivatives have to be derived for the First-Order approximation, the second advantage compared to the complete Laplacian approximation is the possibility to specify two approximations of the Fisher information in nonlinear mixed effects models with normally distributed random effects in a closed form:

Remark 5.7. *The approximated conditional distribution of β_i given y_i resulting from a First-Order-Linearization in β is of the form*

$$\beta_i|_{Y_i=y_i} \overset{app.}{\sim} \mathcal{N}\Big(\beta + M_\beta^{-1} F_\beta^T \big(y_i - \eta(\beta,\xi_i)\big), \sigma^2 M_\beta^{-1}\Big).$$

With $V_\beta := I_{m_i} + F_\beta D F_\beta^T$ an approximation of the Fisher information by an approximation of the conditional mean results in:

$$\frac{1}{\sigma^4} D^{-1} Cov\big(E(\beta_i|Y_i)\big) D^{-1} \quad \approx \quad \frac{1}{\sigma^4} F_\beta^T V_\beta^{-1} Cov(Y_i) V_\beta^{-1} F_\beta.$$

Proof: The result for the approximated conditional distribution is a direct consequence of theorem 5.5 with $\widehat{\beta}_i = \beta$ and the approximation of the Fisher information follows since

$$\begin{aligned}\frac{1}{\sigma^4} D^{-1} Cov\big(E(\beta_i|Y_i)\big) D^{-1} \quad &\approx \quad \frac{1}{\sigma^4} D^{-1} M_\beta^{-1} F_\beta^T Cov(Y_i) F_\beta M_\beta^{-1} D^{-1} \\ &= \quad \frac{1}{\sigma^4} F_\beta^T V_\beta^{-1} Cov(Y_i) V_\beta^{-1} F_\beta,\end{aligned}$$

where the equation follows since for V_β regular and M_β as in theorem 5.5:

$$\begin{aligned}D^{-1}(F_\beta^T F_\beta + D^{-1})^{-1} F_\beta^T \quad &= \quad (I_p - F_\beta^T V_\beta^{-1} F_\beta D) F_\beta^T \\ &= \quad F_\beta^T (V_\beta^{-1} V_\beta - V_\beta^{-1} F_\beta D F_\beta^T) = F_\beta^T V_\beta^{-1}.\end{aligned}$$

□

The approximation of the variance of the conditional expectation will be here denoted by $\mathbf{M}_{3,\beta}$:

$$\mathbf{M}_{3,\beta}(\xi_i) := \frac{1}{\sigma^4} F_\beta^T V_\beta^{-1} Cov(Y_i) V_\beta^{-1} F_\beta.$$

A further approximation of the Fisher information matrix is obtained by approximating the conditional variance of the individual parameters for given observations:

Remark 5.8. *The approximated conditional distribution of β_i given y_i resulting from a First-Order-Linearization in β is of the form*

$$\beta_i|_{Y_i=y_i} \overset{app.}{\sim} \mathcal{N}\Big(\beta + M_\beta^{-1} F_\beta^T \big(y_i - \eta(\beta,\xi_i)\big), \sigma^2 M_\beta^{-1}\Big).$$

With $V_\beta := I_{m_i} + F_\beta D F_\beta^T$ an approximation of the Fisher information by an approximation of the conditional variance results in:

$$\frac{1}{\sigma^4} D^{-1} \left[Cov(\beta_i) - E\big(Cov(\beta_i|Y_i)\big)\right] D^{-1} \quad \approx \quad \frac{1}{\sigma^2} F_\beta^T V_\beta^{-1} F_\beta.$$

Proof: The result for the approximated conditional distribution is a direct consequence of theorem 5.5 with $\widehat{\beta}_i = \beta$ and the approximation is given by

$$\begin{aligned}\frac{1}{\sigma^4} D^{-1} \left[Cov(\beta_i) - E\big(Cov(\beta_i|Y_i)\big)\right] D^{-1} \quad &\approx \quad \frac{1}{\sigma^2} D^{-1} - \frac{1}{\sigma^2} D^{-1} (F_\beta^T F_\beta + D^{-1})^{-1} D^{-1} \\ &= \quad \frac{1}{\sigma^2} F_\beta^T V_\beta^{-1} F_\beta,\end{aligned}$$

with an application of a matrix inversion formula for M_β in the last equation. □

Remark 5.8 yields an alternative motivation for the linear mixed effects model approximation:

$$\mathbf{M}_{1,\beta}(\xi_i) := \frac{1}{\sigma^2} F_\beta^T V_\beta^{-1} F_\beta$$

of the preceding chapter, as the obtained information matrices for $\beta_0 = \beta$ coincide. Further approximations can be deduced by other support points of the Taylor approach. However, closed form representations of the approximations are generally only possible, if the support point of the approach does not depend on the observations y_i.
A possibly more refined approximation can be derived by taking the distribution of the observations into account:

$$\begin{aligned} E\big(Cov(\beta_i|Y_i)\big) &= \int_{\mathbb{R}^{m_i}} Cov(\beta_i|Y_i = y_i) \int_{\mathbb{R}^p} \phi_{Y_i|\mathrm{b}_i}(y_i)\phi_{\beta_i}(\mathrm{b}_i) d\mathrm{b}_i dy_i \\ &= \int_{\mathbb{R}^p}\int_{\mathbb{R}^{m_i}} Cov(\beta_i|Y_i = y_i)\phi_{Y_i|\mathrm{b}_i}(y_i)\phi_{\beta_i}(\mathrm{b}_i) dy_i d\mathrm{b}_i. \end{aligned}$$

The covariance of β_i for given observations y_i was with a support point $\widehat{\beta}_i$ in theorem 5.5 approximated by

$$Cov(\beta_i|Y_i = y_i) \approx \sigma^2 M_{\widehat{\beta}_i}^{-1} \;\; := \;\; \sigma^2 (F_{\widehat{\beta}_i}^T F_{\widehat{\beta}_i} + D^{-1})^{-1},$$

such that this approximation in the support point $\widehat{\beta}_i = \mathrm{b}_i$ yields for integrable $M_{\beta_i}^{-1}$:

$$\begin{aligned} E\big(Cov(\beta_i|Y_i)\big) &= \int_{\mathbb{R}^p}\int_{\mathbb{R}^{m_i}} Cov(\beta_i|Y_i = y_i)\phi_{Y_i|\mathrm{b}_i}(y_i)\phi_{\beta_i}(\mathrm{b}_i) dy_i d\mathrm{b}_i \\ &\approx \int_{\mathbb{R}^p} \sigma^2 M_{\mathrm{b}_i}^{-1}\phi_{\beta_i}(\mathrm{b}_i) d\mathrm{b}_i = \sigma^2 E(M_{\beta_i}^{-1}), \end{aligned}$$

where the approximation holds by the argument, that the solution of the penalized least squares problems should be not too far located from the true individual parameter vector, which in the integration is given by b_i. With the same transformations as in remark 5.8 then follows

$$\begin{aligned} \frac{1}{\sigma^4} D^{-1}\big[Cov(\beta_i) - E\big(Cov(\beta_i|Y_i)\big)\big]D^{-1} &\approx \frac{1}{\sigma^4} D^{-1}[Cov(\beta_i) - \sigma^2 E(M_{\beta_i}^{-1})]D^{-1} \\ &= \frac{1}{\sigma^2} E\big(D^{-1} - D^{-1} M_{\beta_i}^{-1} D^{-1}\big) \\ &= \frac{1}{\sigma^2} E(F_{\beta_i}^T V_{\beta_i}^{-1} F_{\beta_i}) \end{aligned}$$

and this approximation will be here defined as $\mathbf{M}_{4,\beta}$:

$$\mathbf{M}_{4,\beta}(\xi_i) := \frac{1}{\sigma^2} E(F_{\beta_i}^T V_{\beta_i}^{-1} F_{\beta_i}).$$

This approximation has to be calculated in nonlinear mixed effects models numerically, as the expectation generally cannot be represented in a closed form. The information approximation $\mathbf{M}_{4,\beta}$ might be of special interest with regard to the two-stage and Lindstrom-Bates estimators, discussed in the fourth chapter. The covariance matrices of these two estimators behave as

$$\sigma^2 \left(\sum_{i=1}^{N} F_{\widehat{\beta}_i}^T V_{\widehat{\beta}_i}^{-1} F_{\widehat{\beta}_i} \right)^{-1}$$

for some estimates $\widehat{\beta}_i$ of the individual parameter vectors. For population studies with big numbers of individuals N, the covariance should hence behave similar to the expectation of the individual information matrices with the distribution of the individual estimates.

The accuracy of the proposed approximations generally depends on the individual sample size m_i and the variance of the individual parameter vectors $\sigma^2 D$. For a bounded inter-individual variance matrix $\sigma^2 D$ and big individual sample sizes m_i, Tierney *et al.* (1989) state in their work on the fully exponential Laplace approximation to ratios of integrals that the accuracy for the approximation of the posterior mean $E\big(h(\beta_i)\big|Y_i = y_i)$ of a function function $h(\beta_i)$ is of order $\mathcal{O}(m_i^{-2})$, while the accuracy for the approximation of the posterior variance is of order $\mathcal{O}(m_i^{-3})$. Following Tierney and Kadane (1986), the proposed accuracy for sufficiently large individual sample sizes m_i is attained already, when using two steps of a Newton iteration for localizing β_i^* instead of the fully exponential Laplace approximation.
The penalized sum of squares:

$$\tilde{l}(\beta_i, \theta; y_i) = \big(y_i - \eta(\beta_i, \xi_i)\big)^T \big(y_i - \eta(\beta_i, \xi_i)\big) + (\beta_i - \beta)^T D^{-1} (\beta_i - \beta),$$

is for individual sample sizes $m_i \to \infty$ dominated by the least squares term

$$\big(y_i - \eta(\beta_i, \xi_i)\big)^T \big(y_i - \eta(\beta_i, \xi_i)\big),$$

what causes the decreasing influence of the distribution of the individual parameter vector β_i on the likelihood of observations y_i for a growing sample size m_i. The approximation of the conditional variance of β_i for given realizations y_i results for an individual experimental design $\xi_i = (x_{i1}, ..., x_{im_i})$, a corresponding design matrix $F_{\widehat{\beta}_i}$ of full column rank with a given vector $\widehat{\beta}_i$ and $n \to \infty$ replications of this experimental design ξ_i within one individual in:

$$\frac{1}{n}(F_{\widehat{\beta}_i}^T F_{\widehat{\beta}_i} + \frac{1}{n} D^{-1})^{-1} \to 0 \ (n \to \infty),$$

such that for the approximation of the Fisher information matrix by the approximation of the conditional variance as in remark 5.8 follows

$$\frac{1}{\sigma^2} D^{-1} - \frac{1}{n\sigma^2} D^{-1} (F_{\widehat{\beta}_i}^T F_{\widehat{\beta}_i} + \frac{1}{n} D^{-1})^{-1} D^{-1} \to \frac{1}{\sigma^2} D^{-1} \ (n \to \infty).$$

The individual parameter vectors β_i are in the population normally distributed with variance $\sigma^2 D$ and can be identified for $n \to \infty$. The approximation $\mathbf{M}_{1,\beta}$ and the Fisher information coincide in this limiting case. The same result can be derived for the information approximation $\mathbf{M}_{4,\beta}$. Unfortunately a generalization of the limiting behavior for $\mathbf{M}_{3,\beta}$ does not hold. Further problems regarding the information approximation $\mathbf{M}_{3,\beta}$ will occur in an example in the next section.

In practical scenarios bounded individual sample sizes m_i are met. For small sample sizes m_i, the impact of the inter-individual variance $\sigma^2 D$ might yield relatively poor approximations, what will be illustrated in the next section.

Notice that all presented approximations coincide for linear response functions η with the true Fisher information.

5.3 Example

Three different approximations of the Fisher information matrix were presented in this chapter. Analytical results on the accuracy of the proposed approximations can unfortunately only be obtained in specific situations. In this section we illustrate the behavior of five different approximations of the Fisher information in the case of the nonlinear mixed effects model presented in example 4.1:

Example 5.9. *Let the individual observations be described by*

$$Y_i = \exp(\beta_i) + \epsilon_i$$

with scalar valued random effects:

$$\beta_i \sim \mathcal{N}(\beta, d) \quad \textit{and } \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

5.3.1 Approximations in the Example

The response function in this example is the exponential function $\eta(\beta_i) = \exp(\beta_i)$, such that the design matrix is given by

$$F_{\beta_0} := \frac{\partial \eta(\beta_i, \xi_i)}{\partial \beta_i^T}|_{\beta_i=\beta_0} = \exp(\beta_0)$$

and the linear approximation of the variance results with $m_i = 1$ and $D = d$ in

$$V_\beta := I_{m_i} + F_\beta D F_\beta^T = \sigma^2 + d\exp(2\beta).$$

The approximations $\mathbf{M}_{1,\beta}$ and $\mathbf{M}_{2,\beta}$ of the Fisher information were introduced in the preceding chapter based on linearizations of the model functions and the resulting statistical model under the assumption of negligible linearization errors:

$$\begin{aligned}
\mathbf{M}_{1,\beta} &:= F_\beta^T V_\beta^{-1} F_\beta = \frac{\exp(2\beta)}{\sigma^2 + d\exp(2\beta)} \text{ and} \\
\mathbf{M}_{2,\beta} &:= F_\beta^T V_\beta^{-1} F_\beta + \frac{1}{2} V_\beta^{-1} \frac{\partial V_\beta}{\partial \beta} V_\beta^{-1} \frac{\partial V_\beta}{\partial \beta} \\
&= \frac{\exp(2\beta)}{\sigma^2 + d\exp(2\beta)} + \frac{2d^2\exp(4\beta)}{\sigma^4 + 2d\sigma^2\exp(2\beta) + d^2\exp(4\beta)}.
\end{aligned}$$

The first approximation $\mathbf{M}_{1,\beta}$ was alternatively motivated in remark 5.8 by an approximation of the conditional variance of the individual parameter vector β_i for given observations y_i. A third approximation was given in remark 5.7 by

$$\mathbf{M}_{3,\beta} = F_\beta^T V_\beta^{-1} Cov(Y_i) V_\beta^{-1} F_\beta.$$

The distribution assumptions on β_i and ϵ_i yield:

$$Cov(Y_i) = Cov\big(\exp(\beta_i)\big) + Cov(\epsilon_i) = \exp(2\beta + d)\big(\exp(d) - 1\big) + \sigma^2,$$

such that with $F_\beta = \exp(\beta)$ follows:

$$\mathbf{M}_{3,\beta} = \frac{\exp(2\beta)\big(\exp(2\beta + d)[\exp(d) - 1] + \sigma^2\big)}{\big(\sigma^2 + d\exp(2\beta)\big)^2}.$$

For the approximation

$$\mathbf{M}_{4,\beta} := E(F_{\beta_i}^T V_{\beta_i}^{-1} F_{\beta_i})$$

it was stated in the last section, that no closed form representation exists. We thus used Monte-Carlo integration for displaying the dependence of $\mathbf{M}_{4,\beta}$ on the variance parameters. A fifth and not yet discussed approximation of the Fisher information is given by the quasi-information. Quasi-likelihood functions are used in generalized linear models if not enough information on the probabilistic model of observations Y_i is given to construct the likelihood function. When the analytical dependence of the first two moments of the observation vector Y_i on the parameter β is known, the quasi-score function for observations y_i is defined by

$$\frac{\partial E(Y_i)^T}{\partial \beta} Cov(Y_i)^{-1}\big(y_i - E(Y_i)\big),$$

as described by Wedderburn (1974) and McCullagh and Nelder (1997). The covariance matrix of the quasi-score function is defined as the quasi-information and results in

$$\mathbf{M}_{5,\beta} := \frac{\partial E(Y_i)^T}{\partial \beta} Cov(Y_i)^{-1} \frac{\partial E(Y_i)}{\partial \beta^T}.$$

For the present example this yields with $E(Y_i) = \exp(\beta + \frac{1}{2}d)$ the fifth approximation:

$$\mathbf{M}_{5,\beta} = \frac{\exp(2\beta + d)}{\exp(2\beta + d)[\exp(d) - 1] + \sigma^2}.$$

5.3.2 Evaluation of the Fisher information

The influence of different parameter values β, d and σ^2 on the accuracy of the approximations was in this example of interest. The Fisher information was examined for 5 different values of the location parameter: $\beta \in \{-2, -1, 0, 1, 2\}$. The variance parameters were parameterized by ρ_d and $\rho_\sigma \in (0, 1)$ in order to observe the dependence of the quality of approximations on the variance parameter d and $\sigma^2 \in \mathbb{R}^+$:

$$\sigma^2 = \frac{\rho_\sigma}{1 - \rho_\sigma} \quad \Rightarrow \quad \rho_\sigma = \frac{\sigma^2}{1 + \sigma^2},\ \rho_\sigma \in (0, 1),$$
$$d = \frac{\rho_d}{1 - \rho_d} \quad \Rightarrow \quad \rho_d = \frac{d}{1 + d},\ \rho \in (0, 1)$$

and the Fisher information was simulated for 1250 different values each of ρ_d for given $\sigma^2 = 1$ and ρ_σ for given $d = 1$ on the interval $(0, 1)$ with 10000 simulated observations per experimental setting $\rho_d,\ \rho_\sigma \in (0, 1)$.

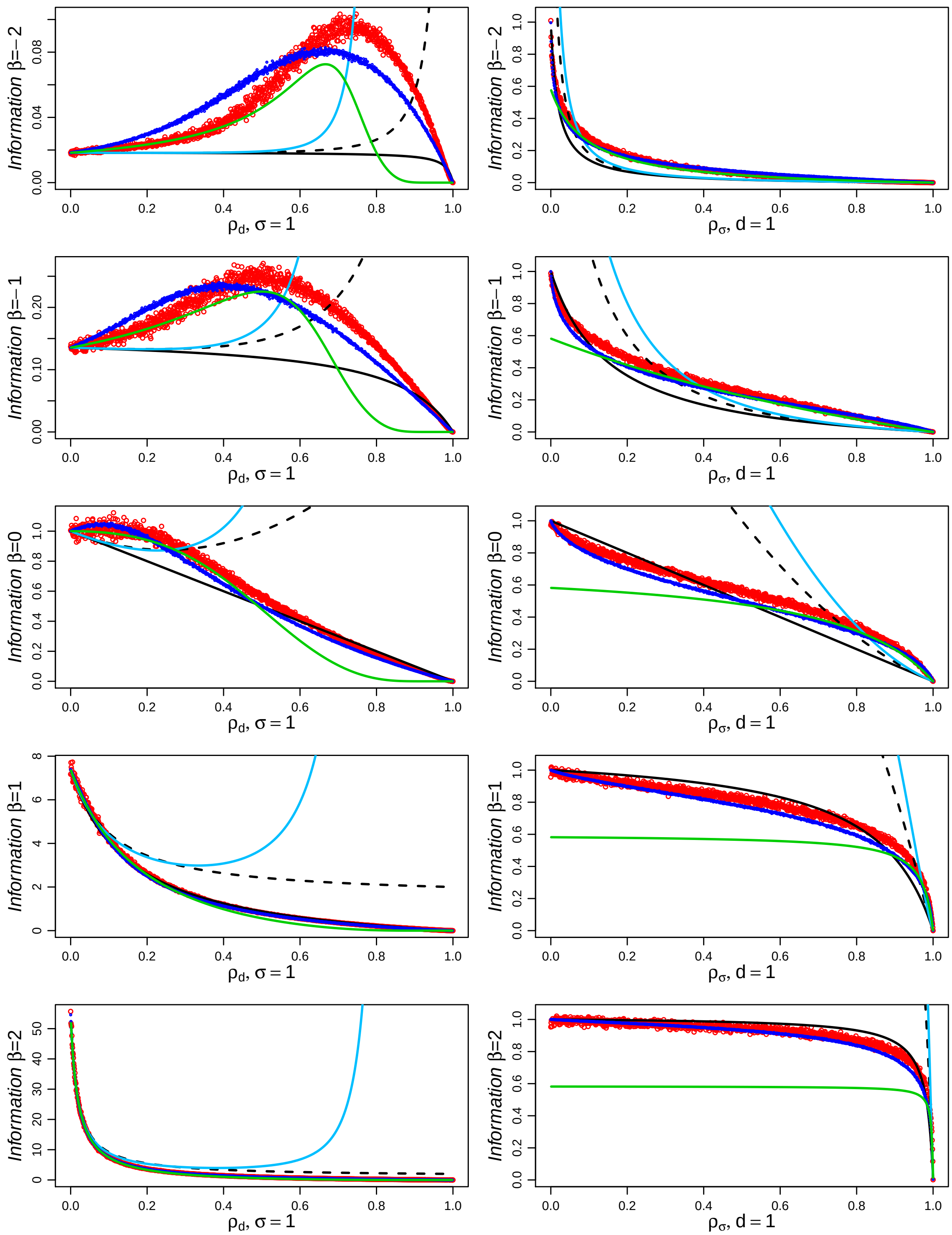

Figure 5.1: Fisher information in dependence on the variance parameters.
Solid: $\mathbf{M}_{1,\beta}$; Dashed: $\mathbf{M}_{2,\beta}$; Light-blue: $\mathbf{M}_{3,\beta}$; Dark-blue: $\mathbf{M}_{4,\beta}$; Green: $\mathbf{M}_{5,\beta}$; Red: $\mathfrak{M}_{ind}^{\beta}$.

5.3.3 Results

Figure 5.1 illustrates the dependence of the Fisher information in this example on the variance of the random effects. The figure shows that all approximations yield for small inter-individual variances d similar results and that no approach approximates the Fisher information satisfactory on the whole positive real line for d.

Specially the often used linear mixed effects approximation $\mathbf{M}_{1,\beta}$ poorly approximates the true Fisher information for moderate values of the inter-individual variance and a negative population parameter β in this example. The linear mixed effects approximation $\mathbf{M}_{1,\beta}$ seems only on the borders of the design region to work well. The results of the approximation $\mathbf{M}_{1,\beta}$ in dependence on the variance parameter σ^2 are more promising.

The approximation $\mathbf{M}_{2,\beta}$ of the Fisher information by the information matrix of a nonlinear heteroscedastic normal model behaves still worse than the approximation $\mathbf{M}_{1,\beta}$. Note that specially the limit for $d \to \infty$ is bigger than zero:

$$\lim_{d\to\infty} \mathbf{M}_{2,\beta} = 2 > 0 = \lim_{d\to\infty} \mathfrak{M}^{\beta}_{ind},$$

where the last limit follows with lemma 5.1 on the upper bound of the Fisher information. The behavior of the Fisher information matrix in dependence on the intra-individual variance σ^2 is not well described by this approximation.

A similar problem is met for the approximation of the Fisher information by the approximation of the conditional mean $\mathbf{M}_{3,\beta}$. The numerator of

$$\mathbf{M}_{3,\beta} = \frac{\exp(2\beta)\big(\exp(2\beta+d)[\exp(d)-1]+\sigma^2\big)}{\big(\sigma^2+d\exp(2\beta)\big)^2}.$$

causes the divergence of this approximation for $d \to \infty$.

In this example the best approximation was given by the mean of the information matrices:

$$\mathbf{M}_{4,\beta} := E(F^T_{\beta_i} V^{-1}_{\beta_i} F_{\beta_i}).$$

Problematic for this approximation is the missing closed form representation, such that the use for experimental design would be restricted to numerical design optimization or optimization based on approximations of this mean information. The approximation $\mathbf{M}_{4,\beta}$ tends to overestimate the Fisher information, similarly to the nonlinear heteroscedastic information $\mathbf{M}_{2,\beta}$, such that it should be used with caution.

The quasi-information $\mathbf{M}_{5,\beta}$ yields reasonable approximations for moderate inter-individual variances, but poorly approximated the Fisher information matrix for big inter-individual variances. However, it can be seen as an robust alternative approach for approximating the Fisher information in nonlinear mixed effects models, as estimators of β based on the quasi-score function should be asymptotically unbiased (Niaparast (2010)) and hence a lower bound of the Fisher information is given by the quasi-information. Note that the quasi-information can only be constructed, when the first two moments of the observation vector are known. This is in most nonlinear mixed effects model not the case, such that one usually would use linear approximations of the response functions in order to approximate the expectation and variance, what leads one back to an information approximation similar to $\mathbf{M}_{1,\beta}$.

5.4 Further Considerations on the Fisher Information

The considerations in the preceding sections were based on the assumptions of a known observation error variance σ^2 and positive definite variance matrix D. Further considerations for the case of non-negative definite variance matrices D and unknown variance parameters σ^2 and α will be briefly undertaken in this section.

5.4.1 Non-negative Definite Matrix D

With the assumption on the positive definiteness of the matrix D the score function for β was calculated as:

$$\frac{\partial l(\theta; y_i)}{\partial \beta} = \frac{1}{\sigma^2} D^{-1}\big(E(\beta_i|Y_i = y_i) - \beta\big),$$

This representation cannot straightforwardly be generalized for singular covariance matrices D. Note that a symmetric matrix D of rank $r < p$ can be expressed with orthogonal matrices H of eigenvectors and the diagonal matrix of corresponding eigenvalues Λ by:

$$D = H\Lambda H^T \quad \text{and} \quad H^T D H = \Lambda,$$

such that:

$$H^T(\beta_i - \beta) \sim \mathcal{N}(0, \sigma^2\Lambda), \text{ with } \Lambda_{j,j} = 0, j > r.$$

By splitting the matrix $H = (H_1, H_2)$ in a $p \times r$ matrix H_1 and $p \times (p - r)$ matrix H_2, the individual parameter vector can be characterized as

$$\beta_i = \beta + H_1 b_i \text{ where } b_i \sim \mathcal{N}(0, \sigma^2\widetilde{\Lambda}) \text{ and } \widetilde{\Lambda} = diag(\Lambda_{1,1}, ..., \Lambda_{r,r}). \tag{5.3}$$

The probability density of the observation vector Y_i then results in

$$f_{Y_i}(y_i) = \int_{\mathbb{R}^r} \phi_{Y_i|\mathbf{b}_i}(y_i)\phi_{b_i}(\mathbf{b}_i)d\mathbf{b}_i$$

with normal densities

$$\begin{aligned}
\phi_{Y_i|\mathbf{b}_i}(y_i) &= \sqrt{2\pi\sigma^2}^{\,-m_i} \exp\left[-\frac{1}{2\sigma^2}\big(y_i - \eta(\beta + H_1\mathbf{b}_i, \xi_i)\big)^T\big(y_i - \eta(\beta + H_1\mathbf{b}_i, \xi_i)\big)\right] \quad \text{and} \\
\phi_{b_i}(\mathbf{b}_i) &= \sqrt{2\pi\sigma^2}^{\,-r}\sqrt{\parallel \widetilde{\Lambda} \parallel}^{\,-1} \exp\left[-\frac{1}{2\sigma^2}\mathbf{b}_i^T\widetilde{\Lambda}^{-1}\mathbf{b}_i\right].
\end{aligned}$$

The representation of the score function as the conditional expectation of β_i cannot be obtained for singular matrices D, since with

$$\frac{\partial \eta(\beta_i, \xi_i)^T}{\partial \mathbf{b}_i} = H_1^T \frac{\partial \eta(\beta_i, \xi_i)^T}{\partial \beta}$$

follows

$$\begin{aligned}
\frac{\partial l(\theta; y_i)}{\partial \beta} &= \frac{1}{\sigma^2} E\big(F_{\beta_i}^T[y_i - \eta(\beta_i, \xi_i)]|Y_i = y_i\big) \\
&= \frac{1}{\sigma^2} H_1\widetilde{\Lambda}^{-1}E(b_i|Y_i = y_i) + \frac{1}{\sigma^2} H_2 H_2^T E\big(F_{\beta_i}^T[y_i - \eta(\beta_i, \xi_i)]|Y_i = y_i\big).
\end{aligned}$$

This representation of the score function otherwise motivates the earlier derived form of the Fisher information matrix

$$\begin{aligned}\mathfrak{M}^{\beta}_{ind}(\xi_i) &= \frac{1}{\sigma^4} Cov\big[E\big(F^T_{\beta_i}[Y_i - \eta(\beta_i,\xi_i)]|Y_i\big)\big] \\ &= \frac{1}{\sigma^2}E(F^T_{\beta_i}F_{\beta_i}) - \frac{1}{\sigma^4}E\big[Cov\big(F^T_{\beta_i}[Y_i - \eta(\beta_i,\xi_i)]|Y_i\big)\big]\end{aligned}$$

with individual parameter vectors as in (5.3). The equality

$$\frac{1}{\sigma^2}E(F^T_{\beta_i}F_{\beta_i}) - \frac{1}{\sigma^4}E\big[Cov\big(F^T_{\beta_i}[Y_i - \eta(\beta_i,\xi_i)]|Y_i\big)\big] = \frac{1}{\sigma^4}D^{-1}Var(E(\beta_i|Y_i))D^{-1}$$

was given in section 5.1 for regular matrices D. The applicability of the proposed approximations $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}_3$ in section 5.2 does however not depend on the regularity of the matrix D and might hence be applied for the approximation of

$$\frac{1}{\sigma^2}E(F^T_{\beta_i}F_{\beta_i}) - \frac{1}{\sigma^4}E\big[Cov\big(F^T_{\beta_i}[Y_i - \eta(\beta_i,\xi_i)]|Y_i\big)\big]$$

in the case of singular inter-individual variance matrices D as well.

Notice that the nonlinear mixed effects model might collapse in dependence on the inter-individual variance matrix to a simpler model, as described in the following example:

Example 5.10. *Let the individual observations be described by*

$$Y_i = \beta_{i;1}\exp(\beta_{i;2}) + \epsilon_i$$

with random effects:

$$\beta_i = (\beta_{i;1},\beta_{i;2})^T \sim \mathcal{N}\big((\beta_1,\beta_2)^T, D\big) \;\; D = diag(d_1, 0) \quad \textit{and } \epsilon_i \sim \mathcal{N}(0,\sigma^2).$$

Then

$$Y_i \sim \mathcal{N}\big(\beta_1\exp(\beta_2), \sigma^2 + d_1\exp(2\beta_2)\big).$$

The true Fisher information matrix $\mathfrak{M}^{\beta}_{ind}$ and the information approximation $\mathbf{M}_{2,\beta}$ in this example coincide:

$$\mathfrak{M}^{\beta}_{ind} \;=\; \mathbf{M}_{2,\beta} = \frac{\exp(2\beta_2)}{\sigma^2 + d_1\exp(2\beta_2)}\begin{pmatrix} 1 & \beta_1 \\ \beta_1 & \beta_1^2 + \frac{2d_1^2\exp(2\beta_2)}{\sigma^2+d_1\exp(2\beta_2)}\end{pmatrix}.$$

The approximations $\mathbf{M}_{1,\beta}$, $\mathbf{M}_{3,\beta}$ and $\mathbf{M}_{5,\beta}$ take here identical values and result in

$$\mathbf{M}_{1,\beta} = \mathbf{M}_{3,\beta} = \mathbf{M}_{5,\beta} = \frac{\exp(2\beta_2)}{\sigma^2 + d_1\exp(2\beta_2)}\begin{pmatrix} 1 & \beta_1 \\ \beta_1 & \beta_1^2\end{pmatrix},$$

and the approximation given by the expectation of the linear mixed effects approximation is calculated as

$$\mathbf{M}_{4,\beta} = \frac{\exp(2\beta_2)}{\sigma^2 + d_1\exp(2\beta_2)}\begin{pmatrix} 1 & \beta_1 \\ \beta_1 & \beta_1^2 + d_1\end{pmatrix}$$

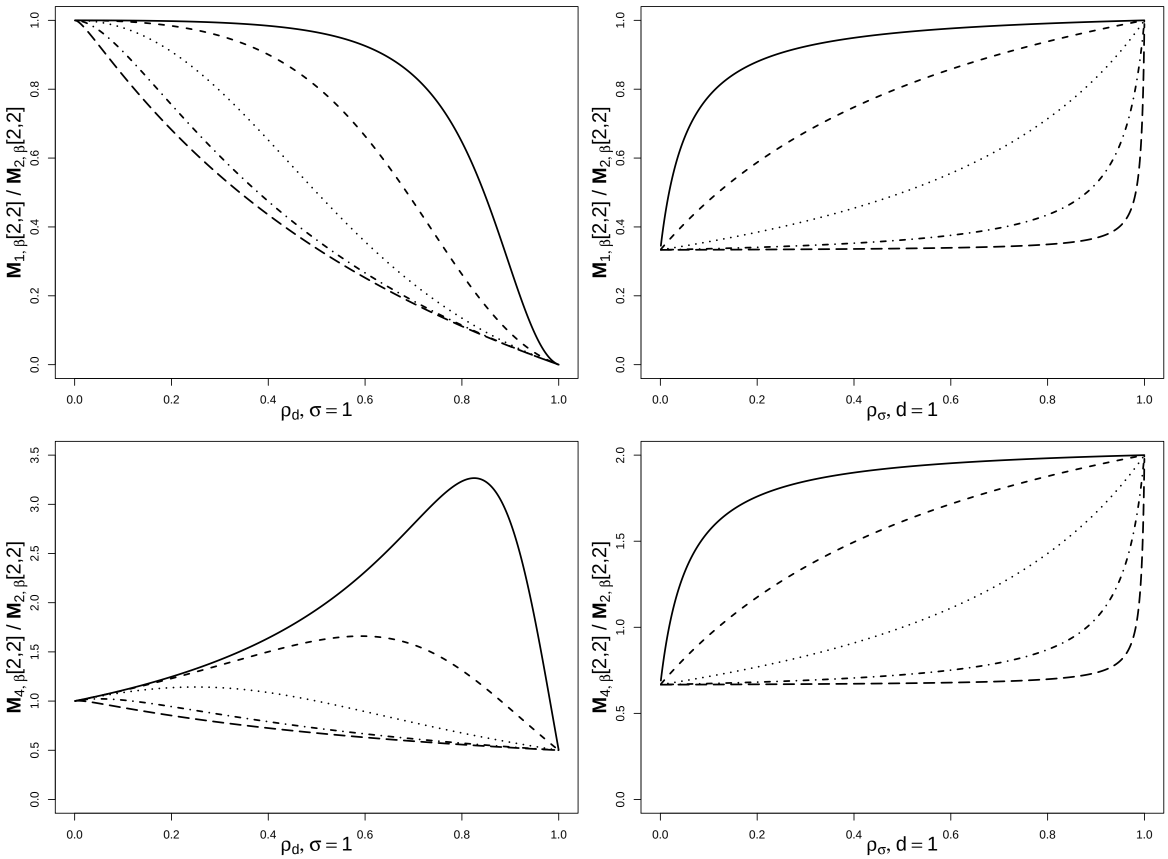

Figure 5.2: Ratios of informations in dependence on β_2;
solid: $\beta_2 = -2$, dashed: $\beta_2 = -1$, dotted: $\beta_2 = 0$, dot-dash: $\beta_2 = 1$, longdash: $\beta_2 = 2$

Only the component for the parameter β_2 of the information matrices differs and specially for $\beta_1 = 1$, example 4.1 illustrates the difference of the components. In figure 5.2 the ratios

$$\frac{(\mathbf{M}_{1,\beta})_{2,2}}{(\mathbf{M}_{2,\beta})_{2,2}} \quad \text{and} \quad \frac{(\mathbf{M}_{4,\beta})_{2,2}}{(\mathbf{M}_{2,\beta})_{2,2}}$$

for the component β_2 were plotted with $\beta_1 = 1$ and the parameterizations $\rho_d = \frac{d_1}{1+d_1}$ (left) and $\rho_\sigma = \frac{\sigma^2}{1+\sigma^2}$ (right). The information approximation $\mathbf{M}_{4,\beta}$ tends to overestimate the true Fisher information in the present example, whereas the linear mixed effects information $\mathbf{M}_{1,\beta}$ underestimates it. This example seemingly contradicts the considerations on the influence of the individual sample size m_i on the accuracy of the information approximations $\mathbf{M}_{1,\beta}$ and $\mathbf{M}_{4,\beta}$ given at the end of section 5.2. These findings were however based on the identifiability of the individual parameter vectors, which in the present example of one observation per individual is not given. Example 5.10 illustrates that approximations of the Fisher information have to be used with caution if the inter-individual variance matrix $\sigma^2 D$ is singular.

5.4.2 Information for the Variance Parameters

The variance parameters σ^2 and α were in the first subsections of the recent chapter considered to be known. Approximations of the variance components of the Fisher information were given by the linear mixed effects and the heteroscedastic normal model approximations in section 4.3. Both approximations yield the same information matrices for the variance parameters:

$$\begin{aligned}
\mathbf{M}_{i,\sigma^2} &:= \tfrac{m_i}{2\sigma^4} \\
\left(\mathbf{M}_{i,\sigma^2,\alpha}\right)_j &:= \tfrac{1}{2\sigma^2}\mathrm{tr}\left[F_\beta^T V_\beta^{-1} F_\beta \frac{\partial D}{\partial \alpha_j}\right],\ j = 1, ..., \nu \\
\left(\mathbf{M}_{i,\alpha}\right)_{j,k} &:= \tfrac{1}{2}\mathrm{tr}\left[F_\beta^T V_\beta^{-1} F_\beta \frac{\partial D}{\partial \alpha_j} F_\beta^T V_\beta^{-1} F_\beta \frac{\partial D}{\partial \alpha_k}\right],\ j,k = 1, ..., \nu,
\end{aligned}$$

for $i = 1, 2$. Estimates of the location parameter β and variance parameters σ^2 and α are in linear mixed effects models uncorrelated:

$$\begin{aligned}
\left(\mathbf{M}_{1,\beta,\sigma^2}\right)_j &:= 0,\ j = 1, ..., p \\
\left(\mathbf{M}_{1,\beta,\alpha}\right)_{j,k} &:= 0,\ j = 1, ..., p,\ k = 1, ...\nu.
\end{aligned}$$

The nonlinear heteroscedastic normal approximation additionally takes the estimated dependence of the location parameter vector β on the variance with the components

$$\begin{aligned}
\left(\mathbf{M}_{2,\beta,\sigma^2}\right)_j &:= \tfrac{1}{2\sigma^2}\mathrm{tr}\left[V_\beta^{-1}\frac{\partial V_\beta}{\partial \beta_j}\right],\ j = 1, ..., p, \\
\left(\mathbf{M}_{2,\beta,\alpha}\right)_{j,k} &:= \tfrac{1}{2}\mathrm{tr}\left[V_\beta^{-1}\frac{\partial V_\beta}{\partial \beta_j} V_\beta^{-1} F_\beta \frac{\partial D}{\partial \alpha_k} F_\beta^T\right],\ j = 1, ..., p,\ k = 1, ..., \nu
\end{aligned}$$

into account.

A generalization of the score-function based approach is for the variance parameters possible, but results in complicated functionals of conditional terms for which the applications of Laplace approximations generally yields no closed form representation. Alternatively, considerations on the distributions of the obtained conditional terms might be undertaken. One example for this case is given by the information matrix component for the variance parameter σ^2. The score function for the parameter σ^2 results in

$$\begin{aligned}
\frac{\partial l(\theta; y_i)}{\partial \sigma^2} &= E\left[\frac{1}{2\sigma^4}\tilde{l}(\beta_i, \theta; y_i)|Y_i = y_i\right] - \frac{m_i + p}{2\sigma^2} \\
&= E\left[\frac{1}{2\sigma^4}\left(y_i - \eta(\beta_i, \xi_i)\right)^T\left(y_i - \eta(\beta_i, \xi_i)\right)|Y_i = y_i\right] \\
&\quad + E\left[\frac{1}{2\sigma^4}(\beta_i - \beta)^T D^{-1}(\beta_i - \beta)|Y_i = y_i\right] - \frac{m_i + p}{2\sigma^2} \\
&= E\left[\frac{1}{2\sigma^4}\left(y_i - \eta(\beta_i, \xi_i) + F_{\beta_i}(\beta_i - \beta)\right)^T\left(y_i - \eta(\beta_i, \xi_i)\right)|Y_i = y_i\right] - \frac{m_i}{2\sigma^2},
\end{aligned}$$

such that the approximation given in the last chapter

$$\mathbf{M}_{1,\sigma^2} = \frac{m_i}{2\sigma^4}$$

can be obtained by assuming the conditional expectation

$$E\left[\frac{1}{\sigma^2}\big(Y_i-\eta(\beta_i,\xi_i)+F_{\beta_i}(\beta_i-\beta)\big)^T\big(Y_i-\eta(\beta_i,\xi_i)\big)|Y_i=y_i\right]$$

to be $\chi^2_{m_i}$-distributed. This idea is motivated by the expectation of the conditional expression:

$$E\left(E\left[\frac{1}{\sigma^2}\big(Y_i-\eta(\beta_i,\xi_i)+F_{\beta_i}(\beta_i-\beta)\big)^T\big(Y_i-\eta(\beta_i,\xi_i)\big)|Y_i\right]\right)=m_i$$

and the term in the conditional expectation, which looks similar to a sum of squared deviations. For the variance of the score function under the distribution assumption then follows

$$\left(\frac{1}{2\sigma^2}\right)^2 Cov\left(E\left[\frac{1}{\sigma^2}\big(Y_i-\eta(\beta_i,\xi_i)+F_{\beta_i}(\beta_i-\beta)\big)^T\big(Y_i-\eta(\beta_i,\xi_i)\big)|Y_i\right]\right)\approx\frac{Cov(\chi^2_{m_i})}{4\sigma^4}=\frac{m_i}{2\sigma^4}.$$

This approach depends heavily on the assumed distribution of the conditional term, such that further considerations on the appropriateness of the $\chi^2_{m_i}$-assumptions have to be undertaken. Similar approaches might be of interest for approximating other variance components of the information matrix. However, as the linear mixed effects approximation and nonlinear heteroscedastic normal approximation of the components for the variance parameters coincide, we will approximate in the following the components of the Fisher information matrix for the variance parameters by the form given in section 4.3.

6 Optimal Designs in Mixed Effects Models

Optimal experimental designs for mixed effects models are discussed in the literature under various aspects. The designs in mixed effects models depend on the measurement replication structure, the optimality criteria, the variance model of the individual parameter vectors, the variance model of the individual observation vectors and the design region. An additional complexity is introduced in nonlinear mixed effects models by the non-satisfactory knowledge of the distribution of the parameter estimates. Designs are in these models based on approximations of the Fisher information and hence additionally depend on the used approximations.

The equivalence theorems as presented in the second chapter are the key instruments for optimizing experimental designs on convex design sets. The population designs and sets of information matrices will be introduced in the first section of this chapter in order to generalize the equivalence theorems to mixed effects models. Special emphasis is laid in subsection 6.1.2 on the properties of the sets of information matrices for different approximations of the Fisher information. Topic of the second section are optimality criteria and equivalence theorems in mixed effects models. The chapter closes with a brief description of numerical approaches for designing experiments in mixed effects models.

6.1 Population Designs and Information Matrices

The observations in the considered mixed effects models depend on individual wise varying parameter vectors, such that two design stages are of interest for the optimization of designs. The individual experimental settings are here described by exact individual designs

$$\xi_i = (x_{i1}, ..., x_{im_i}), \text{ with } x_{ij} \in \mathcal{X},\ j = 1, ..., m_i,$$

where the design space $\mathcal{X}$ is assumed to be a compact set. The proportions ω_i of different individual designs ξ_i in the population are described by the population designs ζ:

$$\zeta = \begin{pmatrix} \xi_1 & ... & \xi_k \\ \omega_1 & ... & \omega_k \end{pmatrix}.$$

We will here concentrate on population designs with identical individual sample sizes $m_i = m$, $i = 1, ..., N$. The population designs can then be defined analogously to approximate designs in ordinary regression models:

Definition 6.1. *The measure* $\zeta = \sum_{i=1}^{k} \omega_i \delta_{\xi_i}$ *with support points* $\xi_i \in \mathcal{X}^m$ *and weights* $\omega_i \geq 0$, $i = 1, ..., k$; $\sum_{i=1}^{k} \omega_i = 1$, *for some* $k \in \mathbb{N}$; *is called population design on* $\mathcal{X}^m$.
The set of all population designs with identical individual sample size m *is denoted by* Ξ^m.

The set Ξ^m results as the convex hull of the set $\mathcal{X}^m$ of exact individual designs of size m and with the compactness of $\mathcal{X}^m$ follows the compactness of Ξ^m. The sets of population designs with m observation under which certain aspects are identifiable might be similarly defined as in definition 2.5. Note however, that the identifiability of aspects depends on the used estimators and the resulting information matrices. This problem will be illustrated by an example in the next chapter.

6.1.1 The Set of Population Information Matrices

The inverse of the covariance matrix of an estimator will be called the information matrix, such that the definition of the term information matrix always depends on the applied estimators. The Fisher information is used for planning experiments, when only few knowledge of the stochastic behavior of estimators is given, as its inverse yields under some regularity conditions a lower bound of the variance of any unbiased estimator. The weighted sum of approximations of the individual Fisher information matrices can be applied for approximating the population Fisher information in mixed effects models:

$$\mathfrak{M}_{pop}(\zeta) = \sum_{i=1}^{k} \omega_i \mathfrak{M}_{ind}(\xi_i) \approx \sum_{i=1}^{k} \omega_i \mathbf{M}_{\cdot;ind}(\xi_i) =: \mathbf{M}_{\cdot;pop}(\zeta),$$

with the index "$\cdot$" representing the used information approximation. The sets of all population information matrices are defined by

$$\mathcal{M}_{\cdot;m} := \{\mathbf{M}_{\cdot;pop}(\zeta), \zeta \in \Xi^m\},$$

where the index $(\cdot; m)$ now illustrates the dependence on the individual sample size and the representation of the information matrices $\mathbf{M}_{\cdot;pop}$. With the assumption on the compactness of the sets of individual information matrices

$$\{\mathbf{M}_{\cdot;ind}(\xi), \xi \in \mathcal{X}^m\}$$

follows the compactness and the convexity of the sets $\mathcal{M}_{\cdot;m}$, since for matrices $\mathbf{M}_{\cdot;pop}(\zeta_1)$ and $\mathbf{M}_{\cdot;pop}(\zeta_2) \in \mathcal{M}_{\cdot;m}$ with population designs ζ_1 and ζ_2 holds for arbitrary $\alpha \in [0, 1]$

$$\begin{aligned}(1-\alpha)\mathbf{M}_{\cdot;pop}(\zeta_1) + \alpha\mathbf{M}_{\cdot;pop}(\zeta_2) &= (1-\alpha)\sum_{i=1}^{k_1} \omega_{i,1}\mathbf{M}_{\cdot;ind}(\xi_{i,1}) + \alpha \sum_{i=1}^{k_2} \omega_{i,2}\mathbf{M}_{\cdot;ind}(\xi_{i,2}) \\ &= \sum_{i=1}^{k_3} \omega_{i,3}\mathbf{M}_{\cdot;ind}(\xi_{i,3}) = \mathbf{M}_{\cdot;pop}(\zeta_3),\end{aligned}$$

where the population design ζ_3 is of the form

$$\zeta_3 = \begin{pmatrix} \xi_{1,1} & \dots & \xi_{k_1,1} & \xi_{1,2} & \dots & \xi_{k_2,2} \\ (1-\alpha)\omega_{1,1} & \dots & (1-\alpha)\omega_{k_1,1} & \alpha\omega_{1,2} & \dots & \alpha\omega_{k_2,2} \end{pmatrix}.$$

Hence the compactness of the set of individual information matrices is the critical point for the generalization of the design theory from the second chapter to mixed effects models. Specially the continuity of individual information matrices as vector valued functions from the compact design region $\mathcal{X}^m$ is sufficient for the compactness of the set $\mathcal{M}_{\cdot;m}$.

6.1.2 Information Matrices in Mixed Effects Models

Throughout this subsection we assume the observations Y_i of individuals i to follow a mixed effects model:

$$Y_i = \eta(\beta_i, \xi) + \epsilon_i, \quad \beta_i \sim \mathcal{N}(\beta, \sigma^2 D) \text{ and } \epsilon_i \sim \mathcal{N}(0, \sigma^2 I_m)$$

with known variance parameters $\sigma^2 > 0$ and D and a vector valued response function η, which is differentiable for $\beta_i \in \mathbb{R}^p$. The design matrix $F_\beta(\xi)$ and the approximation $V_\beta(\xi)$ of the true observation variance $Cov(Y_i)$ are defined as in the preceding chapters by

$$F_\beta(\xi) := \frac{\partial \eta(\beta_i, \xi)}{\partial \beta_i^T}|_{\beta_i=\beta} \quad \text{and} \quad V_\beta(\xi) := I_m + F_\beta(\xi) D F_\beta(\xi)^T.$$

The properties of the response function η can imply the continuity of the individual information matrices and with this the compactness of the set $\mathcal{M}_{\cdot;m}$. The function η has therefor to fulfill characteristics depending on the used approximations of the information matrix.

Corollary 6.2. *Let $\eta(\beta_i, \xi)$ be differentiable in $\beta_i = \beta$ and let $F_\beta(\xi)$ be continuous on $\mathcal{X}^m$. Then the components of the matrix*

$$\mathbf{M}_{1,\beta}(\xi) \quad := \quad \frac{1}{\sigma^2} F_\beta(\xi)^T V_\beta(\xi)^{-1} F_\beta(\xi)$$

are continuous functions of $\xi \in \mathcal{X}^m$.

Proof: The result readily follows since $V_\beta(\xi)$ is regular and $V_\beta(\xi)^{-1}$ is continuous on $\mathcal{X}^m$, such that all components of $\mathbf{M}_{1,\beta}(\xi)$ are sums of products of continuous functions on $\mathcal{X}^m$. □

For the matrix $\mathbf{M}_{2,\beta}(\xi)$ additionally the second derivatives of $\eta(\beta_i, \xi)$ with respect to β_i have to be continuous functions on $\mathcal{X}^m$:

Corollary 6.3. *Let $\eta(\beta_i, \xi)$ be twice differentiable in $\beta_i = \beta$ and let*

$$\frac{\partial^2 \eta(\beta_i, \xi)}{\partial \beta_{i;j} \partial \beta_{i;k}}|_{\beta_i=\beta} \quad \textit{and} \quad F_\beta(\xi)$$

be continuous on $\mathcal{X}^m$. Then the components

$$(\mathbf{M}_{2,\beta}(\xi))_{j,k} \quad := \quad (\mathbf{M}_{1,\beta}(\xi))_{j,k} + \frac{1}{2}\mathrm{tr}\left[V_\beta(\xi)^{-1}\frac{\partial V_\beta(\xi)}{\partial \beta_j} V_\beta(\xi)^{-1}\frac{\partial V_\beta(\xi)}{\partial \beta_k}\right], \ j,k = 1,...,p$$

of the matrix $\mathbf{M}_{2,\beta}(\xi)$ are continuous functions of $\xi \in \mathcal{X}^m$.

Proof: With corollary 6.2 follows the continuity of $\left(\mathbf{M}_{1,\beta}(\xi)\right)_{j,k}$. The second term is continuous as

$$\frac{\partial V_\beta(\xi)}{\partial \beta_j} = \frac{\partial F_\beta(\xi)}{\partial \beta_j} D F_\beta(\xi)^T + F_\beta(\xi) D \frac{\partial F_\beta(\xi)^T}{\partial \beta_j}$$

is with the conditions of the corollary a continuous function on $\mathcal{X}^m$. □

The continuity of the conditional expectation based approximation $\mathbf{M}_{3,\beta}(\xi)$ additionally depends on the continuity of the covariance of observations Y_i:

Corollary 6.4. *Let $\eta(\beta_i,\xi)$ be differentiable in β_i and square integrable with respect to $\mathcal{N}(\beta,\sigma^2 D)$ on $\mathcal{X}^m$. For $F_\beta(\xi)$ and $\eta(\beta_i,\xi)$ continuous on $\mathcal{X}^m$, the components of the matrix*

$$\mathbf{M}_{3,\beta}(\xi) \quad := \quad \frac{1}{\sigma^4} F_\beta(\xi)^T V_\beta(\xi)^{-1} Cov(Y_i) V_\beta(\xi)^{-1} F_\beta(\xi)$$

are continuous functions of $\xi \in \mathcal{X}^m$.

Proof: The continuity of $Cov(Y_i)$ follows, since $\eta(\beta_i,\xi)$ is square integrable and a continuous function on $\mathcal{X}^m$. With the same arguments as in corollary 6.2 then follows the continuity of $\mathbf{M}_{3,\beta}(\xi)$ on $\mathcal{X}^m$. □

For the approximation $\mathbf{M}_{4,\beta}(\xi)$ the integrability of the matrix $F_{\beta_i}(\xi)^T V_{\beta_i}(\xi)^{-1} F_{\beta_i}(\xi)$ is needed for the desired continuity:

Corollary 6.5. *Let $\eta(\beta_i,\xi)$ be differentiable in β_i and $F_{\beta_i}(\xi)$ be continuous on $\mathcal{X}^m$. If*

$$F_{\beta_i}(\xi)^T V_{\beta_i}(\xi)^{-1} F_{\beta_i}(\xi)$$

is on $\mathcal{X}^m$ $\mathcal{N}(\beta,\sigma^2 D)$ integrable, the components of the expectation with respect to the individual parameter vectors β_i:

$$\mathbf{M}_{4,\beta}(\xi) \quad := \quad \frac{1}{\sigma^2} E\left(F_{\beta_i}(\xi)^T V_{\beta_i}(\xi)^{-1} F_{\beta_i}(\xi)\right)$$

are continuous functions of $\xi \in \mathcal{X}^m$.

Proof: The continuity of $F_{\beta_i}(\xi)^T V_{\beta_i}(\xi)^{-1} F_{\beta_i}(\xi)$ follows with corollary 6.2 and with the integrability then follows the continuity of $\mathbf{M}_{4,\beta}(\xi)$. □

The quasi-information is continuous under the assumption of continuous covariance matrices of observations and a continuous matrix of derivatives of the expectation of observations:

Corollary 6.6. *Let $\eta(\beta_i,\xi)$ be square integrable with respect to $\mathcal{N}(\beta,\sigma^2 D)$ on $\mathcal{X}^m$ and the expectation be differentiable in β. For $\eta(\beta_i,\xi)$ and*

$$\frac{\partial}{\partial \beta} E\left(Y_i\right)^T$$

continuous on $\mathcal{X}$, the components of the matrix

$$\mathbf{M}_{5,\beta}(\xi) \quad := \quad \frac{\partial E\left(Y_i\right)^T}{\partial \beta} Cov(Y_i)^{-1} \frac{\partial E\left(Y_i\right)}{\partial \beta^T}.$$

are continuous functions of $\xi \in \mathcal{X}^m$.

Proof: With the continuity and the square integrability of $\eta(\beta_i, \xi)$ follows the continuity of the covariance matrix. The covariance matrix is positive definite, since with $\sigma^2 > 0$

$$Cov(Y_i) = E\left[Cov(Y_i|\beta_i)\right] + Cov\left[E\big(Y_i|\beta_i\big)\right] = \sigma^2 I_m + Cov\big(\eta(\beta_i, \xi)\big) > 0.$$

The result follows, as $\frac{\partial}{\partial\beta} E\big(Y_i\big)^T$ is continuous on $\mathcal{X}^m$. □

Similar results can be deduced for an unknown intra-individual variance σ^2 and by α parameterized inter-individual variance matrix D, which additionally depend on the partial derivatives of the inter-individual variance matrix D with respect to α.

With the compactness of the set of information matrices, every population information matrix $\mathbf{M}_{\cdot;pop} \in \mathcal{M}_{\cdot;m}$ can be represented as a convex combination of information matrices of population designs consisting of one supporting individual design $\xi \in \mathcal{X}^m$ only. Hence Caratheodory's theorem can be applied for limiting the number of supporting individual designs as in ordinary regression models. In the special case of mixed effects models with known variance parameters σ^2 and α, every population information matrix can be represented as the weighted sum of at most $\frac{1}{2}p(p+1)+1$ individual information matrices.

6.2 Optimality Criteria

The same optimality criteria as in ordinary linear and nonlinear regression models can be applied in mixed effects models for comparing the quality of population designs. Generally the information matrix for the whole parameter vector $\theta = (\beta^T, \sigma^2, \alpha^T)^T$ is of interest. If however the variance parameters σ^2 and α are assumed to be known and β is the only parameter to be estimated, one can reduce the optimization problem to the information matrix for the parameter vector β.

The Fisher information matrix in mixed effects models is often approximated with matrices of block diagonal structure, as described in section 4.3.1:

$$\mathfrak{M}_{pop}(\cdot) \approx \begin{pmatrix} \mathfrak{M}^{\beta}_{pop}(\cdot) & 0 & 0 \\ 0 & \mathfrak{M}^{\sigma^2}_{pop}(\cdot) & \mathfrak{M}^{\sigma^2,\alpha}_{pop}(\cdot) \\ 0 & \mathfrak{M}^{\sigma^2,\alpha}_{pop}(\cdot)^T & \mathfrak{M}^{\alpha}_{pop}(\cdot) \end{pmatrix}, \tag{6.1}$$

what simplifies the design optimization. The true Fisher information in nonlinear mixed effects models will generally be not of a block diagonal structure, such that the whole information matrix has to be taken into account in the case of unknown variance parameters σ^2 and α, even if interest lies only in the estimation of β.

The optimality criteria in linear models were defined with respect to some sets Ξ on which some aspects $\psi(\beta) = L_\psi(\beta)$ were identifiable. The same definitions can be applied in linear mixed effects models with the parameter vector θ. In nonlinear mixed effects models the identifiability of aspects $\psi(\theta)$ in general depends on the identifiability of the parameter vectors θ and with this on the used approximation $\mathbf{M}_{\cdot;ind}$ of the Fisher information matrix. One particular example of this dependence is given for the case of known variance parameters in chapter 7, where a D-optimal design with respect to the approximation $\mathbf{M}_{4,\beta}$ is singular with respect to the approximation $\mathbf{M}_{1,\beta}$.

Definition 6.7. *$\Xi^m(\mathbf{M}_{\cdot;pop})$ denotes the set of all population designs ζ on Ξ^m with positive definite information matrix $\mathbf{M}_{\cdot;pop}(\zeta)$.*

6.2.1 Optimality Criteria in Mixed Effects Models

All optimality criteria in nonlinear mixed effects models depend on the used approximations $\mathbf{M}_{\cdot;pop}$ of the Fisher information matrix. The D-optimality criterion in mixed effects models is of the same form as for ordinary regression models in chapter 2:

Definition 6.8. *A design ζ^* is called locally D-optimal with respect to a certain approximation if*

$$\| \mathbf{M}_{\cdot;pop}(\zeta^*) \| \geq \| \mathbf{M}_{\cdot;pop}(\zeta) \|, \ \forall \ \zeta \in \Xi^m.$$

The design ζ^* minimizes $\Phi_{D;\cdot}\big(\mathbf{M}_{\cdot;pop}(\zeta)\big) := -\log\big(\| \mathbf{M}_{\cdot;pop}(\zeta) \| \big)$ on Ξ^m.

D-optimal designs maximize the determinant of the whole information matrix, even if only the location parameter vector β is of interest. This problem might be circumvented with the D_A-optimality as given in Silvey (1980). A design is called D_A-optimal for the estimation of some linear combinations $A^T\theta$ of the parameter vector θ, if it minimizes the expression

$$\Phi_{D_A;\cdot}\big(\mathbf{M}_{\cdot;pop}(\zeta)\big) := \log\big(\| A^T\mathbf{M}_{\cdot;pop}(\zeta)^{-1}A \| \big) \text{ on } \Xi^m(\mathbf{M}_{\cdot;pop}).$$

With the approximation of the Fisher information matrix by a block-diagonal matrix as in (6.1), the optimality criterion $\Phi_{D_A;\cdot}$ with a matrix $A^T = (I_p, 0)$ corresponds to the D-optimality criterion for the reduced information matrix $A^T\mathbf{M}_{\cdot;pop}A$. These criteria however differ, if the block-diagonality of the information matrix is not given.

Schmelter (2007a) minimizes with the G-criterion the maximal variance of the prediction of the response of a typical individual. The response function of typical individuals is given by $\eta\big(E(\beta_i), x\big)$, $x \in \mathcal{X}$ and generally does not coincide with the mean response over the population:

$$\eta\big(E(\beta_i), x\big) \neq E\big(\eta(\beta_i, x)\big), \ x \in \mathcal{X}.$$

The prediction of the response is for the typical individual given by $\eta(\widehat{\beta}, x)$, where $\widehat{\beta}$ is an estimate of the true population location parameter vector β. For consistent and asymptotic normally distributed estimates $\widehat{\theta}^N$ of θ and N observed individuals under a population design ζ with population information matrix $\mathbf{M}_{pop}(\zeta)$:

$$\sqrt{N}(\widehat{\theta}^N - \theta) \xrightarrow{\mathfrak{L}} \mathcal{N}\big(0, \mathbf{M}_{pop}(\zeta)^{-1}\big) \quad (N \to \infty),$$

the Delta-method (Rao (1973, pp. 385)) can be applied if $\eta(\beta, x)$ is differentiable in β with the gradient:

$$f_\theta(x) := \frac{\eta(\beta, x)}{\partial\theta} = \left(\frac{\eta(\beta, x)}{\partial\beta^T}, 0, 0 \cdot \mathbf{1}_\nu^T\right)^T \neq 0,$$

yielding

$$\sqrt{N}\big(\eta(\widehat{\beta},x)-\eta(\beta,x)\big) \xrightarrow{\mathfrak{L}} \mathcal{N}\big(0, f_\theta(x)^T \mathbf{M}_{pop}(\zeta)^{-1} f_\theta(x)\big) \quad (N \to \infty).$$

The derivatives of η with respect to the variance parameters σ^2 and α vanish, as the response function η depends only on the experimental settings and the location parameter vector. The G-criterion is with an information matrix $\mathbf{M}_{\cdot;pop}(\zeta)$ hence defined as follows:

Definition 6.9. *A design ζ^* is called locally G-optimal with respect to a certain approximation if*

$$\max_{x\in\mathcal{X}} f_\theta(x)^T \mathbf{M}_{\cdot;pop}(\zeta^*)^{-1} f_\theta(x) \;\leq\; \max_{x\in\mathcal{X}} f_\theta(x)^T \mathbf{M}_{\cdot;pop}(\zeta)^{-1} f_\theta(x), \quad \forall \zeta \in \Xi^m(\mathbf{M}_{\cdot;pop}).$$

The design ζ^* minimizes $\Phi_{G;\cdot}\big(\mathbf{M}_{\cdot;pop}(\zeta)\big) := \max_{x\in\mathcal{X}} f_\theta(x)^T \mathbf{M}_{\cdot;pop}(\zeta)^{-1} f_\theta(x)$ on $\Xi^m(\mathbf{M}_{\cdot;pop})$.

Linear criteria were introduced in chapter 2 and based on symmetric non-negative definite matrices L. The size of the symmetric non-negative definite matrix L in the here proposed mixed effects models depends on the size of the information matrix. Generally $(p+1+\nu)\times(p+1+\nu)$ matrices L have to be considered for an unknown intra-individual variance σ^2 and a ν-dimensional parametrization α of the matrix D. In the case of known variance parameters σ^2 and α, the information matrix and with this the matrix L are given by $p\times p$ matrices.

Definition 6.10. *A design ζ^* is called locally L-optimal with respect to a certain approximation if*

$$\operatorname{tr}\,\big(L\mathbf{M}_{\cdot;pop}(\zeta^*)^{-1}\big) \;\leq\; \operatorname{tr}\,\big(L\mathbf{M}_{\cdot;pop}(\zeta)^{-1}\big),\; \forall\,\zeta \in \Xi^m(\mathbf{M}_{\cdot;pop}).$$

The design ζ^* minimizes $\Phi_{L;\cdot}\big(\mathbf{M}_{\cdot;pop}(\zeta)\big) := \operatorname{tr}\,\big(L\mathbf{M}_{\cdot;pop}(\zeta)^{-1}\big)$ on $\Xi^m(\mathbf{M}_{\cdot;pop})$.

The Delta-method has to be applied for the derivation of the matrix L in the $IMSE$-criterion. For consistent and asymptotically normally distributed estimators $\widehat{\theta}$ of θ with population information matrix $\mathbf{M}_{pop}(\zeta)$ then follows:

$$\begin{aligned}\int_{\mathcal{X}} E\big([\eta(\widehat{\beta},x)-\eta(\beta,x)]^2\big)\mu(dx) &\approx \int_{\mathcal{X}} f_\theta(x)^T \mathbf{M}_{pop}(\zeta)^{-1} f_\theta(x)\mu(dx)\\ &= \operatorname{tr}\,\Big(\int_{\mathcal{X}} f_\theta(x) f_\theta(x)^T \mu(dx) \mathbf{M}_{pop}(\zeta)^{-1}\Big),\end{aligned}$$

such that the matrix L is given by

$$L := \int_{\mathcal{X}} f_\theta(x) f_\theta(x)^T \mu(dx).$$

The reliable estimation of the response function of the typical individual may be not the only aim of designing experiments. Examples of other functions of interest in the particular case of pharmacokinetics are the area under the curve and the time point of maximal concentration (Holland-Letz (2009, p. 35)). The Delta-method may be applied for arbitrary measurable and differentiable functions of the location parameter θ. The G-criterion and $IMSE$-criterion can be generalized to criteria depending on some real valued and differentiable functions h. This leads in some special cases with

$$c_\theta := \frac{\partial h(\theta)}{\partial \theta}$$

to the c_θ-optimality criterion:

Definition 6.11. *A design ζ^* is called locally c_θ-optimal with respect to a certain approximation if*

$$c_\theta^T \mathbf{M}_{\cdot;pop}(\zeta^*)^{-1} c_\theta \;\leq\; c_\theta^T \mathbf{M}_{\cdot;pop}(\zeta)^{-1} c_\theta, \;\forall\, \zeta \in \Xi^m(\mathbf{M}_{\cdot;pop}).$$

The design ζ^* minimizes $\Phi_{c_\theta;\cdot}\big(\mathbf{M}_{\cdot;pop}(\zeta)\big) := c_\theta^T \mathbf{M}_{\cdot;pop}(\zeta)^{-1} c_\theta$ on $\Xi^m(\mathbf{M}_{\cdot;pop})$.

All here presented criteria are convex optimality criteria and monotone with respect to the Loewner partial ordering on non-negative definite matrices. Moreover the presented D-, L- and c_θ-optimality criteria are differentiable and can be generalized to criteria minimizing the variance matrix of estimators of aspects $\psi(\theta)$ in some sense. Optimal designs however not necessarily exist with the above definitions. One particular example is given by a straight line regression with one observation per individual on the design region $\mathcal{X} = [0,1]$ and random intercepts, where the optimal design for the estimation of the intercept proposes all observations to be taken with individual designs $\xi_i = 0$, what breaks the rule of identifiability of all parameters. A slight modification of the proposed design with a small weight and support point $\omega > 0$ of the design ζ:

$$\zeta = \begin{pmatrix} (0) & (\omega) \\ 1-\omega & \omega \end{pmatrix}$$

yields in this example high efficiencies and satisfies for all $\omega \in (0,1)$ the condition of identifiability.

Analogue definitions for designs in mixed effects models can be given for optimality criteria as the MV- and E-optimality. Note that the G-, MV- and E-optimality criteria are in general not differentiable, what complicates the construction of optimal designs (Torsney and López-Fidalgo (1995), Kiefer (1974)).

6.2.2 Equivalence Theorems in Mixed Effects Models

The assumptions on the convexity and compactness of the sets of population information matrices $\mathcal{M}_{\cdot;m}$ allow the generalization of the design theory from ordinary regression models to designs in mixed effects models. With the Fréchet derivative of optimality criteria Φ of information matrices $\mathbf{M}_1$ in the direction of $\mathbf{M}_2$:

$$F_\Phi(\mathbf{M}_1, \mathbf{M}_2) := \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}\big(\Phi[(1-\varepsilon)\mathbf{M}_1 + \varepsilon \mathbf{M}_2] - \Phi[\mathbf{M}_1]\big),$$

equivalence theorems can be deduced for proving the optimality of information matrices and the corresponding designs in mixed effects models:

Theorem 6.12. *(Silvey (1980)) For convex criteria Φ on $\mathcal{M}_{\cdot;m}$, the design ζ^* is Φ-optimal if and only if*

$$F_\Phi\big(\mathbf{M}_{\cdot;pop}(\zeta^*), \mathbf{M}_{\cdot;pop}(\zeta)\big) \geq 0 \;\; \textit{for all} \;\; \zeta \in \Xi^m.$$

This equivalence theorem does not depend on the differentiability of the optimality criterion Φ in the design ζ^*, such that it can be applied for G-, MV- and E-optimality as well. The linearity of the Fréchet derivative in the second argument for differentiable optimality criteria simplifies the verification of optimal designs:

Theorem 6.13. *(Silvey (1980)) If Φ is convex on $\mathcal{M}_{\cdot;m}$ and differentiable in a design $\zeta^* \in \Xi^m$, then ζ^* is Φ-optimal if and only if*

$$F_\Phi\big(\mathbf{M}_{\cdot;pop}(\zeta^*), \mathbf{M}_{\cdot;ind}(\xi)\big) \geq 0 \;\; \textit{for all} \;\; \xi \in \mathcal{X}^m.$$

Optimal population designs ζ^* with supporting individual designs ξ_i^* and corresponding positive proportions ω_i^* attain for in ζ^* differentiable criteria Φ this lower bound, since by

$$F_\Phi\big(\mathbf{M}_{\cdot;pop}(\zeta^*), \mathbf{M}_{\cdot;ind}(\xi)\big) \quad \geq \quad 0 \;\; \text{for all} \;\; \xi \in \mathcal{X}^m$$

and the linearity of the Fréchet derivative in ζ^* it holds:

$$0 \quad = \quad F_\Phi\big(\mathbf{M}_{\cdot;pop}(\zeta^*), \mathbf{M}_{\cdot;pop}(\zeta^*)\big) = \sum_{i=1}^{k} \omega_i^* F_\Phi\big(\mathbf{M}_{\cdot;pop}(\zeta^*), \mathbf{M}_{\cdot;ind}(\xi_i^*)\big).$$

The computation of the Fréchet derivative is for the most frequently used optimality criteria easily done with the representation:

$$F_\Phi(\mathbf{M}_1, \mathbf{M}_2) := \frac{\partial}{\partial \varepsilon} \Phi[(1-\varepsilon)\mathbf{M}_1 + \varepsilon \mathbf{M}_2]|_{\varepsilon=0}.$$

The results for derivatives of the determinant, the trace and the inverse of a matrix are given in the work on vector differential calculus by Wand (2002):

$$\begin{aligned}
\frac{\partial}{\partial \varepsilon} \parallel \mathbf{M}(\varepsilon) \parallel \quad &= \quad \parallel \mathbf{M}(\varepsilon) \parallel \;\; \text{tr } \mathbf{M}(\varepsilon)^{-1} \frac{\partial}{\partial \varepsilon} \mathbf{M}(\varepsilon) \\
\frac{\partial}{\partial \varepsilon} \text{tr } \mathbf{M}(\varepsilon) \quad &= \quad \text{tr } \frac{\partial}{\partial \varepsilon} \mathbf{M}(\varepsilon) \text{ and} \\
\frac{\partial}{\partial \varepsilon} \mathbf{M}(\varepsilon)^{-1} \quad &= \quad -\mathbf{M}(\varepsilon)^{-1} \frac{\partial}{\partial \varepsilon} \big(\mathbf{M}(\varepsilon)\big) \mathbf{M}(\varepsilon)^{-1}.
\end{aligned}$$

The equivalence theorems for the standard criteria as D- and L-optimality readily follow with the above representation of the Fréchet derivative and were first stated by Fedorov (1972, pp. 209) for designs in the case of simultaneous observations of several random quantities:

Theorem 6.14. *(Fedorov (1972)) The population design ζ^* with $p' \times p'$ information matrix $\mathbf{M}_{\cdot;pop}(\zeta^*)$ minimizes $\Phi_{D;\cdot}$ on Ξ^m if and only if*

$$g_{D;\cdot;\zeta^*}(\xi) := \text{tr } \big(\mathbf{M}_{\cdot;pop}(\zeta^*)^{-1} \mathbf{M}_{\cdot;ind}(\xi)\big) - p' \leq 0 \; \textit{for all} \; \xi \in \mathcal{X}^m.$$

Proof: Since

$$\begin{aligned} F_{\Phi_{D;\cdot}}\big(\mathbf{M}_{\cdot;pop}(\zeta^*), \mathbf{M}_{\cdot;ind}(\xi)\big) &= -\frac{\partial}{\partial\varepsilon}\log[\|\ (1-\varepsilon)\mathbf{M}_{\cdot;pop}(\zeta^*) + \varepsilon\mathbf{M}_{\cdot;ind}(\xi)\ \|]|_{\varepsilon=0} \\ &= -\mathrm{tr}\ \Big(\mathbf{M}_{\cdot;pop}(\zeta^*)^{-1}\frac{\partial}{\partial\varepsilon}[(1-\varepsilon)\mathbf{M}_{\cdot;pop}(\zeta^*) + \varepsilon\mathbf{M}_{\cdot;ind}(\xi)]\Big) \\ &= -\mathrm{tr}\ \big(\mathbf{M}_{\cdot;pop}(\zeta^*)^{-1}[\mathbf{M}_{\cdot;ind}(\xi) - \mathbf{M}_{\cdot;pop}(\zeta^*)]\big), \end{aligned}$$

the equivalence follows with theorem 6.13:

$$\mathrm{tr}\ \big(\mathbf{M}_{\cdot;pop}(\zeta^*)^{-1}\mathbf{M}_{\cdot;ind}(\xi)\big) - p' \le 0 \quad \Leftrightarrow \quad F_{\Phi_{D;\cdot}}\big(\mathbf{M}_{\cdot;pop}(\zeta^*), \mathbf{M}_{\cdot;ind}(\xi)\big) \ge 0. \qquad \square$$

Notice that the equivalence of D- and G-optimal designs is in mixed effects models because of the correlation structure and the heteroscedasticity of the observation vector generally not satisfied. Schmelter (2007a) discusses this issue on the example of a straight line regression model with random slopes. In nonlinear mixed effects models this seems additionally apparent, since the gradient $f_\theta(x)$ has not necessarily a similar connection to the Fisher information as in linear models.

For linear optimality criteria the equivalence theorem takes the following form:

Theorem 6.15. *(Fedorov (1972)) The population design ζ^* minimizes $\Phi_{L;\cdot}$ on $\Xi^m(\mathbf{M}_{\cdot;pop})$ if and only if*

$$g_{L;\cdot;\zeta^*}(\xi) := \mathrm{tr}\ \big(\mathbf{M}_{\cdot;pop}(\zeta^*)^{-1}L\mathbf{M}_{\cdot;pop}(\zeta^*)^{-1}[\mathbf{M}_{\cdot;ind}(\xi) - \mathbf{M}_{\cdot;pop}(\zeta^*)]\big) \le 0 \textit{ for all } \xi \in \mathcal{X}^m.$$

Proof: Since

$$\begin{aligned} & F_{\Phi_{L;\cdot}}\big(\mathbf{M}_{\cdot;pop}(\zeta^*), \mathbf{M}_{\cdot;ind}(\xi)\big) \\ = \ & \frac{\partial}{\partial\varepsilon}\mathrm{tr}\ \big([(1-\varepsilon)\mathbf{M}_{\cdot;pop}(\zeta^*) + \varepsilon\mathbf{M}_{\cdot;ind}(\xi)]^{-1}L\big)|_{\varepsilon=0} \\ = \ & -\mathrm{tr}\ \Big(L\mathbf{M}_{\cdot;pop}(\zeta^*)^{-1}\frac{\partial}{\partial\varepsilon}[(1-\varepsilon)\mathbf{M}_{\cdot;pop}(\zeta^*) + \varepsilon\mathbf{M}_{\cdot;ind}(\xi)]\mathbf{M}_{\cdot;pop}(\zeta^*)^{-1}\Big) \\ = \ & -\mathrm{tr}\ \big(\mathbf{M}_{\cdot;pop}(\zeta^*)^{-1}L\mathbf{M}_{\cdot;pop}(\zeta^*)^{-1}[\mathbf{M}_{\cdot;ind}(\xi) - \mathbf{M}_{\cdot;pop}(\zeta^*)]\big), \end{aligned}$$

the equivalence follows with theorem 6.13:

$$\begin{aligned} & \mathrm{tr}\ \big(\mathbf{M}_{\cdot;pop}(\zeta^*)^{-1}L\mathbf{M}_{\cdot;pop}(\zeta^*)^{-1}[\mathbf{M}_{\cdot;ind}(\xi) - \mathbf{M}_{\cdot;pop}(\zeta^*)]\big) \le 0 \\ \Leftrightarrow \ & F_{\Phi_{L;\cdot}}\big(\mathbf{M}_{\cdot;pop}(\zeta^*), \mathbf{M}_{\cdot;ind}(\xi)\big) \ge 0. \qquad \square \end{aligned}$$

Designs with singular information matrices have to be taken into account for the c-optimality criterion in linear models. An equivalence theorem based on results of Elfving (1952) was for certain structures of the information matrices presented by Holland-Letz (2009). Silvey (1978) states a sufficient condition for the optimality of approximate designs with singular information matrices. Pukelsheim and Titterington (1983) proved that this condition yields an equivalence theorem for designs with singular information matrices, which takes for c_θ-optimality in linear mixed effects models the form as in theorem 6.15, however with a particular generalized inverse of $\mathbf{M}_{\cdot;pop}(\zeta^*)$:

Theorem 6.16. *(Silvey (1978)) A design ζ^* in $\Xi^m_{c_\theta}(\mathbf{M}_{\cdot;pop})$ with $p' \times p'$-information matrix $\mathbf{M}_{\cdot;pop}(\zeta^*)$ of rank $r < p'$ is c_θ-optimal in $\Xi^m_{c_\theta}(\mathbf{M}_{\cdot;pop})$ if and only if there exists a $p' \times (p'-r)$ matrix H, such that*

$$\left(\mathbf{M}_{\cdot;pop}(\zeta^*) + H^T H\right)$$

is regular and for all $\xi \in \mathcal{X}^m$ holds

$$c_\theta^T\left(\mathbf{M}_{\cdot;pop}(\zeta^*) + H^T H\right)^{-1}\left(\mathbf{M}_{\cdot;ind}(\xi) - [\mathbf{M}_{\cdot;pop}(\zeta^*) + H^T H]\right)\left(\mathbf{M}_{\cdot;pop}(\zeta^*) + H^T H\right)^{-1} c_\theta \leq 0.$$

For regular population designs ζ^*, this equivalence theorems simplifies to theorem 6.15, as the matrix $M_{\cdot;pop}(\zeta^*)$ is then of rank p'.

Equivalence theorems in mixed effects models can be applied for optimizing population designs based on the behavior of the functions $g_{D;\cdot;\zeta^*}$ and $g_{L;\cdot;\zeta^*}$. Some examples on the use of the equivalence theorems for analytically solving optimal design problems in mixed effects models can be found under the assumption of known variance parameters σ^2 and α in the publications by Cheng (1995), Graßhoff *et al.* (2012) or Mielke (2009).

6.2.3 Design Algorithms in Mixed Effects Models

The representation of population designs as approximate designs on the design space $\mathcal{X}^m$ allows the application of the optimization procedures mentioned in the second chapter. The population design $\zeta_n \in \Xi^m$ is improved by adding weight $\alpha_n \in (0,1)$ to an individual design $\xi_n \in \mathcal{X}^m$ fulfilling $g_{\cdot;\cdot;\zeta_n}(\xi_n) > 0$:

$$\zeta_{n+1} := (1-\alpha_n)\zeta_n + \alpha_n \delta_{\xi_n}.$$

Typically the individual design ξ_n^* with

$$\xi_n^* := \operatorname*{argmax}_{\xi \in \mathcal{X}^m} g_{\cdot;\cdot;\zeta_n}(\xi)$$

is chosen in order to approach the steepest decent. Note that in every step an m-dimensional optimization problem has to be solved for localizing the optimal individual design ξ_n^*. The algorithm might be quickened by considering other sequences of individual designs ξ_n.

A second numerical approach is based on the design space and given by ordinary optimization algorithms minimizing

$$\Phi\left(\mathbf{M}_{\cdot;pop}(\zeta)\right) \to \min_{\zeta \in \widetilde{\Xi}^m} \text{ with } \widetilde{\Xi}^m := \left\{\mathcal{X}^{km} \times [0,1]^{k-1},\ \sum_{i=1}^{k-1} \omega_i \leq 1\right\},$$

where the number of supporting individual designs k can be limited by the dimension of the information matrix with an application of Caratheodory's theorem:

$$k \leq \frac{1}{2}p' \times (p'+1) + 1.$$

Derivatives of optimality criteria and of the approximations $\mathbf{M}_{1;ind}(\xi)$ and $\mathbf{M}_{2;ind}(\xi)$ of the Fisher information matrix with respect to the experimental settings can be readily computed, as was shown on the approximation $\mathbf{M}_{1;\beta}$ of the Fisher information for the parameter vector β by Mielke (2011a). In comparison to the Fedorov and Wynn algorithms, the there applied $BFGS$-algorithm omits the computation of local maxima in each iteration. Weights and design points in the $BFGS$-algorithm are updated together in each iteration. This optimization algorithm might however stop in a local maximum. Equivalence theorems can after convergence of the algorithm be used in order to verify the optimality of designs.

An optimization algorithm based on information matrices instead of designs was proposed for approximate designs in linear regression models by Gaffke and Heiligers (1996), which might be generalized to mixed effects models. Note however, that the algorithm returns at convergence an optimal information matrix and the corresponding design then has to be determined.

Throughout this chapter we assumed the individual sample size to be given by some number m. One of the biggest problems for the determination of optimal designs in mixed effects models is this individual sample size. The derivation of optimal individual sample sizes m depends on cost constraints, physical limitations and ethical considerations. Moreover, the efficiency of population designs with respect to the invented optimality criteria depends for a given number of total measurements m_T heavily on the number of samples m per individual. Designs with many observations $m >> p$ on few individuals tend under these settings to be less efficient for the estimation of the population location parameter than designs with few observations on many individuals. Another problem is the verification of optimal designs, which is for big individual sample sizes and convex design regions $\mathcal{X}$ not trivial. The optimality results of Schmelter (2007b) on approximate individual designs might help finding good initial population designs for arbitrary individual sample sizes m.

7 Information Approximation and Designs

Different approximations of the Fisher information matrix in nonlinear mixed effects models were discussed in the fifth chapter. In the preceding chapter optimality criteria and equivalence theorems for mixed effects models were derived, which will be applied in the present chapter for illustrating the influence of different information approximations on designs of population studies.

Optimal experimental designs in nonlinear mixed effects models are in the literature usually derived either with the linear mixed effects approximation (e.g. Schmelter (2007a)) or with the nonlinear heteroscedastic normal approximation as in Retout and Mentré (2003). The impact of information approximations on the design is however not well discussed in the literature. Merlé and Tod (2001) and Bazzoli *et al.* (2009) discuss the appropriateness of certain approximations, but implications to the design of population studies and comparisons with other information approximations are not given by the authors. Recently, Mielke (2011a, b) computes optimal designs and compares the resulting sampling schemes for different approximations in pharmacokinetic models.

Optimal experimental designs are deduced in this chapter for two pharmacokinetic models. Compartment models are briefly introduced in the first section with the here used information approximations. Topic of the second section is the one-compartment model without absorption. Designs are computed for different Fisher information approximations and compared to the optimal design resulting from a simulation-based approximation of the Fisher information matrix. In the third section the impact of information approximations on designs in a one-compartment model with first-order absorption is examined.

7.1 Compartment Models

Pharmacokinetics describes the absorption, distribution, metabolism and excretion of a drug in a body. Compartmental models are often applied in order to estimate the time-course of a drug concentration in a body. Seber and Wild (2003, ch. 8) and Schmelter (2007a) present the ideas and theory of compartment models, such that we will not go into the details of compartment-modeling for pharmacokinetic studies. The concentration-time profile in one-compartment models is defined by the three parameters:

- k_a: the absorption rate constant, here β_1;

- Cl: clearance, as the amount of plasma, which is cleared within one unit of time, here β_2;

- V_c: the volume of distribution in the central compartment, here β_3.

The one-compartment model without absorption is used in pharmacokinetic studies with intravenous bolus administrations of the drug. The drug is immediately injected into the compartment, such that the model for the concentration-time profile is given by an exponential decay depending on an elimination rate k_e and the volume V_c of distribution in the central compartment, where the elimination rate k_e is defined in terms of the clearance Cl and the volume of distribution by

$$k_e = \frac{Cl}{V_c}.$$

The concentration-time profile is given in dependence on the dose $\mathbf{D}$ by

$$\eta_1(\beta, x) := \frac{\mathbf{D}}{\beta_3} \exp\left(-\frac{\beta_2}{\beta_3}x\right)$$

with the location parameter vector $\beta := (\beta_2, \beta_3)^T$.

Additionally the absorption of the drug in the body has to be taken into account for orally administered drugs, such that the one compartment model with first order absorption depends on an absorption rate constant k_a. The concentration-time profile is described by the bi-exponential model:

$$\eta_2(\beta, x) := \frac{\mathbf{D}\beta_1}{\beta_3\beta_1 - \beta_2} \left[\exp\left(-\frac{\beta_2}{\beta_3}x\right) - \exp\left(-\beta_1 x\right)\right], \text{ where } \beta := (\beta_1, \beta_2, \beta_3)^T.$$

Observations of the concentration are in general obtained by taking blood samples at a time x after the drug administration. The observations are influenced by sample-wise varying errors and individual-wise varying parameter vectors. Complicated variance structures for the individual observation errors as in Bazzoli *et al.* (2009) may be assumed. However, we will restrict ourselves in this chapter to intra-individual models as in Mentré *et al.* (1997):

$$Y_{ij} = \eta_k(\beta_i, x_{ij}) \exp(\epsilon_{ij}) \quad \text{with} \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \quad k = 1, 2,$$

as even for these simpler error structures no sufficient results are known.

The individual parameter vectors are assumed to be log-normally distributed with

$$\beta_{i;l} = \beta_l \exp(b_{i;l}),\ l = 1, 2, 3 \ \text{ where } b_i = (b_{i;1}, b_{i;2}, b_{i;3})^T \sim \mathcal{N}(0, \sigma^2 D),\ D = diag(d_1, d_2, d_3).$$

Figure 7.1 illustrates possible time-concentration profiles of six different individuals in the presented one-compartment models with the according model of the individual parameter vectors.

The log-concentration $\widetilde{Y}_{ij} := \log(Y_{ij})$ follows with $\widetilde{\beta}_i := \log(\beta) + b_i$ and with the response function

$$\widetilde{\eta}_k(\widetilde{\beta}_i, x) := \log\left[\eta_k\big(\exp(\widetilde{\beta}_i), x\big)\right],\ x \in \mathcal{X} \tag{7.1}$$

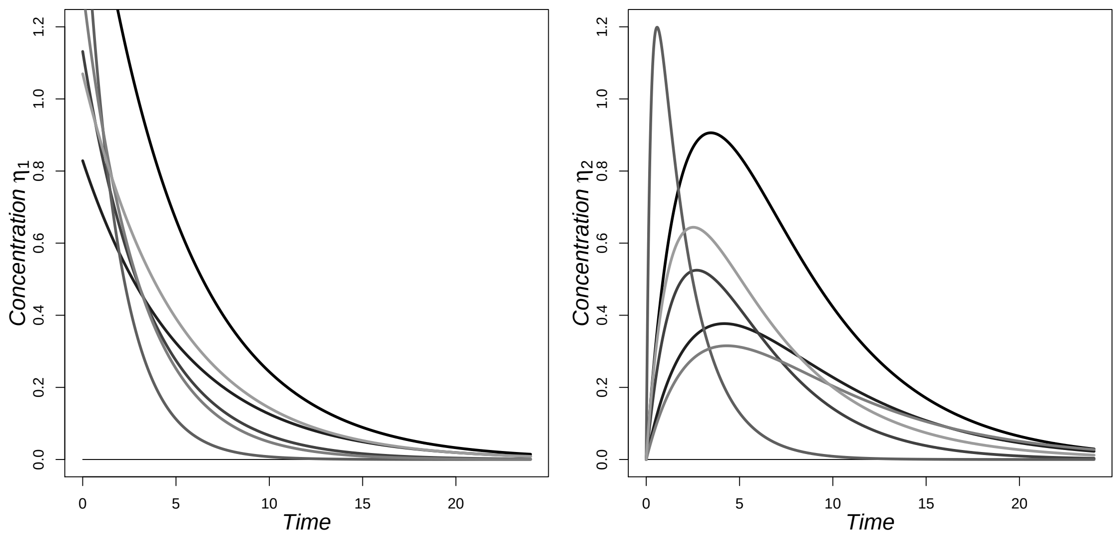

Figure 7.1: Concentration-Time profiles in one-compartment models:
Left: η_1 - without absorption; Right: η_2 - with absorption

a nonlinear mixed effects models as in corollary 5.2. The design matrix is denoted as in chapter 5 for an exact individual design $\xi_i = (x_{i1}, ..., x_{im})$ and with the now vector valued response function $\widetilde{\eta}_k$ by

$$F_{\beta_0}(\xi_i) := \frac{\partial \widetilde{\eta}_k(\widetilde{\beta}_i, \xi_i)}{\partial \widetilde{\beta}_i^T}|_{\widetilde{\beta}_i = \beta_0}$$

and the approximation V_{β_0} of the variance matrix is given by

$$\sigma^2 V_{\beta_0}(\xi_i) := \sigma^2 \big(I_{m_i} + F_{\beta_0}(\xi_i) D F_{\beta_0}(\xi_i)^T\big).$$

We apply the approximations presented in chapter 5 in order to approximate the Fisher information matrix. The resulting information approximation $\mathbf{M}_{1,\beta}$ is with an application of corollary 5.2 for the function $\gamma(\beta) := \big(\log(\beta_1), \log(\beta_2), \log(\beta_3)\big)^T$ and the matrix

$$G(\beta) := \begin{pmatrix} \frac{1}{\beta_1} & 0 & 0 \\ 0 & \frac{1}{\beta_2} & 0 \\ 0 & 0 & \frac{1}{\beta_3} \end{pmatrix}$$

of the form

$$\mathbf{M}_{1,\beta}(\xi_i) \quad = \quad \frac{1}{\sigma^2} G(\beta) F_{\gamma(\beta)}(\xi_i)^T V_{\gamma(\beta)}(\xi_i)^{-1} F_{\gamma(\beta)}(\xi_i) G(\beta).$$

The information approximation $\mathbf{M}_{2,\beta}$ for an individual design ξ_i is derived as described in section 4.3.2 with an application of a first order Taylor approach:

$$\widetilde{Y}_i = \widetilde{\eta}_k(\widetilde{\beta}_i, \xi_i) + \epsilon_i \approx \widetilde{\eta}_k\big(\gamma(\beta), \xi_i\big) + F_{\gamma(\beta)}(\xi_i)\big(\widetilde{\beta}_i - \gamma(\beta)\big) + \epsilon_i,$$

and for the model of $\widetilde{Y}_i$ under the assumption of negligible linearization errors follows

$$\widetilde{Y}_i \overset{app.}{\sim} \mathcal{N}\Big(\widetilde{\eta}_k\big(\gamma(\beta), \xi_i\big), \sigma^2 V_{\gamma(\beta)}\big(\xi_i\big)\Big).$$

The information approximation is hence of the form

$$\begin{aligned} \mathbf{M}_{2,\beta}(\xi_i) &= \mathbf{M}_{1,\beta}(\xi_i) + \frac{1}{2}S(\xi_i) \quad \text{with} \\ \left(S(\xi_i)\right)_{j,l} &= \text{tr} \left[V_{\gamma(\beta)}(\xi_i)^{-1} \frac{\partial V_{\gamma(\beta)}(\xi_i)}{\partial \beta_j} V_{\gamma(\beta)}(\xi_i)^{-1} \frac{\partial V_{\gamma(\beta)}(\xi_i)}{\partial \beta_l} \right], \; j,l = 1, ..., 3. \end{aligned}$$

For the approximations $\mathbf{M}_{3,\beta}$, $\mathbf{M}_{4,\beta}$ and $\mathbf{M}_{5,\beta}$ the moments of interest generally have to be approximated numerically. Specially for $\mathbf{M}_{3,\beta}$ and $\mathbf{M}_{4,\beta}$ the function $\gamma(\beta)$ has to be taken into account:

$$\begin{aligned} \mathbf{M}_{3,\beta}(\xi_i) &= \frac{1}{\sigma^4} G(\beta) F_{\gamma(\beta)}(\xi_i)^T V_{\gamma(\beta)}(\xi_i)^{-1} Cov(Y_i) V_{\gamma(\beta)}(\xi_i)^{-1} F_{\gamma(\beta)}(\xi_i) G(\beta) \quad \text{and} \\ \mathbf{M}_{4,\beta}(\xi_i) &= \frac{1}{\sigma^2} G(\beta) E\big[F_{\widetilde{\beta}_i}(\xi_i)^T V_{\widetilde{\beta}_i}(\xi_i)^{-1} F_{\widetilde{\beta}_i}(\xi_i)\big] G(\beta). \end{aligned}$$

The Quasi-information is obtained with the derivatives of the expectation and the variance matrix of the observation vector:

$$\mathbf{M}_{5,\beta}(\xi_i) = \frac{\partial E(Y_i)^T}{\partial \beta} Cov(Y_i)^{-1} \frac{\partial E(Y_i)}{\partial \beta^T}.$$

Specific results on locally optimal designs for the introduced information approximations in two one-compartment models are given in the two following sections.

7.2 One Compartment Without Absorption

The model of the log-concentration in the one compartment model without absorption results with the transformations of the preceding section for the j-th observation of the i-th individual in

$$\begin{aligned} \widetilde{Y}_{ij} &= \widetilde{\eta}_1(\widetilde{\beta}_i, x_{ij}) + \epsilon_{ij} \\ &= \log(\mathbf{D}) - \widetilde{\beta}_{i;3} - x_{ij} \exp(\widetilde{\beta}_{i;2} - \widetilde{\beta}_{i;3}) + \epsilon_{ij} \end{aligned}$$

where the random effects are assumed to be normally distributed:

$$\widetilde{\beta}_i = (\widetilde{\beta}_{i;2}, \widetilde{\beta}_{i;3})^T \sim \mathcal{N}\big((\log[\beta_2], \log[\beta_3])^T, \sigma^2 diag(d_2, d_3)\big) \text{ and } \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2).$$

Expectation and variance are for the j-th observation of the i-th individual under experimental settings x_{ij} specified in a closed form by

$$\begin{aligned} E(\widetilde{Y}_{ij}) &= \log(\mathbf{D}) - \log(\beta_3) - x_{ij} \frac{\beta_2}{\beta_3} \exp\big(\frac{1}{2}\sigma^2(d_2 + d_3)\big) , \\ Var(\widetilde{Y}_{ij}) &= \sigma^2(1 + d_3) + x_{ij}^2 \frac{\beta_2^2}{\beta_3^2} \exp\big(\sigma^2(d_2 + d_3)\big)\big(\exp[\sigma^2(d_2 + d_3)] - 1\big) \\ &\quad -2x_{ij}\sigma^2 d_3 \frac{\beta_2}{\beta_3} \exp\big(\frac{1}{2}\sigma^2(d_2 + d_3)\big) \end{aligned}$$

and observations within one individual at experimental settings x_{ij} and $x_{ij'}$ are correlated:

$$\begin{aligned} Cov(\widetilde{Y}_{ij}, \widetilde{Y}_{ij'}) &= \sigma^2 d_3 + x_{ij} x_{ij'} \frac{\beta_2^2}{\beta_3^2} \exp\big(\sigma^2(d_2+d_3)\big)\big(\exp[\sigma^2(d_2+d_3)] - 1\big) \\ &\quad -(x_{ij} + x_{ij'})\sigma^2 d_3 \frac{\beta_2}{\beta_3} \exp\big(\frac{1}{2}\sigma^2(d_2+d_3)\big), \end{aligned}$$

such that the information approximations $\mathbf{M}_{3,\beta}$ and $\mathbf{M}_{5,\beta}$ can be readily calculated.

For analyzing the influence of the approximations on the design, we assume the variance parameters to be known and optimize the population designs in the model with the numbers for the parameters as given in Schmelter (2007a, exp. 8.7):

$$\beta_2 = 25,\ \beta_3 = 88,\ \sigma^2 = 0.01,\ d_2 = 12.5,\ d_3 = 9.0.$$

The individual experimental design in this model describes the time-points $x_{ij} \in \mathcal{X}$ after drug administration at which blood samples are to be taken from individuals. D- and $IMSE$-optimal designs for the information approximations $\mathbf{M}_{1,\beta}$, $\mathbf{M}_{2,\beta}$, $\mathbf{M}_{3,\beta}$ and $\mathbf{M}_{5,\beta}$ were numerically derived on the design region $\mathcal{X} = [0.1, 24]$ in Mielke (2011a) for individual designs consisting of $m = 1, 2, 3$ and 4 observations per individual. D-optimal designs for the approximations $\mathbf{M}_{1,\beta}$ and $\mathbf{M}_{5,\beta}$ can alternatively be analytically constructed for individual sampling schemes with one observation per individual:

Theorem 7.1. *Let the information matrix for an individual design $\xi = (x)$ with $x \in [x_l, x_u] \subseteq \mathbb{R}$ be given by*

$$M(x) := \frac{1}{a_2 + b_2 x + c_2 x^2} \begin{pmatrix} x^2 a_1^2 & x a_1(b_1 + c_1 x) \\ x a_1(b_1 + c_1 x) & (b_1 + c_1 x)^2 \end{pmatrix}$$

and let the polynomial $v(x) := a_2 + b_2 x + c_2 x^2$ be positive for all $x \in [x_l, x_u]$. Then every population design

$$\zeta^* := \begin{pmatrix} (x_1) & (x_2) \\ 0.5 & 0.5 \end{pmatrix}$$

with support points x_1, x_2 in $[x_l, x_u]$ fulfilling

$$\kappa(x_1, x_2) := 2a_2 + b_2(x_1 + x_2) + 2c_2 x_1 x_2 = 0 \tag{7.2}$$

is D-optimal on Ξ^1.
If no such pair of supports points exists, the design ζ^* with $x_1 = x_l$ and $x_2 = x_u$ is D-optimal on Ξ^1.

Proof: Let $M_{pop}(\zeta)$ define the population information for a regular population design ζ and $\widetilde{g}_{\zeta;D}(x)$ define a modified sensitivity function:

$$\widetilde{g}_{\zeta;D}(x) := (a_2 + b_2 x + c_2 x^2)\big(\text{tr}\,[M_{pop}(\zeta)^{-1} M(x)] - 2\big) \parallel M_{pop}(\zeta) \parallel .$$

For D-optimal designs ζ and arbitrary individual designs $\xi = (x) \in [x_l, x_u]$ with information matrices $M(x)$ holds with theorem 6.14:

$$\text{tr } M_{pop}(\zeta)^{-1}M(x) \leq 2 \iff \widetilde{g}_{\zeta;D}(x) \leq 0.$$

Assume that the population designs consists of two different supporting individual designs (x_1) and (x_2). Then the optimal weight is 0.5 on each individual design. The modified sensitivity function results in

$$\widetilde{g}_{\zeta;D}(x) = \frac{a_1^2 b_1^2 \big(2a_2 + 2c_2 x_1 x_2 + b_2(x_2 + x_1)\big)(x - x_2)(x - x_1)}{2(a_2 + b_2 x_1 + c_2 x_1^2)(a_2 + b_2 x_1 + c_2 x_1^2)},$$

such that the population design with supporting individual designs (x_l) and (x_u) is D-optimal if

$$\kappa(x_l, x_u) := 2a_2 + 2c_2 x_l x_u + b_2(x_l + x_u) \geq 0.$$

The following equivalence holds with $v(x) := a + bx + cx^2 > 0$ for all $x \in [x_l, x_u]$ in the case of negative $\kappa(x_l, x_u)$:

$$\kappa(x_l, x_u) = 2a_2 + 2c_2 x_l x_u + b_2(x_l + x_u) < 0 \iff v(x_l) + v(x_u) - c(x_u - x_l)^2 < 0.$$

Further note that $\kappa(x_l, x_l) = 2v(x_l) > 0$ and that $\kappa(x_l, \cdot)$ is a continuous function in the second argument. Hence a point $x^* \in (x_l, x_u]$ exists with

$$\kappa(x_l, x^*) = 0 \Longrightarrow \tilde{g}_{\zeta^*;D}(x) = 0 \ \forall x \in [x_l, x_u]$$

for the population design ζ^* with weights 0.5 on individual designs (x_l) and (x^*). □

Optimal designs for the approximations $\mathbf{M}_{1,\beta}$ and $\mathbf{M}_{5,\beta}$ with one observation per individual can with the above theorem easily be constructed as the solutions of the equations

$$\begin{aligned} \sigma^2 + \sigma^2 F_{\gamma(\beta)}(x_1) D F_{\gamma(\beta)}(x_2)^T &= 0 \ \text{ for } \mathbf{M}_{1,\beta} \text{ and} \\ \sigma^2 + Cov(\widetilde{Y}_{i1}, \widetilde{Y}_{i2}) &= 0 \ \text{ for } \mathbf{M}_{5,\beta}. \end{aligned}$$

The Fisher information matrix was approximated in the example of one observation per individual based on 10000 simulated observations in 1000 different experimental settings on the design region $\mathcal{X} = [0.1, 24]$. Similarly the matrix $\mathbf{M}_{4,\beta}$ was approximated on the design region $\mathcal{X}$. The resulting dependences of the components of the information matrices on the experimental settings are illustrated in figure 7.2. The simulation results were used for a nonlinear least squares fit of the individual Fisher information matrix $\mathfrak{M}_{ind}^{\beta}$ and the matrix $\mathbf{M}_{4,\beta}$ by a matrix $M(x)$ of the form as in theorem 7.1. The Fisher information matrix is with a population design

$$\zeta = \begin{pmatrix} \xi_1 & \dots & \xi_k \\ \omega_1 & \dots & \omega_k \end{pmatrix}$$

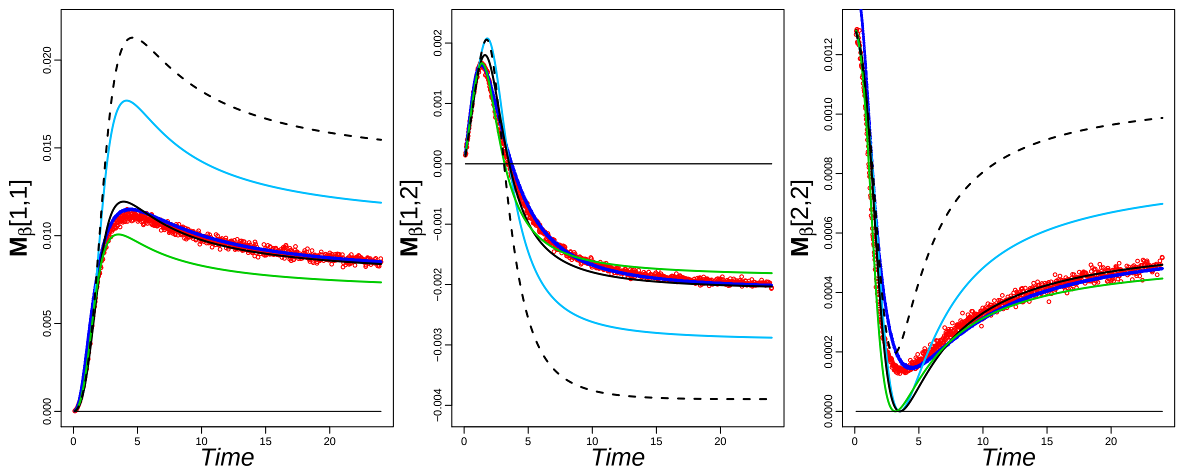

Figure 7.2: Components of the Fisher information in dependence on the time:
Solid: $\mathbf{M}_{1,\beta}$; Dashed: $\mathbf{M}_{2,\beta}$; Light-blue: $\mathbf{M}_{3,\beta}$; Dark-blue: $\mathbf{M}_{4,\beta}$; Green: $\mathbf{M}_{5,\beta}$; Red: $\mathfrak{M}^{\beta}_{ind}$.

for the location parameters β then given by the weighted sum of the individual information matrices:

$$\mathfrak{M}^{\beta}_{pop}(\zeta) = \sum_{i=1}^{k} \omega_i \mathfrak{M}^{\beta}_{ind}(\xi_i).$$

D- and $IMSE$-optimal designs were computed for the estimated dependence of the information matrices on the experimental settings. The resulting designs with one observation per individual and the D- and $IMSE$-efficiency in terms of the Fisher information:

$$\delta_{F;D}(\zeta) := \left(\frac{\| \mathfrak{M}^{\beta}_{pop}(\zeta) \|}{\| \mathfrak{M}^{\beta}_{pop}(\zeta^*) \|} \right)^{1/2}, \quad \delta_{F;IMSE}(\zeta) := \left(\frac{\text{tr } \mathfrak{M}^{\beta}_{pop}(\zeta^*)^{-1} L}{\text{tr } \mathfrak{M}^{\beta}_{pop}(\zeta)^{-1} L} \right)$$

are given in table 7.1. The designs in the table were derived numerically. Notice that the D-optimal designs for the approximations $\mathbf{M}_{1,\beta}$ and $\mathbf{M}_{5,\beta}$ are with theorem 7.1 not uniquely defined. The design

$$\widetilde{\zeta}_1^* := \begin{pmatrix} (0.10) & (\ 4.09) \\ 0.50 & 0.50 \end{pmatrix}$$

is in the class of D-optimal designs with respect to $\mathbf{M}_{1,\beta}$ the most efficient in terms of $\delta_{F;D}$:

$$\delta_{F;D}(\widetilde{\zeta}_1^*) = 0.9962.$$

Although the information matrices $\mathbf{M}_{2,\beta}$ and $\mathbf{M}_{3,\beta}$ poorly approximate the Fisher information in this example, the resulting designs are not inefficient. The similar dependence of the components of the information matrix on the experimental settings here causes similar designs and relatively high efficiencies. However, this example shows as well, that designs with the usual information approximations $\mathbf{M}_{1,\beta}$ and $\mathbf{M}_{2,\beta}$ should be handled with care. As all

Table 7.1: D- and $IMSE$-optimal Designs for the proposed information approximations

	$\zeta^*_{j;D}$	$\delta_{F;D}(\cdot)$	$\zeta^*_{j;IMSE}$	$\delta_{F;IMSE}(\cdot)$
$\mathbf{M}_{1,\beta}$	$\begin{pmatrix} (1.13) & (11.98) \\ 0.50 & 0.50 \end{pmatrix}$	0.9336	$\begin{pmatrix} (2.16) & (23.82) \\ 0.09 & 0.91 \end{pmatrix}$	0.9918
$\mathbf{M}_{2,\beta}$	$\begin{pmatrix} (0.96) & (\ 6.00) \\ 0.47 & 0.53 \end{pmatrix}$	0.9264	$\begin{pmatrix} (0.62) & (\ 9.35) \\ 0.08 & 0.92 \end{pmatrix}$	0.8953
$\mathbf{M}_{3,\beta}$	$\begin{pmatrix} (1.70) & (24.00) \\ 0.50 & 0.50 \end{pmatrix}$	0.9136	$\begin{pmatrix} (2.49) & (24.00) \\ 0.10 & 0.90 \end{pmatrix}$	0.9894
$\mathbf{M}_{4,\beta}$	$\begin{pmatrix} (0.10) & (\ 2.70) & (24.00) \\ 0.35 & 0.46 & 0.19 \end{pmatrix}$	0.9871	$\begin{pmatrix} (0.10) & (2.70) & (24.00) \\ 0.03 & 0.09 & 0.88 \end{pmatrix}$	0.9974
$\mathbf{M}_{5,\beta}$	$\begin{pmatrix} (0.93) & (11.99) \\ 0.50 & 0.50 \end{pmatrix}$	0.9409	$\begin{pmatrix} (1.98) & (23.99) \\ 0.10 & 0.90 \end{pmatrix}$	0.9902
$\mathfrak{M}^{\beta}_{ind}$	$\begin{pmatrix} (0.10) & (\ 4.10) \\ 0.45 & 0.55 \end{pmatrix}$	1.0000	$\begin{pmatrix} (0.10) & (4.14) & (24.00) \\ 0.04 & 0.11 & 0.85 \end{pmatrix}$	1.0000

proposed designs are locally optimal designs and based on approximations of the Fisher information matrix, efficiency in the proposed approximations not necessarily implies efficiency with respect to the true Fisher information. Possible problems regarding the computation of the efficiency of approximations of the matrices $\mathbf{M}_{4,\beta}$ and $\mathfrak{M}^{\beta}_{ind}$ will be met in the following example on designs in compartment models with absorption.

7.3 One Compartment With First Order Absorption

In the one compartment model with first order absorption, the j-th observation of the log-concentration of the i-th individual is modeled by

$$\begin{aligned} \widetilde{Y}_{ij} &= \widetilde{\eta}_2(\widetilde{\beta}_i, x_{ij}) + \epsilon_{ij} \\ &= \log(\mathbf{D}) + \log\left(\frac{\exp(-x\exp[\widetilde{\beta}_{i;2} - \widetilde{\beta}_{i;3}]) - \exp(-x\exp[\widetilde{\beta}_{i;1}])}{\exp(\widetilde{\beta}_{i;3}) - \exp(\widetilde{\beta}_{i;2} - \widetilde{\beta}_{i;1})}\right) + \epsilon_{ij}. \end{aligned}$$

with normally distributed random variables

$$\widetilde{\beta}_i \sim \mathcal{N}\big((\log[\beta_1], \log[\beta_2], \log[\beta_3]), \sigma^2 diag(d_1, d_2, d_3)\big) \text{ and } \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2).$$

Mean and covariance of observations within one individual cannot be represented in a closed form, what complicates the derivation of the approximations $\mathbf{M}_{3,\beta}$ and $\mathbf{M}_{5,\beta}$. The Fisher information matrix $\mathfrak{M}^{\beta}_{ind}$ can in the case of one observation per individual again be approximated using simulations. Simulations are also applied for computing the dependence of the matrices $\mathbf{M}_{3,\beta}$, $\mathbf{M}_{4,\beta}$ and $\mathbf{M}_{5,\beta}$ on the experimental settings. The results for the components of the Fisher information in this model are illustrated in figure 7.3 for the numbers given in Schmelter (2007a, exp. 8.7):

$$\beta_1 = 0.61,\ \beta_2 = 25,\ \beta_3 = 88,\ \sigma^2 = 0.01,\ d_1 = 89.3,\ d_2 = 12.5,\ d_3 = 9.0.$$

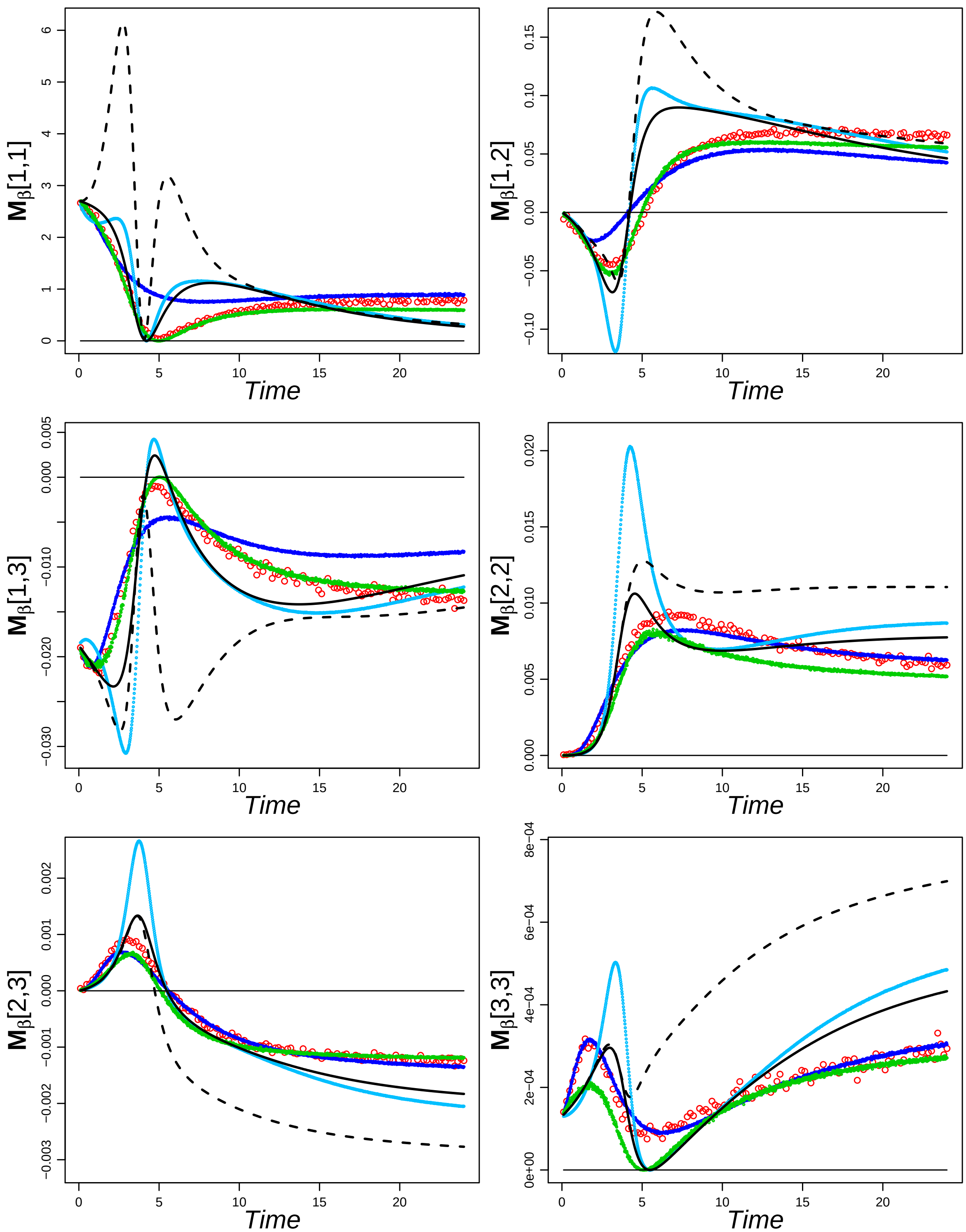

Figure 7.3: Components of the Fisher information in dependence on the time:
Solid: $\mathbf{M}_{1,\beta}$; Dashed: $\mathbf{M}_{2,\beta}$; Light-blue: $\mathbf{M}_{3,\beta}$; Dark-blue: $\mathbf{M}_{4,\beta}$; Green: $\mathbf{M}_{5,\beta}$; Red: $\mathfrak{M}^{\beta}_{ind}$.

Table 7.2: D- and $IMSE$-optimal Designs for the proposed information approximations

	$\zeta^*_{j;D}$	$\zeta^*_{j;IMSE}$
$\mathbf{M}_{1,\beta}$	$\begin{pmatrix} (0.10) & (\ 4.18) & (24.00) \\ 0.33 & 0.33 & 0.33 \end{pmatrix}$	$\begin{pmatrix} (0.10) & (\ 5.12) & (24.00) \\ 0.16 & 0.16 & 0.68 \end{pmatrix}$
$\mathbf{M}_{2,\beta}$	$\begin{pmatrix} (3.10) & (\ 5.18) & (24.00) \\ 0.61 & 0.09 & 0.30 \end{pmatrix}$	$\begin{pmatrix} (2.98) & (\ 5.39) & (24.00) \\ 0.08 & 0.25 & 0.67 \end{pmatrix}$
$\mathbf{M}_{3,\beta}$	$\begin{pmatrix} (0.10) & (\ 3.28) & (\ 4.47) & (24.00) \\ 0.21 & 0.20 & 0.28 & 0.31 \end{pmatrix}$	$\begin{pmatrix} (0.10) & (\ 4.80) & (24.00) \\ 0.19 & 0.13 & 0.68 \end{pmatrix}$
$\mathbf{M}_{4,\beta}$	$\begin{pmatrix} (2.85) & (24.00) \\ 0.41 & 0.59 \end{pmatrix}$	$\begin{pmatrix} (3.26) & (24.00) \\ 0.12 & 0.88 \end{pmatrix}$
$\mathbf{M}_{5,\beta}$	$\begin{pmatrix} (0.10) & (\ 4.47) & (22.20) \\ 0.32 & 0.35 & 0.33 \end{pmatrix}$	$\begin{pmatrix} (0.10) & (\ 4.73) & (22.02) \\ 0.16 & 0.16 & 0.68 \end{pmatrix}$
$\mathfrak{M}^{\beta}_{ind}$	$\begin{pmatrix} (2.32) & (\ 6.40) & (24.00) \\ 0.61 & 0.01 & 0.38 \end{pmatrix}$	$\begin{pmatrix} (2.24) & (24.00) \\ 0.17 & 0.83 \end{pmatrix}$

Continuously differentiable quadratic regression splines were fitted to the observed data in order to approximately describe the dependence of the simulated information matrices on the experimental settings. With the restriction on continuous approximations of the Fisher information matrix, the design theory of the preceding chapter is applicable for the derivation of D- and $IMSE$-optimal design with one observation per individual. The resulting designs are given in table 7.2. Table 7.3 illustrates the D- and $IMSE$-efficiencies of the proposed designs, compared to each other with the definitions:

$$\delta_{j;D}(\zeta) := \left(\frac{\| \mathbf{M}_{j;pop}(\zeta) \|}{\| \mathbf{M}_{j;pop}(\zeta^*_j) \|} \right)^{1/3},\ j = 1, ..., 5\ ,\ \delta_{F;D}(\zeta) := \left(\frac{\| \mathfrak{M}^{\beta}_{pop}(\zeta) \|}{\| \mathfrak{M}^{\beta}_{pop}(\zeta^*) \|} \right)^{1/3} \quad \text{and}$$

$$\delta_{j;IMSE}(\zeta) := \left(\frac{\text{tr } \mathbf{M}_{j;pop}(\zeta^*)^{-1} L}{\text{tr } \mathbf{M}_{j;pop}(\zeta)^{-1} L} \right),\ j = 1, ..., 5\ , \quad \delta_{F;IMSE}(\zeta) := \left(\frac{\text{tr } \mathfrak{M}^{\beta}_{pop}(\zeta^*)^{-1} L}{\text{tr } \mathfrak{M}^{\beta}_{pop}(\zeta)^{-1} L} \right).$$

Of special interest are here the designs with respect to the approximation $\mathbf{M}_{4,\beta}$. The high efficiency of optimal designs for the approximation $\mathbf{M}_{4,\beta}$ in terms of the simulation based Fisher information motivates the use of this approximation for the construction of optimal designs, as already in the preceding section. One disadvantage is however given by the small number of supporting individual designs, what leads to zero efficiency in terms of the information approximation $\mathbf{M}_{1,\beta}$. Although the simulated Fisher information matrix for the proposed design is positive definite, the estimation cannot be done straightforwardly with the in 4.2 presented estimators. Another problem is apparent for the spline-approximations. The resulting information matrices are with the proposed approximations not guaranteed to be non-negative definite. Negative efficiencies were in the table 7.3 set to zero and highlighted with an asterisk. All information approximations theoretically provide non-negative definite matrices, such that the observed failures are based on the simulations and spline approximations. Note specially, that the true efficiency of the designs $\zeta^*_{4;\cdot}$ with respect to the criteria

Table 7.3: D- and $IMSE$-efficiencies of the proposed designs

	$\zeta^*_{1;D}$	$\zeta^*_{2;D}$	$\zeta^*_{3;D}$	$\zeta^*_{4;D}$	$\zeta^*_{5;D}$	$\zeta^*_{F;D}$
$\delta_{1;D}$	1.00	0.66	0.96	0.00	0.97	0.37
$\delta_{2;D}$	0.55	1.00	0.75	0.95	0.59	0.77
$\delta_{3;D}$	0.96	0.79	1.00	0.35	0.94	0.00*
$\delta_{4;D}$	0.84	0.98	0.91	1.00	0.83	0.98
$\delta_{5;D}$	0.98	0.00*	0.90	0.00*	1.00	0.53
$\delta_{F;D}$	0.83	0.88	0.87	0.95	0.82	1.00

	$\zeta^*_{1;IMSE}$	$\zeta^*_{2;IMSE}$	$\zeta^*_{3;IMSE}$	$\zeta^*_{4;IMSE}$	$\zeta^*_{5;IMSE}$	$\zeta^*_{F;IMSE}$
$\delta_{1;IMSE}$	1.00	0.83	0.99	0.00	0.97	0.37
$\delta_{2;IMSE}$	0.85	1.00	0.79	0.68	0.83	0.59
$\delta_{3;IMSE}$	0.99	0.90	1.00	0.03	0.97	0.00*
$\delta_{4;IMSE}$	0.88	0.91	0.87	1.00	0.87	0.98
$\delta_{5;IMSE}$	0.95	0.13	0.95	0.00*	1.00	0.00*
$\delta_{F;IMSE}$	0.91	0.87	0.90	0.95	0.86	1.00

of the approximations $\mathbf{M}_{1,\beta}$, $\mathbf{M}_{3,\beta}$ and $\mathbf{M}_{5,\beta}$ is equal to zero, as these approximations are defined as products of matrices with rank one.

The approximations $\mathbf{M}_{1,\beta}$ and $\mathbf{M}_{2,\beta}$ are usually applied for optimizing experimental designs in the literature. Although the supporting individual designs and optimal weights for the approximations $\mathbf{M}_{1,\beta}$ and $\mathbf{M}_{2,\beta}$ differ, the comparison of the presented results shows no big difference in the efficiency with respect to the Fisher information.

Similar considerations might be undertaken for bigger individual sampling schemes $m > 1$. The computational effort for accurate approximations of the dependence of the Fisher information matrix on the experimental settings makes the problem for growing individual sample sizes m however intractable. The computation of the matrix $\mathbf{M}_{4,\beta}$ is less complex than the computation of the Fisher information matrix, but the approximation of the dependence of $\mathbf{M}_{4,\beta}$ on the experimental settings is for big individual sample sizes also not readily possible. The problematic term in the derivation of the information approximation $\mathbf{M}_{3,\beta}$ and the quasi information matrix $\mathbf{M}_{5,\beta}$ is given by the covariance matrix of the observation vector. Even for individual sample sizes $m > 3$ only the covariance of pairs of observations has to be approximated for gaining an insight in the dependence of the matrix $\mathbf{M}_{3,\beta}$ on the experimental settings. For the quasi-information $\mathbf{M}_{5,\beta}$ additionally the derivative of the expectation has to be calculated, such that both approximations might be applied for optimizing designs.

The big benefit of the approximations $\mathbf{M}_{1,\beta}$ and $\mathbf{M}_{2,\beta}$ remains their closed from dependence on the experimental settings, which readily allows the computation of optimal designs. The most efficient designs in terms of the simulated Fisher information were however obtained with the novel approximation $\mathbf{M}_{4,\beta}$. Both examples show that efficiencies of derived optimal designs can and should be compared with respect to an evaluated Fisher information under the according experimental settings, in order to gain more insight in the quality of the proposed designs.

8 Discussion and Outlook

Locally optimal experimental designs for the estimation of population parameters in models with nonlinear mixed effects were introduced and discussed within this thesis. The presented results were however bound to some prior knowledge of the parameters of interest, such that derived design might be inefficient, when the parameters are not correctly specified in the planning stage of the experiment. Different approaches on optimal designs of experiments are briefly described in this chapter before a short summary and discussion of the derived results closes this thesis.

8.1 Further Considerations

Aim of this thesis was the construction of optimal designs in nonlinear mixed effects models. Different approximations of the Fisher information matrix were proposed in order to approach this aim. These approximations were mainly based on the assumptions of normally distributed random effects with known variances and some prior knowledge of the true location parameter β. All these assumptions are in practice not necessarily fulfilled and should be taken into account when designing experiments.

Information for the Variance parameters

The design optimization in mixed effects models was for known variance parameters based on $p \times p$ information matrices with the dimension p of the parameter vector β. In chapter 6 it was stated that the whole information matrix has to be taken into account for planning experiments in the case of unknown variance parameters. We here approximated the Fisher information for the variance parameters by the information matrix resulting from linear mixed effects models. The construction of optimal designs for the estimation of aspects of the parameter vector β then simplifies with the block-diagonality of the given approximation to the design problem with respect to the reduced information matrix of the case with known variance parameters. Although estimators for $\theta = (\beta^T, \sigma^2, \alpha^T)^T$ with the desired block-diagonal structure of the information matrix might exist, further work is needed in order to gain more insight in the Fisher information for the variance parameters.

One particular approximation of the Fisher information matrix without this block-diagonal structure is given by the nonlinear heteroscedastic normal model approximation as presented in section 4.3.2. Optimal designs for the estimation of β in the case of unknown variance parameters then depend on the whole information matrix. The D_A-optimal design with $A^T = (I_p, 0)$ for the estimation of β in the case of unknown variance parameters results in the example of the one-compartment model without absorption in chapter 7 in:

$$\zeta^*_{D_A;2} := \begin{pmatrix} (0.10) & (2.30) & (24.00) \\ 0.33 & 0.36 & 0.31 \end{pmatrix}$$

and hence differs from the optimal design in the case of known variance parameters:

$$\zeta^*_{D;2} := \begin{pmatrix} (0.96) & (6.00) \\ 0.47 & 0.53 \end{pmatrix}.$$

The efficiencies of both designs are calculated as 0.93 in comparison to each other. Further studies and approximations on the off-diagonal elements of the Fisher information are for the construction of optimal designs in nonlinear mixed effects models with unknown variance parameters of big interest.

Local Optimality

One problem for the here studied locally optimal designs of experiments in nonlinear models is the dependence of the information matrices on the prior unknown parameters θ. Bayesian and minimax approaches might alternatively be applied in order to compute designs, which are more robust with respect to parameter misspecifications. The uncertainty on the true value of the parameter θ is modeled in Bayesian optimality criteria with the help of probability distributions (Pronzato and Walter (1985)). Schmelter (2007a) briefly describes different possible optimality criteria for the construction of Bayesian optimal designs. Minimax approaches are applied in order to minimize the maximum value of a criterion over a set of possible parameter values. The main problem for both approaches remains the missing closed form representation of the Fisher information. For the construction of locally optimal designs the functional dependence of the Fisher information on the design settings was of interest. Insight in the dependence of the information matrices on the parameter values is additionally needed for optimally designing experiments with Bayesian and minimax approaches. This closed form relationship is in nonlinear mixed effects models generally only given by the information approximations $\mathbf{M}_{1,\beta}$ and $\mathbf{M}_{2,\beta}$.

Often some prior knowledge on the true underlying distribution of the individual parameters is present in clinical trials. A third alternative for reducing the loss of efficiency in the case of misspecified parameter values is given by adaptively designed studies. The study designs are then stage-wise applied and optimized, making it possible to use estimation results of earlier study stages for the design of recent stages. Problems for adaptive designs in nonlinear mixed effects models are given by the non-satisfactory knowledge of the distribution of estimates in the case of bounded samples sizes and by the individual wise varying parameter vectors, which imply correlated observations within each individual. Examinations of the convergence behavior of estimators in stage-wise planned designs are necessary for the study of adaptive designs in nonlinear mixed effects models.

Locally optimal designs propose the observation of different individuals under certain specified experimental settings. It is practically not possible to take samples exactly under the designed settings and it will be generally not possible to observe all individuals the same number of times. Pronzato (2002) includes random effects in the design settings in order to address the uncertainty on the exact experimental settings, whereas Bogacka *et al.* (2008) propose the construction of sampling windows, which are based on the sensitivity function of D-optimal designs. Other approaches on sampling windows are based on a prior defined efficiency which is to be achieved with respect to locally optimal designs (Duffull *et al.* (2001)). Optimal sampling windows for nonlinear mixed effects models can be constructed with the proposed information approximations $\mathbf{M}_{1,\beta}$ and $\mathbf{M}_{2,\beta}$. Considerations for the other here presented information approximations and studies on the appropriateness of different sampling window approaches should be undertaken in order to design studies more robust with respect to the realized sampling times.

Further Approximations

The problem of the missing closed form representation of the Fisher information for the design of experiments in nonlinear mixed effects models was in this work circumvented by the use of different approximations of the Fisher information matrix. The computed designs for different approximations were relatively efficient, although no proposed information matrix approximated the Fisher information matrix entirely well. The presented approximations were mainly based on the normality of the random effects, such that further considerations on the quality of approximations in models with other than normal distributions would be of interest. Insight in the quasi-information $\mathbf{M}_{5,\beta}$ is for general models of the observation vector Y desired, as it is not based on distributional assumptions, but on the first two moments of the observation vector Y, making it more robust with respect to distribution misspecifications. Another problem was given in chapter 7.3 by the missing closed form representations of the information approximations $\mathbf{M}_{3,\beta}$, $\mathbf{M}_{4,\beta}$ and $\mathbf{M}_{5,\beta}$. Reliable analytical approximations of these matrices should be derived and examined for the construction of locally and Bayesian optimal designs with respect to these approximations.

8.2 Summary and Discussion

The foundations of the optimal design theory in mixed effects models were built in the second chapter with the introduction on optimal experimental designs in ordinary regression models. This topic is well developed in the literature and generalizations of the design theory from ordinary regression models to linear mixed effects models readily follow with the considerations in Schmelter (2007a) and the equivalence theorems of Fedorov (1972, ch. 5). The linear mixed effects model results as a linear model with a special covariance structure, making it possible to apply the same estimation theory as in ordinary linear models. A similar generalization of nonlinear models to nonlinear mixed effects models unfortunately does not hold. The properties of estimators in nonlinear mixed effects models with bounded individual sample sizes are not sufficiently known, such that the Fisher information is of interest for optimally planning experiments, which however cannot be represented in a closed form in the examined models. Linearizations of the model are in the literature usually applied in order to transform the complicated nonlinear mixed effects models, to models of a simpler structure, such as linear mixed effects (Schmelter (2007a)) or nonlinear normal models with heteroscedastic errors (Retout and Mentré (2003)). The Fisher information in the resulting models is then used for the construction of optimal designs. Examples in section 4.3 and in the publication by Mielke and Schwabe (2010) illustrated the difference of in the literature proposed approximations and motivated further studies on the behavior of the Fisher information matrix in nonlinear mixed effects models.

The representation of the Fisher information as the covariance of the conditional expectation for given observations implied in chapter 5 novel approximations of the Fisher information matrix in nonlinear mixed effects models. Five different approximations of the Fisher information matrix in nonlinear mixed effects models with individual observation vectors:

$$Y_i = \eta(\beta_i, \xi_i) + \epsilon_i,$$

exact individual designs $\xi_i = (x_{i1}, ..., x_{im_i}) \in \mathcal{X}^{m_i}$ and normally distributed random effects

$$\beta_i \sim \mathcal{N}(\beta, \sigma^2 D),\ \epsilon_i \sim \mathcal{N}(0, \sigma^2 I_{m_i}),$$

were examined in the following. The design matrix F_β and the linear approximation of the covariance matrix of the observation vector Y_i were defined by

$$F_\beta := \frac{\partial \eta(\beta_i, \xi_i)}{\partial \beta_i^T}|_{\beta_i=\beta} \quad \text{and} \quad V_\beta := I_{m_i} + F_\beta D F_\beta^T.$$

An example on the sum of a log-normally and a normally distributed random variable provided some insight in the accuracy of the proposed approximations in dependence on the variance parameters in section 5.3. The generalization of the design theory in chapter 6 to mixed effects models allowed the application of the convex design theory for the introduced approximations in two pharmacokinetic models in the seventh chapter. All approximations implied in these examples relatively efficient designs when compared to a simulation based Fisher information matrix. Differences where however given by the specification of the dependence of the approximations on the experimental settings and model parameters, what carries forward to the applicability of the approximations in studies with more observations.

Approximation $\mathbf{M}_{1,\beta}$

The approximation $\mathbf{M}_{1,\beta}$ is probably the most frequently used approximation in the literature on designs of experiments in nonlinear mixed effects models. The matrix

$$\mathbf{M}_{1,\beta} := \frac{1}{\sigma^2} F_\beta^T V_\beta^{-1} F_\beta$$

results either as the Fisher information matrix of an approximating linear mixed effects model or as an approximation of the Fisher information via an approximation of the expectation of the conditional variance of individual parameter vectors for given observations. The big benefit of this approximation is the known dependence of the matrix $\mathbf{M}_{1,\beta}$ on the experimental settings and the parameter vector θ, such that the convex design theory can be readily applied for the design construction. In section 5.3 it could be seen that the accuracy of the approximation might be not entirely satisfying in dependence on the inter-individual variance matrix. Therefor, the Fisher information was not overestimated and in dependence on the value of the location parameter vector reasonable approximations were obtained. Small problems occurred in an example on a singular inter-individual variance matrix in section 5.4, where the Fisher information matrix was correctly specified by the approximation $\mathbf{M}_{2,\beta}$. The limiting behavior of $\mathbf{M}_{1,\beta}$ for big inter-individual variances coincides for positive definite inter-individual variance matrices with the behavior of the Fisher information, what was evident with the construction of an upper bound of the Fisher information matrix.

Approximation $\mathbf{M}_{2,\beta}$

The approximation $\mathbf{M}_{2,\beta}$ is based on a linearization of the model function and the assumptions of negligible linearization errors, which shall imply the normality of the observation vector:

$$\begin{aligned}\mathbf{M}_{2,\beta}(\xi_i) &:= \mathbf{M}_{1,\beta}(\xi_i) + \frac{1}{2}S(\xi_i) \text{ with} \\ (S(\xi_i))_{j,k} &= \text{tr}\left[V_{\gamma(\beta)}(\xi_i)^{-1}\frac{\partial V_{\gamma(\beta)}(\xi_i)}{\partial\beta_j}V_{\gamma(\beta)}(\xi_i)^{-1}\frac{\partial V_{\gamma(\beta)}(\xi_i)}{\partial\beta_k}\right],\ j,k = 1,...,p.\end{aligned}$$

The limiting behavior of the approximation does generally not coincide with that of the Fisher information matrix. An example by Mielke and Schwabe (2010) and the example in section 5.3 illustrated this problem, which carries forward to the overestimation of the Fisher information, as illustrated in the plots of the dependence of the Fisher information on the experimental settings in chapter 7. Note that the additional matrix term $\frac{1}{2}S$ is a consequence of the assumed normality of the observation vector Y_i and the dependence of the linearized variance matrix V_β on the population parameter vector β. Two misspecifications of the model are hence committed in order to increase the information by the term $\frac{1}{2}S$. The information approximation $\mathbf{M}_{2,\beta}$ is nevertheless of big interest for design purposes, as the dependence on the experimental settings and parameters is given in a closed form, what makes the design theory readily applicable.

Approximation $\mathbf{M}_{3,\beta}$

The approximation $\mathbf{M}_{3,\beta}$ was motivated as an approximation of the variance of the conditional expectation of the individual parameters for given observations:

$$\mathbf{M}_{3,\beta} := \frac{1}{\sigma^4}F_\beta^T V_\beta^{-1} Cov(Y_i) V_\beta^{-1} F_\beta.$$

The results for the approximation $\mathbf{M}_{3,\beta}$ in the presented example of section 5.3 where not satisfactory. The divergence of the approximation $\mathbf{M}_{3,\beta}$ in dependence on the inter-individual variance caused a complete misspecification of the asymptotic behavior of the true Fisher information. Further inaccuracies of the approximation were obvious in the examples on pharmacokinetic models, what however did not lead to inefficient designs in these examples.

An additional problem is given by the missing closed form representation of the matrix $\mathbf{M}_{3,\beta}$ in the case of a one compartment model with absorption. The approximations of the matrix by regression splines did not solve this problem entirely. Although designs could be constructed with the so calculated dependence of the information on the experimental settings, new problems occurred as the resulting information matrices were not necessarily non-negative definite. For the construction of optimal designs in nonlinear mixed effects models we hence do not recommend to use only the matrix $\mathbf{M}_{3,\beta}$.

Approximation $\mathbf{M}_{4,\beta}$

Two possible motivations were given for the approximation $\mathbf{M}_{4,\beta}$:

$$\mathbf{M}_{4,\beta} \quad := \quad \frac{1}{\sigma^2} E(F_{\beta_i}^T V_{\beta_i}^{-1} F_{\beta_i}).$$

The matrix $\mathbf{M}_{4,\beta}$ follows as an approximation of the expectation of the conditional variance of individual parameter vectors for given observations. Otherwise it was mentioned, that the matrix $\mathbf{M}_{4,\beta}$ behaves similar to the theoretical information matrix of two-stage and Lindstrom and Bates estimators in nonlinear mixed effects models. The approximation $\mathbf{M}_{4,\beta}$ modeled the behavior of the Fisher information matrix in dependence on the variance and the experimental settings in all presented examples relatively well. Two problems were however evident when working with the matrix $\mathbf{M}_{4,\beta}$. The Fisher information was overestimated in some cases and no closed form representation of the dependence of the approximation on the experimental settings or the parameter vector exists, what complicates the constructions of optimal designs. Optimal designs were computed in the examples on pharmacokinetic models with the help of simulation based approximations of the matrix $\mathbf{M}_{4,\beta}$ and resulted in both examples as the most efficient designs in terms of the computed Fisher information matrix. Despite this efficiency with respect to the Fisher information, further problems occurred with respect to the information matrices $\mathbf{M}_{1,\beta}$, $\mathbf{M}_{3,\beta}$ and $\mathbf{M}_{5,\beta}$ in the example of a one compartment model with absorption. The number of different supporting individual designs of the optimal design with respect to $\mathbf{M}_{4,\beta}$ fell below the number of parameters to be estimated, such that the estimation might be problematic with these sparse sampling schemes. Further note that the approximation of the dependence of the matrix $\mathbf{M}_{4,\beta}$ on the experimental settings is for big individual sample sizes not easily done. Numerical algorithms for the computation of optimal designs can however circumvent this problem of the missing complete knowledge of the functional dependence of $\mathbf{M}_{4,\beta}$ on the design settings.

Approximation $\mathbf{M}_{5,\beta}$

The quasi-information was briefly described in section 5.3 and motivated the approximation

$$\mathbf{M}_{5,\beta} \quad := \quad \frac{\partial E(Y_i)^T}{\partial \beta} Cov(Y_i)^{-1} \frac{\partial E(Y_i)}{\partial \beta^T}.$$

The theory on quasi-likelihood estimation in nonlinear mixed effects models is not well developed and further studies are needed in order to verify the asymptotic theory on quasi-likelihood estimators in nonlinear mixed effects models. The quasi-likelihood estimation goes here back to an iteratively reweighted nonlinear least squares problem for a vector of observations $Y = (Y_1^T, ..., Y_N^T)^T$ with realizations $y = (y_1^T, ..., y_N^T)^T$:

$$L_{QL}(\beta; y) := \sum_{i=1}^{N} (y_i - E_\beta(Y_i))^T Cov_{\beta_0}(Y_i)^{-1}(y_i - E_\beta(Y_i)) \to \min_{\beta \in \mathbb{R}^p},$$

where β_0 is chosen as the minimizing argument $\widehat{\beta}$. The quasi-information $\mathbf{M}_{5,\beta}$ serves with the asymptotic results on nonlinear least squares estimators and the unbiasedness of the estimating equation as a lower bound of the Fisher information matrix and is hence for the construction of optimal designs of interest. The generally missing closed form representation of the information matrix $\mathbf{M}_{5,\beta}$ as a function of the experimental settings is here not as problematic as for the matrix $\mathbf{M}_{4,\beta}$. Only the covariance function of two observations and the derivative of the mean with respect to the individual parameter vectors have to be approximated in order to specify a dependence of the matrix $\mathbf{M}_{5,\beta}$ on arbitrary experimental designs $\xi \in \mathcal{X}^m$. The quasi-information provided in the presented examples reasonable approximations of the simulation based Fisher information matrix. Moreover, the quasi-information is of big interest for planning experimental designs in nonlinear mixed effects models, as it is not based on the normality of the random effects and might be hence more robust with respect to distribution misspecifications.

The Fisher information $\mathfrak{M}^\beta$

The examples in section 5.3 and chapter 7 illustrated the influence of the variance parameters and the experimental settings on the Fisher information matrix. Optimal designs could be constructed for samples sizes of one observation per individual with the help of simulations in chapter 7. The results might theoretically be generalized to bigger individual sample sizes, what is practically relatively intractable. The computational burden for the evaluation of the Fisher information and approximation of its dependence on the experimental settings grows quickly with the individual sample size, such that optimal designs with respect to the true Fisher information remain an unsolved area of interest for future studies. However, the Fisher information can be used as a benchmark for the computation of efficient designs in nonlinear mixed effects models. Optimal designs for different information approximations should be compared with respect to the computed Fisher information matrix in order to determine efficient sampling schemes.

Although no overall optimal approximation was presented in this thesis, new motivations for approximations and more insight in the Fisher information in nonlinear mixed effects models were given. All approximations provided in the illustrative examples relatively efficient designs. Specially the resulting approximations and designs with respect to the novel approximation $\mathbf{M}_{4,\beta}$ show great promise for the computation of experimental designs in nonlinear mixed effects models. We recommend the construction of optimal designs with the here presented approximations and the comparison of resulting candidate designs with respect to a common criteria.

Nomenclature

$\mathcal{X}$	design region	5
$x;\ x_j;\ x_{ij}$	elements of the design region	5
Ξ^m	set of approximate designs on $\mathcal{X}^m$	64
$m_i;\ m$	individual sample size	6
m_T	total sample size	21
N	number of individuals	17
ξ_i	individual design i	23
ζ	population design	24
δ_ξ	one-point measure on $\xi \in \mathcal{X}^m$	10
ω_i	weight of design settings i in approximate designs	10
$\epsilon;\ \epsilon_i$	observation errors	5
Y	vector of all observations	6
Y_i	vector of observations in the i-th individual	18
Y_{ij}	j-th observation in the i-th individual	17
$E(\cdot);\ E(\cdot\|*)$	expectation of $\cdot$ (conditional on given $*$)	6
$Cov(\cdot);\ Cov(\cdot\|*)$	covariance matrix of $\cdot$ (conditional on given $*$)	6
η	response function	5
$\tilde{l}(\beta_i, \theta; y_i)$	penalized sum of squares	30
$L.(\cdot; y)$	least squares objective functions	6
$L(\theta; y);\ l(\theta; y)$	(log)-likelihood function	8
$F_{\beta_0}(\xi)$	design matrix evaluated at design ξ and $\beta_i = \beta_0$	7
I_m	m-dimensional identity matrix	6
$V_\beta := I_m + F_\beta^T D F_\beta$	covariance matrix in the linearized model	18
β	(population) location parameter vector	18

β_i	location parameter vector of i-th individual	18
$\beta_{i;l}$	l-th component of the vector β_i	18
b_i	individual random effects	18
$\widehat{\beta}.$	estimator for β	6
$\widehat{\beta}_{\cdot,i}$; $\widehat{\beta}_i$	estimator for β_i	20
β_i^*	penalized least squares estimate for β_i	34
σ^2	variance of observation errors	5
$\sigma^2 D$	variance of individual parameter vectors	18
α	ν-dimensional parameterization of D	21
$\theta = (\beta^T, \sigma^2, \alpha^T)$	whole parameter vector	21
$\widehat{\sigma}^2$; $\widehat{\sigma^2 D}$; $\widehat{\alpha}$	estimates of variance parameters	21
ψ; $\widehat{\psi}$	linear aspect of β and estimate	7
$\Phi.$	optimality criterion	11
$F_{\Phi.}(\cdot,\cdot)$	Fréchet derivative	14
$g.(\cdot)$	sensitivity function	15
ϕ	density of the normal distribution	30
$a \overset{app.}{\sim} N(\cdot,\cdot)$	distribution of a is approximated by a normal distribution	33
$\mathfrak{M}$	Fisher information matrix	8
$\mathfrak{M}_{pop}$	normalized population Fisher information matrix	25
$\mathfrak{M}_{ind}$	individual Fisher information matrix	25
$\mathfrak{M}_{ind}^{\beta}$	individual Fisher information matrix for the parameter vector β	42
$M_{\widehat{\beta}_i}$	approximation of the conditional variance	49
$\mathbf{M}$	not necessarily closer specified information matrices	10
$\mathbf{M}_{1,\beta}$	approximation of the individual Fisher information for β	36
$\mathbf{M}_{2,\beta}$	approximation of the individual Fisher information for β	37
$\mathbf{M}_{3,\beta}$	approximation of the individual Fisher information for β	51
$\mathbf{M}_{4,\beta}$	approximation of the individual Fisher information for β	52
$\mathbf{M}_{5,\beta}$	approximation of the individual Fisher information for β	55

Bibliography

Atkinson, A. and Bailey, R. (2001). One hundred years of the design of experiments on and off the pages of Biometrika. *Biometrika*, **88**, 53–97.

Atkinson, A. and Cook, R. (1995). D-Optimum designs for heteroscedastic linear models. *Journal of the American Statistical Association*, **90**, 204–212.

Bandemer, H. and Bellmann, A. (1994). *Statistische Versuchsplanung.* Teubner, Leipzig.

Bazzoli, C., Retout, S., and Mentré, F. (2009). Fisher information matrix for nonlinear mixed effects multiple response models: Evaluation of the appropriateness of the first order linearization using a pharmacokinetic/pharmacodynamic model. *Statistics in Medicine*, **28**, 1940–1956.

Beal, S. and Sheiner, L. (1998). NONMEM Users Guide - Part VII: Conditional Estimation Methods. NONMEM Project Group, University of California, San Francisco.

Bogacka, B., Johnson, P., Jones, B., and Volkov, O. (2008). D-efficient window experimental designs. *Journal of Statistical Planning and Inference*, **138**, 160–168.

Cheng, C. (1995). Optimal regression designs under random block effects. *Statistica Sinica*, **5**, 485–497.

Davidian, M. and Carroll, R. (1987). Variance function estimation. *Journal of the American Statistical Association*, **82**, 1079–1091.

Davidian, M. and Giltinan, D. (1995). *Nonlinear Models for Repeated Measurement Data.* Chapman & Hall, London.

Demidenko, E. (2005). *Mixed Effects Models: Theory and Applications.* Wiley, New Jersey.

Duffull, S., Mentré, F., and Aarons, L. (2001). Optimal design of a population pharmacodynamic experiment for Ivabradine. *Pharmaceutical Research*, **18**, 83–89.

Elfving, G. (1952). Optimum allocation in linear regression theory. *The Annals of Mathematical Statistics*, **23**, 225–262.

EMEA (2006). 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.

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

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

Gaffke, N. and Heiligers, B. (1996). Second order methods for solving extremum problems from optimal linear regression design. *Optimization*, **36**, 41–57.

Gladitz, J. and Pilz, J. (1982). Construction of optimal designs in random coefficient regression models. *Mathematische Operationsforschung und Statistik*, **13**, 371–385.

Graßhoff, U., Doebler, A., Holling, H., and Schwabe, R. (2012). Optimal design for linear regression models in the presence of heteroscedasticity caused by random coefficients. *Journal of Statistical Planning and Inference*, **142**, 1108–1113.

Großmann, C. and Terno, J. (1997). *Numerik der Optimierung.* Teubner, Stuttgart.

Holland-Letz, T. (2009). *Bestimmung c-Optimaler Versuchspläne in Modellen mit Zufälligen Effekten, mit Anwendungen in der Pharmakokinetik.* Ph.D. thesis, TU Dortmund.

Jennrich, R. (1969). Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, **40**, 633–643.

Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *The Annals of Statistics*, **74**, 849–879.

Kiefer, J. and Wolfowitz, J. (1959). Optimum designs in regression problems. *The Annals of Mathematical Statistics*, **30**, 271–294.

Kuhn, E. and Lavielle, M. (2001). Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, **49**, 1020–1038.

Laird, N. and Ware, J. (1982). Random effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Lindstrom, M. and Bates, D. (1990). Nonlinear mixed-effects models for repeated measures data. *Biometrics*, **46**, 673–687.

Liski, E., Mandal, N., Shah, K., and Sinha, B. (2002). *Topics in Optimal Design.* Lecture Notes in Statistics. Springer, New York.

McCullagh, P. and Nelder, J. (1997). *Generalized Linear Models.* Chapman & Hall, London. (Reprinted).

Mentré, F., Burtin, P., Merlé, Y., van Bree, J., Mallet, A., and Steimer, J.-L. (1995). Sparse-sampling optimal designs in pharmacokinetics and toxicokinetics. *Drug Information Journal*, **29**, 997–1019.

Mentré, F., Mallet, A., and Baccar, D. (1997). Optimal design in random-effects regression models. *Biometrika*, **84**, 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**, 363–388.

Mielke, T. (2009). Sparse sampling D-optimal designs in quadratic regression with random effects. In S. M. Ermakov, V. B. Melas, and A. N. Pepelyshev, editors, *Proceedings of the 6th St. Petersburg Workshop on Simulation*, pages 1066–1071.

Mielke, T. (2011a). Approximation of the Fisher information and design in nonlinear mixed effects models. Preprint 22-11, Fakultät für Mathematik, Otto-von-Guericke University Magdeburg.

Mielke, T. (2011b). Nonlinear mixed effects models: Approximation of the Fisher information and design. In V. Melas, G. Nachtmann, and D. Rasch, editors, *Optimal Design of Experiments - Theory and Applications*, pages 108–115. BOKU, Vienna, Austria.

Mielke, T. and Schwabe, R. (2010). Some considerations on the Fisher information in nonlinear mixed effects models. In A. Giovagnoli, A. Atkinson, B. Torsney, and C. May, editors, *mODa9 - Advances in Model-Oriented Design and Analysis*, pages 129–136. Physica, Heidelberg.

Niaparast, M. (2010). *Optimal Designs for Mixed Effects Poisson Regression Models*. Ph.D. thesis, Otto-von-Guericke University Magdeburg.

Nie, L. (2007). Convergence rate of MLE in generalized linear and nonlinear mixed-effects models: Theory and applications. *Journal of Statistical Planning and Inference*, **137**, 1787–1804.

Pinheiro, J. and Bates, D. (1995). Approximation to the loglikelihood function in the nonlinear mixed effects model. *Journal of Computational and Graphical Statistics*, **4**, 12–35.

Pinheiro, J. and Bates, D. (2000). *Mixed-Effects Models in S and S-Plus*. Springer, New York.

Pronzato, L. (2002). Information matrices with random regressors. Application to experimental design. *Journal of Statistical Planning and Inference*, **108**, 189–200.

Pronzato, L. (2010). Asymptotic properties of adaptive penalized optimal designs over a finite space. In A. Giovagnoli, A. Atkinson, B. Torsney, and C. May, editors, *mODa 9 - Advances in Model-Oriented Design and Analysis*, pages 165–172. Physica, Heidelberg.

Pronzato, L. and Walter, E. (1985). Robust experimental design via stochastic approximation. *Mathematical Biosciences*, **75**, 103–120.

Prus, M. and Schwabe, R. (2011). Optimal designs for individual prediction in random coefficient regression models. In V. Melas, G. Nachtmann, and D. Rasch, editors, *Optimal*

Design of Experiments - Theory and Applications, pages 122–129. BOKU, Vienna, Austria.

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

Pukelsheim, F. and Titterington, D. (1983). General differential and Lagrangian theory for optimal experimental design. *The Annals of Statistics*, **11**, 1060–1068.

Rao, C. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, **52**, 447–458.

Rao, C. (1973). *Linear Statistical Inference and its Applications.* John Wiley & Sons, New York. Second edition.

Retout, S. and Mentré, F. (2003). Further developments of the Fisher information matrix in nonlinear mixed effects models. *Journal of Biopharmaceutical Statistics*, **13**, 209–227.

Retout, S., Duffull, 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.

Schmelter, T. (2007a). *Experimental Design For Mixed Models With Application to Population Pharmacokinetic Studies.* Ph.D. thesis, Otto-von-Guericke University Magdeburg.

Schmelter, T. (2007b). The optimality of single-group designs for certain mixed models. *Metrika*, **65**, 183–193.

Schmelter, T., Benda, N., and Schwabe, R. (2007). Some curiosities in optimal designs for random slopes. In J. López-Fidalgo, J. Rodríguez-Díaz, and B. Torsney, editors, *mODa8 - Advances in Model-Oriented Design and Analysis*, pages 189–195. Physica, Heidelberg.

Schott, J. (1997). *Matrix Analysis for Statistics.* John Wiley & Sons, New York.

Searle, S. (1971). *Linear Models.* John Wiley & Sons, New York.

Seber, G. and Wild, C. (2003). *Nonlinear Regression.* Wiley & Sons New York.

Sheiner, L. and Beal, S. (1980). Evaluation of methods for estimating population pharmacokinetic parameters. I. Michaelis-Menten model: Routine clinical pharmacokinetic data. *Journal of Pharmacokinetics and Biopharmaceutics*, **8**(6), 553–571.

Sheiner, L. and Beal, S. (1981). Evaluation of methods for estimating population pharmacokinetic parameters II. Biexponential model and experimental pharmacokinetic data. *Journal of Pharmacokinetics and Biopharmaceutics*, **9**(5), 635–651.

Sheiner, L. and Beal, S. (1983). Evaluation of methods for estimating population pharmacokinetic parameters. III. Monoexponential model: Routine clinical pharmacokinetic data. *Journal of Pharmacokinetics and Biopharmaceutics*, **11**(3), 303–319.

Sheiner, L., Rosenberg, B., and Melmon, K. (1972). Modeling of individual pharmacokinetics for computer-aided drug dosage. *Computers and Biomedical Research*, **5**, 441–459.

Silvey, S. (1978). Optimal design measures with singular information matrices. *Biometrika*, **65**, 553–559.

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

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

Tierney, L. and Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of American Statistical Association*, **81**, 82–86.

Tierney, L., Kass, R., and Kadane, J. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, **84**, 710–716.

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 Biopharmaceutics*, **26**, 689–715.

Torsney, B. and López-Fidalgo, J. (1995). MV-optimization in simple linear regression. In C. Kitsos and G. W. Müller, editors, *mODa4 - Advances in Model-Oriented Data and Analysis*, pages 67–79. Physica, Heidelberg.

Verberke, G. and Molenberghs, G. (2001). *Linear Mixed Models for Longitudinal Data*. Springer, New York.

Vonesh, E. (1996). A note on the use of Laplace's approximation for nonlinear mixed-effects models. *Biometrika*, **83**, 447–452.

Vonesh, E. and Carter, R. (1992). Mixed effects nonlinear regression for unbalanced repeated measures. *Biometrics*, **48**, 1–18.

Wand, M. (2002). Vector differential calculus in statistics. *The American Statistician*, **56**, 55–62.

Wedderburn, W. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**, 439–447.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika*, **80**, 791–795.

Wolfinger, R. and Lin, X. (1997). Two Taylor-series approximation methods for nonlinear mixed models. *Computational Statistics & Data Analysis*, **25**, 465–490.

Wynn, H. (1970). The sequential generation of D-optimum experimental designs. *The Annals of Mathematical Statistics*, **41**, 1655–1664.