Approximation of the Fisher Information and Design in Nonlinear Mixed Effects Models

T. Mielke

Abstract The missing closed form representation of the joint probability density of the observations is one main problem in the analysis of nonlinear mixed effects models. Often local approximations based on linearized models are then applied to approximately describe the properties of estimators. These local approximations are used for designing the experiments. The presentation of alternative motivations of Fisher information approximations are the aim of the present paper. Some locally optimal designs for a pharmacokinetic model are derived with the proposed approximations.

1 Introduction

Nonlinear mixed effects models are frequently used in the analysis of grouped data. Specially in pharmacological studies the observed individuals usually share a common response structure, such that information from individual responses might be merged to obtain efficient estimates. In the statistical model of these studies, the difference in the observations of different individuals are assumed to depend on an

T. Mielke , Otto-von-Guericke University Magdeburg, e-mail: tobias.mielke@ovgu.de

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

2000 Mathematics Subject Classification: 62K05, 62J02.

Keywords: Fisher information, mixed effects models, nonlinear models, optimal design

observation wise varying observation error and on individual parameter vectors. The mixed effects models can be used to model population studies by assuming the individual parameter vectors to be realizations of independent and identically distributed random variables, what yields for nonlinear response functions of the individual parameters nontrivial models. A vast literature describes nonlinear mixed effects models, with special regards on normally distributed random effects. As no closed form of the likelihood exists, estimators based on weighted sums of squares (e.g. Pinheiro and Bates[9], Davidian and Giltinan[1]) or on stochastic approximation (e.g. Kuhn and Lavielle[6]) are proposed to study the typical population behavior. Following the literature, most proposed estimators are consistent with an information matrix behaving as in linear mixed effects models or nonlinear normal heteroscedastic models. The limitations of the stochastic behavior of the estimators were discussed on the monoexponential model by Demidenko[2]. Davidian and Giltinan[1] discuss the problems occuring in likelihood estimation in the case of distribution missspecifications. However, the linearized model builds the foundations for optimal experimental designs in mixed effects models. The impact of the model linearization was dicussed under the assumption of normally distributed parameters by Merle and Tod[7]. Mielke and Schwabe[8] discuss the problem of two approximations based on model linearizations in either the true population location parameter vector or in a guess of this vector on a simple example. The approximations of the information matrix in the present article will be motivated by Bayesian models. Therefor the mixed effects model formulation is described in the second section. Tierney et. al. ([14],[15]) described fully exponential Laplace approximations as an accurate method for approximating posterior moments and densities. The here presented approximations of the Fisher information, obtained by approximations of conditional moments with a similar heuristic, are described in section 3. In section 4 the impact of different approximations on the design of experiments is compared in a one compartment open model.

2 Mixed Effects Models

Let the j-th observation of the i-th individual under experimental setting $x_{ij} \in X$ be described by

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

with a real valued response function η, a p dimensional individual parameter vector β_i and an observation error $\varepsilon_{ij} \in \mathbb{R}$. To avoid difficulties, the real valued response function η is assumed to be differentiable in β_i and x_{ij}.
The individual experimental design $\xi_i = (x_{i1}, ..., x_{im_i})$ with $x_{ij} \in X$ 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

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

For a given parameter vector β_i, the vector of the m_i observations within the i-th individual are completely described up to the unknown normally distributed observation error vector $\varepsilon_i = (\varepsilon_{i1}, ..., \varepsilon_{im_i})^T$ by the response function η and the individual experimental settings ξ_i, such that the intraindividual statistical model results in

$$Y_i \sim N(\eta(\beta_i, \xi_i), \sigma^2 I_{m_i}). \tag{1}$$

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

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

inducing the interindividual variation. An alternative but equivalent description of β_i as $\beta + b_i$ yields the interpretation, that all individuals share some fixed parameter β, which is just influenced by individual effects b_i. Observation errors ε_i and individual parameter vectors β_i are assumed to be stochastically independent and the observations of different individuals are stochastically independent as well. Throughout this paper we assume the variance parameter $\theta := (\sigma^2, D)$ to be known.

The assumptions on the normality of the random effects ε_i and β_i yield for linear response functions η the normality of the random variable Y_i. For a nonlinear response function, $\sigma^2 > 0$ and a positive definite matrix D, no closed form representation of the probability density $f_{Y_i}(y_i)$ of Y_i exists. The marginal density of observations y_i results in integral form in

$$L(y_i, \beta, \theta) := f_{Y_i}(y_i) = \int_{\mathbb{R}^p} \phi_{Y_i|\beta_i}(y_i)\phi_{\beta_i}(\beta_i)d\beta_i.$$

The parameters β and θ influence the likelihood by the normal densities $\phi_{Y_i|\beta_i}$ and ϕ_{β_i} with mean and variance as in equations (1) and (2):

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

with $\| \cdot \|$ describing the determinant.

Linear approximations of the response functions are common approaches in the design theory to circumvent the problem of the missing closed form representation of the likelihood. Retout et. al.[11] approximate the model by a linearization of the response function around the true population mean β, such that for a design matrix defined as

$$F_\beta(\xi_i) := (\frac{\partial \eta(\beta_i, \xi_i)}{\partial \beta_i^T}|_{\beta_i = \beta}) \qquad (3)$$

follows

$$Y_i \approx \eta(\beta, \xi_i) + F_\beta(\xi_i)(\beta_i - \beta) + \varepsilon_{ij}. \qquad (4)$$

The distribution assumption on β_i and ε_{ij} implies normally distributed vectors of observations with heteroscedastic errors:

$$Y_i \sim N(\eta(\beta, \xi_i), \sigma^2[I_{m_i} + F_\beta(\xi_i)DF_\beta(\xi_i)^T]).$$

An alternative linearization in a guess β_0 of the true parameter vector yields as approximation a linear mixed effects model[10]

$$Y_i \sim N(\eta(\beta_0, \xi_i) + F_{\beta_0}(\xi_i)(\beta - \beta_0), \sigma^2[I_{m_i} + F_{\beta_0}(\xi_i)DF_{\beta_0}(\xi_i)^T]).$$

Note that both approximations yield for linear response functions η the true linear mixed effects model, as $F_\beta(\xi_i) = F_{\beta_0}(\xi_i)$ is then independent of β. The dependence of the variance in the nonlinear heteroscedastic model formulation (4) on the population location vector β might imply additional information as described by Mielke and Schwabe[8]. For designing experiments, the proposed approximations are used to derive the inverse of the Fisher information as a lower bound of the variance of any unbiased estimator. However, the approximations are at the first sight approximations based on the functional dependence of the observations on the unknown location parameter vector, yielding approximations of the Fisher information which are based on the likelihood functions implied by the approximated model.

For nonlinear response functions, the Fisher information can be described in terms of conditional moments. Let therefor

$$f_{\beta_i|Y_i=y_i}(\beta_i) := \frac{\phi_{Y_i|\beta_i}(y_i)\phi_{\beta_i}(\beta_i)}{f_{Y_i}(y_i)} \tag{5}$$

denote the conditional probability density of β_i for given observations y_i. The conditional moments of β_i for given observations y_i are here written as

$$E_{y_i}(\beta_i) := E(\beta_i|Y_i = y_i) \text{ and } Var_{y_i}(\beta_i) = Var(\beta_i|Y_i = y_i).$$

With the log-Likelihood-function $l(y_i, \beta, \theta) = \log[L(y_i, \beta, \theta)]$ and the assumption of interchangeability of differentiation and integration, the score function for β results in

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

The Fisher information describes the covariance of the score function, such that it can be represented in the form

$$\begin{aligned}\mathfrak{M}_\beta(\xi_i) &= E\Big(\frac{\partial l(Y_i,\beta,\theta)}{\partial\beta}\frac{\partial l(Y_i,\beta,\theta)}{\partial\beta^T}\Big)\\ &= \frac{1}{\sigma^2}D^{-1}E[(E_{Y_i}(\beta_i)-\beta)(E_{Y_i}(\beta_i)-\beta)^T]D^{-1}\frac{1}{\sigma^2}\\ &= \frac{1}{\sigma^2}D^{-1}Var(E_{Y_i}(\beta_i))D^{-1}\frac{1}{\sigma^2}\\ &= \frac{1}{\sigma^2}D^{-1}-\frac{1}{\sigma^2}D^{-1}E(Var_{Y_i}(\beta_i))D^{-1}\frac{1}{\sigma^2},\end{aligned}$$

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

$$Var(\beta_i) = E(Var_{Y_i}(\beta_i)) + Var(E_{Y_i}(\beta_i)) = \sigma^2 D.$$

Following this representation, only knowledge of the conditional moments of the individual parameter vector, given the observations is of interest for deriving the information matrix.

The presented results can be readily generalized for any differentiable dependence of the mean $g(\beta)$ of the individual parameter vector β_i on the population location parameter β:

Remark 1. Let for $j = 1,...,m_i$

$$Y_{ij} = \eta(\beta_i,x_{ij}) + \varepsilon_{ij}, \text{ with } \varepsilon_{ij} \sim N(0,\sigma^2) \text{ and } \beta_i \sim N_p(g(\beta),\sigma^2 D),$$

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

$$\begin{aligned}\mathfrak{M}_\beta(\xi_i) &= G(\beta)^T\frac{1}{\sigma^2}D^{-1}Var(E_{Y_i}(\beta_i))D^{-1}\frac{1}{\sigma^2}G(\beta)\\ &= G(\beta)^T\Big(\frac{1}{\sigma^2}D^{-1}-\frac{1}{\sigma^2}D^{-1}E(Var_{Y_i}(\beta_i))D^{-1}\frac{1}{\sigma^2}\Big)G(\beta).\end{aligned}$$

As the moments of interest cannot be represented in a closed form, reliable approximations of conditional moments should lead to reliable approximations of the Fisher information. Standard approaches for estimating the conditional moments are Monte-Carlo methods and quadrature rules. However, the computational burden for deriving an approximation of the analytical dependence of the information matrix on the experimental settings is already for relatively small dimensions of the parameter vector and for discrete individual designs of size bigger than one intensive. Laplace approximations are used in Bayesian-Theory for approximating posterior densities of the parameters of interest. In the third section we take use of similar approaches to approximate the Fisher information.

3 Approximations of the Fisher Information

Without knowing the true value of the likelihood of observations y_i, a common approach for its optimization is the maximization of each individual integrand $\phi_{Y_i|\beta_i}(y_i)\phi_{\beta_i}(\beta_i)$ with respect to β_i, $i = 1,...,N$, as well as the maximization of the joint likelihood

$$\prod_{i=1}^{N} \phi_{Y_i|\beta_i}(y_i)\phi_{\beta_i}(\beta_i)$$

with respect to the common mean β of the individual parameter vectors β_i. Obviously this leads to the penalized least squares step as in the Lindstrom and Bates algorithm[9].

The idea of the maximization of the individual integrand $\phi_{Y_i|\beta_i}(y_i)\phi_{\beta_i}(\beta_i)$ is used in the Laplace approximation as well. Up to some constant $-(2\sigma^2)^{-1}$ let $\tilde{l}$ describe the exponent of the normal densities:

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

In the Laplace approximation the function $\tilde{l}$ is approximated by a second order Taylor approximation around a value β_i^* minimizing $\tilde{l}$. With this approach the linear term in the Taylor approximation vanishes, such that the approximated exponent of the integrand is a quadratic form of β_i, yielding a normal density and with this:

$$\int_{\mathbb{R}^p} \phi_{Y_i|\beta_i}(y_i)\phi_{\beta_i}(\beta_i)d\beta_i \approx \frac{1}{c} \cdot \exp(-\frac{1}{2\sigma^2}\tilde{l}(y_i,\beta_i^*,\beta,\theta)), \text{ where}$$

$$c = \sqrt{2\pi\sigma^2}^{\,m_i} \sqrt{\| D \| \| \frac{1}{2} \frac{\partial^2 \tilde{l}(y_i,\beta_i,\beta,\theta)}{\partial\beta_i \partial\beta_i^T}|_{\beta_i=\beta_i^*} \|}.$$

Note that the support point for the Taylor approach β_i^* is a function depending on the observation y_i and the population parameters β and θ. Instead of using the whole Hesse-Matrix of $\tilde{l}$ with respect to β_i an often used simplification is to ignore the term induced by the second derivatives of η:

$$\begin{aligned}\frac{\partial^2\tilde{l}(y_i,\beta_i,\beta,\theta)}{\partial\beta_i\partial\beta_i^T} &= -2\frac{\partial^2\eta(\beta_i,\xi_i)}{\partial\beta_i\partial\beta_i^T}(y_i-\beta_i)+2\frac{\partial\eta(\beta_i,\xi_i)^T}{\partial\beta_i}\frac{\partial\eta(\beta_i,\xi_i)}{\partial\beta_i^T}+2D^{-1}\\ &\approx 2\frac{\partial\eta(\beta_i,\xi_i)^T}{\partial\beta_i}\frac{\partial\eta(\beta_i,\xi_i)}{\partial\beta_i^T}+2D^{-1},\end{aligned}$$

what is similar to the result when using just a linear approximation of the function η in β_i^* for approximating $\tilde{l}$, as will be shown in Remark 3.

In order to approximate the conditional distribution of β_i for given observations y_i, the same approach is applied for the numerator $f_{Y_i}(y_i)$ of (5), such that for a support point of the Taylor approach $\hat{\beta}_i$ these approximations yield a normal density:

Theorem 1. *Let $Y_i = \eta(\beta_i,\xi_i)+\varepsilon_i$, with $\beta_i \sim N(\beta,\sigma^2 D)$ and $\varepsilon_i \sim N(0,\sigma^2 I_{m_i})$ stochastically independent and let*

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

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

$$\beta_i|_{Y_i=y_i} \overset{app.}{\sim} N(\hat{\beta}_i - M_{\hat{\beta}_i}^{-1}\tilde{F}_{\hat{\beta}_i}, \sigma^2 M_{\hat{\beta}_i}^{-1}).$$

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

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

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

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

The representation of the conditional density (5) yields with the above approximation of the integral and an analogue approximation of the numerator

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

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

Note that the presented result in no way claims normality of the individual parameter vectors under given observations, but just specifies an approximating distribution when using for both numerator and denominator of the conditional density the same Taylor approach.

Often a closed form representation of a point β_i^* maximizing $\tilde{l}$ as a function of the observations y_i and the population parameters cannot be obtained. An other problem is met for sparse individual sampling schemes, as the function $\tilde{l}$ might then be multimodal. When applying the Taylor approach in the mode of $\tilde{l}$, one obtains for the approximation the following result:

Remark 2. The approximated conditional distribution of β_i given y_i is for the Laplacian approximation of the form

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

For Maximum-Likelihood-Estimation and for approximating the information matrix, interest lies in the conditional mean of the individual parameter vector β_i for given observations y_i. However, Remark 1 shows, that by the Laplacian approximation, the conditional mean of the individual parameter vector is approximated by its mode. Due to the nonlinearity of the function η in the parameter vector β_i, the random variable $\beta_i|_{Y_i=y_i}$ is not symmetrically distributed around β_i^*, such that mode and mean need not coincide. Hence the estimate may become biased. Despite the bias in the estimate of the conditional mean, one might use the derived approximation of the conditional mean or the conditional variance for the calculation of the Fisher information. Tierney and Kadane[14] presented a method for more accurate approximations of posterior moments by applying Laplace approximations to numerator and denominator with different support points for the Taylor expansion. Generally 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 ([14]). Problems for the presented approaches occur as the approximated conditional moments usually nonlinearly depend on the observations y_i, such that the derivation of the expectation or the variance of these conditional moments would require a second level of approximations.

Alternatively, First-Order Approximations might be used for approximating the function $\tilde{l}$. Instead of approximating the whole function $\tilde{l}$, simply the nonlinear function η is approximated in an estimate $\hat{\beta}_i$ of the individual parameter vector β_i:

Theorem 2. *Let $Y_i = \eta(\beta_i, \xi_i) + \varepsilon_i$, with $\beta_i \sim N(\beta, \sigma^2 D)$ and $\varepsilon_i \sim N(0, \sigma^2 I_{m_i})$ stochastically independent, and let*

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

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

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

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

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

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

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

As in the proof of Theorem 1 one obtains for the approximation of the integral with a constant $c := \sqrt{2\pi\sigma^2}^{\,m_i+p}\sqrt{\| D \|}$:

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

Applying the same approximation to the numerator of (5) yields for the conditional density

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

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

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

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

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

$$\begin{aligned} & -F^T_{\beta_i^*}(y_i - \eta(\beta_i^*, \xi)) + D^{-1}(\beta_i^* - \beta) = 0 \\ \Leftrightarrow\ & D^{-1}\beta_i^* = F^T_{\beta_i^*}(y_i - \eta(\beta_i^*, \xi)) + D^{-1}\beta, \end{aligned}$$

such that

$$\begin{aligned} & M^{-1}_{\beta_i^*}(F^T_{\beta_i^*}(y_i - \eta(\beta_i^*, \xi_i) + F_{\beta_i^*}\beta_i^*) + D^{-1}\beta) \\ = {} & M^{-1}_{\beta_i^*}(F^T_{\beta_i^*}(y_i - \eta(\beta_i^*, \xi_i)) + D^{-1}\beta + F^T_{\beta_i^*}F_{\beta_i^*}\beta_i^*) \\ = {} & M^{-1}_{\beta_i^*}(D^{-1}\beta_i^* + F^T_{\beta_i^*}F_{\beta_i^*}\beta_i^*) = \beta_i^*. \end{aligned}$$

□

This result motivates the earlier presented simplified approximation of the Hesse matrix of $\tilde{l}$. However, 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 must 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 4. The approximated conditional distribution of β_i given y_i resulting from a First-Order-Linearization in β is of the form

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

With $V := \sigma^2(I_{m_i} + F_\beta D F_\beta^T)$ the approximated Fisher information results in:

$$\begin{aligned} \mathfrak{M}_{1,\beta}(\xi_i) &:= F_\beta^T V^{-1} Var(Y_i) V^{-1} F_\beta \\ \mathfrak{M}_{2,\beta}(\xi_i) &:= F_\beta^T V^{-1} F_\beta. \end{aligned}$$

Proof: The result for the approximated conditional distribution is a direct consequence of Theorem 2 with $\hat{\beta}_i = \beta$.

The first approximation of the information matrix follows since

$$\begin{aligned}\mathfrak{M}_{1,\beta}(\xi_i) &= \frac{D^{-1}}{\sigma^2} Var(E_{Y_i}(\beta_i)) \frac{D^{-1}}{\sigma^2} \\ &= \frac{D^{-1}}{\sigma^2} M_\beta^{-1} F_\beta^T Var(Y_i) F_\beta M_\beta^{-1} \frac{D^{-1}}{\sigma^2} \\ &= F_\beta^T V^{-1} Var(Y_i) V^{-1} F_\beta,\end{aligned}$$

where the last equation follows since for V regular and M_β

$$\begin{aligned}\frac{1}{\sigma^2} D^{-1} (F_\beta^T F_\beta + D^{-1})^{-1} F_\beta^T &= \frac{1}{\sigma^2} (I_p - \sigma^2 F_\beta^T V^{-1} F_\beta D) F_\beta^T \\ &= \frac{1}{\sigma^2} F_\beta^T (V^{-1} V - \sigma^2 V^{-1} F_\beta D F_\beta^T) = F_\beta^T V^{-1}.\end{aligned}$$

The second approximation follows analogiously:

$$\begin{aligned}\mathfrak{M}_{2,\beta}(\xi_i) &= \frac{D^{-1}}{\sigma^2} [Var(\beta_i) - E(Var_{Y_i}(\beta_i))] \frac{D^{-1}}{\sigma^2} \\ &= \frac{D^{-1}}{\sigma^2} - \sigma^2 \frac{D^{-1}}{\sigma^2} (F_\beta^T F_\beta + D^{-1})^{-1} \frac{D^{-1}}{\sigma^2} \\ &= F_\beta^T V^{-1} F_\beta,\end{aligned}$$

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

Remark 4 yields an alternative motivation for the approximation of the Fisher information by approximating the model by a First-Order-Linearization in an estimate β_0, as the obtained information matrices coincide in both cases.

The accuracy of the proposed approximations depends on the individual sample size m_i and the variance of the individual parameter vectors $\sigma^2 D$. Tierney, Kass and Kadane[15] state in their work on the fully exponential Laplace approximation to ratios of integrals that the accuracy for the approximation of the mean is of order $O(m_i^{-2})$, while the accuracy for the variance is of order $O(m_i^{-3})$. Instead of using the fully exponential Laplace approximation Tierney[14] comments that already two steps of a Newton iteration for localizing β_i^* are sufficient for the proposed accuracy. Following these results, we construct optimal designs and examine the proposed approximations of the Fisher information $\mathfrak{M}_{1,\beta}$ and $\mathfrak{M}_{2,\beta}$ in the fifth section for small individual sample sizes on an example. Of special interest is the influence of the different approximations on the design of the experiments.

4 Population Designs

Optimal designs for population studies are in many disciplines from ethical and economical point of view of great interest. The informations from different individuals offer the possibility to estimate the population parameters even in cases when individual sampling schemes yield individual singular information matrices. With a compact design space X and experimental settings $x_{ij} \in X$, the designs of individuals are described by m_i-tupels

$$\xi_i := (x_{i1}, ..., x_{im_i}),$$

such that discrete designs are target of the optimization on the individual level. We here assume that the number of observations is identical for all individuals, so that $m_i = m$ for all ξ_i. The independence of the observations of different individuals yields for the population information matrix and N individuals

$$\mathfrak{M}_{pop,\beta} := \sum_{i=1}^{N} \mathfrak{M}_{\beta}(\xi_i).$$

When normalizing this population information matrix, population designs are completely described by

$$\zeta := (\xi_1, ..., \xi_k, \omega_1, ..., \omega_k) \text{ with } \omega_i \geq 0,\ \sum_{i=1}^{k-1} \omega_i \leq 1 \text{ and } \omega_k = 1 - \sum_{i=1}^{k-1} \omega_i,$$

yielding a continuous optimization problem on the set of approximate population designs with a resulting normalized information matrix

$$\mathfrak{M}_{pop,\beta}(\zeta) := \sum_{i=1}^{k} \omega_i \mathfrak{M}_{\beta}(\xi_i).$$

Target of the optimization problem is the maximization of the information. As for $p > 1$ the information is a matrix, real valued functionals of the information matrices are used for comparing the quality of designs. In the case of $p > 1$ and for mixed effects models, different optimality criteria generally imply different optimal designs. With the asymptotic normality of the maximum likelihood estimator, the content of the confidence ellipsoid is inverse proportional to the determinant of the information matrix, yielding the D-optimality criterion:

$$\Phi_D(\zeta) := -\log(\| \mathfrak{M}_{pop,\beta}(\zeta) \|).$$

An other often used class of optimality criteria are the linear criteria:

$$\Phi_L(\zeta) := Tr(L\mathfrak{M}_{pop,\beta}^{-1}(\zeta)).$$

Designs can be optimized under different aspects in dependence of the matrix L. The A- and $IMSE$-criterion are examples for linear criteria. The A-criterion ($L = I_p$) minimizes the mean of the diagonal entries of the inverted information and hence minimizes the arithmetic mean of the variance of the estimates. The $IMSE$-criterion is based on point-wise predictions on the design region, minimizing the integrated mean squared error over the design region X. The matrix L is for the $IMSE$-criteria depending on the response function η of the observations.

A result of Schmelter[12] motivates the use of population designs consisting of one group only. However, as the result is presented for approximate individual designs, it usually cannot be applied in more realistic scenarios with limited numbers of individual observations. Even worse: when comparing the efficiency of population designs with the same number of total observations $N \times m$, population designs with individual designs consisting of big numbers of observations $m \gg p$ lose efficiency compared to designs with smaller numbers of individual observations $0 < \tilde{m} < m$. Note that for X as a subset of $\mathbb{R}$ the population information matrices induced by the individual designs ξ_i build in vector notation a convex subset of the $\mathbb{R}^{\frac{p(p+1)}{2}}$, such that the optimal population design as a boundary point of this set can be represented by Caratheodory's Theorem as a convex combination of at most $p(p+1)/2$ individual designs ([3]). Fedorov's equivalence theorems for designs of experiments in the case of simultaneous observation of several quantities can be applied for the construction of optimal designs for various design criteria. In each iteration of the proposed Fedorov-Wynn algorithm the maxima in X^m of the sensitivity function

$$g_{\zeta,D}(\xi) := Tr[\mathfrak{M}^{-1}_{pop,\beta}(\zeta)\mathfrak{M}_\beta(\xi)] \text{ for } D\text{-optimality and}$$
$$g_{\zeta,L}(\xi) := Tr[\mathfrak{M}^{-1}_{pop,\beta}(\zeta)L\mathfrak{M}^{-1}_{pop,\beta}(\zeta)(\mathfrak{M}_\beta(\xi)-I_p)] \text{ for linear criteria}$$

are sought. Instead of using this steepest descent method, alternatively Quasi-Newton algorithms might be applied to solve the optimization problem for design spaces X as subsets of $\mathbb{R}$. The Quasi-Newton algorithms yields the advantage that the stepwise maximization can be omitted. The gradients of the criterion functions with respect to the experimental settings can be relatively easy determined, when knowing the derivative of the individual Fisher information with respect to experimental settings $x_{ij} \in X$. For the approximations of the Fisher information deduced in section 3, the derivatives can be straightforwardly computed. Specially for $\mathfrak{M}_{2,\beta}$, the derivatives take a simple functional form:

Theorem 3. *Let $Y_i = \eta(\beta_i,\xi)+\varepsilon_i$ with $\beta_i \sim N(\beta,\sigma^2 D)$, $\varepsilon_i \sim N(0,\sigma^2 I_m)$ and*

$$\xi := (x_1,...,x_m) \text{ with } x_j \in X,$$
$$F_\beta := \tfrac{\partial\eta(\beta_i,\xi)}{\partial\beta_i^T}|_{\beta_i=\beta}, \text{ and } F'_{\beta,j} := \tfrac{\partial}{\partial x_j}F_\beta(\xi).$$

With $V := (I_m + F_\beta D F_\beta^T)$ the derivative with respect to x_j of the information-approximation $\mathfrak{M}_{2,\beta}(\xi)$ results in

$$\frac{\partial}{\partial x_j}\mathfrak{M}_{2,\beta}(\xi) = \frac{1}{\sigma^2}[(I_p-\mathfrak{M}_{2,\beta}(\xi)D)F'^T_{\beta,j}V^{-1}F_\beta + F_\beta^T V^{-1}F'_{\beta,j}(I_p - D\mathfrak{M}_{2,\beta}(\xi))].$$

Proof: Note that with

$$\frac{\partial}{\partial x_j}V^{-1} = -V^{-1}(F'_{\beta,j}DF_\beta^T + F_\beta D F'^T_{\beta,j})V^{-1}$$

follows

$$\begin{aligned}\frac{\partial}{\partial x_j}\mathfrak{M}_{2,\beta}(\xi) = \frac{1}{\sigma^2}[&F'^T_{\beta,j}V^{-1}F_\beta - F_\beta^T V^{-1}F_\beta D F'^T_{\beta,j}V^{-1}F_\beta \\ &+F_\beta^T V^{-1}F'_{\beta,j} - F_\beta^T V^{-1}F'_{\beta,j}DF_\beta^T V^{-1}F_\beta]\end{aligned}$$

$$= \frac{1}{\sigma^2}[(I_p - \mathfrak{M}_{2,\beta}(\xi)D)F'^T_{\beta,j}V^{-1}F_\beta + F_\beta^T V^{-1}F'_{\beta,j}(I_p - D\mathfrak{M}_{2,\beta}(\xi))]. \qquad \square$$

With the knowledge of the derivatives of the approximated Fisher information $\mathfrak{M}_{2,\beta}$, the derivatives of optimality criteria can be easily specified:

Theorem 4. *Let $Y_i = \eta(\beta_i, \xi_i) + \varepsilon_i$ with $\beta_i \sim N(\beta, \sigma^2 D)$, $\varepsilon_i \sim N(0, \sigma^2 I_m)$ and let*

$$\begin{aligned}\xi_i &:= (x_{i1}, ..., x_{im}) \text{ with } x_{ij} \in X,\ i = 1, ..., k,\\ \zeta &:= (\xi_1, ..., \xi_k, \omega_1, ..., 1 - \sum_{i=1}^{k-1} \omega_i) \text{ with } \omega_i \geq 0,\ \sum_{i=1}^{k-1} \omega_i \leq 1.\end{aligned}$$

The derivatives of the criterion function result for D-optimality in

$$\begin{aligned}\frac{\partial \Phi_D(\zeta)}{\partial x_{ij}} &= -\frac{2\omega_i}{\sigma^2} Tr[(I_p - \mathfrak{M}_{2,\beta}(\xi_i)D)\mathfrak{M}^{-1}_{pop,\beta}(\zeta)F_\beta(\xi_i)^T V_\beta^{-1}(\xi_i)\frac{\partial}{\partial x_{ij}}F_\beta(\xi_i)]\\ \frac{\partial \Phi_D(\zeta)}{\partial \omega_i} &= -Tr[\mathfrak{M}^{-1}_{pop,\beta}(\zeta)(\mathfrak{M}_{2,\beta}(\xi_i) - \mathfrak{M}_{2,\beta}(\xi_k))].\end{aligned}$$

Proof: Since for real t, $p \times p$ on t depending matrices A and symmetric $p \times p$ matrices M

$$\frac{\partial}{\partial t}[-log(det(A))] = -Tr(A^{-1}\frac{\partial}{\partial t}A) \text{ and } Tr(MA) = Tr(MA^T),$$

the result follows with Theorem 3 as

$$\begin{aligned}\frac{\partial}{\partial x_{ij}}\mathfrak{M}_{pop,\beta}(\zeta) &= \omega_i \frac{\partial}{\partial x_{ij}}\mathfrak{M}_{2,\beta}(\xi_i) \text{ and}\\ \frac{\partial}{\partial \omega_i}\mathfrak{M}_{pop,\beta}(\zeta) &= \mathfrak{M}_{2,\beta}(\xi_i) - \mathfrak{M}_{2,\beta}(\xi_k). \qquad \square\end{aligned}$$

The same approach can be applied for linear optimality criteria. The resulting derivatives are of a similar form as in the case of D-optimality:

Theorem 5. *Let $Y_i = \eta(\beta_i, \xi_i) + \varepsilon_i$ with $\beta_i \sim N(\beta, \sigma^2 D)$, $\varepsilon_i \sim N(0, \sigma^2 I_m)$ and let*

$$\begin{aligned}\xi_i &:= (x_{i1}, ..., x_{im}) \text{ with } x_{ij} \in X,\ i = 1, ..., k,\\ \zeta &:= (\xi_1, ..., \xi_k, \omega_1, ..., 1 - \sum_{i=1}^{k-1} \omega_i) \text{ with } \omega_i \geq 0,\ \sum_{i=1}^{k-1} \omega_i \leq 1.\end{aligned}$$

The derivatives of the criterion function result for L-optimality in

$$\frac{\partial \Phi_L(\zeta)}{\partial x_{ij}} = -\frac{2}{\sigma^2} Tr[(I_p - \mathfrak{M}_{2,\beta}(\xi_i)D)\mathfrak{M}^{-1}_{pop,\beta}(\zeta)L\mathfrak{M}^{-1}_{pop,\beta}(\zeta)F_\beta(\xi_i)^T V_\beta^{-1}(\xi_i)\frac{\partial}{\partial x_{ij}}F_\beta(\xi_i)]$$
$$\frac{\partial \Phi_L(\zeta)}{\partial \omega_i} = -Tr[\mathfrak{M}^{-1}_{pop,\beta}(\zeta)L\mathfrak{M}^{-1}_{pop,\beta}(\zeta)(\mathfrak{M}_{2,\beta}(\xi_i) - \mathfrak{M}_{2,\beta}(\xi_k))].$$

Proof: Since for real t, regular on t depending $p \times p$ matrices A

$$\frac{\partial}{\partial t}Tr(LA^{-1}) = -Tr(LA^{-1}(\frac{\partial}{\partial t}A)A^{-1}),$$

the result follows as in Theorem 4. □

For the proposed approximation $\mathfrak{M}_{1,\beta}$ similar results can be obtained, which additionally depend on the derivative of the variance function. The derivatives were used in an *BFGS*-algorithm ([5]) for optimizing the experimental designs in a pharmacokinetic model. The constraints on the experimental designs were included in the criterion function via twice continuously differentiable exact penalty functions, such that the achieved designs are generally not optimal designs under the original optimality criteria. However, when constraining the achieved designs on the design region, efficient designs are obtained.

5 Example

In pharmacological studies often the typical behavior of populations is of big interest. Specially in the analysis of pharmacokinetics nonlinear mixed effects models are frequently met. In this section we examine different approximations of the Fisher information and construct locally optimal designs for a one compartment models. We restrict ourselves to designs with relatively sparse individual sampling schemes.

The time-concentration relationship of a drug in a compartment with intravenous bolus injection is characterized by the clearance Cl, the volume V of the compartment and the dose $\mathbf{D}$, yielding the response function

$$\eta_1(x) = \frac{\mathbf{D}}{V}\exp(-\frac{Cl}{V}x).$$

The random effects are assumed to influence the observations proportional, such that observed concentrations and individual parameters with negative values are sets of probability zero:

$$Cl_i = Cl \times \exp(b_{1,i}),\ V_i = V \times \exp(b_{2,i}),\ \text{with } (b_{1,i}, b_{2,i}) \sim N_2((0,0), \sigma^2 D),$$

$$Y_{ij} = \tfrac{\mathbf{D}}{V_i} \exp(-\tfrac{Cl_i}{V_i} x_{ij}) \exp(\varepsilon_{ij}),\ \text{where } \varepsilon_{ij} \sim N(0, \sigma^2).$$

The proposed model is not of the form presented in the previous sections. However, let

$$\beta_{1,i} := log(Cl) + b_{1,i} \text{ and } \beta_{2,i} := log(V) + b_{2,i},\ \text{ such that}$$

$$\beta_i = (\beta_{1,i}, \beta_{2,i})^T \sim N_2((log(Cl), log(V))^T, \sigma^2 D)$$

and let the response function be defined as

$$\eta(\beta_i, x_{ij}) := log(\mathbf{D}) - \beta_{2,i} - x_{ij} \times \exp(\beta_{1,i} - \beta_{2,i}),$$

then the logarithm of the observations $\tilde{y}_i := log(y_i)$ is a nonlinear mixed effects model as previously described. The approximations derived in section 3 can be applied to obtain the Fisher information for the mean β of the vector β_i. With an application of Remark 1 we then obtain the approximated Fisher information of the population location parameters clearance Cl and volume V.

As a simplification, we assume the random effects to be stochastically independent, such that the matrix D results in a diagonal matrix with entries d_1 and d_2 for Cl and V respectively. Note that expectation and variance can be specified in the proposed model in a closed form

$$\begin{aligned} E(\tilde{Y}(x_{ij})) &= log(\mathbf{D}) - log(V) - x_{ij} \times \frac{Cl}{V} \exp(\frac{1}{2}\sigma^2(d_1 + d_2))\ , \\ Var(\tilde{Y}(x_{ij})) &= \sigma^2(1 + d_2) + x_{ij}^2 \frac{Cl^2}{V^2} \exp(\sigma^2(d_1 + d_2))(\exp(\sigma^2(d_1 + d_2)) - 1) \\ &\quad -2x_{ij}\sigma^2 d_2 \frac{Cl}{V} \exp(\frac{1}{2}\sigma^2(d_1 + d_2)) \end{aligned}$$

and that observations within one individual are correlated with

$$Cov(\tilde{Y}(x_{ij}),\tilde{Y}(x_{ij'})) = \sigma^2 d_2 + x_{ij}x_{ij'}\frac{Cl^2}{V^2}\exp(\sigma^2(d_1+d_2))(\exp(\sigma^2(d_1+d_2))-1)$$
$$-(x_{ij}+x_{ij'})\sigma^2 d_2 \frac{Cl}{V}\exp(\frac{1}{2}\sigma^2(d_1+d_2)),$$

offering the Quasi-information $\mathfrak{M}_{3,\beta}$ as in generalized linear mixed models to be used as an approximation of the information:

$$\mathfrak{M}_{3,\beta}(\xi) := \frac{\partial E(Y)^T}{\partial \beta} Var(Y)^{-1} \frac{\partial E(Y)}{\partial \beta^T}$$

A fourth approximation of the Fisher information is used to show the effect of the information approximation on the design of experiments by a nonlinear heteroscedastic normal model approximation as in (4), leading with F_β and V as in Theorem 3 to an information matrix of the form ([11])

$$\mathfrak{M}_{4,\beta}(\xi) := \frac{1}{\sigma^2} F_\beta^T V^{-1} F_\beta + \frac{1}{2}\tilde{S}_\beta,$$

where for i and $j = 1,...,p$:

$$(\tilde{S}_\beta)_{(i,j)} = Tr[\frac{\partial V}{\partial \beta_i} V^{-1} \frac{\partial V}{\partial \beta_j} V^{-1}].$$

The resulting approximations in the present situation of a one compartment open model are shown in figure 1 for a time horizon of 24 hours and specified values for β and θ as in example 8.7 in Schmelter[13]:

$$\beta_1 = log(25),\ \beta_2 = log(88),\ \sigma^2 = 0.01,\ d_1 = 12.5,\ d_2 = 9.0\ .$$

In the figure the dependence of the entries of the information matrices on the time are plotted, together with values obtained by simulating the expectation of the conditional variance of the individual parameter vectors for the situation of one observation per individual. The figures are based on 10000 simulated observations per time point in order to estimate the conditional mean of the parameter vector for simulated observations y_i. The resulting figure shows the discussed problem of the missing accuracy of the information approximation based on the conditional expectation $\mathfrak{M}_{1,\beta}$ (dotdash). Note that the only difference compared to $\mathfrak{M}_{2,\beta}$ (solid) is induced by the ratio of the variance in the linearized model and in the original model.

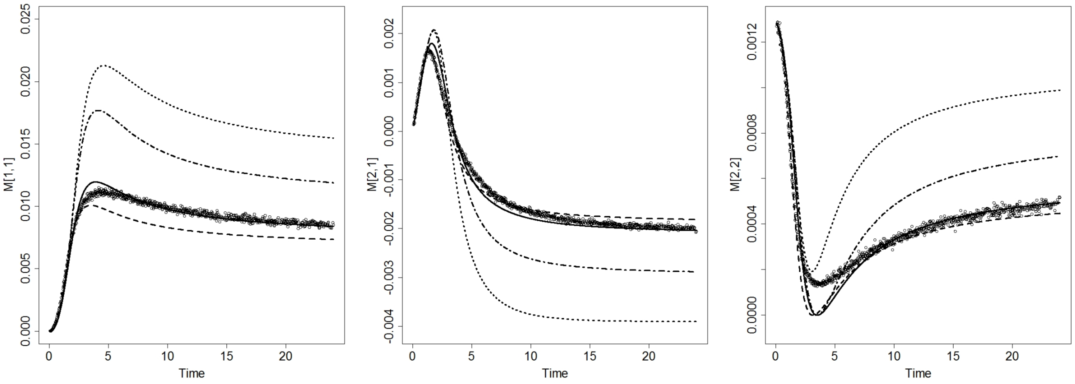

Fig. 1 Components of the information matrix - Simulations and Approximations: $\mathfrak{M}_{1,\beta}$(dotdash), $\mathfrak{M}_{2,\beta}$(solid), $\mathfrak{M}_{3,\beta}$(dashed) and $\mathfrak{M}_{4,\beta}$(dotted)

The possible impact of an approximation of the nonlinear mixed effects model by a nonlinear normal model with heteroscedastic errors was discussed by Schwabe and Mielke[8]. The figure shows the additional information predicted by the authors. However, the figure shows as well, that minima and maxima of the different approximations are located close to each other, such that optimal designs will probably not differ much. Similarly to the case of straight line regression discussed by Graßhoff et al.[4] with one observation per individual and heteroscedastic observations, one can show for the proposed model that the D-optimal one observation designs for the information matrices $\mathfrak{M}_{2,\beta}$ and $\mathfrak{M}_{3,\beta}$ are characterized by every pair (x_1,x_2) of experimental settings fulfilling

$$\sigma^2 + Cov(\tilde{Y}(x_1), \tilde{Y}(x_2)) = 0$$

with the covariance of the linearized model for $\mathfrak{M}_{2,\beta}$ and the exact covariance for $\mathfrak{M}_{3,\beta}$. For approximating the $IMSE$-optimality criterion, we applied a linear approximation of the model function η in the true population location parameter β, such that for the design region $X = [0.1, 24]$:

$$L := \int_{[0.1,24]} \frac{\partial \eta(\beta_i, x)^T}{\partial \beta_i}|_{\beta_i=\beta} \frac{\partial \eta(\beta_i, x)}{\partial \beta_i^T}|_{\beta_i=\beta} dx.$$

Efficient designs were numerically derived with a Quasi-Newton algorithm as described in section 4. The obtained results for designs with less than 5 observations per individual are listed in table 1 for *D*-optimality and in table 2 for *IMSE*-optimality. For both optimality criteria and for $m > 2$ observations the proposed designs consist of individual designs containing measurement replications. This might cause problems considering the independence of the observations. As the sensitivity function is continuous, one might obtain efficient designs without measurement replications in experimental settings by simply adding an additional observation very close to the proposed measurement instead of the proposed replication. As predicted, the resulting designs for the *D*-optimality criterion and the different approximations are similar:

Table 1 *D*-efficient Designs for the proposed information approximations

$\mathfrak{M}_{j,\beta}(\zeta)$	$m=1$	$m=2$	$m=3$	$m=4$
$\mathfrak{M}_{1,\beta}$	$[(1.70),(24.00),0.50,0.50]$	$[(0.10,8.06),1.00]$	$[(0.10,0.10,11.83),1.00]$	$[(0.10,0.10,0.10,15.58),1.00]$
$\mathfrak{M}_{2,\beta}$	$[(1.13),(11.98),0.50,0.50]$	$[(0.10,7.66),1.00]$	$[(0.10,0.10,11.38),1.00]$	$[(0.10,0.10,0.10,15.89),1.00]$
$\mathfrak{M}_{3,\beta}$	$[(0.93),(11.99),0.50,0.50]$	$[(0.10,6.97),1.00]$	$[(0.10,0.10,10.32),1.00]$	$[(0.10,0.10,0.10,13.70),1.00]$
$\mathfrak{M}_{4,\beta}$	$[(0.96),(\ 6.00),0.47,0.53]$	$[(0.10,7.83),1.00]$	$[(0.10,0.10,11.52),1.00]$	$[(0.10,0.10,0.10,15.46),1.00]$

For the *IMSE*-criterion more differences in the designs were observed. Specially for $\mathfrak{M}_{1,\beta}$ and $\mathfrak{M}_{2,\beta}$ more-group designs can slightly improve the efficiency of the studies. Despite the differences in the experimental designs, most proposed designs yield a high efficiency (> 0.99) when compared to each other. The only design with a lower efficiency (0.91) in terms of the *IMSE*-criterion for $\mathfrak{M}_{2,\beta}$ is induced by $\mathfrak{M}_{4,\beta}$

for an individual sample size $m = 1$.

This relatively high efficiency for different experimental designs carries forward to the optimization algorithm. In dependence of the starting values of the iterations, problems occured when optimally weighting the individual designs in the population. Specially for $\mathfrak{M}_{2,\beta}$ with 3 observations per individual the convergence to a two-group design was heavily depending on the starting values of the *BFGS*-algorithm.

Table 2 *IMSE*-efficient Designs for the proposed information approximations

$\mathfrak{M}_{j,\beta}$	m	$\zeta = [\xi_1, ...\xi_k, \omega_1, ..., \omega_k]$
$\mathfrak{M}_{1,\beta}$	1	$[(2.49), (24.00), 0.10, 0.90]$
	2	$[(0.10, 24.00), (20.35, 20.35), 0.64, 0.36]$
	3	$[(0.10, 24.00, 24.00), (20.09, 20.09, 20.09), 0.64, 0.36]$
	4	$[(0.10, 0.10, 24.00, 24.00), (20.00, 20.01, 20.01, 20.01), 0.59, 0.41]$
$\mathfrak{M}_{2,\beta}$	1	$[(2.17), (24.00), 0.09, 0.91]$
	2	$[(0.10, 24.00), 1.00]$
	3	$[(0.10, 0.10, 24.00), (0.10, 24.00, 24.00), 0.49, 0.51]$
	4	$[(0.10, 0.10, 24.00, 24.00), 1.00]$
$\mathfrak{M}_{3,\beta}$	1	$[(1.98), (23.99), 0.10, 0.90]$
	2	$[(0.10, 24.00), 1.00]$
	3	$[(0.10, 21.55, 21.55), 1.00]$
	4	$[(0.10, 0.10, 24.00, 24.00), 1.00]$
$\mathfrak{M}_{4,\beta}$	1	$[(0.62), (\ 9.35), 0.08, 0.92]$
	2	$[(0.10, 15.77), 1.00]$
	3	$[(0.10, 0.10, 23.09), 1.00]$
	4	$[(0.10, 0.10, 0.10, 24.00), 1.00]$

Note that in the situation of more than 2 observations per individual, the verification of the optimality cannot be easily done using the equivalence theorem. For big individual sample sizes, an optimality result on approximate individual designs can be applied to demonstrate the quality of designs. Schmelter ([13], p.65) presents equivalence theorems for this situation when only one-group designs are allowed for an information matrix as $\mathfrak{M}_{2,\beta}$. The inter-individual variance $\sigma^2 D$ has for linear criteria with regular individual information matrices and approximate individual designs no influence on the design. Obviously this leads in the present situation to *IMSE*-optimal approximate individual one-group designs with equal weights on $t_1 = 0.1$ and $t_2 = 24$, as for $\sigma^2 D = 0$ the situation simplifies to a straight line regression problem.

6 Discussion

Different approximations of the Fisher information in nonlinear mixed effects models were presented in the present article and the effects of the approximations on the design of experiments were studied on a simple example. The derivation of the Fisher information in terms of conditional moments allowed the estimation of the dependence of the components of the information matrix on the experimental settings by simulations in the example of a One-Compartment model with intravenous bolus injection and one observation per individual. In this situation an approximation of the information matrix by the Fisher information of a linear mixed effects model, resulting from a first order linearization of the response function in the population mean β, yielded the best results. For the case of more observations per individual the computational effort to obtain reasonable results for similar simulations is very high. The similar structure of the derivatives for D- and linear optimality criteria can be applied in Quasi-Newton algorithms for optimizing the experimental designs. The present example unfortunately yielded problems for the proposed algorithm, when computing more group designs. One might solve the malfunction by implementing after convergence of the proposed algorithm a step of a standard design algorithm to ensure that the *BFGS*-routine will include more group designs as well.

The similar dependence of the components of the informations matrices for the different approximations on the experimental settings induces similar designs. Specially for D-optimality and more than one observation per individual, the designs differed just marginally. The influence of the approximations on the populations designs for the *IMSE*-optimality criteria was stronger. For the approximations $\mathfrak{M}_{1,\beta}$ and $\mathfrak{M}_{2,\beta}$

more group designs improved the efficiency. The additional term in the information approximation $\mathfrak{M}_{4,\beta}$ yields more information of the observations on the left border of the design region, such that balanced individual designs are for $\mathfrak{M}_{4,\beta}$ suboptimal for even numbers of individual observations. All designs yielded a similar efficiency in the proposed example. The impact of different information approximations on the designs of experiments and specially their efficiency will be of great interest for getting more insight in the behavior of optimal designs in nonlinear mixed effects models.

Acknowledgment The author is very grateful for the help by the project SKAVOE of the BMBF and to Rainer Schwabe (Univ. of Magdeburg).

References

1. Davidian, M., Giltinan, D.M. (1995). *Nonlinear Models for Repeated Measurement Data*. Chapman & Hall, London.
2. Demidenko, E. (2005). *Mixed Effects Models: Theory and Applications*. Wiley, New Jersey.
3. Fedorov, V.V. (1972). *Theory of optimal experiments*. Academic Press, New York.
4. Graßhoff, U., Doebler, A., Holling, H., Schwabe, R. (2009). Optimal design for linear regression models in the presence of heteroscedasticity caused by random coefficients. *Journal of Statistical Planning and Inference*(to appear).
5. Großmann, C., Terno, J. (1997). *Numerik der Optimierung*. Teubner, Stuttgart.
6. Kuhn, E., Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics* & *Data Analysis*, 49:1020-1038.
7. Merle, Y., Tod, M. (2001). Impact of Pharmacokinetic-Pharmacodynamix Model Linearization on the Accuracy of Population Information Matrix and Optimal Design. *Journal of Pharmacokinetics and Pharmacodynamics*, 28:363-388.
8. Mielke, T., Schwabe, R. (2010). Some Considerations on the Fisher Information in Nonlinear Mixed Effects Models. In *mODa9 - Advances in Model-Oriented Design and Analysis* (Giovagnoli, A. ,Atkinson, A.C., Torsney, B., May, C. (eds.)), Physica, Heidelberg: 129-136.
9. Pinheiro, J.C., Bates, D.M. (2000). *Mixed-Effects Models in S and S-Plus*. Springer, New York.
10. Retout, S., Duffull, S., Mentre, 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.

11. Retout, S., Mentre, F. (2003). Further Developments of the Fisher Information Matrix in Nonlinear Mixed Effects Models with Evaluation in Population Pharmacokinetics. *Journal of Biopharmaceutical Statistics*, 13:209-227.
12. Schmelter, T. (2007). The optimality of single-group designs for certain mixed models. *Metrika*, 65:183-193.
13. Schmelter, T. (2007). *Experimental Design For Mixed Models With Application to Population Pharmacokinetic Studies*. PhD Thesis, Otto-von-Guericke University Magdeburg.
14. Tierney, L., Kadane, J.B. (1986). Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of American Statistical Association*, 81:82-86.
15. Tierney, L., Kass, R.E., Kadane, J.B. (1989). Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions. *Journal of American Statistical Association*, 84:710-716.