Optimal Designs for the Prediction of Individual Parameters in Hierarchical Models

Maryna Prus*, Rainer Schwabe

Otto-von-Guericke University Magdeburg, Institute for Mathematical Stochastics,
PF 4120, D-39016 Magdeburg, Germany

Abstract: Characterizations of optimal designs are derived for the prediction of individual response curves within the framework of hierarchical linear mixed models. It is shown that the so obtained optimal designs may differ substantially from those propagated in the literature so far and that the latter may become useless in terms of their performance.

Keywords: Linear Mixed Model, Hierarchical Model, Random Coefficient Regression, Prediction, Individual Design.

1 Introduction

During the last decades hierarchical models such as random coefficient regression have become popular in many statistical applications. These models account for individual effects of subjects or observational units under investigation by the assumption that these subjects or units come form a larger population and that their individual variability is governed by a random source. Originally such models were introduced in the biosciences, in particular, in plant and animal breeding (see for example Henderson (1984)) for selection purposes. More recently, working groups evolved in medical and pharmaceutical research (PAGE: Population Approach Group in Europe `www.page-meeting.org/`, PODE: Population Optimum Design of Experiments `www.maths.qmul.ac.uk/ bb/PODE/PODE.html`), which are mainly concerned with the mean (*population*) parameters in hierarchical models. Moreover, this hierarchical approach is also knwon as panel models in sociological research and econometrics and may be found in the context of structural equation modelling in psychology. Even in official statistics this concept appears in the methodology of small area estimation (see e. g. Rao (2003)).

The problem of estimation of the mean population parameters ("typical response") has been widely considered in the literature (see Henderson *et al.* (1959), Isotalo *et al.* (2011)). For the estimation ("prediction") of the individual random effects the common solution is a Bayesian approach (see for example Pukelsheim (1993)), if the population parameters are assumed to be knwon a priori. This assumption is quite unlikely to be appropriate for most populations in real appplications, where the population parameters will be unknown as well.

The lacking knowledge of the population parameters makes the analysis substantially more difficult. A straightforward possibility is to use empirical Bayes estimators as proposed by Bryk and Raudenbush (1992) (see also Molenberghs and Verbeke (2001)). Alternatively the individual random parameters may be estimated using Henderson's mixed-model equations (see Henderson (1963)), which can occasionally coincide with the previous approach.

While there is a reasonable literature on optimal designs for population mean parameters (see e. g. Entholzner *et al.* (2005), Fedorov and Hackl (1997), Liski *et al.* (2002)) hardly any work has been done to obtain designs, which are optimal for the prediction of the individual parameters: In their pioneering paper Gladitz and Pilz (1982) established that Bayesian optimal designs are also optimal for the prediction of individual effects. Their result required the prior knowledge of the "typical" response mentioned before. However, this crucial assumption seems to have been

*Maryna Prus: maryna.prus@ovgu.de

overlooked in the aftermath, which led to the common believe that concerning optimal designs for prediction everything had been solved by Gladitz and Pilz (1982).

As a rare exception Fedorov and Hackl (1997) made a clear distinction between the two situations with and without prior knowledge of the population mean parameters. In their monograph they claim that the design, which is optimal in the associated model without random effects, retains its optimality for the estimation of the individual parameters in the hierarchical model. This statement seems to be motivated by a conditional approach, in which the individual response is estimated by the observations of the corresponding individual only, neglecting the fact that the individual effects are assumed to be random and stem from a common population. Some other related findings are given by Candel (2009), who considered optimal designs for prediction of random effects in the particular case of polynomial growth curves.

As often in live the truth lies somewhere in between: In the following we will see that the prediction of the individual parameters leads to compound criteria, which constitute a compromise between the Bayesian optimal design and the optimal design for the model without random effects.

Here we will mainly focus on the IMSE (integrated mean squared error) criterion, which aims at minimizing the expected quadratic deviation of the predicted response curves from the corresponding actual individual curves, averaged over all subjects or units. As an alternative we also introduce a modification of the most popular D-criterion to adapt it to the needs of prediction, if the model may contain also some parameters, which are fixed across all individuals. The modified D-criterion can then be interpreted to aim at minimizing the volume of the prediction ellipsoid within the relevant subspace. In both cases we end up with a compound criterion (see Cook and Wong (1994)), i. e. a weighted average of the fully Bayesian criterion and the criterion related to the model without random effects.

The paper is organized as follows: In the second section we specify the model, present the prediction for the individual random effects and develop the mean squared error matrix for the prediction. In the third section design criteria are introduced and analytical results are presented for the characterization of optimal designs for prediction. These results are illustrated by an example in Section 4, which ahows that the performance of the Bayesian optimal design may break down in some situations. Section 5 provides some tools for the construction of optimal designs by means of in- and equivariance considerations and some illustrative examples. We conclude the paper with some discussion in the final Section 5. Proofs are deferred to an Appendix.

2 Model Specification and Prediction

For random coefficient regression the observations are specified by a hierarchical model. First, on the individual level the jth observation Y_{ij} of individual i is given by

$$Y_{ij} = \mathbf{f}(x_{ij})^\top \boldsymbol{\beta}_i + \varepsilon_{ij} \tag{1}$$

for $j = 1, .., m_i$ and $i = 1, .., n$, where m_i is the number of observations at individual i, n is the number of individuals, and $\mathbf{f} = (f_1, ..., f_p)^\top$ is a set of known regression functions. The values x_{ij} of the explanatory variables may range over the experimental region $\mathcal{X}$. The observational errors ε_{ij} are assumed to be centered with zero mean and homoscedastic and uncorrelated with common variance $\sigma^2 > 0$.

Second, on the population level it is supposed that the individual parameters $\boldsymbol{\beta}_i = (\beta_{i1}, ..., \beta_{ip})^\top$ are realizations from a common distribution with unknown population mean $\mathrm{E}(\boldsymbol{\beta}_i) = \boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\top$ and a $p \times p$ population covariance matrix $\mathrm{Cov}(\boldsymbol{\beta}_i) = \sigma^2 \mathbf{D}$.

Moreover, all individual parameters $\boldsymbol{\beta}_i$ and all observational errors ε_{ij} are assumed to be uncorrelated.

To simplify the presentation we will restrict ourselves here to random coefficient regression models, in which all individuals are observed under the same regime, i. e. all individuals i have the same number $m_i = m$ of observations at the same values $x_{ij} = x_j$ of the explanatory variables. Then the individual vector $\mathbf{Y}_i = (Y_{i1}, ..., Y_{im})^\top$ of observations can be specified by the individual linear model equation in vector notation,

$$\mathbf{Y}_i = \mathbf{F}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i \tag{2}$$

with identical individual design matrices $\mathbf{F} = (\mathbf{f}(x_1), ..., \mathbf{f}(x_m))^\top$ across all individuals. Here $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, ..., \varepsilon_{im})^\top$ is the individual vector of observational errors with covariance matrix $\mathrm{Cov}\,(\boldsymbol{\varepsilon}_i) = \sigma^2\mathbf{I}_m$, where $\mathbf{I}_k$ denotes the $k \times k$ identity matrix. By the hierarchical structure $\mathrm{Cov}\,(\mathbf{Y}_i) = \mathrm{Cov}\,(\mathbf{F}\boldsymbol{\beta}_i) + \mathrm{Cov}\,(\boldsymbol{\varepsilon}_i) = \sigma^2(\mathbf{I}_m + \mathbf{F}\mathbf{D}\mathbf{F}^\top)$, which produces correlations between the observations within an individual and can account for the specific properties of each individual.

Alternatively, by separating the random effects from the population mean the individual vector $\mathbf{Y}_i$ of observations can be written as

$$\mathbf{Y}_i = \mathbf{F}\boldsymbol{\beta} + \mathbf{F}\mathbf{b}_i + \boldsymbol{\varepsilon}_i\,, \tag{3}$$

where $\mathbf{b}_i = \boldsymbol{\beta}_i - \boldsymbol{\beta}$ is the individual effect compared to the population mean. The contribution of the last two terms in (3) can then be interpreted as variance components.

The full vector $\mathbf{Y} = (\mathbf{Y}_1^\top, ..., \mathbf{Y}_n^\top)^\top$ of the observations of all individuals can finally be expressed by the model equation

$$\mathbf{Y} = (\mathbf{I}_n \otimes \mathbf{F})\mathbf{B} + \boldsymbol{\varepsilon} = (\mathbf{1}_n \otimes \mathbf{F})\boldsymbol{\beta} + (\mathbf{I}_n \otimes \mathbf{F})\mathbf{b} + \boldsymbol{\varepsilon}\,, \tag{4}$$

where $\mathbf{B} = (\boldsymbol{\beta}_1^\top, ..., \boldsymbol{\beta}_n^\top)^\top$, $\mathbf{b} = (\mathbf{b}_1^\top, ..., \mathbf{b}_n^\top)^\top$ and $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1^\top, ..., \boldsymbol{\varepsilon}_n^\top)^\top$ are the vectors of all individual coefficients, all individual effects and all observatinal errors, respectively, "$\otimes$" is the common Kronecker product of matrices or vectors, and $\mathbf{1}_k$ denotes the vector of length k with all entries equal to 1. By the independence of the individuals the covariance matrix of the full observational vector $\mathbf{Y}$ is block diagonal with $\mathrm{Cov}\,(\mathbf{Y}) = \sigma^2\mathbf{I}_n \otimes (\mathbf{I}_m + \mathbf{F}\mathbf{D}\mathbf{F}^\top)$.

Using Gauss-Markov theory the best linear unbiased estimator $\hat{\boldsymbol{\beta}}$ of the population parameter $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \Big((\mathbf{1}_n \otimes \mathbf{F})^\top(\mathbf{I}_n \otimes (\mathbf{I}_m + \mathbf{F}\mathbf{D}\mathbf{F}^\top))^{-1}(\mathbf{1}_n \otimes \mathbf{F})\Big)^{-1}(\mathbf{1}_n \otimes \mathbf{F})^\top(\mathbf{I}_n \otimes (\mathbf{I}_m + \mathbf{F}\mathbf{D}\mathbf{F}^\top))^{-1}\mathbf{Y}\,, \tag{5}$$

which obviously simplifies to $\hat{\boldsymbol{\beta}} = (\mathbf{F}^\top(\mathbf{I}_m + \mathbf{F}\mathbf{D}\mathbf{F}^\top)^{-1}\mathbf{F})^{-1}\mathbf{F}^\top(\mathbf{I}_m + \mathbf{F}\mathbf{D}\mathbf{F}^\top)^{-1}\bar{\mathbf{Y}}$, where the term $\bar{\mathbf{Y}} = \frac{1}{n}\sum_{i=1}^n \mathbf{Y}_i$ denotes the averaged response across the individuals.

Employing some matrix algebra this estimator is readily seen to coincide with the ordinary least squares estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top\bar{\mathbf{Y}} \tag{6}$$

and can be rewritten as the average $\hat{\boldsymbol{\beta}} = \frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\beta}}_{i;\mathrm{ind}}$ of the individualized estimates $\hat{\boldsymbol{\beta}}_{i;\mathrm{ind}} = (\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top\mathbf{Y}_i$ of the individual parameters $\boldsymbol{\beta}_i$ based on the observations $\mathbf{Y}_i$ of subject i only (see e. g. Entholzner *et al.* (2005)). It is worth-while mentioning that the best linear unbiased estimator for the population parameter $\boldsymbol{\beta}$ does not depend on the dispersion matrix $\mathbf{D}$.

In the following we will, however, not be interested in estimating the population mean itself but in the prediction ("estimation") of the individual parameters $\boldsymbol{\beta}_i$ and the individual responses $\mu_i(x) = \mathbf{f}(x)^\top\boldsymbol{\beta}$. The subsequent theorem provides a representation of these predictions.

Theorem 1. *In the case of identical individual design matrices* $\mathbf{F}$ *the best linear unbiased prediction*

$$\hat{\boldsymbol{\beta}}_i = \mathbf{D}((\mathbf{F}^\top\mathbf{F})^{-1} + \mathbf{D})^{-1}\hat{\boldsymbol{\beta}}_{i;\mathrm{ind}} + (\mathbf{F}^\top\mathbf{F})^{-1}((\mathbf{F}^\top\mathbf{F})^{-1} + \mathbf{D})^{-1}\hat{\boldsymbol{\beta}}, \tag{7}$$

of the individual parameter $\boldsymbol{\beta}_i$ *is a weighted average of the individualized estimate* $\hat{\boldsymbol{\beta}}_{i;\mathrm{ind}}$ *based on the observations of subject* i *only and the estimator* $\hat{\boldsymbol{\beta}}$ *for the population parameter.*

Note that the contribution of the estimated population mean $\hat{\boldsymbol{\beta}}$ to the individual prediction $\hat{\boldsymbol{\beta}}_i$ essentially diminishes with an increasing number of observations m per individual included in the individual information matrix $\mathbf{F}^\top\mathbf{F}$.

For a regular dispersion matrix $\mathbf{D}$ the representation of the predictors may be simplified.

Corollary 1. *If* $\mathbf{D}$ *is regular, then the best linear unbiased prediction* $\hat{\boldsymbol{\beta}}_i$ *of the individual parameter* $\boldsymbol{\beta}_i$ *is given by*

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{F}^\top\mathbf{F} + \mathbf{D}^{-1})^{-1}(\mathbf{F}^\top\mathbf{F}\,\hat{\boldsymbol{\beta}}_{i;\mathrm{ind}} + \mathbf{D}^{-1}\hat{\boldsymbol{\beta}})\,. \tag{8}$$

As a consequence of Theorem 1 and Corollary 1 the individual prediction may be calculated as a weighted average of the individual observational vector $\mathbf{Y}_i$ and the averaged response $\bar{\mathbf{Y}}$

$$\hat{\boldsymbol{\beta}}_i = \mathbf{D}((\mathbf{F}^\top\mathbf{F})^{-1} + \mathbf{D})^{-1}(\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top\mathbf{Y}_i + (\mathbf{F}^\top\mathbf{F})^{-1}((\mathbf{F}^\top\mathbf{F})^{-1} + \mathbf{D})^{-1}(\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top\bar{\mathbf{Y}} \tag{9}$$

in general and

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{F}^\top\mathbf{F} + \mathbf{D}^{-1})^{-1}(\mathbf{F}^\top\mathbf{Y}_i + \mathbf{D}^{-1}(\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top\bar{\mathbf{Y}})\,, \tag{10}$$

if the dispersion matrix $\mathbf{D}$ is regular.

In any case the best linear unbiased prediction for the individual response $\mu_i(x) = \mathbf{f}(x)^\top\boldsymbol{\beta}_i$ is obtained by $\hat{\mu}_i(x) = \mathbf{f}(x)^\top\hat{\boldsymbol{\beta}}_i$.

Example 1. We consider the particular case of random intercepts (random block effects), where an explicit individual constant term is included for each individual, $f_1(x) \equiv 1$, say. Then the dispersion matrix $\mathbf{D}$ can be written as $\mathbf{D} = d_1\mathbf{e}_1\mathbf{e}_1^\top$, where $\mathbf{e}_1 = (1, 0, ..., 0)^\top$ denotes the first unit vector of length p. Hence, $\mathbf{D}$ has rank one and is not regular, when there are effects of the explanatory variable ($p \geq 2$).

By the alternative representation $\hat{\boldsymbol{\beta}}_i = \mathbf{D}((\mathbf{F}^\top\mathbf{F})^{-1} + \mathbf{D})^{-1}\left(\hat{\boldsymbol{\beta}}_{i;\mathrm{ind}} - \hat{\boldsymbol{\beta}}\right) + \hat{\boldsymbol{\beta}}$ derived from (7) the prediction simplifies to

$$\hat{\boldsymbol{\beta}}_i = \frac{d_1}{1 + md_1}\mathbf{1}_m^\top\left(\mathbf{Y}_i - \bar{\mathbf{Y}}\right)\mathbf{e}_1 + \hat{\boldsymbol{\beta}} \tag{11}$$

in this particular situation. It is easy to verify that only the predicted intercepts differ for different individuals, while the estimates of the others parameters are constant across the individuals. Similarly the predicted responses $\hat{\mu}_i(x)$ only differ by an individual specific constant independent of the value x of the explanatory variables.

The performance of the prediction is measured in terms of the mean squared error matrix $\mathrm{MSE} = \mathrm{Cov}\,(\hat{\mathbf{B}} - \mathbf{B})$ of the predictor $\hat{\mathbf{B}} = (\hat{\boldsymbol{\beta}}_1^\top, ..., \hat{\boldsymbol{\beta}}_n^\top)^\top$ for the vector $\mathbf{B}$ of all individual coefficients. The following theorem provides a useful representation of the mean squared error.

Theorem 2. *The mean squared error of the prediction of the individual coefficients* $\mathbf{B}$ *is given by*

$$\mathrm{MSE} = \sigma^2\left((\mathbf{I}_n - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top) \otimes (\mathbf{D} - \mathbf{D}((\mathbf{F}^\top\mathbf{F})^{-1} + \mathbf{D})^{-1}\mathbf{D}) + \tfrac{1}{n}(\mathbf{1}_n\mathbf{1}_n^\top) \otimes (\mathbf{F}^\top\mathbf{F})^{-1}\right)\,. \tag{12}$$

In the case that the dispersion matrix $\mathbf{D}$ is regular, the mean squared error matrix for the predictors $\hat{\mathbf{B}}$ simplifies to a weighted average of the Bayesian covariance matrix $\sigma^2(\mathbf{F}^\top\mathbf{F}+\mathbf{D}^{-1})^{-1}$ propagated by Gladitz and Pilz (1982) and the covariance matrix $\sigma^2(\mathbf{F}^\top\mathbf{F})^{-1}$ for $\boldsymbol{\beta}$ in the model without random effects ($\mathbf{D}=\mathbf{0}$).

Corollary 2. *If* $\mathbf{D}$ *is regular, then*

$$\mathrm{MSE} = \sigma^2 \left((\mathbf{I}_n - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top) \otimes (\mathbf{F}^\top\mathbf{F} + \mathbf{D}^{-1})^{-1} + \tfrac{1}{n}(\mathbf{1}_n\mathbf{1}_n^\top) \otimes (\mathbf{F}^\top\mathbf{F})^{-1} \right). \tag{13}$$

Example 1 (cont.). In the case of random intercepts the MSE matrix (12) has the form

$$\mathrm{MSE} = \sigma^2 \left(\frac{d_1}{1+md_1} (\mathbf{I}_n - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top) \otimes (\mathbf{e}_1\mathbf{e}_1^\top) + \tfrac{1}{n}(\mathbf{1}_n\mathbf{1}_n^\top) \otimes (\mathbf{F}^\top\mathbf{F})^{-1} \right). \tag{14}$$

3 Optimal Design

The mean squared error matrix derived in Theorem 2 and Corollary 2 depends on the design of the experiment, i. e. on the choice of the experimental settings $x_1, ..., x_m$ for each individual. The quality of an experiment measured by the mean squared error matrix does not depend on the order of the observations. Thus we may rewrite the mean squared error matrix in terms of distinct settings $x_1, ..., x_k$, say, and their respective numbers $m_1, ..., m_k$ of replications ($\sum_{j=1}^k m_j = m$). The individual design can then be defined as

$$\xi = \begin{pmatrix} x_1 & , ..., & x_k \\ m_1 & , ..., & m_k \end{pmatrix}. \tag{15}$$

In this paper we focus on the concept of approximate (continuous) designs in the sense of Kiefer (1974), for which the requirement of integer values for the replication numbers m_j is dropped and only the conditions $m_j \geq 0$ and $\sum_{j=1}^k m_j = m$ have to be satisfied.

Then for any approximate design ξ the individual information matrix is defined by

$$\mathbf{M}(\xi) = \sum_{j=1}^{k} m_j \mathbf{f}(x_j)\mathbf{f}(x_j)^\top, \tag{16}$$

which coincides with the individual information matrix $\mathbf{F}^\top\mathbf{F}$ in the case of an exact design, where each setting x_j is replicated m_j times.

With this notation we can define the mean squared error matrix of an approximate design by

$$\mathrm{MSE}(\xi) = (\mathbf{I}_n - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top) \otimes (\mathbf{D} - \mathbf{D}(\mathbf{M}(\xi)^{-1} + \mathbf{D})^{-1}\mathbf{D}) + \tfrac{1}{n}(\mathbf{1}_n\mathbf{1}_n^\top) \otimes \mathbf{M}(\xi)^{-1} \tag{17}$$

in accordance with the representation given in Theorem 2 for an exact design, when we suppress the constant factor σ^2. As in Corollary 2 this formula may be simplified to

$$\mathrm{MSE}(\xi) = (\mathbf{I}_n - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top) \otimes (\mathbf{M}(\xi) + \mathbf{D}^{-1})^{-1} + \tfrac{1}{n}(\mathbf{1}_n\mathbf{1}_n^\top) \otimes \mathbf{M}(\xi)^{-1} \tag{18}$$

in the case of a regular dispersion matrix $\mathbf{D}$.

As common in experimental design the full mean squared error matrix cannot be minimized with respect to the Loewner ordering in the non-negative definite sense. Therefore we will deal in the following with two optimality criteria, which describe the quality of a design by a real valued

functional of the mean squared error matrix. The first criterion ("IMSE") is rather appealing and aims at minimizing the expected mean squared (L_2-)distance of the predicted from the actual individual response function μ_i over the design region $\mathcal{X}$. The second criterion ("D-criterion") is prepared to extend the determinant criterion, which is most commonly used in optimal design theory because of its nice analytical properties and which aims at minimizing the volume of the prediction ("confidence") ellipsoid in the case of Gaussian errors. Special care has to be taken with this extension in the case of a non-regular dispersion matrix $\mathbf{D}$, because then the prediction ellipsoid is concentrated in a lower-dimensional subspace and minimization of the volume has to be done with respect to this subspace.

3.1 The IMSE-criterion

The version of the Integrated Mean Squared Error ("IMSE") of prediction, which will be used here, is defined as the sum over all individuals

$$\mathrm{IMSE}_{\mathrm{pred}} = \sum_{i=1}^{n} \mathrm{E}\left(\int_{\mathcal{X}} (\hat{\mu}_i(x) - \mu_i(x))^2 \nu(\mathrm{d}x) \right) \tag{19}$$

of the expected (L_2-)distances of the predicted response $\hat{\mu}_i = \mathbf{f}^\top \hat{\boldsymbol{\beta}}_i$ from the individual response $\mu_i = \mathbf{f}^\top \boldsymbol{\beta}_i$ with respect to some suitable measure ν on the experimental region $\mathcal{X}$, which is typically chosen to be uniform on $\mathcal{X}$ with $\nu(\mathcal{X}) = 1$. Similar to the integrated mean squared error of estimation this magnitude may be rearranged to a linear criterion

$$\mathrm{IMSE}_{\mathrm{pred}} = \mathrm{tr}\left(\mathrm{MSE} \cdot (\mathbf{I}_n \otimes \mathbf{V})\right) \tag{20}$$

in the MSE matrix, where $\mathbf{V} = \int_{\mathcal{X}} \mathbf{f}(x)\mathbf{f}(x)^\top \nu(\mathrm{d}x)$ is the "information matrix" for the weight distribution ν in the fixed effects model and "tr" denotes the trace of the matrix. With this relation the IMSE-criterion for prediction will be defined for an approximate design ξ by

$$\mathrm{IMSE}_{\mathrm{pred}}(\xi) = \mathrm{tr}\left(\mathrm{MSE}\,(\xi) \cdot (\mathbf{I}_n \otimes \mathbf{V})\right). \tag{21}$$

By Theorem 2 it follows that

$$\mathrm{IMSE}_{\mathrm{pred}}(\xi) = \mathrm{tr}\left((\mathbf{M}(\xi)^{-1} + (n-1)(\mathbf{D} - \mathbf{D}(\mathbf{M}(\xi)^{-1} + \mathbf{D})^{-1}\mathbf{D})\mathbf{V}\right). \tag{22}$$

This simplifies by Corollary 2 to

$$\mathrm{IMSE}_{\mathrm{pred}}(\xi) = \mathrm{tr}\left(\mathbf{M}(\xi)^{-1}\mathbf{V}\right) + (n-1)\,\mathrm{tr}\left((\mathbf{M}(\xi) + \mathbf{D}^{-1})^{-1}\mathbf{V}\right), \tag{23}$$

which is proportional to a weighted average of the IMSE-criterion $\mathrm{tr}\left(\mathbf{M}(\xi)^{-1}\mathbf{V}\right)$ in the model without random effects and the corresponding Bayesian IMSE-criterion $\mathrm{tr}\left((\mathbf{M}(\xi) + \mathbf{D}^{-1})^{-1}\mathbf{V}\right)$. Thus the IMSE-criterion for prediction can be interpreted as a compound criterion, which can equivalently be identified with a certain constrained criterion according to Cook and Wong (1994).

The IMSE-criterion for prediction is easily seen to be convex and, hence, the optimal designs can be cahracterized by means of a standard equivalence theorem.

Theorem 3. *The approximate design ξ^* is IMSE-optimal for prediction if and only if*

$$\begin{aligned} &\mathbf{f}(x)^\top \mathbf{M}(\xi^*)^{-1}\mathbf{V}\mathbf{M}(\xi^*)^{-1}\mathbf{f}(x) \\ &\quad + (n-1)\,\mathbf{f}(x)^\top \mathbf{M}(\xi^*)^{-1}(\mathbf{M}(\xi^*)^{-1} + \mathbf{D})^{-1}\mathbf{D}\mathbf{V}\mathbf{D}(\mathbf{M}(\xi^*)^{-1} + \mathbf{D})^{-1}\mathbf{M}(\xi^*)^{-1}\mathbf{f}(x) \\ \leq\ &\frac{1}{m}(\mathrm{tr}\left(\mathbf{M}(\xi^*)^{-1}\mathbf{V}\right) + (n-1)\,\mathrm{tr}\left(\mathbf{D}(\mathbf{M}(\xi^*)^{-1} + \mathbf{D})^{-1}\mathbf{M}(\xi^*)^{-1}(\mathbf{M}(\xi^*)^{-1} + \mathbf{D})^{-1}\mathbf{D}\mathbf{V}\right)) \end{aligned} \tag{24}$$

for all $x \in \mathcal{X}$. *Moreover, for every support point* x_j *of* ξ^* *with positive weight* $(m_j > 0)$ *equality holds in* (24).

For a regular dispersion matrix $\mathbf{D}$ the condition (24) of Theorem 3 simplifies to that of a compound criterion.

Corollary 3. *If* $\mathbf{D}$ *is regular, the approximate design* ξ^* *is IMSE-optimal for prediction if and only if*

$$\begin{aligned} &\mathbf{f}(x)^\top \mathbf{M}(\xi^*)^{-1}\mathbf{V}\mathbf{M}(\xi^*)^{-1}\mathbf{f}(x) + (n-1)\,\mathbf{f}(x)^\top (\mathbf{M}(\xi^*)+\mathbf{D}^{-1})^{-1}\mathbf{V}(\mathbf{M}(\xi^*)+\mathbf{D}^{-1})^{-1}\mathbf{f}(x) \\ &\quad \leq \frac{1}{m}(\operatorname{tr}(\mathbf{M}(\xi^*)^{-1}\mathbf{V}) + (n-1)\operatorname{tr}((\mathbf{M}(\xi^*)+\mathbf{D}^{-1})^{-1}\mathbf{M}(\xi^*)(\mathbf{M}(\xi^*)+\mathbf{D}^{-1})^{-1}\mathbf{V})) \end{aligned} \quad (25)$$

for all $x \in \mathcal{X}$. *Moreover, for any support point* x_j *of* ξ^* *with positive weight* $(m_j > 0)$ *equality holds in* (25).

Example 1 (cont.). In the case of random intercepts the IMSE-criterion

$$\mathrm{IMSE}_{\mathrm{pred}}(\xi) = \operatorname{tr}(\mathbf{M}(\xi)^{-1}\mathbf{V}) + \frac{(n-1)d_1}{1+md_1}\nu(\mathcal{X}), \quad (26)$$

depends on the dispersion matrix only through an additive constant term.

Corollary 4. *The IMSE-optimal design in the fixed effects model is IMSE-optimal for prediction in the random intercepts model.*

However, if there is a more general influence of the random coefficients on terms, which include effects of the explanatory variables x, the IMSE-optimal design for prediction may depend substantially on the dispersion matrix $\mathbf{D}$, as will be exposed in section 4.

3.2 The *D*-criterion

For estimation the most popular design criterion is the D-criterion, which aims at maximizing the determinant of the information matrix or, equivalently, at minimizing the determinant of the covariance matrix. For prediction we adopt this criterion by minimizing the determinant of the mean squared error (MSE) matrix. However, this direct approach only makes sense in the case of a regular dispersion matrix $\mathbf{D}$. As usual a logarithmic version of the criterion will be employed to retain convexity,

$$\mathrm{D}_{\mathrm{pred}}(\xi) = \ln\det \mathrm{MSE}(\xi). \quad (27)$$

This D-criterion may be represented in the form

$$\mathrm{D}_{\mathrm{pred}}(\xi) = \ln\det(\mathbf{M}(\xi)^{-1}) + (n-1)\ln\det((\mathbf{M}(\xi)+\mathbf{D}^{-1})^{-1}). \quad (28)$$

For a general dispersion matrix $\mathbf{D}$ this criterion has to be adjusted properly: While from an interpretation point of view the D-criterion aims at minimizing the volume of a prediction ellipsoid for the individual parameters, it has to be taken into account that for a singular dispersion matrix $\mathbf{D}$ the prediction ellipsoid is concentrated in a lower dimensional subspace. Therefore we will make use of the property that for positive definite symmetric matrix the determinant is equal to the product of the eigenvalues. Thus for a subspace we have to consider the relevant (non-zero) eigenvalues corresponding to the eigenvectors spanning that subspace. If the dispersion matrix $\mathbf{D}$ is a non-zero matrix ($\mathbf{D} \neq \mathbf{0}$) with rank q, $1 \leq q \leq p$, than for the individual parameters $\mathbf{B}$ the prediction ellipsoid is of dimension $(n-1)q+p$. Consequently, we define the

(modified) D-criterion for prediction as the logarithm of the product of the $(n-1)q+p$ largest eigenvalues of the MSE matrix

$$\mathrm{D}_{\mathrm{pred}}(\xi) = \ln\det(\mathbf{M}(\xi)^{-1}) + (n-1)\ln\prod_{l=1}^{q}\lambda_l(\xi,\mathbf{D})\,, \tag{29}$$

where $\lambda_1(\xi,\mathbf{D}),...,\lambda_q(\xi,\mathbf{D})$ are the q largest eigenvalues of the matrix $\mathbf{D}-\mathbf{D}(\mathbf{M}(\xi)^{-1}+\mathbf{D})^{-1}\mathbf{D}$. Note that for a regular individual information matrix $\mathbf{M}(\xi)$ these eigenvalues are positive. In the regular case ($q=p$) definition (29) is in accordance with (28).

As for the IMSE-criterion the D-criterion can be recognized as a compound criterion, which is a weighted average of the D-criterion $-\ln\det\mathbf{M}(\xi)$ in the fixed effects model and the Bayesian D-criterion $-\ln\det(\mathbf{M}(\xi)+\mathbf{D}^{-1})$. Then a D-optimal design can be once again characterized by means of an equivalence theorem.

Theorem 4. *The approximate design ξ^* is D-optimal for prediction if and only if*

$$\begin{aligned}&\mathbf{f}(x)^\top\mathbf{M}(\xi^*)^{-1}\mathbf{f}(x) + (n-1)\,\mathbf{f}(x)^\top(\mathbf{D}-\mathbf{D}(\mathbf{M}(\xi^*)^{-1}+\mathbf{D})^{-1}\mathbf{D})\mathbf{f}(x)\\ &\quad\le\ \frac{1}{m}(p+(n-1)\,\mathrm{tr}\,(\mathbf{D}(\mathbf{M}(\xi^*)^{-1}+\mathbf{D})^{-1}))\end{aligned} \tag{30}$$

for all $x\in\mathcal{X}$. Moreover, for any support point x_j of ξ^* with positive weight ($m_j>0$) equality holds in (30).

For regular dispersion matrices $\mathbf{D}$ condition (30) of Theorem 4 simplifies,

Corollary 5. *If $\mathbf{D}$ is regular, the approximate design ξ^* is D-optimal for prediction if and only if*

$$\begin{aligned}&\mathbf{f}(x)^\top\mathbf{M}(\xi^*)^{-1}\mathbf{f}(x) + (n-1)\,\mathbf{f}(x)^\top(\mathbf{D}^{-1}+\mathbf{M}(\xi^*))^{-1}\mathbf{f}(x)\\ &\quad\le\ \frac{1}{m}(p+(n-1)\,\mathrm{tr}\,((\mathbf{D}^{-1}+\mathbf{M}(\xi^*))^{-1}\mathbf{M}(\xi^*)))\end{aligned} \tag{31}$$

for all $x\in\mathcal{X}$. Moreover, for any support point x_j of ξ^* with positive weight ($m_j>0$) equality holds in (31).

Example 1 (cont.). In the particular case of random intercepts the D-criterion simplifies to

$$\mathrm{D}_{\mathrm{pred}}(\xi) = \ln\det(\mathbf{M}(\xi)^{-1}) + (n-1)\ln\frac{d_1}{1+md_1}\,, \tag{32}$$

which establishes the following result.

Corollary 6. *The D-optimal design in the fixed effects model is D-optimal for prediction in the random intercepts model.*

4 Example: Linear Regression with Random Slope

While for models with random intercepts the dispersion matrix $\mathbf{D}$ has no influence on the optimal design for the criteria considered, the impact of the dispersion matrix $\mathbf{D}$ may become crucial in situations, where the random coefficients are associated with the effects of explanatory variables. This will be illustrated by the simple case of linear regression with random slope,

$$Y_{ij} = \beta_{i1} + \beta_{i2}x_j + \varepsilon_{ij}\,, \tag{33}$$

on the experimental region $\mathcal{X} = [0, 1]$. For the IMSE-criterion we will consider a uniform weighting $\nu = \lambda|_{[0,1]}$ throughout, i. e. $\mathbf{V} = \int_0^1 \mathbf{f}(x)\mathbf{f}(x)^\top \mathrm{d}x$. It has to be noted that the optimal designs obtained in this section require a prior guess of the dispersion matrix $\mathbf{D}$ and are, hence, only locally optimal.

By the Theorems 3 and 4 the sensitivity functions of the IMSE- and the (modified) D-criterion are parabolas with positive leading coefficients, and, hence, the optimal designs ξ^* have only the two support points $x = 0$ and $x = 1$ on the boundary of the interval and are of the general form

$$\xi = \begin{pmatrix} 0 & 1 \\ m - m_1 & m_1 \end{pmatrix}, \tag{34}$$

where m_1 denotes the (generalized) number of replications at $x = 1$. For design optimality only the optimal number m_1^* of replications at $x = 1$ has to be determined.

For the model without random effects ($\mathbf{D} = \mathbf{0}$) it is well-known that the design, which assigns equal weights $m_1^* = m/2$ and $m_0^* = 1 - m_1^* = m/2$ to both endpoints, is optimal for both criteria under consideration. Because of Corollary 4 and 6 this remains true for prediction in models with solely random intercepts.

In the following we consider the case of random slopes for illustrative purposes, where we assume for simplicity a constant intercept across the individuals, i. e. $\beta_{i1} \equiv \beta_1$ and $\mathbf{D} = d_2 \mathbf{e}_2 \mathbf{e}_2^\top$, where $\mathbf{e}_2 = (0, 1)^\top$ is the second unit vector of length $p = 2$. This model assumption may seem to be inadequate for practical applications, but it highlights the behavior, when the variability of the slopes (effect sizes) is large compared to the variability of the intercepts (baselines). According to the criterion functions (22) and (29), the corresponding IMSE- and D-criteria are given by the following formulas:

$$\mathrm{IMSE}_{\mathrm{pred}}(\xi) = \frac{1}{3}\left(\frac{m}{m_1(m - m_1)} + (n - 1)\frac{d_2}{1 + m_1 d_2}\right), \tag{35}$$

$$\mathrm{D}_{\mathrm{pred}}(\xi) = \ln \frac{1}{m_1(m - m_1)} + (n - 1) \ln \frac{d_2}{1 + m_1 d_2}. \tag{36}$$

It is noteworthy that the optimal numbers m_1^* as well as the optimal proportions $w_1^* = m_1^*/m$ do not only depend on the dispersion matrix $\mathbf{D}$ but also on the number m of observations per individual and even on the sample size n. For our numerical calculations we fix the sample size to $n = 100$ individuals and the intra-individual number of observations per subject to $m = 10$. Then the optimal values for m_1^* can be determined in dependence on the variability d_2 of the slopes. These values are depicted in Figure 1 against the standardized variance parameter $\rho = d_2/(1 + d_2)$, which mimics the intra-class correlation and has been chosen instead of the raw slope variance d_2 to cover the whole range of possible values by a finite interval ($0 \leq \rho < 1$). Note that ρ is a monotonically increasing transformation of d_2, which preserves qualitative behaviors.

For the IMSE-criterion the optimal number m_1^* (solid line) increases in d_2 from $m_1^* = m/2 = 5$ for $d_2 = 0$ ($\rho = 0$) to $m_1^* \approx 0.91 \cdot m = 9.1$ for $d_2 \to \infty$ ($\rho \to 1$). The D-optimal designs assign an optimal number m_1^* (dashed line) increasing from $m_1^* \approx m/2$ for small values of the slope variance ($d_2 \to 0$) to $m_1^* \approx 0.99 \cdot m = 9.9$ for $d_2 \to \infty$. (Note that $d_2 = 0$ would lead to a different criterion.)

The next graphic (Figure 2) exhibits the efficiency of the equi-replicated design ($m_0 = m_1 = m/2$) for varying values of ρ. For $d_2 = 0$ (resp. $d_2 \to 0$) the efficiency equals 1 for both criteria, and the efficiency decreases, as the dispersion d_2 increases, with limiting values 0.60 for the IMSE-criterion (solid line) and 0.53 for the modified D-criterion (dashed line) when $d_2 \to \infty$. The latter efficiency is computed as

$$\mathrm{eff}_{\mathrm{D}}(\xi) = \left(\frac{\exp(\mathrm{D}_{\mathrm{pred}}(\xi^*))}{\exp(\mathrm{D}_{\mathrm{pred}}(\xi))}\right)^{\frac{1}{(n-1)q+p}} = \left(\frac{\det(\mathbf{M}(\xi^*))^{-1}(\prod_{k=1}^q \lambda_k(\xi^*, \mathbf{D}))^{n-1}}{\det(\mathbf{M}(\xi))^{-1}(\prod_{k=1}^q \lambda_k(\xi, \mathbf{D}))^{n-1}}\right)^{\frac{1}{(n-1)q+p}}, \tag{37}$$

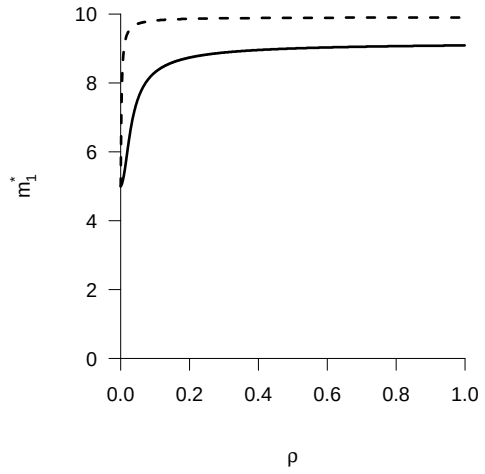

Figure 1: Optimal numbers m_1^*: IMSE-criterion (solid line) D-criterion (dashed line)

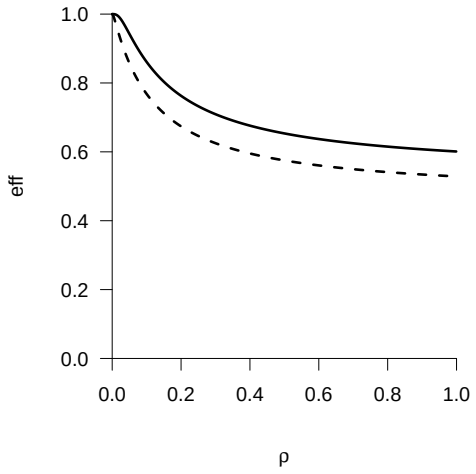

Figure 2: Efficiency of the equi-replicated design ($m_1 = m/2$): IMSE-criterion (solid line) D-criterion (dashed line)

while for the IMSE-criterion the standard formula

$$\text{eff}_{\text{IMSE}}(\xi) = \frac{\text{IMSE}_{\text{pred}}(\xi^*)}{\text{IMSE}_{\text{pred}}(\xi)} \tag{38}$$

is used. Note that these efficiencies are designed to be homogeneous such that the efficiency of a design ξ describes the proportion of the sample size required for the optimal design ξ^* compared to ξ to obtain the same precision measured in terms of the corresponding design criterion. Or, equivalently, the deficiency of a design $\text{def}(\xi) = 1/\text{eff}(\xi) - 1$, i. e. the surplus of the reciprocal over 100 %, measures the additional proportion in sample size required for ξ compared to the optimal design ξ^*.

5 Some Invariance Considerations in Design for Prediction

In classical design theory for fixed effect models a commonly proposed advantage of the D- and the IMSE-criterion is their invariance with respect to reparameterizations. This fact may also be used to cover relabeling ang rescaling of the explanatory variables, where appropriate. In particular, rescaling can be used to derive optimal designs for some suitable standardized experimental region and map them to general regions by linear transformations, which are compatible with the regression functions. For example, this approach works for complete polynomial regression models, in which any monomial term is accompanied by all monomials of lower degree and an intercept, or for analysis of variance models, in which any interaction term comes with its lower degree interactions and the corresponding main effects. However, for mixed linear models of the random coefficient regression type considered here an additional induced transformation of the dispersion matrix $\mathbf{D}$ has to be taken into account. This is true for the problem of estimating the population parameters $\boldsymbol{\beta}$ (see Graßhoff *et al.* (2012)) and similarly for the present problem of prediction of the individual parameters $\boldsymbol{\beta}_i$ or the individual responses μ_i.

To see this we consider a one-to-one transformation $g : \mathcal{X} \to \mathcal{X}_1$ of the experimental region $\mathcal{X}$ onto its image $\mathcal{X}_1 = g(\mathcal{X})$, and we assume that the regression function $\mathbf{f}$ is defined simultaneously

on both experimental regions $\mathcal{X}$ and $\mathcal{X}_1$. Further we require the existence of a regular $p \times p$ matrix $\mathbf{Q}_g$ such that

$$\mathbf{f}(g(x)) = \mathbf{Q}_g \, \mathbf{f}(x) \tag{39}$$

for all $x \in \mathcal{X}$. Then the regression function $\mathbf{f}$ will be said to be linearly equivariant with respect to the transformation g (see Schwabe (1996), chapter 3).

Next we define the corresponding transformation for designs

$$g : \xi = \begin{pmatrix} x_1, ..., x_k \\ m_1 \, , ..., \, m_k \end{pmatrix} \to \xi^g = \begin{pmatrix} g(x_1), ..., g(x_k) \\ m_1 \, , ..., \, m_k \end{pmatrix}, \tag{40}$$

which transports the frequencies m_j to the image $g(x_j)$ of the associated design point x_j. Note that ξ^g is the measure theoretic image of ξ under the mapping g. For this induced design ξ^g it follows from the definition (17) of the mean squared error that

$$\mathrm{MSE}(\xi^g, \mathbf{D}_g) = (\mathbf{I}_n \otimes \mathbf{Q}_g^{-\top}) \, \mathrm{MSE}(\xi, \mathbf{D}) \, (\mathbf{I}_n \otimes \mathbf{Q}_g^{-1}), \tag{41}$$

where $\mathbf{D}_g = \mathbf{Q}_g^{-\top} \mathbf{D} \mathbf{Q}_g^{-1}$ is the corresponding induced dispersion matrix and $\mathbf{Q}_g^{-\top} = (\mathbf{Q}_g^{\top})^{-1} = (\mathbf{Q}_g^{-1})^{\top}$ denotes the inverse of $\mathbf{Q}_g^{\top}$ for brevity. In (41) we have added explicitly the dispersion matrix $\mathbf{D}$ and $\mathbf{D}_g$, respectively, to the definition of the mean squared error in order to indicate the influence of the dispersion matrix on the MSE.

Theorem 5. *If the design ξ^* is D-optimal for prediction for the dispersion matrix $\mathbf{D}$ on the experimental region $\mathcal{X}$, then the design $(\xi^*)^g$ is D-optimal for prediction for the dispersion matrix $\mathbf{D}_g$ on the experimental region $\mathcal{X}_1 = g(\mathcal{X})$.*

For the IMSE-criterion also the weighting measure ν has to be transformed to its image ν^g. Denote by $\mathbf{V}_g = \int_{g(\mathcal{X})} \mathbf{f}(x)\mathbf{f}(x)^{\top} \nu^g(\mathrm{d}x)$ the "information matrix" for the weight distribution ν^g in the induced fixed effects model on $\mathcal{X}_1 = g(\mathcal{X})$. Then $\mathbf{V}_g = \mathbf{Q}_g \mathbf{V} \mathbf{Q}_g^{\top}$ and by (41) the IMSE is preserved, $\mathrm{IMSE}_{\mathrm{pred}}(\xi^g, \mathbf{D}_g, \nu^g) = \mathrm{IMSE}_{\mathrm{pred}}(\xi, \mathbf{D}, \nu)$, where we have added the weighting measure ν resp. ν^g for clarification to the notation of the IMSE. This obviously establishes the following result.

Theorem 6. *If the design ξ^* is IMSE-optimal for prediction for the dispersion matrix $\mathbf{D}$ on the experimental region $\mathcal{X}$ with respect to ν, then the design $(\xi^*)^g$ is IMSE-optimal for prediction for the dispersion matrix $\mathbf{D}_g$ on the experimental region $\mathcal{X}_1 = g(\mathcal{X})$ with respect to ν^g.*

By formula (40) the induced optimal design $(\xi^*)^g$ has the form

$$\xi_g^* = \begin{pmatrix} g(x_1^*), ..., g(x_k^*) \\ m_1^* \, , ..., \, m_k^* \end{pmatrix}, \tag{42}$$

where $x_1^*, ..., x_k^*$ and $m_1^*, ..., m_k^*$ are the design points and the associated frequencies of the optimal design ξ^* on the experimental region $\mathcal{X}$.

In general the original and the induced dispersion matrix will differ (see Example 4 below), but for random intercepts the situation simplifies.

Example 1 (cont.). In the case of random intercepts the original and the induced dispersion matrix coincide ($\mathbf{D}_g = \mathbf{D} = d_1 \mathbf{e}_1 \mathbf{e}_1^{\top}$). This is in accordance with the findings that for fixed effects models the IMSE- and D-optimal designs, which are also optimal for random intercepts, can be obtained by transformation.

The following example of a scale transformation illustrates the possibility to obtain an optimal design for a more general experimental region from known results for a standard one.

Example 2. We consider again the model (33) of simple linear regression ($Y_{ij} = \beta_{i1} + \beta_{i2} x_j + \varepsilon_{ij}$) on the experimental region $\mathcal{X} = [0, 1]$. For the IMSE we consider again uniform weighting, $\nu = \lambda|_{[0,1]}$. According to Section 4, the optimal design ξ^* is of the form (34) for every dispersion matrix $\mathbf{D}$ and, hence, concentrated on the endpoints 0 and 1 of the experimental region with corresponding optimal weights $m_0^* = m - m_1^*$ and m_1^*, respectively, depending on $\mathbf{D}$. We further assume that both individual parameters β_{i1} and β_{i2} are random and independent of each other, i. e. the dispersion matrix $\mathbf{D} = \operatorname{diag}(d_1, d_2)$ is diagonal, where the diagonal entries d_1 and d_2 are the variances of the intercepts and the slopes, respectively. Optimal designs can be obtained numerically for any given d_1, d_2, m, and n.

Next we consider the same simple linear regression model (33), but now on the more general experimental region $\mathcal{X}_1 = [0, a]$ for some $a > 0$, which can be obtained from $[0, 1]$ by rescaling, $g(x) = ax$. The induced dispersion matrix $\mathbf{D_g} = \operatorname{diag}(d_1, d_2/a^2)$ is also diagonal. Moreover the uniform weighting ν is transformed to $\nu^g = \lambda|_{[0,a]}/a$, which is itself uniform (up to the scaling factor $1/a$).

Vice versa, if we consider a general diagonal dispersion matrix $\mathbf{D}_a = \operatorname{diag}(d_{a1}, d_{a2})$ on $[0, a]$, this can be obtained as the induced dispersion matrix, if we start with $\mathbf{D} = \operatorname{diag}(d_{a1}, a^2 d_{a2})$ on the standardized region. It follows from Theorems 5 and 6 that the designs, which are IMSE- and D-optimal for prediction in (33) on the experimental region $\mathcal{X} = [0, a]$, are concentrated on the endpoints 0 and a and that the corresponding optimal frequencies can be obtained as those in the standardized case for $\mathbf{D} = \operatorname{diag}(d_{a1}, a^2 d_{a2})$.

As in fixed effects models one may be tempted to employ symmetry to characterize optimal designs, if there is a (finite) group G of transformations $g : \mathcal{X} \to \mathcal{X}$ of the experimental region onto itself, which satisfy (39) (see Schwabe (1996), chapter 3) and, additionally, leave the dispersion matrix invariant, i. e. $\mathbf{D}_g = \mathbf{D}$ for all $g \in G$.

Theorem 7. *Let G be a finite group of transformations on $\mathcal{X}$, for which the regression function $\mathbf{f}$ is linearly equivariant and for which the dispersion matrix $\mathbf{D}$ is invariant. If ξ^* is D-optimal on $\mathcal{X}$ for prediction with dispersion $\mathbf{D}$, then the design $\bar{\xi}^* = \frac{1}{\#G} \sum_{g \in G} (\xi^*)^g$ is also D-optimal on $\mathcal{X}$ for prediction with dispersion $\mathbf{D}$.*

For the IMSE-criterion we additionally need the invariance of the weighting measure, i. e. $\nu^g = \nu$ for all $g \in G$.

Theorem 8. *Let G be a finite group of transformations on $\mathcal{X}$, for which the regression function $\mathbf{f}$ is linearly equivariant and for which the dispersion matrix $\mathbf{D}$ and the weighting measure ν are invariant. If ξ^* is IMSE-optimal for prediction for dispersion $\mathbf{D}$ on $\mathcal{X}$ with respect to ν, then the design $\bar{\xi}^* = \frac{1}{\#G} \sum_{g \in G} (\xi^*)^g$ is also IMSE-optimal for prediction for dispersion $\mathbf{D}$ on $\mathcal{X}$ with respect to ν.*

Obviously the symmetrized design $\bar{\xi}^*$ is invariant with respect to G, i. e. $\bar{\xi}^* = (\bar{\xi}^*)^g$ for all $g \in G$. This result allows to reduce the search for an optimal design to the essentially complete class of invariant designs, if the conditions of Theorems 7 and 8 are met.

Example 3. Here we consider the simple linear regression model (33) on the standard symmetric experimental region $[-1, 1]$ and we will make use of the symmetry to determine optimal designs. For this the dispersion matrix $\mathbf{D}$ of the random parameters is required to be diagonal.

Similar argumentation as in Section 4 can be employed to show that the IMSE- and D-optimal designs have the form

$$\xi = \begin{pmatrix} -1 & 1 \\ m - m_1 & m_1 \end{pmatrix}. \tag{43}$$

First we consider the group $G = \{id, g\}$ induced by the sign change $g(x) = -x$, and $id(x) = x$ denotes the identity. With $\mathbf{Q_g} = \mathrm{diag}\,(1, -1)$ and $\mathbf{Q}_{id}$ equal to the identity matrix the dispersion matrix is invariant ($\mathbf{D_g} = \mathbf{D}$), if $\mathbf{D}$ is diagonal. The symmetrization of any design of the required form (43) is again of the same form and results in the invariant design

$$\xi^* = \begin{pmatrix} -1 & 1 \\ m/2 & m/2 \end{pmatrix}, \tag{44}$$

which is D- and IMSE-optimal because of Theorems 7 resp. 8, when ν is uniform.

This result can now be used to obtain optimal designs on other intervals for specific dispersion matrices.

Example 4. As in Example 2 it can be derived from Example 3 that the design $\xi^* = \begin{pmatrix} -a & a \\ m/2 & m/2 \end{pmatrix}$ is IMSE- (uniform ν) and D-optimal for prediction in the simple linear regression model (33) on the experimental region $[-a, a]$ for any $a > 0$, if the dispersion matrix of the random parameters is diagonal.

Moreover, if we transform the model (33) from the interval $[-1, 1]$ to some more general experimental region $[b, c]$ with $c > b$, then the design $\xi_g^* = \begin{pmatrix} b & c \\ m/2 & m/2 \end{pmatrix}$ will be IMSE- (uniform ν) and D-optimal for prediction, if the dispersion matrix has the specific form

$$\mathbf{D}_g = \begin{pmatrix} d_1 + d_2(c+b)^2/(c-b)^2 & -2d_2(c+b)/(c-b)^2 \\ -2d_2(c+b)/(c-b)^2 & 4d_2/(c-b)^2 \end{pmatrix}, \tag{45}$$

which is induced by $g(x) = \frac{b+c}{2} + \frac{c-b}{2}x$ from $\mathbf{D} = \mathrm{diag}\,(d_1, d_2)$. Note that in the special case $b = 0$ and $c = a$ this induced dispersion matrix of the random parameters can be written as

$$\mathbf{D}_g = \begin{pmatrix} d_1 + d_2 & -2d_2/a \\ -2d_2/a & 4d_2/a^2 \end{pmatrix}, \tag{46}$$

which simplifies to

$$\mathbf{D}_g = \begin{pmatrix} d_1 + d_2 & -2d_2 \\ -2d_2 & 4d_2 \end{pmatrix} \tag{47}$$

in the case $a = 1$ for the standard interval $[0, 1]$. This means that the design $\xi^* = \begin{pmatrix} 0 & 1 \\ m/2 & m/2 \end{pmatrix}$ is IMSE- and D-optimal for prediction in (33) on $[0, 1]$ if the dispersion matrix is of the form (47).

Finally we extend the results to a slightly more complicated model.

Example 5. For the quadratic regression model with random parameters

$$Y_{ij} = \beta_{i1} + \beta_{i2}x_j + \beta_{i3}x_j^2 + \varepsilon_{ij} \tag{48}$$

we confine our interest to optimal designs for prediction on the standard symmetric design region $\mathcal{X} = [-1, 1]$ with a chessboard structured dispersion matrix

$$\mathbf{D} = \begin{pmatrix} d_{11} & 0 & d_{13} \\ 0 & d_{22} & 0 \\ d_{13} & 0 & d_{33} \end{pmatrix}. \tag{49}$$

Again we make use of the sign change group introduced in Example 3, where now $\mathbf{Q}_g = \text{diag}\,(1, -1, 1)$, which leaves both the dispersion matrix $\mathbf{D}$ and the uniform weighting ν invariant.

By standard arguments we can conclude from Theorems 3 and 4 that the IMSE- and D-optimal designs are supported by, at most, three design points including the two endpoints of the interval, and by Theorems 7 and 8 that the symmetric optimal design has the form

$$\xi^* = \begin{pmatrix} -1 & 0 & 1 \\ m^* & m - 2m^* & m^* \end{pmatrix}, \tag{50}$$

where the optimal frequency m^* at $x = 1$ and $x = -1$ depends on the design criterion as well as on the entries of the dispersion matrix $\mathbf{D}$ and on the number of individuals n.

As an easy consequence the IMSE- (with uniform ν) and D-optimal designs for prediction on any symmetric interval $[-a, a]$ will be of the form

$$\xi^* = \begin{pmatrix} -a & 0 & a \\ m^* & m - 2m^* & m^* \end{pmatrix}. \tag{51}$$

6 Discussion and Conclusions

The main aim of this paper is to clarify the problem of optimal designs for prediction in hierarchical linear mixed models. To this end we derive the mean squared error matrix for the prediction (estimation) of the individual effects (individual parameters) to characterize the performance of an experimental design. It turns out that this matrix is a matrix weighted combination of the corresponding mean squared error matrices in a Bayesian setup and the related model without random effects. It is worth-while mentioning that the mean squared error matrix differs from that for the individual deviations from the population mean, which only depends on the Bayesian part. As a consequence the corresponding optimal designs are in general not the same. In particular, as the Bayesian criterion may result in a singular optimal design, this is not appropriate for prediction because of lack of estimability.

As a criterion of design optimality we use here the the integrated mean squared error and propose a modified version of the determinant criterion, which is an extension of the standard D-criterion in the situation of a singular dispersion matrix of the random effects. Both criteria can be simplified if the dispersion matrix is regular. For both criteria the objective function results in a compromise between the Bayesian and the standard fixed effects counterparts. Thus the criteria can be considered as compound criteria (see e. g. Cook and Wong (1994)) with weights depending on the number of experimental units. As a by-product it can be seen that in the special case of random intercepts the optimal design in the fixed effects model remains optimal for prediction, whatever the variance ratio and the number of observations is. For the other cases we formulate the optimality conditions in the form of equivalence theorems and obtain results, which differ from those proposed in the literature. Finally we give some constructions of optimal designs by considerations of invariance and equivariance in situtations, where transformations of the design region are compatible with the structure of the dispersion matrix. The theoretical results are illustrated by some examples of straight line and quadratic regression.

Although our results presented here are obtained in the framework of approximate designs, which are in general not directly applicable, highly efficient exact designs can be determined using appropriate rounding of the optimal approximate designs. Related results are under preparation.

A further problem for applications is the fact that the optimal designs depend on the dispersion parameters and are, hence, only locally optimal. Sensitivity and robustness with respect to the variance parameters is another topic of further research.

Alternative maningful design criteria can also be considered. We expect that the qualitative results will be similar to those for the IMSE- and (modified) D-criterion, but weights may differ.

Furthermore, the more general model, where different individual designs are allowed, will have to be investigated in future. This will, in particular, become relevant for small numbers of observations (sparse sampling), which typically occur e. g. in pharmacological studies in children.

A Appendix

A.1 Proof of Theorem 1 and Theorem 2

To utilize the theoretical results available for the estimation of individual parameters (prediction) we will recognize model (1) as a special case of the general linear mixed model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{52}$$

with specific design matrices $\mathbf{X}$ and $\mathbf{Z}$ for the fixed effects and the random effects, respectively. Here $\boldsymbol{\beta}$ denotes again the fixed effect (population parameter), and $\boldsymbol{\gamma}$ are the random effects. These random effects and the observational errors $\boldsymbol{\varepsilon}$ are assumed to have zero mean and to be all uncorrelated with corresponding full rank covariance matrices $\operatorname{Cov}(\boldsymbol{\gamma}) = \mathbf{G}$ and $\operatorname{Cov}(\boldsymbol{\varepsilon}) = \mathbf{R}$, respectively.

According to Henderson *et al.* (1959) and Christensen (2002), the mixed model equations

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^\top\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^\top\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}^\top\mathbf{R}^{-1}\mathbf{Y} \\ \mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{Y} \end{pmatrix} \tag{53}$$

yield the *Best Linear Unbiased Estimator* $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ and the *Best Linear Unbiased Predictor* $\hat{\boldsymbol{\gamma}}$ for $\boldsymbol{\gamma}$ provided the fixed effects design matrix $\mathbf{X}$ has full column rank. Henderson (1975) established that then the joint mean squared error matrix for both $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ is of the form

$$\operatorname{Cov}\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^\top\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^\top\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix}^{-1}. \tag{54}$$

Now our model (4) for the full observational vector $\mathbf{Y}$ will be adapted to the more general case (52). First we note that all error terms ε_{ij} are homoscedastic and uncorrelated, and, hence, $\mathbf{R} = \operatorname{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_{n\cdot m}$. With $\mathbf{X} = \mathbf{1}_n \otimes \mathbf{F}$, $\mathbf{Z} = \mathbf{I}_n \otimes \mathbf{F}$ and $\boldsymbol{\gamma} = \mathbf{b}$ the model equation (4) has the form of the general model (52). If the covariance matrix $\mathbf{D}$ for the random parameters $\boldsymbol{\beta}_i$ is non-singular, the matrix $\mathbf{G} = \operatorname{Cov}(\boldsymbol{\gamma}) = \mathbf{I}_n \otimes (\mathbf{F}\mathbf{D}\mathbf{F}^\top)$ is also non-singular, as long as designs with full rank design matrices $\mathbf{F}$ are considered, as is required here throughout.

Special care has to be taken of the fact that the covariance matrix $\mathbf{D}$ for the random parameters may be singular, which allows for some individual parameters to be non-random. For this let q be the rank of $\mathbf{D}$. Then there exists a $p \times q$ matrix $\mathbf{K}$ with $\mathbf{D} = \mathbf{K}\mathbf{K}^\top$ and $\operatorname{rank}(\mathbf{K}) = q$ such that $\mathbf{K}^\top\mathbf{K}$ is non-singular. With this notation we introduce the random variables

$$\mathbf{c}_i = (\mathbf{K}^\top\mathbf{K})^{-1}\mathbf{K}^\top(\boldsymbol{\beta}_i - \boldsymbol{\beta}), \tag{55}$$

for which holds $\boldsymbol{\beta}_i - \boldsymbol{\beta} = \mathbf{K}\mathbf{c}_i$ almost surely. Then the model equation (1) can be written as

$$Y_{ij} = \mathbf{f}(x_j)^\top\boldsymbol{\beta} + \mathbf{f}(x_j)^\top\mathbf{K}\mathbf{c}_i + \varepsilon_{ij}, \tag{56}$$

which results in vector notation in

$$\mathbf{Y}_i = \mathbf{F}\boldsymbol{\beta} + \mathbf{F}\mathbf{K}\mathbf{c}_i + \boldsymbol{\varepsilon}_i \tag{57}$$

for the observations $\mathbf{Y}_i$ of individual i and, finally, in the complete observation vector

$$\mathbf{Y} = (\mathbf{1}_n \otimes \mathbf{F})\boldsymbol{\beta} + (\mathbf{I}_n \otimes (\mathbf{F}\mathbf{K}))\mathbf{c} + \boldsymbol{\varepsilon} \tag{58}$$

for all individuals, where $\mathbf{c} = (\mathbf{c}_1^\top, .., \mathbf{c}_n^\top)^\top$ is a vector of random effects. Now by replacing $\mathbf{I}_n \otimes \mathbf{F}$ by $\mathbf{Z} = \mathbf{I}_n \otimes (\mathbf{FK})$ and $\mathbf{b}$ by $\boldsymbol{\gamma} = \mathbf{c}$ the model equation (58) is of the general form (52) with non-singular covariance matrices $\mathbf{R} = \sigma^2 \mathbf{I}_n \otimes \mathbf{I}_m$ and $\mathbf{G} = \sigma^2 \mathbf{I}_n \otimes \mathbf{I}_q$, respectively.

With this notation the mean squared error matrix (54) becomes

$$\begin{aligned} \operatorname{Cov} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{c}} - \mathbf{c} \end{pmatrix} &= \sigma^2 \begin{pmatrix} n\mathbf{F}^\top\mathbf{F} & \mathbf{1}_n^\top \otimes (\mathbf{F}^\top\mathbf{FK}) \\ \mathbf{1}_n \otimes (\mathbf{K}^\top\mathbf{F}^\top\mathbf{F}) & \mathbf{I}_n \otimes (\mathbf{K}^\top\mathbf{F}^\top\mathbf{FK}) + \mathbf{I}_n \otimes \mathbf{I}_q \end{pmatrix}^{-1} \\ &= \frac{\sigma^2}{n} \begin{pmatrix} (\mathbf{F}^\top\mathbf{F})^{-1} + \mathbf{K}\mathbf{K}^\top & -\mathbf{1}_n^\top \otimes \mathbf{K} \\ -\mathbf{1}_n \otimes \mathbf{K}^\top & (n\mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n^\top) \otimes (\mathbf{K}^\top\mathbf{F}^\top\mathbf{FK} + \mathbf{I}_q)^{-1} + (\mathbf{1}_n\mathbf{1}_n^\top) \otimes \mathbf{I}_q \end{pmatrix}, \end{aligned} \tag{59}$$

where the second equality follows by a standard inversion formula for partitioned matrices. By (53) this matrix is also relevant for the calculation of the estimator

$$\hat{\boldsymbol{\beta}} = (\tfrac{1}{n}\mathbf{1}_n^\top \otimes (\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top)\mathbf{Y} = (\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top\bar{\mathbf{Y}} \tag{60}$$

of the fixed effects $\boldsymbol{\beta}$ and the predictor

$$\hat{\mathbf{c}} = ((\mathbf{I}_n - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top) \otimes (\mathbf{I}_q - \mathbf{K}^\top((\mathbf{F}^\top\mathbf{F})^{-1} + \mathbf{K}\mathbf{K}^\top)^{-1}\mathbf{K})\,\mathbf{K}^\top\mathbf{F}^\top)\mathbf{Y} \tag{61}$$

of the random effects $\mathbf{c}$, where we have made use of the standard inversion formula for the sum of matrices

$$(\mathbf{K}^\top\mathbf{F}^\top\mathbf{FK} + \mathbf{I}_q)^{-1} = \mathbf{I}_q - \mathbf{K}^\top((\mathbf{F}^\top\mathbf{F})^{-1} + \mathbf{K}\mathbf{K}^\top)^{-1}\mathbf{K}\,. \tag{62}$$

This leads to

$$\begin{aligned} \hat{\mathbf{c}}_i &= (\mathbf{I}_q - \mathbf{K}^\top((\mathbf{F}^\top\mathbf{F})^{-1} + \mathbf{K}\mathbf{K}^\top)^{-1}\mathbf{K})\mathbf{K}^\top\mathbf{F}^\top(\mathbf{Y}_i - \bar{\mathbf{Y}}) \\ &= \mathbf{K}^\top((\mathbf{F}^\top\mathbf{F})^{-1} + \mathbf{K}\mathbf{K}^\top)^{-1}(\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top(\mathbf{Y}_i - \bar{\mathbf{Y}}) \\ &= \mathbf{K}^\top((\mathbf{F}^\top\mathbf{F})^{-1} + \mathbf{D})^{-1}(\hat{\boldsymbol{\beta}}_{i;\mathrm{ind}} - \hat{\boldsymbol{\beta}}) \end{aligned} \tag{63}$$

on the individual level.

The best linear unbiased predictor for the random parameters $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{K}\mathbf{c}_i$ is the corresponding linear combination $\hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\beta}} + \mathbf{K}\hat{\mathbf{c}}_i$ of the estimators for the fixed and random effects. The representation (7) of Theorem 1 for the predictors $\hat{\boldsymbol{\beta}}_i$ follows from (63).

Thus the best linear unbiased predictor for the full vector $\mathbf{B}$ of all individual parameters equals $\hat{\mathbf{B}} = (\mathbf{1}_n \otimes \mathbf{I}_p)\hat{\boldsymbol{\beta}} + (\mathbf{I}_n \otimes \mathbf{K})\hat{\mathbf{c}}$ with mean squared error matrix

$$\begin{aligned} \operatorname{Cov}(\hat{\mathbf{B}} - \mathbf{B}) &= (\mathbf{1}_n \otimes \mathbf{I}_p \mid \mathbf{I}_n \otimes \mathbf{K}) \operatorname{Cov} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{c}} - \mathbf{c} \end{pmatrix} \begin{pmatrix} \mathbf{1}_n^\top \otimes \mathbf{I}_p \\ \mathbf{I}_n \otimes \mathbf{K}^\top \end{pmatrix} \\ &= \sigma^2 \left((\tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top) \otimes (\mathbf{F}^\top\mathbf{F})^{-1} + (\mathbf{I}_n - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top) \otimes (\mathbf{K}(\mathbf{K}^\top\mathbf{F}^\top\mathbf{FK} + \mathbf{I}_q)^{-1}\mathbf{K}^\top) \right). \end{aligned} \tag{64}$$

according to (59). The mean squared error matrix can be recognized as a "weighted combination" of the covariance matrix $\sigma^2(\mathbf{F}^\top\mathbf{F})^{-1}$ for $\boldsymbol{\beta}$ in the model $\mathbf{Y} = \mathbf{F}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ without random effects and dhe mean squared error matrix $\sigma^2\mathbf{K}(\mathbf{K}^\top\mathbf{F}^\top\mathbf{FK} + \mathbf{I}_q)^{-1}\mathbf{K}^\top$ for $\boldsymbol{\beta}_i = \mathbf{K}\mathbf{c}_i$ in the model $\mathbf{Y}_i = \mathbf{FK}\mathbf{c}_i + \boldsymbol{\varepsilon}_i$ with $\operatorname{Cov}(\mathbf{c}_i) = \sigma^2\mathbf{I}_q$ for the zero mean random parameter $\mathbf{c}_i$.

The result (12) of Theorem 2 then follows imnmediately by (62).

A.2 Proof of Theorem 3 and Theorem 4

Here we make use of the general equivalence theorem (see e. g. Silvey (1980)): A design ξ^* is optimal with respect to a convex criterion function Φ to be minimized if and only if the directional derivative $\mathrm{F}_\Phi(\mathbf{M}(\xi^*); \mathbf{M}(\xi_x))$ of the criterion function at $\mathbf{M}(\xi^*)$ in the direction of

$\mathbf{M}(\xi_x) = m\,\mathbf{f}(x)\mathbf{f}(x)^\top$ is nonnegative on the whole experimental region ($x \in \mathcal{X}$), where ξ_x denotes the one-point design in x. Moreover, this directional derivative equals zero for all support points x_j^* of the optimal design ξ^*.

The IMSE-criterion for prediction is a compound criterion, i. e. it is a weighted sum of the IMSE-criterion in the model without random efects and the corresponding Bayesian IMSE-criterion in the "reduced" model $\mathbf{Y}_i = \mathbf{FKc}_i + \boldsymbol{\varepsilon}_i$ introduced in section A.1. As a consequence the IMSE-criterion for prediction inherits convexity from its components and its directional derivative is also a weighted sum of the counterparts of the components. Thus the directional derivative at $\mathbf{M}_1$ in the direction of $\mathbf{M}_2$ is

$$- \operatorname{tr}\left(((n-1)\mathbf{D}(\mathbf{I}_p + \mathbf{M}_1\mathbf{D})^{-1}(\mathbf{M}_2 - \mathbf{M}_1)(\mathbf{I}_p + \mathbf{D}\mathbf{M}_1)^{-1}\mathbf{D} + \mathbf{M}_1^{-1}(\mathbf{M}_2 - \mathbf{M}_1)\mathbf{M}_1^{-1})\mathbf{V}\right), \quad (65)$$

which establishes Theorem 3.

To determine the directional derivative in the case of the modified D-criterion we represent this criterion in the form

$$\mathrm{D}_{\mathrm{pred}}(\xi) = (n-1)(\ln\det(\mathbf{K}^\top\mathbf{K}) + \ln\det((\mathbf{K}^\top\mathbf{M}(\xi)\mathbf{K} + \mathbf{I}_q)^{-1})) + \ln\det(\mathbf{M}(\xi)^{-1}). \quad (66)$$

Hence, also this criterion can be identified as a compound criterion, which provides convexity and the following form of the directional derivative

$$- \operatorname{tr}\left((n-1)(\mathbf{D} - \mathbf{D}(\mathbf{M}_1^{-1} + \mathbf{D})^{-1}\mathbf{D})(\mathbf{M}_2 - \mathbf{M}_1) + \mathbf{M}_1^{-1}(\mathbf{M}_2 - \mathbf{M}_1)\right), \quad (67)$$

which proves Theorem 4.

A.3 Proof of Theorem 5

By (66) the criterion function of the modified D-criterion can be written as

$$\mathrm{D}_{\mathrm{pred}}(\xi, \mathbf{D}) = (n-1)(\ln\det(\mathbf{K}^\top\mathbf{K}) - \ln\det(\mathbf{K}^\top\mathbf{M}(\xi)\mathbf{K} + \mathbf{I}_q)) - \ln\det(\mathbf{M}(\xi)). \quad (68)$$

Let $\mathbf{K}_g = \mathbf{Q}_g^{-\top}\mathbf{K}$. Then $\mathbf{K}_g$ has full column rank q and satisfies $\mathbf{K}_g\mathbf{K}_g^\top = \mathbf{D}_g$. The modified D-criterion for the transformed design ξ^g on the experimental region $\mathcal{X}_1$ may be represented as

$$\mathrm{D}_{\mathrm{pred}}(\xi^g, \mathbf{D}_g) = (n-1)(\ln\det(\mathbf{K}_g^\top\mathbf{K}_g) - \ln\det(\mathbf{K}_g^\top\mathbf{M}_g(\xi^g)\mathbf{K}_g + \mathbf{I}_q)) - \ln\det(\mathbf{M}_g(\xi^g)), \quad (69)$$

where $\mathbf{M}_g(\xi^g) = \mathbf{Q}_g\mathbf{M}(\xi)\mathbf{Q}_g^\top$ is the information matrix in the transformed model without random effects. This can be transformed to

$$\begin{aligned}&\mathrm{D}_{\mathrm{pred}}(\xi^g, \mathbf{D}_g)\\ &\quad = \mathrm{D}_{\mathrm{pred}}(\xi, \mathbf{D}) + (n-1)(\ln\det(\mathbf{K}^\top\mathbf{Q}_g^{-1}\mathbf{Q}_g^{-\top}\mathbf{K}) - \ln\det(\mathbf{K}^\top\mathbf{K})) - 2\ln|\det(\mathbf{Q}_g)|, \quad (70)\end{aligned}$$

where on the right hand side only the first term depends on the design.

A.4 Proof of Theorem 6

By (20) we represent the IMSE-criterion for the transformed design ξ_g on the experimental region $\mathcal{X}_1$ as

$$\mathrm{IMSE}_{\mathrm{pred}}(\xi^g, \mathbf{D}_g, \nu^g) = \operatorname{tr}\left(\mathrm{MSE}(\xi^g, \mathbf{D}_g)(\mathbf{I}_n \otimes \mathbf{V}_g)\right), \quad (71)$$

where $\mathbf{V}_g = \int_{\mathcal{X}_1} \mathbf{f}(x)\mathbf{f}(x)^\top\nu^g(\mathrm{d}x) = \int_{\mathcal{X}} \mathbf{f}(g(x))\mathbf{f}(g(x))^\top\nu(\mathrm{d}x)$, which can be rewritten as

$$\mathbf{V}_g = \int_{\mathcal{X}} \mathbf{Q}\,\mathbf{f}(x)\mathbf{f}(x)^\top\mathbf{Q}^\top\nu(\mathrm{d}x) = \mathbf{Q}\mathbf{V}\mathbf{Q}^\top \quad (72)$$

according to (39). Then it follows from (41) that the criterion functions for the designs ξ and ξ_g coincide

$$\mathrm{IMSE}_{\mathrm{pred}}(\xi^g, \mathbf{D}_g, \nu^g) = \mathrm{IMSE}_{\mathrm{pred}}(\xi, \mathbf{D}, \nu). \quad (73)$$

A.5 Proof of Theorem 7

Let the design ξ^* be IMSE-optimal on the experimental region $\mathcal{X}$. Then by Theorem 6 the transformed designs ξ_g^* of the form (40) are also IMSE-optimal on $\mathcal{X}$ for all transformations g from the group G, i. e.

$$\mathrm{IMSE}_{\mathrm{pred}}(\xi^*, \mathbf{D}) = \mathrm{IMSE}_{\mathrm{pred}}(\xi_g^*, \mathbf{D}), \text{ for all } g \in G. \tag{74}$$

As the criterion function is convex, it immediately follows that

$$\mathrm{IMSE}_{\mathrm{pred}}(\bar{\xi}^*, \mathbf{D}) \leq \mathrm{IMSE}_{\mathrm{pred}}(\xi^*, \mathbf{D}), \tag{75}$$

which gives the IMSE-optimality of the design $\bar{\xi}^*$.

To proof the D-optimality of $\bar{\xi}^*$ we use Theorem 5 instead of Theorem 6.

Acknowledgment

This research was partially supported by grant MÄQNU 05M10NMB of the German Federal Ministry of Education and Research (BMBF). Part of the work was done during a visit at the International Newton Institute in Cambridge.

References

Bryk, A. S. and Raudenbush, S. W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage, Newbury Park.

Candel, M. J. (2009). Optimal designs for empirical bayes estimators of individual linear and quadratic growth curves in linear mixed models. *Statistical Methods in Medical Research*, **18**, 397–419.

Christensen, R. (2002). *Plane Answers to Complex Questions: The Theory of Linear Models*. Springer, New York.

Cook, R. D. and Wong, W. K. (1994). On the equivalence of constrainer and compound designs. *Journal of the American Statistical Association*, **89**, 687–692.

Entholzner, M., Benda, N., Schmelter, T., and Schwabe, R. (2005). A note on designs for estimating population parameters. *Biometrical Letters - Listy Biometryczne*, **42**, 25–41.

Fedorov, V. and Hackl, P. (1997). *Model-Oriented Design of Experiments*. Springer, New York.

Gladitz, J. and Pilz, J. (1982). Construction of optimal designs in random coefficient regression models. *Mathematische Operationsforschung und Statistik, Series Statistics*, **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.

Henderson, C. R. (1963). *Selection index and expected genetic advance. In: Statistical Genetic and Plant Breeding*. National Academy of Science - National Research Council.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, **31**, 423–477.

Henderson, C. R. (1984). *Applications of Linear Models in Animal Breeding.* Guelph, Canada.

Henderson, C. R., Kempthorne, O., Searle, S. R., and von Krosigk, C. M. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, **15**, 192–218.

Isotalo, J., Puntanen, S., and Styan, G. P. H. (2011). *Matrix Tricks for Linear Statistical Models.* Springer, Berlin.

Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *Annals of Statistics*, **2**, 849–879.

Liski, E. P., Mandal, N. K., Shah, K. R., and Sinha, B. K. (2002). *Topics in Optimal Design.* Springer, New York.

Molenberghs, G. and Verbeke, G. (2001). *Linear Mixed Models for Longitudinal Data.* Springer, New York.

Pukelsheim, F. (1993). *Optimal Design of Experiments.* J.Wiley, New York, Philadelphia.

Rao, J. N. K. (2003). *Small Area Estimation.* Wiley, New York.

Schwabe, R. (1996). *Optimum Designs for Multi-Factor Models. Lecture Notes in Statistics.* Springer, New York.

Silvey, S. D. (1980). *Optimal Design.* Chapman & Hall, London.