Journal of Statistical Research *https://doi.org/10.3329/jsr.v56i2.67463*
2022, Vol. 56, No. 2, pp. 101-114 *ISSN 0256 - 422 X*

OPTIMAL ALLOCATION SCHEMES IN MIXED ANCOVA MODELS FOR LONGITUDINAL DATA

XIAOJIAN XU*
Department of Mathematics and Statistics, Brock University, St. Catharines, Ontario, Canada
Email: xxu@brocku.ca

SANJOY K. SINHA
School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada
Email: sinha@math.carleton.ca

SUMMARY

We discuss the construction of optimal allocation schemes for the linear mixed model with clustered outcomes or repeated measurements often encountered in longitudinal studies. We consider both treatment and covariate effects in the mixed model, where latent processes are used to describe random cluster or subject effects. A goal of optimal design schemes is to determine proportions of sample units allocated to each treatment for a given total sample size. We develop the optimal designs in a general setting using both D- and A-optimal design criteria. Specifically, we propose a two-stage design approach to deal with unknown parameters in the linear mixed model, where the variances of the random effects across the treatment groups are considered different. We study the empirical properties of the proposed designs using Monte Carlo simulations. An application is also provided using actual clinical data from a longitudinal study.

Keywords and phrases: Longitudinal data; Mixed model; Optimal design of experiments; Random effects; Two-stage design

1 Introduction

The choice of an optimal design is an important issue in many clinical experiments. A common goal of the design of experiment is to develop a predictive model as precisely as possible, while optimizing the cost associated with measuring the outcomes. Optimal designs for generalized linear and nonlinear regression models have been considered by many authors in recent years (e.g., Chaudhuri and Mykland, 1993; Wiens, 1994; Sinha and Wiens, 2002; Sinha, 2013; Xu and Sinha, 2020). Recently, Hore, Dewanji, and Chatterjee (2014) discussed design problems for ANCOVA models, where optimal allocation of two treatment groups was considered for given covariates, but with no random effects in the model.

In a case-control clinical study, subjects may be measured repeatedly over a fixed period of time. Repeated outcomes from a given subject are correlated by nature. Latent random effects in a mixed

* Corresponding author

model setting are often used to describe the correlations among the repeated outcomes in the longitudinal experiment. The random effects variance components in the mixed model can vary across different treatment groups. In this paper, we present an optimal design strategy that can accommodate the correlations among repeated observations and also can account for unequal random effects variance components for different treatment groups, so that efficient estimates of the model parameters can be attained for reliable predictions. We formulate our problem of interest using a similar ANCOVA framework, but in a mixed model setting, where random effects are used to describe the correlation structure among clustered data or repeated measurements. In particular, we consider a set of design problems and extend the previous work by: (i) developing an optimal design scheme with heteroscedastic random effects in the ANCOVA model; (ii) presenting the optimal allocation theory for longitudinal experiments; and (iii) addressing the problem of missing observations due to drop-outs in follow-up times.

The rest of the paper is organized as follows. Section 2 introduces the model and notation for analyzing longitudinal continuous data in the framework of the linear mixed model. Section 3 reviews the maximum likelihood method for estimating the regression parameters and variance components. Sections 4 presents the proposed optimal design criteria for the linear mixed model. Section 5 investigates the performance of the proposed design based on a simulation study. Section 6 presents an application using actual clinical data from a longitudinal study. Section 7 provides some concluding remarks.

2 Model and Notation

Suppose there are k treatments in a longitudinal study, where the i^{th} treatment ($i = 1, \ldots, k$) is given to the j^{th} individual ($j = 1, \ldots, n_i$), with each individual being observed at a possible varied set of T_{ij} time points ($t = 1, \ldots, T_{ij}$). Consider a linear mixed model for the longitudinal response y_{ijt}, given by

$$y_{ijt} = \mathbf{z}_j^T \boldsymbol{\alpha} + \mathbf{x}_{ijt}^T \boldsymbol{\beta} + u_{ij} + \varepsilon_{ijt}, \tag{2.1}$$

where $\mathbf{x}_{ijt}$ represents a $p \times 1$ vector of covariates observed with the response y_{ijt}, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ a vector of fixed covariate effects, $\mathbf{z}_j = (z_{1j}, \ldots, z_{k,j})^T$ a vector of k treatment indicators, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)^T$ a vector of corresponding fixed treatment effects, and u_{ij} a random interaction effect between the i^{th} treatment and j^{th} individual that is assumed to follow an independent normal distribution with mean 0 and unequal variance σ_i^2. The random error term ε_{ijt} is assumed to follow an independent normal distribution with mean 0 and a common variance σ_ε^2, independent of the random effects u_{ij}.

Model (2.1) can be written in a matrix form as

$$\mathbf{y}_{ij} = \mathbf{Z}_{ij}\boldsymbol{\alpha} + \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{1}_{ij}u_{ij} + \boldsymbol{\varepsilon}_{ij}, \tag{2.2}$$

where $\mathbf{y}_{ij} = (y_{ij1}, \ldots, y_{ijT_{ij}})^T$, $\mathbf{1}_{ij}$ is a $T_{ij} \times 1$ vector of 1's, $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ are $T_{ij} \times p$ and $T_{ij} \times k$ design matrices with their t^{th} rows being equal to $\mathbf{x}_{ijt}^T$ and $\mathbf{z}_j^T$, respectively, and the vector of random errors $\boldsymbol{\varepsilon}_{ij} = (\varepsilon_{ij1}, \ldots, \varepsilon_{ijT_{ij}})^T$. Note that in the variance components literature, this model is also called a linear model with nested error structure (Christenson, 1996; Wang and Ma, 2002).

From (2.2), the marginal distribution of $\mathbf{y}_{ij}$ is multivariate normal with the mean vector

$$\boldsymbol{\mu}_{ij} = E(\mathbf{y}_{ij}) = \mathbf{Z}_{ij}\boldsymbol{\alpha} + \mathbf{X}_{ij}\boldsymbol{\beta},$$

and variance-covariance matrix

$$\text{Var}(\mathbf{y}_{ij}) = \sigma_\varepsilon^2\mathbf{I} + \sigma_i^2\mathbf{1}_{ij}\mathbf{1}_{ij}^T = \sigma_\varepsilon^2\left(\mathbf{I} + d_i\mathbf{1}_{ij}\mathbf{1}_{ij}^T\right) = \sigma_\varepsilon^2\mathbf{V}_{ij},$$

where $\mathbf{V}_{ij} = \mathbf{I} + d_i\mathbf{1}_{ij}\mathbf{1}_{ij}^T$ and $d_i = \sigma_i^2/\sigma_\varepsilon^2$.

Our goal is to adopt an optimal design strategy that can select the number of subjects n_i assigned to the i^{th} treatment. The optimal choice may depend on the values of unequal variances σ_i^2 of the random interaction effects u_{ij}. Another goal may be to choose both the design points $\mathbf{x}_{ijt}$ and sample sizes n_i using an optimal strategy. We can adopt a sequential D-optimal design, as the Fisher information involves the model parameters that are not known a priori.

We define the partition matrix $\mathbf{W}_{ij} = (\mathbf{Z}_{ij}|\mathbf{X}_{ij})$ and the partition vector $\boldsymbol{\theta}^T = \left(\boldsymbol{\alpha}^T|\boldsymbol{\beta}^T\right)$. Then the marginal distribution of $\mathbf{y}_{ij}$ is given by

$$\mathbf{y}_{ij} \sim N\left(\mathbf{W}_{ij}\boldsymbol{\theta}, \sigma_\varepsilon^2\mathbf{V}_{ij}\right).$$

By the Gauss-Markov Theorem, the BLUE of $\boldsymbol{\theta}$ may be obtained as

$$\hat{\boldsymbol{\theta}} = \left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mathbf{W}_{ij}^T\mathbf{V}_{ij}^{-1}\mathbf{W}_{ij}\right)^{-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mathbf{W}_{ij}^T\mathbf{V}_{ij}^{-1}\mathbf{y}_{ij}, \tag{2.3}$$

which is also the MLE and generalized least squares (GLS) estimator of $\boldsymbol{\theta}$, provided that the inverse $(\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mathbf{W}_{ij}^T\mathbf{V}_{ij}^{-1}\mathbf{W}_{ij})^{-1}$ exists. We note that (2.3) depends on possible unknown variance components d_i. In the next section, we discuss the likelihood method for estimating the variance components.

3 The ML Estimation

Recall the linear mixed model (2). The total number of observations from all $n = \sum_{i=1}^{k} n_i$ experimental units is given by $N = \sum_{i=1}^{k} N_i = \sum_{i=1}^{k}\sum_{j=1}^{n_i} T_{ij}$, where $N_i = \sum_{j=1}^{n_i} T_{ij}$ denotes the total number of observations from the i^{th} treatment $(i = 1, \ldots, k)$.

The log-likelihood function (without the constant term) is given by

$$l(\boldsymbol{\theta}, \sigma_\varepsilon^2, d_i) = -\frac{1}{2}\left[N\ln\sigma_\varepsilon^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\ln|\mathbf{V}_{ij}| + \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\mathbf{y}_{ij} - \mathbf{W}_{ij}\boldsymbol{\theta}\right)^T\mathbf{V}_{ij}^{-1}\left(\mathbf{y}_{ij} - \mathbf{W}_{ij}\boldsymbol{\theta}\right)}{\sigma_\varepsilon^2}\right]. \tag{3.1}$$

Taking the derivative with respect to σ_ε^2, the log-likelihood can be maximized at

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\mathbf{y}_{ij} - \mathbf{W}_{ij}\boldsymbol{\theta}\right)^T\mathbf{V}_{ij}^{-1}\left(\mathbf{y}_{ij} - \mathbf{W}_{ij}\boldsymbol{\theta}\right). \tag{3.2}$$

Now, consider the variance-profile log-likelihood function (Lindstrom and Bates, 1988) obtained by substituting (3.2) into (3.1), i.e., by replacing σ_ε^2 with $\hat{\sigma}_\varepsilon^2$. The MLEs of $(\boldsymbol{\theta}, d_i)$ can be obtained by maximizing the profile log-likelihood (without the constant term)

$$l\left(\boldsymbol{\theta}, d_i\right) = -\frac{1}{2}\left[N \ln\left\{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\mathbf{y}_{ij} - \mathbf{W}_{ij}\boldsymbol{\theta}\right)^T \mathbf{V}_{ij}^{-1}\left(\mathbf{y}_{ij} - \mathbf{W}_{ij}\boldsymbol{\theta}\right)\right\} + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\ln\left|\mathbf{V}_{ij}\right|\right]. \tag{3.3}$$

Let $\mathbf{e}_{ij} = \mathbf{y}_{ij} - \mathbf{W}_{ij}\hat{\boldsymbol{\theta}} = \mathbf{y}_{ij} - \mathbf{W}_{ij}\{(\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mathbf{W}_{ij}^T\mathbf{V}_{ij}^{-1}\mathbf{W}_{ij})^{-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mathbf{W}_{ij}^T\mathbf{V}_{ij}^{-1}\mathbf{y}_{ij}\}$. As the information on d_i is only contained within the i^{th} treatment observations, we may maximize the "partial" likelihood function for the i^{th} treatment observations. The "partial" profile log-likelihood function of each d_i (without constant terms) is given by

$$l\left(d_i\right) = -\frac{1}{2}\left[N \ln\left\{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mathbf{e}_{ij}^T\left(\mathbf{I} + d_i\mathbf{1}_{ij}\mathbf{1}_{ij}^T\right)^{-1}\mathbf{e}_{ij}\right\} + \sum_{j=1}^{n_i}\ln\left|\mathbf{I} + d_i\mathbf{1}_{ij}\mathbf{1}_{ij}^T\right|\right]. \tag{3.4}$$

Since

$$\begin{aligned}\left|\mathbf{I} + d_i\mathbf{1}_{ij}\mathbf{1}_{ij}^T\right| &= 1 + T_{ij}d_i \quad \text{and} \\ \left(\mathbf{I} + d_i^T\mathbf{1}_{ij}\mathbf{1}_{ij}^T\right)^{-1} &= \mathbf{I} - \frac{d_i}{1 + T_{ij}d_i}\mathbf{1}_{ij}\mathbf{1}_{ij}^T,\end{aligned}$$

the log-likelihood (7) can be rewritten as

$$l\left(d_i\right) = -\frac{1}{2}\left[N \ln\left\{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mathbf{e}_{ij}^T\left(\mathbf{I} - \frac{d_i}{1 + T_{ij}d_i}\mathbf{1}_{ij}\mathbf{1}_{ij}^T\right)\mathbf{e}_{ij}\right\} + \sum_{j=1}^{n_i}\ln\left(1 + T_{ij}d_i\right)\right]. \tag{3.5}$$

This leads to a simplified estimator for (2.3) similar to the approach utilized in Demidenko (2004):

$$\hat{\boldsymbol{\theta}} = \left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\mathbf{W}_{ij}^T\mathbf{W}_{ij} - \frac{T_{ij}^2 d_i}{1 + T_{ij}d_i}\bar{\mathbf{w}}_{ij}\bar{\mathbf{w}}_{ij}^T\right)\right]^{-1}\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\mathbf{W}_{ij}^T\mathbf{y}_{ij} - \frac{T_{ij}^2 d_i}{1 + T_{ij}d_i}\bar{\mathbf{w}}_{ij}\bar{y}_{ij}\right)\right], \tag{3.6}$$

where

$$\bar{\mathbf{w}}_{ij} = \frac{1}{T_{ij}}\sum_{t=1}^{T_{ij}}\mathbf{w}_{ijt} \text{ and } \bar{y}_{ij} = \frac{1}{T_{ij}}\sum_{t=1}^{T_{ij}}y_{ijt}.$$

Additionally, we have

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\mathbf{y}_{ij} - \mathbf{W}_{ij}\boldsymbol{\theta}\right)^T\left(\mathbf{I} - \frac{d_i}{1 + T_{ij}d_i}\mathbf{1}_{ij}\mathbf{1}_{ij}^T\right)\left(\mathbf{y}_{ij} - \mathbf{W}_{ij}\boldsymbol{\theta}\right), \tag{3.7}$$

and

$$\mathbf{e}_{ij} = \mathbf{y}_{ij} - \mathbf{W}_{ij}\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\mathbf{W}_{ij}^T\mathbf{W}_{ij} - \frac{T_{ij}^2 d_i}{1+T_{ij}d_i}\bar{\mathbf{w}}_{ij}\bar{\mathbf{w}}_{ij}^T\right)\right]^{-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\mathbf{W}_{ij}^T\mathbf{y}_{ij} - \frac{T_{ij}^2 d_i}{1+T_{ij}d_i}\bar{\mathbf{w}}_{ij}\bar{y}_{ij}^T\right).$$

Let

$$\begin{aligned} S &= S(\boldsymbol{\theta}) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\|\mathbf{y}_{ij} - \mathbf{W}_{ij}\boldsymbol{\theta}\|^2, \\ h_{ij} &= \frac{1}{T_{ij}}\mathbf{e}_{ij}^T\mathbf{1}_{ij} = \frac{1}{T_{ij}}\sum_{t=1}^{T_{ij}}\left(y_{ijt} - \mathbf{w}_{ijt}^T\boldsymbol{\theta}\right) = \bar{y}_{ij} - \boldsymbol{\theta}^T\bar{\mathbf{w}}_{ij}. \end{aligned}$$

Then (3.5) becomes

$$l(d_i) = -\frac{1}{2}\left[N\ln\left\{S - \sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{T_{ij}^2 d_i}{1+T_{ij}d_i}h_{ij}^2\right\} + \sum_{j=1}^{n_i}\ln(1+T_{ij}d_i)\right]. \tag{3.8}$$

3.1 Iterative solutions

In the previous section, we presented the likelihood function and its maximization in a general setting. For longitudinal studies or repeated measurement designs, the number of observations for each individual are often fixed, i.e., $T_{ij} = T$. Even with a less restricted condition for which the number of observations per individual stays the same within a treatment group, numerical iterative procedures may be adopted for estimating the variance components.

Let T_i denote the number of observations from each of the n_i individuals within the i^{th} treatment group. Then (3.8) becomes

$$l(d_i) = -\frac{1}{2}\left[N\ln\left\{S - \sum_{i=1}^{k}\frac{T_i^2 d_i}{1+T_i d_i}A_i\right\} + n_i\ln(1+T_i d_i)\right], \tag{3.9}$$

where

$$A_i = \sum_{j=1}^{n_i}h_{ij}^2.$$

Denote

$$\Delta = \sum_{i=1}^{k}\frac{T_i^2 d_i}{1+T_i d_i}A_i.$$

Then taking the partial derivative of $l(d_i)$ in (12) with respect to d_i and equating it to zero gives an

iterative equation for d_i in the form

$$\begin{aligned} d_i^{(r+1)} &= \frac{NA_i}{n_i(S-\Delta^{(r)})} - \frac{1}{T_i} \\ &= \frac{N\sum_{j=1}^{n_i}\left(\bar{y}_{ij} - \boldsymbol{\theta}^T\bar{\mathbf{w}}_{ij}\right)^2}{n_i\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}\|\mathbf{y}_{ij} - \mathbf{W}_{ij}\boldsymbol{\theta}\|^2 - \sum_{i=1}^{k}\frac{T_i^2 d_i^{(r)}}{1+T_i d_i^{(r)}}A_i\right]} - \frac{1}{T_i}, \end{aligned} \tag{3.10}$$

for $r = 0, 1, 2, \ldots$. We discuss the choice of initial values for d_i in Section 4.1.

4 Optimal Designs

From (2.3), we can have the variance-covariance matrix of the ML estimator $\hat{\boldsymbol{\theta}}$ in the form

$$\text{Var}(\hat{\boldsymbol{\theta}}) = \sigma_\varepsilon^2\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mathbf{W}_{ij}^T\mathbf{V}_{ij}^{-1}\mathbf{W}_{ij}\right]^{-1}.$$

Often we are interested in the differences between treatment effects in the ANCOVA framework. Here we aim to estimate $\boldsymbol{\beta}$ and $\alpha_i - \alpha_{i-1}$, for $i = 2, \ldots, k$. Let

$$\begin{aligned} \mathbf{C} &= \begin{pmatrix} -1 & 1 & & \\ & -1 & 1 & \\ & & \ldots & \ldots \\ & & & -1 \quad 1 \end{pmatrix}_{(k-1)\times(k)}, \\ \mathbf{B} &= \begin{pmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_p \end{pmatrix}_{(p+k-1)\times(p+k)}. \end{aligned}$$

Then

$$\text{Var}(\mathbf{B}\hat{\boldsymbol{\theta}}) = \sigma_\varepsilon^2\mathbf{B}\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mathbf{W}_{ij}^T\mathbf{V}_{ij}^{-1}\mathbf{W}_{ij}\right]^{-1}\mathbf{B}^T.$$

In this paper, we consider D-optimal and A-optimal designs, which can be attained by minimizing the determinant and trace of $\text{Var}(\mathbf{B}\hat{\boldsymbol{\theta}})$, respectively. The loss functions are given by

$$\begin{aligned} L_D &= \left|\mathbf{B}\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\mathbf{W}_{ij}^T\mathbf{W}_{ij} - \frac{T_{ij}^2 d_i}{1+T_{ij}d_i}\bar{\mathbf{w}}_{ij}\bar{\mathbf{w}}_{ij}^T\right)\right]^{-1}\mathbf{B}^T\right|, \\ L_A &= \text{tr}\left(\mathbf{B}\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\mathbf{W}_{ij}^T\mathbf{W}_{ij} - \frac{T_{ij}^2 d_i}{1+T_{ij}d_i}\bar{\mathbf{w}}_{ij}\bar{\mathbf{w}}_{ij}^T\right)\right]^{-1}\mathbf{B}^T\right). \end{aligned}$$

It is also sensible to optimally select the n_i by minimizing

$$L_{D_1} = \left| \mathbf{B}_1 \left[\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\mathbf{W}_{ij}^T \mathbf{W}_{ij} - \frac{T_{ij}^2 d_i}{1 + T_{ij} d_i} \bar{\mathbf{w}}_{ij} \bar{\mathbf{w}}_{ij}^T \right) \right]^{-1} \mathbf{B}_1^T \right|,$$

or

$$L_{A_1} = \text{tr} \left(\mathbf{B}_1 \left[\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\mathbf{W}_{ij}^T \mathbf{W}_{ij} - \frac{T_{ij}^2 d_i}{1 + T_{ij} d_i} \bar{\mathbf{w}}_{ij} \bar{\mathbf{w}}_{ij}^T \right) \right]^{-1} \mathbf{B}_1^T \right),$$

where

$$\mathbf{B}_1 = \begin{pmatrix} \mathbf{C} & \mathbf{0} \end{pmatrix}_{(k-1)\times(p+k)},$$

and optimally select $\mathbf{x}_{ijt}$ by minimizing

$$L_{D_2} = \left| \mathbf{B}_2 \left[\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\mathbf{W}_{ij}^T \mathbf{W}_{ij} - \frac{T_{ij}^2 d_i}{1 + T_{ij} d_i} \bar{\mathbf{w}}_{ij} \bar{\mathbf{w}}_{ij}^T \right) \right]^{-1} \mathbf{B}_2^T \right|,$$

or

$$L_{A_2} = \text{tr} \left(\mathbf{B}_2 \left[\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\mathbf{W}_{ij}^T \mathbf{W}_{ij} - \frac{T_{ij}^2 d_i}{1 + T_{ij} d_i} \bar{\mathbf{w}}_{ij} \bar{\mathbf{w}}_{ij}^T \right) \right]^{-1} \mathbf{B}_2^T \right),$$

where

$$\mathbf{B}_2 = \begin{pmatrix} \mathbf{0} & \mathbf{I}_p \end{pmatrix}_{p\times(p+k)}.$$

4.1 A proposed two-stage design

For simplicity, we consider the specific case $T_{ij} = T$ for all (i, j) to describe our two-stage design process. Let $\mathbf{1}$ be a $T \times 1$ vector of 1's. Suppose we have n experimental units available. The proposed design involves the following steps:

Step 1. Consider an initial balanced design of n_0 units (subjects) randomly assigned to k treatments with an equal number of $n^* = (n_0/k)$ units per treatment. Initial values of $\mathbf{x}_{ij}$ $(i = 1, \ldots, k; j = 1, \ldots, n^*)$ are selected uniformly from a design space and remain the same for T time points in the longitudinal setting.

Step 2. Observe longitudinal response y_{ijt} $(t = 1, \ldots, T)$ at T time points for each $(i, j)^{\text{th}}$ unit. Then obtain initial estimates of the model parameters as follows.

Let $\bar{y} = [1/(n_0 T)] \sum_{i=1}^{k} \sum_{j=1}^{n^*} \sum_{t=1}^{T} y_{ijt}$ be the overall mean of all observations. Consider

elements of the initial $\mathbf{V}_i^{(0)}$ obtained from

$$\begin{aligned}
\sigma_\varepsilon^{2(0)} &= \frac{1}{n_0 T}\sum_{i=1}^{k}\sum_{j=1}^{n^*}\sum_{t=1}^{T}(y_{ijt}-\bar{y})^2,\\
\sigma_i^{2(0)} &= \frac{1}{n^* T}\sum_{j=1}^{n^*}\sum_{t=1}^{T}(y_{ijt}-\bar{y}_i)^2-\sigma_\varepsilon^{2(0)},\\
d_i^{(0)} &= \frac{\sigma_i^{2(0)}}{\sigma_\varepsilon^{2(0)}}.
\end{aligned}$$

These lead to

$$\begin{aligned}
\mathbf{V}_i^{(0)} &= \mathbf{I}+d_i^{(0)}\mathbf{1}\mathbf{1}^T,\\
\mathbf{V}_i^{(0)-1} &= \mathbf{I}-\frac{d_i^{(0)}}{1+Td_i^{(0)}}\mathbf{1}\mathbf{1}^T.
\end{aligned}$$

Then obtain initial estimates of $\boldsymbol{\theta}$ and Δ as

$$\boldsymbol{\theta}^{(0)} = \left(\sum_{i=1}^{k}\sum_{j=1}^{n^*}\mathbf{W}_{ij}^T\mathbf{V}_i^{(0)-1}\mathbf{W}_{ij}\right)^{-1}\sum_{i=1}^{k}\sum_{j=1}^{n^*}\mathbf{W}_{ij}^T\mathbf{V}_i^{(0)-1}\mathbf{y}_{ij}, \tag{4.1}$$

$$\Delta^{(0)} = \sum_{i=1}^{k}\left[\frac{T^2 d_i^{(0)}}{1+Td_i^{(0)}}\sum_{j=1}^{n^*}\left(\bar{y}_{ij}-\boldsymbol{\theta}^{(0)T}\bar{\mathbf{w}}_{ij}\right)^2\right]. \tag{4.2}$$

Step 3: Update estimate of d_i by $d_i^{(r)}$ for $r=1$ using the iterative equation (3.10).
Step 4: Update estimates of $\boldsymbol{\theta}$ and Δ by the iterative equations

$$\boldsymbol{\theta}^{(r)} = \left[\sum_{i=1}^{k}\sum_{j=1}^{n^*}\mathbf{W}_{ij}^T\left(\mathbf{I}-\frac{d_i^{(r)}}{1+Td_i^{(r)}}\mathbf{1}\mathbf{1}^T\right)\mathbf{W}_{ij}\right]^{-1}\sum_{i=1}^{k}\sum_{j=1}^{n^*}\mathbf{W}_{ij}^T\left(\mathbf{I}-\frac{d_i^{(r)}}{1+Td_i^{(r)}}\mathbf{1}\mathbf{1}^T\right)\mathbf{y}_{ij} \tag{4.3}$$

$$\Delta^{(r)} = \sum_{i=1}^{k}\left[\frac{T^2 d_i^{(r)}}{1+Td_i^{(r)}}\sum_{j=1}^{n^*}\left(\bar{y}_{ij}-\boldsymbol{\theta}^{(r)T}\bar{\mathbf{w}}_{ij}\right)^2\right], \tag{4.4}$$

for $r=1$. Iterate between Steps 3 and 4 using $r=2,3,\ldots$ to obtain desirable "first stage estimates" $\hat{\boldsymbol{\theta}}$ and $\hat{d}_i$'s from the initial design.

Step 5: Conduct the second stage of the design (optimal design stage) with the first stage estimates. Denote $m_i = n_i - n^*$, with fixed $\mathbf{x}_{ij}$ as specified in Step 1. Optimally choose m_i^* with $\sum_{i=1}^{k} m_i^* = n - n_0$ by minimizing L_{D_1} or L_{A_1}.

5 Simulation Study

Here we investigate the performance of the proposed design using a simulation study. Consider a longitudinal experiment with only two treatment groups (treatment vs. placebo) and a single

baseline covariate x. The response y_{it} from the ith subject at the tth time point is described by the linear mixed model

$$y_{it} = \alpha_0 + \alpha_1 z_i + \beta_1 x_i + \beta_2 t + u_i + \epsilon_{it}, \tag{5.1}$$

for $i = 1, \ldots, n$, and $t = 1, 2, \ldots, T$, where z_i is a binary indicator of the treatment group, x_i represents the value of a baseline covariate from $N(1, 1)$, u_i are assumed independent $N(0, (1 - z_i)\sigma_0^2 + z_i\sigma_1^2)$, and ϵ_i are assumed independent $N(0, \sigma_\epsilon^2)$. Also, u_i and ϵ_{it} are assumed independent of each other. Consider only $T = 3$ time points.

We assume that initially there are n_0 subjects under study, with an equal number of $n_{01} = (n_0/2)$ subjects in the control group and $n_{02} = (n_0/2)$ subjects in the treatment group. Consider assigning n_1 new subjects to the control group and n_2 new subjects to the treatment group, so that $n = (n_{01} + n_{02}) + (n_1 + n_2)$. Our goal is to find an optimal allocation of n_1 and n_2 subjects by (i) maximizing the determinant of the Fisher information or by (ii) minimizing the trace of the variance-covariance matrix of $\hat{\boldsymbol{\theta}} = (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\beta}_1, \hat{\beta}_2)^T$. For the estimation of the model parameters and optimal allocation of sample sizes, we follow the algorithm described earlier in Section 4.

Figure 1 exhibits plots for three representative data sets, where the determinants of the Fisher information are shown for different values of the sample size n_1. The regression parameters in model (18) were fixed at $\alpha_0 = 2, \alpha_1 = 1$, $\beta_1 = 0.5$, and $\beta_2 = 0.5$, and the error variance was fixed at $\sigma_\epsilon = 2$. Three combinations of the random effects variances were chosen as $(\sigma_0, \sigma_1) = (2, 2)$ (left panel), $(\sigma_0, \sigma_1) = (2, 1)$ (middle panel), and $(\sigma_0, \sigma_1) = (1, 2)$ (right panel). Initially, a total of $n_0 = 50$ subjects were considered in the experiment. Then the optimal design strategy was used to choose $n_1 + n_2 = 100$ new subjects with n_1 subjects for the control group and n_2 subjects for the treatment group. It appears from the plots in Figure 1 that the optimal value of n_1 depends on the values of the treatment variances (σ_0, σ_1). Typically, when $\sigma_0 = \sigma_1$, the optimum allocation chooses roughly equal values of n_1 and n_2. But when the variances are unequal, the optimal n_1 for the control group appears to be inversely related to the variance component σ_0 for the control group, i.e., n_1 is higher when σ_0 is smaller ($\sigma_0 < \sigma_1$).

We ran a set of simulations based on 1000 replicates of data sets, where each data set was generated using the linear mixed model (18), with the regression parameters fixed at $\alpha_0 = 2, \alpha_1 = 1$, $\beta_1 = 0.5$, $\beta_2 = 0.5$, and the error variance at $\sigma_\epsilon = 1$. Figure 2 displays histograms of the optimum allocations n_1 for 1000 replicates of data sets and for three combinations of random effects variances: $(\sigma_0, \sigma_1) = (2, 2)$ (left panel), $(\sigma_0, \sigma_1) = (3, 1)$ (middle panel), and $(\sigma_0, \sigma_1) = (1, 3)$ (right panel). As before, we chose the initial number of subjects as $n_0 = 50$. Optimal design strategy was used to choose $n_1 + n_2 = 100$ new subjects with n_1 subjects for the control group and n_2 for the treatment group. It is clear from the plots that when the variances are equal, i.e., $(\sigma_0, \sigma_1) = (2, 2)$, the average value of n_1 is roughly 50, indicating an equal allocation of n_1 and n_2 in average. On the other hand, when the variances are unequal, the optimal allocations generally choose different values of n_1 and n_2. For example, the average optimal value of n_1 is roughly 30 when $(\sigma_0, \sigma_1) = (3, 1)$, whereas average n_1 is roughly 70 when $(\sigma_0, \sigma_1) = (1, 3)$.

In the aforementioned simulation study, we also investigated the gain in efficiency from the proposed optimal allocation, as compared to the naive equal allocation ($n_1 = n_2$). The efficiencies were calculated as the ratio of the determinants of the Fisher information under the optimal and

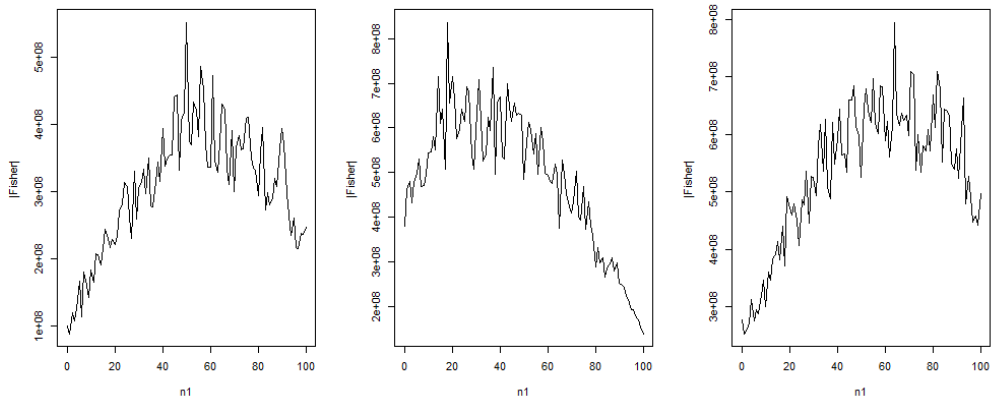

Figure 1: Plots of sample size n_1 against determinant of the Fisher information for three representative data sets. Regression parameters in (18): $\alpha_0 = 2, \alpha_1 = 1$, $\beta_1 = 0.5$, $\beta_2 = 0.5$; error variance: $\sigma_\epsilon = 2$. Random effects variances: $(\sigma_0, \sigma_1) = (2, 2)$ (left panel); $(\sigma_0, \sigma_1) = (2, 1)$ (middle); $(\sigma_0, \sigma_1) = (1, 2)$ (right).

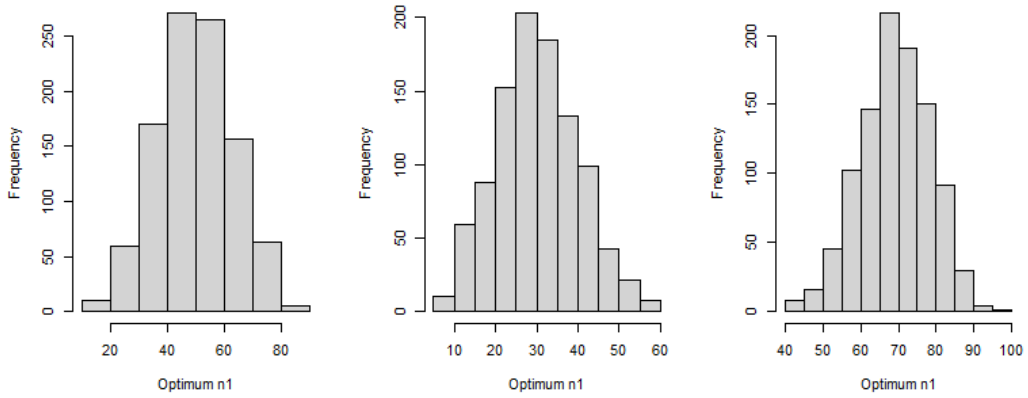

Figure 2: Histogram plots of optimum values of n_1 for 1000 replicates of data sets. Regression parameters in (18): $\alpha_0 = 2, \alpha_1 = 1$, $\beta_1 = 0.5$, $\beta_2 = 0.5$; error variance: $\sigma_\epsilon = 1$. Random effects variances: (left panel) $(\sigma_0, \sigma_1) = (2, 2)$ (left panel); $(\sigma_0, \sigma_1) = (3, 1)$ (middle); $(\sigma_0, \sigma_1) = (1, 3)$ (right).

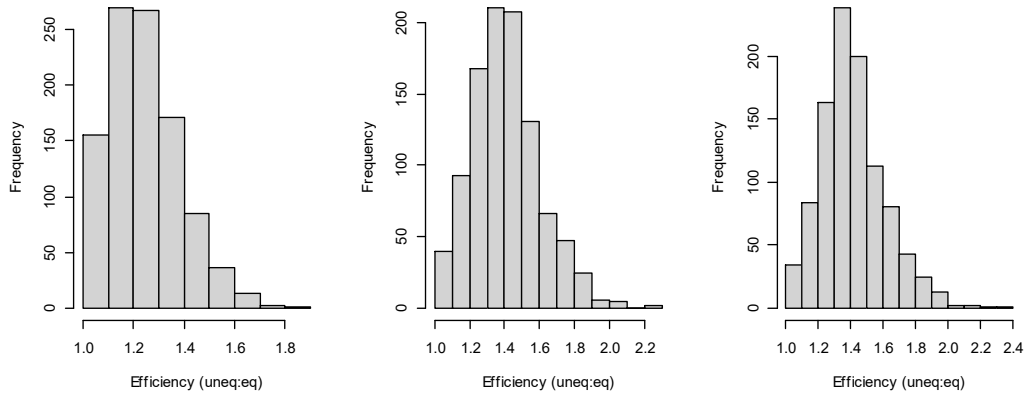

Figure 3: Efficiencies of optimum allocations for 1000 replicates of data sets. Efficiencies are calculated as the ratio of the determinants of the Fisher information under optimal and equal allocations. Regression parameters in (18): $\alpha_0 = 2, \alpha_1 = 1, \beta_1 = 0.5, \beta_2 = 0.5$; error variance: $\sigma_\epsilon = 1$. Random effects variances: $(\sigma_0, \sigma_1) = (2, 2)$ (left panel); $(\sigma_0, \sigma_1) = (3, 1)$ (middle); $(\sigma_0, \sigma_1) = (1, 3)$ (right).

equal allocations. Figure 3 presents histograms of the efficiencies of optimum allocations for 1000 replicates of data sets as considered earlier, for three combinations of random effects variances: $(\sigma_0, \sigma_1) = (2, 2)$ (left panel), $(\sigma_0, \sigma_1) = (3, 1)$ (middle panel), and $(\sigma_0, \sigma_1) = (1, 3)$ (right panel). As before, the proposed optimal design strategy was used to choose $n_1 + n_2 = 100$ new subjects with n_1 subjects for the control group and n_2 for the treatment group. It is clear from the plots that when the treatment variances are different, on average, the optimum allocation provides roughly 40% gain in efficiency as compared to the equal allocation.

6 Application: PANSS Data

Here we present an application of the proposed optimal design strategy using some actual clinical data obtained from a longitudinal trial of drug therapies for schizophrenia, initially studied by Diggle (1998). This randomized clinical trial compares different drug regimes in the treatment of chronic schizophrenia. The study was based on 523 patients who were randomly allocated to placebo and five active agents (treatments). The primary response was a measure of psychiatric disorder determined by the total score on the Positive and Negative Symptom Rating Scale (PANSS). We use a subset of the data available at the link: https://www.lancaster.ac.uk/staff/diggle/APTS-data-sets, which contains longitudinal measurements from 150 patients randomized among three treatment groups (placebo and two active agents) each of size 50, where patients were measured at weeks 0 (baseline), 1, 2, 4, 6, and 8. The longitudinal data contain a number of missing values due to

Table 1: Linear mixed model fit to PANSS data. The initial data contain $n = 150$ patients measured at six follow-up times 0, 1, 2, 4, 6, and 8 (in weeks).

Parameter	Estimate	SE	z-value
α_0	95.0389	2.4038	39.5376
α_1	–8.5368	2.9878	–2.8572
β	–1.1378	0.2120	–5.3677
σ_ϵ^2	176.7771	10.8221	16.3348
σ_0^2	228.4355	55.8720	4.0886
σ_1^2	291.2141	47.4575	6.1363

dropouts. We assume that these values are missing at random (MAR). Under the MAR mechanism, the likelihood approach does not require any missing data model when estimating the model parameters.

We use a linear mixed effects model to fit the initial data as

$$y_{it} = \alpha_0 + \alpha_1 z_i + \beta \text{Time}_{it} + u_i + \epsilon_{it}, \tag{6.1}$$

for $i = 1, \ldots, 150$, and $t = 1, \ldots, 6$, where the binary predictor z_i takes the value 0 for the placebo group and 1 for the treatment (both active agents) group, follow-up time Time_{it} takes the values 0, 1, 2, 4, 6, and 8 (in weeks), random effects u_i are assumed independent $N(0, (1 - z_i)\sigma_0^2 + z_i\sigma_1^2)$, and random errors ϵ_{it} are assumed independent $N(0, \sigma_\epsilon^2)$. The random effects u_i and random errors ϵ_{it} are also assumed mutually independent. We fit model (19) using the classical maximum likelihood (ML) method. Table 1 presents ML estimates, their standard errors and corresponding z-values for the regression parameters and variance components. The estimates of the variance components (σ_0^2, σ_1^2) appear to be different for the placebo and treatment groups.

Now consider augmenting the data by allocating 100 new patients between the placebo and treatment groups, so that there are n_1 new patients in the placebo group and $n_2 = 100 - n_1$ new patients in the treatment group. Figure 4 shows an optimal value of $n_1 = 75$, as determined by the Fisher information shown in the left panel.

For illustrative purposes, we also studied the optimal allocation using a subset of the original data with 60 patients and treating those as initial data. Based on these "initial data", we recalculated the ML estimates of the model parameters, and then allocated 100 new patients between the placebo and treatment groups using the same optimal strategy considered earlier. For this subset of initial data, an optimal value of n_1 is obtained at $n_1 = 54$, as shown in the right panel of Figure 4. This demonstrates that the optimal allocation can depend on the behaviour of the initial data and their model fit.

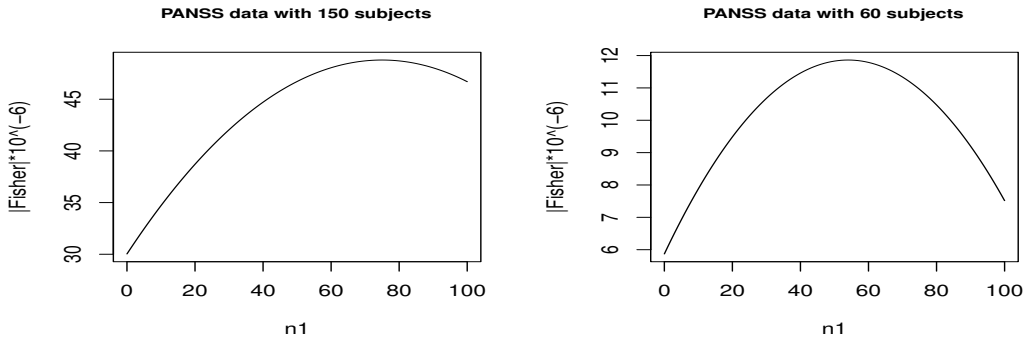

Figure 4: Analysis of PANSS data. The left panel shows the determinant of the Fisher information against sample size n_1 for an initial data set of $n_0 = 150$ patients. An optimal allocation of 100 new patients chooses $n_1 = 75$ patients for the control group and $n_2 = 25$ patients for the treatment group. The right panel shows results for an initial subset of $n_0 = 60$ patients in which the optimal allocation of 100 new subjects chooses $n_1 = 54$ patients for the control group and $n_2 = 46$ patients for the treatment group.

7 Conclusions

We have proposed and explored the construction of optimal allocation schemes for longitudinal studies in the framework of the linear mixed model for clustered data. We have shown that one can achieve considerable gain in efficiency by choosing the sample size for different treatment groups based on the proposed optimal schemes. Although the method is developed in the setting of a longitudinal study, it can also be used in other types of clustered data analysis, where the random cluster effects variances may be linked to demographic variables or some biomarkers.

The proposed optimal allocation schemes may also be extended to nonlinear mixed effects models for clustered data, where the response is still continuous, but the response variable is associated with available covariates and biomarkers by a known nonlinear function. Work remains to be done in this direction.

Acknowledgements

The research of both authors is partially supported by grants from the Natural Sciences and Engineering Research Council of Canada.

References

Chaudhuri, P. and Mykland, P. (1993), "Nonlinear experiments: optimal design and inference based on likelihood," *Journal of the American Statistical Association*, 88, 538–546.

Christenson, R. (1996), *Plane Answers to Complex Questions: The Theory of Linear Models*. New York: Springer-Verlag.

Demidenko, E. (2004). *Mixed models: Theory and applications*, John Wiley & Sons, Canada.

Diggle, P. J. (1998), "Dealing with missing values in longitudinal studies," In *Advances in the statistical analysis of medical data* (ed. B. S. Everitt and G. Dunn), pp. 203–28. Edward Arnold, London.

Hore, S., Dewanji, A., and Chatterjee, A. (2014), "Design issues related to allocation of experimental units with known covariates into two treatment groups," *Journal of Statistical Planning and Inference*, 155, 117–126.

Lindstrom, M.J. and Bates (1988), "Newton-Raphson and EM algorithms for linear mixed-effect models for repeated-measures data," *Journal of the American Statistical Association*, 83, 1014–1022.

Sinha, S. K. (2013), "Robust designs for multivariate logistic regression," *Metron*, 71, 157–173.

Sinha, S. K. and Wiens, D. P. (2002), "Robust sequential designs for nonlinear regression," *The Canadian Journal of Statistics*, 30(4), 601–618.

Wang, S.G. and Ma, W.Q. (2002), "On exact tests of linear hypothesis in linear models with nested error structure," *Journal of Statistical Planning and Inference*, 106, 225–233.

Wiens, D. P. (1994), " Robust designs for approximately linear regression: M-estimated parameters," *Journal of Statistical Planning and Inference*, 40, 135–160.

Xu, X. and Sinha, S. K. (2020), " Robust designs for generalized linear mixed models with possible model misspecification," *Journal of Statistical Planning and Inference*, 210, 20–41.

Received: May 20, 2023

Accepted: June 4, 2023