D-optimal population designs in linear mixed effects models for multiple longitudinal data

Hongyan Jiang & Rongxian Yue

Published online: 12 Feb 2021.

Submit your article to this journal

Article views: 49

View related articles

View Crossmark data

Taylor & Francis
Taylor & Francis Group

Check for updates

D-optimal population designs in linear mixed effects models for multiple longitudinal data

Hongyan Jiang[a] and Rongxian Yue[b]

[a]Department of Mathematics and Physics, Huaiyin Institute of Technology, Huaian, People's Republic of China; [b]Department of Mathematics, Shanghai Normal University, Shanghai People's Republic of China

ABSTRACT

The main purpose of this paper is to investigate D-optimal population designs in multi-response linear mixed models for longitudinal data. Observations of each response variable within subjects are assumed to have a first-order autoregressive structure, possibly with observation error. The equivalence theorems are provided to characterise the D-optimal population designs for the estimation of fixed effects in the model. The semi-Bayesian D-optimal design which is robust against the serial correlation coefficient is also considered. Simulation studies show that the correlation between multi-response variables has tiny effects on the optimal design, while the experimental costs are important factors in the optimal designs.

1. Introduction

In many statistical studies, it is quite common that repeated measured data are collected over time and are often referred to as longitudinal data. Mixed effects models are widely used for the analysis of such correlated data. However, the responses are usually time-structured when they are measured at different time points. Modeling of autocorrelations through use of autoregressive error structure, in addition with random effects terms, may be more appropriate (Chi & Reinsel, 1989). For a list of correlation structures, we also refer to the monographs by Verbeke and Molenberghs (2000) and Molenberghs and Verbeke (2005).

The aim of optimal experimental design is to arrange a reasonable design schedule, so that the experimenter can obtain the maximal information of the experiment at minimal cost. In the case of single-response with correlated errors under the AR(1) structure assumption, Tan and Berger (1999) studied optimal designs for linear mixed effects models with longitudinal data, while Berger et al. (2003) and Azurduy (2009) studied the robust designs in a maximin approach. Nyberg et al. (2012) considered the optimal designs for non-linear mixed effects models with serial correlation. Liu et al. (2019) discussed the D-optimal design for the multi-response linear mixed effects models by modelling the serial correlation with totally mixed effects. For the case of longitudinal data generated from the multi-response models with repeated measurements, correlated errors also play an important role in model description and estimation. Roy (2006) discussed how the correlation coefficient between two response variables was effected by the autocorrelation structure of repeated measures. Wang and Fan (2010) applied the multi-response linear mixed effects model (MLMM) with correlated errors to model repeated continuous data and presented ECM procedure to obtain the maximum likelihood estimates of model parameters. A general description of the statistical analysis of the multi-response longitudinal models can be found in Verbeke et al. (2014).

In this paper, we are interested in deducing the approximate designs of the fixed effects parameters in the multi-response linear mixed effects models with AR(1) autoregressive error structure. The generalised estimating equation (GEE) approach (Liang & Zeger, 1986) is used to estimate the model. The effects of the experimental cost on the optimal population design is also discussed.

The rest of the paper is organised as follows. Section 2 introduces the notation and provides the Fisher information matrices of individual design and population design. Section 3 gives the equivalence theorem for the D-optimal population design with respect to the estimation of the fixed effects and defines the relative D-efficiency by considering a linear experimental cost. The methodology is illustrated with a real example concerning the AIDS Clinical Trials Group Study in Section 4. Concluding remarks are given in Section 5.

CONTACT Rongxian Yue ✉ yue2@shnu.edu.cn 🖃 Department of Mathematics, Shanghai Normal University, Shanghai 200234, People's Republic of China

2. Model description and information matrix

2.1. Assumptions and preliminaries

Suppose there are N independent individuals in a longitudinal study and for each individual there are r response variables observed over time. Although each response variable may be observed under different design schedules, index matrices can be used (Ogungbenro et al., 2007) to describe it. Hence, a balanced design (Schmelter, 2007b), in which all r responses are measured under the same design schedule, is considered.

Let $\boldsymbol{y}_{ij} = (y_{ij1}, y_{ij2}, \ldots, y_{ijr})^T$ be an $r \times 1$ response vector observed at the jth measurement of individual i, and let $\boldsymbol{e}_{ij} = (e_{ij1}, e_{ij2}, \ldots, e_{ijr})^T$ be the within individual observation error vector with zero mean, $j = 1, 2, \ldots, m_i$. Let $\boldsymbol{G}$ be an $r \times r$ covariance matrix describing the correlation of the r different responses. The multi-response linear regression model with mixed effects is defined as

$$\boldsymbol{y}_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}\boldsymbol{b}_i + \boldsymbol{e}_{ij}, \quad (1)$$

where $\boldsymbol{X}_{ij}$ is an $r \times p$ fixed-effects design matrix, $\boldsymbol{Z}_{ij}$ is an $r \times l$ random-effects design matrix, $\boldsymbol{\beta}$ is a p-dimensional fixed-effects vector and $\boldsymbol{b}_i$ is an l-dimensional ($l \leq p$) random-effects vector with mean zero and $l \times l$ variance–covariance matrix $\boldsymbol{D}$ (Molenberghs & Verbeke, 2005). It is assumed that $\boldsymbol{D}$ is a positive definite matrix.

In general, in a longitudinal study, at least three repeated observations of the responses are required over time. We assume $\boldsymbol{t}_i = (t_{i1}, t_{i2}, \ldots, t_{im_i})$ ($m_i \geq 3$) is an m_i-point design (without repetition) for individual i with equal weight $1/m_i$ on each point, $i = 1, 2, \ldots, N$. The observations of different subjects are independent, but those within subject i conditioned on $\boldsymbol{b}_i$ are dependent. Let $\boldsymbol{R}_i$ be a working autocorrelation matrix among all m_i possible observations $y_{i1k}, y_{i2k}, \ldots, y_{im_ik}$ of the kth response variable. The common working correlation structure could be as simple as independent, equicorrelated, first-order autoregressive AR(1), or could be unspecified. In this paper, the autoregressive process of order one, AR(1) is used (Diggle, 1988; Wang & Fan, 2010), and then the autocorrelation matrix of the individual i is $\boldsymbol{R}_i = (\rho^{|t_{ij}-t_{ij'}|})$, where ρ ($0 \leq \rho \leq 1$) is the autocorrelation coefficient that gives the correlation between two time points that are one unit of time apart. This correlation structure implies that repeated measurements are more highly correlated in closer time and become less correlated as the difference in time points increases.

For each individual i, denoting $\boldsymbol{y}_i = (\boldsymbol{y}_{i1}^T, \boldsymbol{y}_{i2}^T, \ldots, \boldsymbol{y}_{im_i}^T)^T$ and $\boldsymbol{e}_i = (\boldsymbol{e}_{i1}^T, \boldsymbol{e}_{i2}^T, \ldots, \boldsymbol{e}_{im_i}^T)^T$, the multi-response linear mixed model in (1) is expressed as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i, \quad (2)$$

where $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}^T, \boldsymbol{X}_{i2}^T, \ldots, \boldsymbol{X}_{im_i}^T)^T$ and $\boldsymbol{Z}_i = (\boldsymbol{Z}_{i1}^T, \boldsymbol{Z}_{i2}^T, \ldots, \boldsymbol{Z}_{im_i}^T)^T$. It is assumed that the random-effects vector $\boldsymbol{b}_i$ is independent with $\boldsymbol{e}_i$. By letting $\boldsymbol{A}_i$ be the diagonal marginal variance matrix of $\boldsymbol{y}_i$, the covariance matrix $\boldsymbol{H}_i$ of $\boldsymbol{y}_i$ in model (2) can be expressed as (Wang & Fan, 2010)

$$\boldsymbol{H}_i = \mathrm{Cov}(\boldsymbol{y}_i) = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \boldsymbol{A}_i^{1/2}(\boldsymbol{R}_i \otimes \boldsymbol{G})\boldsymbol{A}_i^{1/2}. \quad (3)$$

According to the GEE method, the Fisher information matrix of a single individual design $\boldsymbol{t}_i = (t_{i1}, t_{i2}, \ldots, t_{im_i})$ for individual i is

$$\boldsymbol{M}_{\mathrm{ind}}(\boldsymbol{t}_i) = \boldsymbol{X}_i^T\boldsymbol{H}_i^{-1}\boldsymbol{X}_i. \quad (4)$$

The Fisher information matrix in (4) depends on the value of ρ, which is often not explicitly known in an experiment. Assuming that ρ has a range of possible values, we investigate a robust D-optimal design criterion (see Section 3).

2.2. Information matrix of the population design

Consider now a population design comprising Q distinct individual designs with n_q individuals allocated to the design with m_q time points $\boldsymbol{t}_q = (t_{q1}, t_{q2}, \ldots, t_{qm_q})$ for $q = 1, 2, \ldots, Q$ (Schmelter, 2007a, 2007b). Denote this population design as follows:

$$\zeta = \begin{Bmatrix} \boldsymbol{t}_1 & \boldsymbol{t}_2 & \cdots & \boldsymbol{t}_Q \\ w_1 & w_2 & \cdots & w_Q \end{Bmatrix}, \quad (5)$$

where $w_q = n_q/N$. To turn the discrete optimisation problem into a continuous one, we allow approximate designs, where the weights w_q's can be arbitrary numbers in the interval $(0, 1)$ summing up to 1. Then, the Fisher information matrix for the fixed-effects vector $\boldsymbol{\beta}$ at the population design ζ is defined by

$$\boldsymbol{M}(\zeta) = \sum_{q=1}^{Q} w_q\boldsymbol{M}_{\mathrm{ind}}(\boldsymbol{t}_q), \quad (6)$$

where $\boldsymbol{M}_{\mathrm{ind}}(\boldsymbol{t}_q)$ is determined as in (4). By defining $\boldsymbol{X} = (\boldsymbol{X}_1^T, \boldsymbol{X}_2^T, \ldots, \boldsymbol{X}_Q^T)^T$, $\boldsymbol{H} = \mathrm{diag}(\boldsymbol{H}_1, \boldsymbol{H}_2, \ldots, \boldsymbol{H}_Q)$ and $\boldsymbol{W} = \mathrm{diag}(w_1, w_2, \ldots, w_Q)$, the matrix $\boldsymbol{M}(\zeta)$ in (6) can be expressed as

$$\boldsymbol{M}(\zeta) = \boldsymbol{X}^T\boldsymbol{H}^{-1/2}\boldsymbol{W}\boldsymbol{H}^{-1/2}\boldsymbol{X}. \quad (7)$$

3. The *D*-optimal design and robust design

3.1. The general equivalence theorem

A D-optimal population design, ζ^*, for estimating the fixed effects in the MLMM (1) maximises the logarithm of the determinant of the Fisher information matrix $\boldsymbol{M}(\zeta)$ given in (7), i.e.,

$$\zeta^* = \arg\max_{\zeta} \log |\boldsymbol{M}(\zeta)|. \quad (8)$$

The General Equivalence Theorem relating to the approximate D-optimal population design is given in

the following, which is based on very mild assumptions that the set of all population designs ζ is compact on the design space and the D-optimality criterion $\Psi(\zeta) = -\log|\boldsymbol{M}(\zeta)|$ is convex and differentiable. The Frechet derivative of $\Psi(\zeta)$ in the direction of any individual design $\boldsymbol{t}$ is $\phi(\boldsymbol{t}, \zeta) = p - \text{tr}(\boldsymbol{M}^{-1}(\zeta)\boldsymbol{M}_{\text{ind}}(\boldsymbol{t}))$.

Theorem 3.1: *A population design ζ^* is D-optimal for the estimation of the fixed-effects vector $\boldsymbol{\beta}$ if and only if*

(1) *The design ζ^* maximises $\log|\boldsymbol{M}(\zeta)|$.*
(2) *The design ζ^* minimise $\max_{\boldsymbol{t}} \text{tr}(\boldsymbol{M}^{-1}(\zeta)\boldsymbol{M}_{\text{ind}}(\boldsymbol{t}))$.*
(3) *The maximum of $\text{tr}(\boldsymbol{M}^{-1}(\zeta^*)\boldsymbol{M}_{\text{ind}}(\boldsymbol{t}))$ is attained at the support individual designs of ζ^*, and the maximum value is p.*

This theorem can be used to construct or check the optimality of a population design (5), which provides a basis for the following discussion of optimal designs considering the experimental spending.

3.2. Semi-Bayesian D-optimal design

Both the individual Fisher information matrix in (4) and the population Fisher information matrix in (7) depend on ρ, which is always unknown before starting an experiment. Assume that ρ obeys a distribution with probability density function $f(\rho)$ defined on an interval $[a, b]$ $(0 \leq a < b \leq 1)$. A design is said to be semi-Bayesian D-optimal, denoted by ζ_B^*, if it maximises

$$\Phi_B(\zeta) = \int_a^b \log|\boldsymbol{M}(\zeta)|f(\rho)\,\mathrm{d}\rho. \tag{9}$$

It is known that the semi-Bayesian D-optimal design is robust against a variety of ρ.

In order to calculate the integral defined by (9), a set of 901 points is drawn from the prior distribution of ρ. Numerical approximations are applied by sampling parameter values from the prior distribution of $f(\rho)$ and replacing the integral in (9) with the sample mean. All computations are carried out in MATLAB R2010a (64bt) using a computer with 2.40GHz Intel processor and 4GB RAM. The Matlab function 'fmincon' is used to find the maximum of (9) and the optimal design. The aim of Bayesian design criterion is to find a design which maximises $\Phi_B(\zeta)$, which is equivalent to find a design which minimises the opposite of $\Phi_B(\zeta)$. 'Fmincon' just attempts to find a constrained minimum of a scalar function of several variables starting at an initial estimate. The equidistant design is used as the starting points in our program, and the optimal design is not sensitive to the starting points. The computing time to generate a Bayesian optimal design is about 3 hours, and the running time increases with the increasing number of time points.

For longitudinal data, the more measurements per individual are taken, the more additional information provides. However, the number of measurements per individual is usually limited because of the experimental sampling budget or logistical restrictions. The linear cost function defined by Tekle et al. (2008) is expanded to the population design ζ for the model with r responses

$$C = \sum_{q=1}^{Q} n_q(c_1 + c_2 r m_q), \tag{10}$$

where c_1 is the cost of recruiting a new individual, c_2 represents the average cost per measurement, and $\sum_{q=1}^{Q} n_q = N$.

Thus, as proposed in Fedorov et al. (2002), the Fisher information matrix in (7) under the population design ζ of the form (5) should be normalised for the cost constraint in (10), that is

$$\widetilde{\boldsymbol{M}}(\zeta) = \frac{1}{C}\sum_{q=1}^{Q} w_q \boldsymbol{M}_{\text{ind}}(\boldsymbol{t}_q) \triangleq \sum_{q=1}^{Q} \tilde{w}_q \widetilde{\boldsymbol{M}}_{\text{ind}}(\boldsymbol{t}_q), \tag{11}$$

where $\widetilde{\boldsymbol{M}}_{\text{ind}}(\boldsymbol{t}_q) = \boldsymbol{M}_{\text{ind}}(\boldsymbol{t}_q)/(c_1 + c_2 r m_q)$ is the standard information matrix at the individual design $\boldsymbol{t}_q$ and

$$\tilde{w}_q = \frac{n_q(c_1 + c_2 r m_q)}{CN} = \frac{n_q(c_1 + c_2 r m_q)}{N\sum_{q=1}^{Q} n_q(c_1 + c_2 r m_q)}.$$

The optimal population design problem is to identify the optimal weight, and the locations of the time points of each individual design. The value of frequency n_q of the individual design $\boldsymbol{t}_q$ can be obtained by rounding $\tilde{w}_q CN/(c_1 + c_2 r m_q)$ to the nearest integer subject to $\sum_{q=1}^{Q} n_q(c_1 + c_2 r m_q) \leq C$. For comparing the efficiencies of two population designs ζ_1 and ζ_2 under the cost constraint, the D-efficiency of ζ_1 relative to ζ_2 is defined by

$$RE(\zeta_1; \zeta_2) = \left(\frac{|\widetilde{\boldsymbol{M}}(\zeta_1)|}{|\widetilde{\boldsymbol{M}}(\zeta_2)|}\right)^{1/p}. \tag{12}$$

For the normalised Fisher information matrix $\widetilde{\boldsymbol{M}}(\zeta)$, the general equivalence theorem still holds, and the verification is given in Section 4 by a simulation study.

Theorem 3.2: *For a given linear cost function, a population design ζ^* is D-optimal for the estimation of the fixed-effects vector $\boldsymbol{\beta}$ if and only if*

(1) *The design ζ^* maximises $\log|\widetilde{\boldsymbol{M}}(\zeta)|$.*
(2) *The design ζ^* minimise $\max_{\boldsymbol{t}} \text{tr}(\widetilde{\boldsymbol{M}}^{-1}(\zeta)\widetilde{\boldsymbol{M}}_{\text{ind}}(\boldsymbol{t}))$.*
(3) *The maximum of $\text{tr}(\widetilde{\boldsymbol{M}}^{-1}(\zeta)\widetilde{\boldsymbol{M}}_{\text{ind}}(\boldsymbol{t}))$ is attained at the support individual designs of ζ^*, and the maximum value is p.*

4. Numerical studies

In this section, a linear and quadratic mixed effects model is considered to show the effects of autocorrelation coefficient ρ on optimal designs. The optimal designs for the bivariate AIDS model studied in Wang and Fan (2010) is also discussed.

4.1. Simulation example

Consider a linear and quadratic mixed effects model

$$\begin{cases} y_1 = \beta_{10} + \beta_{11}t + b_{10} + e_1, \\ y_2 = \beta_{20} + \beta_{21}t + \beta_{22}t^2 + b_{20} + b_{21}t + e_2. \end{cases} \quad (13)$$

Assuming that each individual is observed under time schedule $t_1 = (t_1, t_2, \ldots, t_m)$. The unknown parameters are given previously to find the optimal population design. Let $\beta = (1.5, -3, -1, 2, 3)^T$, and the covariance matrix of the two responses y_1 and y_2 is

$$G = \begin{pmatrix} 5 & \gamma_{12}\sqrt{10} \\ \gamma_{12}\sqrt{10} & 2 \end{pmatrix}.$$

The covariance matrix of the three mixed effects variables b_{10}, b_{20}, b_{21} is

$$D = \begin{pmatrix} 2 & 0.1 & 0.5 \\ 0.1 & 4 & 0.9 \\ 0.5 & 0.9 & 3 \end{pmatrix}.$$

The design region is $[-1, 1]$.

We first show that the correlation between the two characteristics has slightly effect on the optimal designs. Fixing $\rho = 0.2$, the individual D-optimal design for different values of $m = 3, 4, 5, 6$ and $\gamma_{12} = 0, \pm0.02, \pm0.1, \pm0.3, \pm0.6, \pm0.9$ are shown in Figure 1. All the optimal designs have design points located in the left and right endpoints of the design area, for value of γ_{12} in the neighbourhood of 0 that the optimal designs are almost the same, with the increase of absolute value of γ_{12}, there are some subtle differences between the allocation of the optimal designs. In addition, the relative D-efficiency of the individual optimal designs at $\gamma_{12} \neq 0$ compared to $\gamma_{12} = 0$ are all almost 1, whatever the number of design points is. Hence, the correlation between characteristics has slight effects on the D-optimal designs, which is consistent with the results in Atkinson et al. (2007). Hence, we take $\gamma_{12} = 0.02$ in the model (13) and consider the effects of ρ on D-optimal designs in what follows.

The individual D-optimal allocations of time points for different autocorrelation coefficient ρ ($\rho = 0$, $0.001, 0.01, 0.1, 0.2, \ldots, 0.9$) are presented in Figure 2, the number of design points are $m = 3, 4, 5$, and 6, respectively. By assuming that ρ obeys an uniform distribution on $[0.05, 0.95]$, the Bayesian optimal designs (9) are also calculated, and the results are also shown in Figure 2. The asterisks show the D-optimal

allocations of time points when $\rho = 0$, the filled dots are the design position for $\rho \neq 0$ and the plus signs are the design position of Bayesian optimal designs. It can be observed that when the number of design points is three, the allocation of the optimal designs are almost the same. While for $m \geq 4$, when ρ equals zero, the optimal design points clustered at the left, middle and right endpoints of the design area, which maybe infeasible and less informative in practice. With the increase of ρ, the locations of the D-optimal design points are moderately affected by the autocorrelation coefficient ρ, especially for large values of ρ. And there isn't much difference between the Bayesian optimal design and locally D-optimal design when $\rho > 0.1$.

4.2. Optimal designs for the AIDS model

The bivariate longitudinal AIDS model is based on a set of data in the AIDS Clinical Trials Group Study 175, and only the post-week-12 data were considered, in which the CD4 and CD8 cell counts (per cubic millimetre) of a total of 20 patients were recorded per protocol about every 12 weeks thereafter on each patience visit. Let $y_{ij} = (\log_{10}(CD4)_{ij}, \log_{10}(CD8)_{ij})^T$ denote the $\log_{10}$-transformed CD4 and CD8 cell counts for patient i at the scaled time $t_{ij} = (week_{ij} - 8)/12 \in [1, 12]$, $i = 1, 2, \ldots, 20, j = 1, 2, \ldots, n_i$. The model considered here is assumed to be the following two-response linear mixed model with AR(1) errors

$$\begin{cases} y_{ij1} = \beta_{10} + \beta_{11}t_{ij} + b_{i10} + b_{i11}t_{ij} + e_{i1j}, \\ y_{ij2} = \beta_{20} + \beta_{21}t_{ij} + b_{i20} + b_{i21}t_{ij} + e_{i2j}. \end{cases} \quad (14)$$

According to Table 1 in Wang and Fan (2010), the maximum likelihood estimates of the fixed-effects vector β via the ECM algorithm based on the data is $\hat{\beta} = (2.5708, -0.0236, 2.8627, -0.0110)^T$, the estimate of the autocorrelation coefficient ρ is 0.3821, and the estimate of the covariance matrix of two response variables is $G = \begin{pmatrix} cc0.0083 & 0.0061 \\ 0.0061 & 0.0111 \end{pmatrix}$.

We now consider the D-optimal single-group design for model (14), where all individuals are observed under the same settings, that is,

$$\zeta_m^* = \begin{Bmatrix} t_m^* \\ 1 \end{Bmatrix},$$

where t_m^* is a single sequence of m time points in the design region $[1, 12]$. Suppose the experimental cost to be considered in an experiment. The relative efficiency defined in (12) is calculated for the cost of recruiting a new individual, $c_1 = 100$, and the average cost per measurement, c_2, ranging from 0.01 to 20. The D-optimal designs ζ_m^* are determined by Theorem 3.2. Figure 3 shows the relative efficiencies of ζ_m^* ($3 \leq m \leq 12$) compared to ζ_{12}^* for $c_1 = 100$ and $c_2 \in \{0.01, 0.1, 1, 2, 5, 20\}$. In each curve corresponding to a value of c_2, the optimal number of time points is the number m at which the
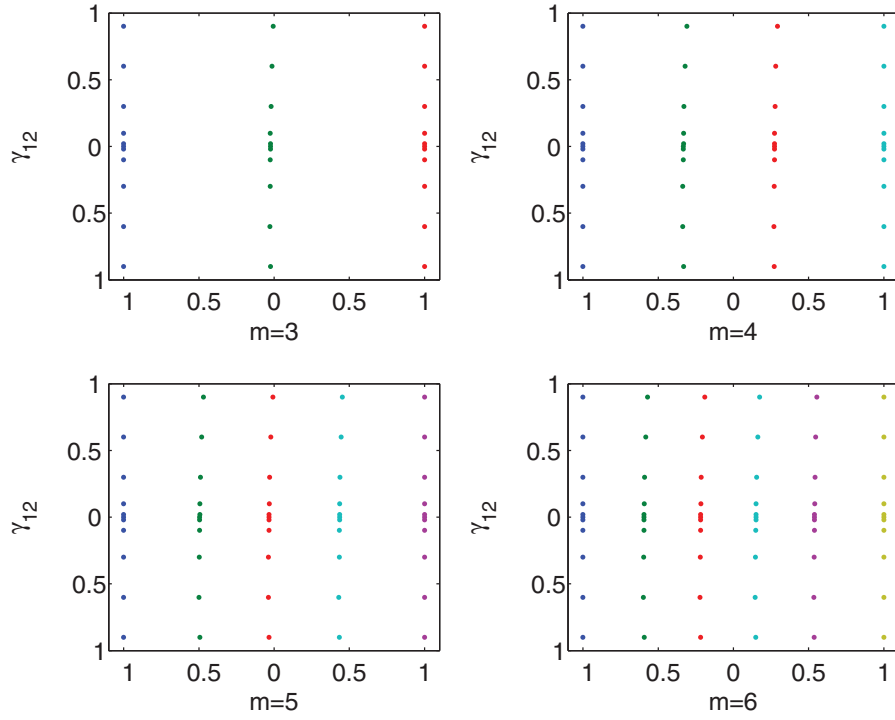
Figure 1. The individual *D*-optimal design allocations of time points for the linear and quadratic mixed effects model (13) for different values of *m* and γ_{12}.
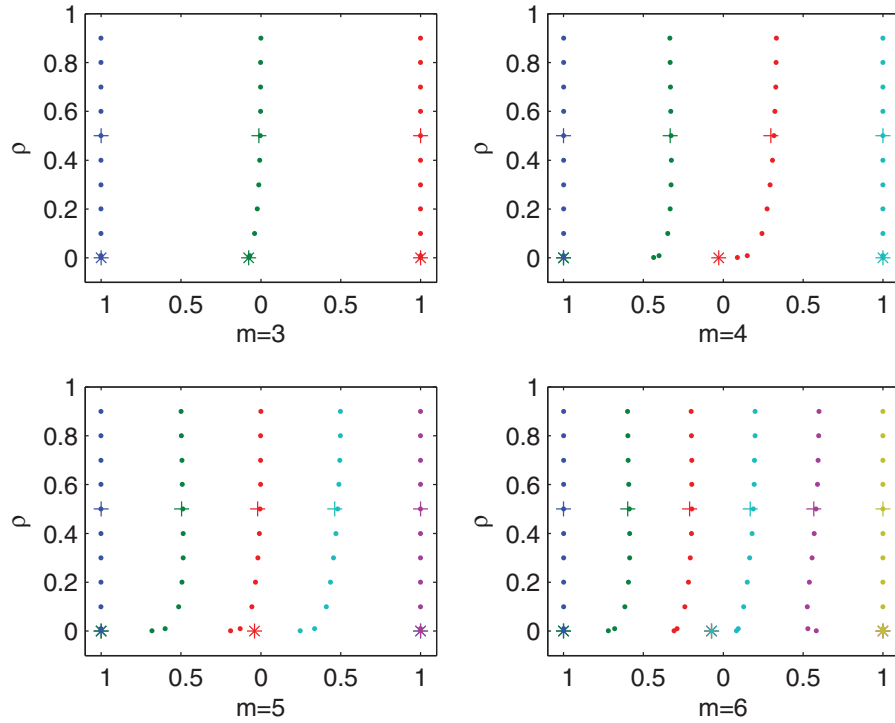

Figure 2. The individual optimal design allocations of time points for the linear and quadratic mixed effects model (13) for different values of *m* and ρ.

relative efficiency is the highest. It is observed that the number of optimal design points is $m = 3$ for $c_2 = 1$, 2, 5, 20. And the performances of the designs, ζ_m^* ($3 \leq m \leq 12$), are almost the same for $c_2 = 0.01$ and 0.1.

We now consider more-group designs. The set of possible sampling times considered is a predefined set *T* (Castañeda & López-Rios, 2016). For example, we let *T* be the 12-point single group balanced *D*-optimal

design for model (14), that is

$$t_{12}^* = \{1, 1.563, 2.162, 2.809, 3.519, 4.319, 5.268, 6.531, $$
$$9.007, 10.249, 11.196, 12\}.$$

In practice, experimenter can choose a set of possible sampling times by their experience or some prior information (Nyberg et al., 2012). We assume that three or
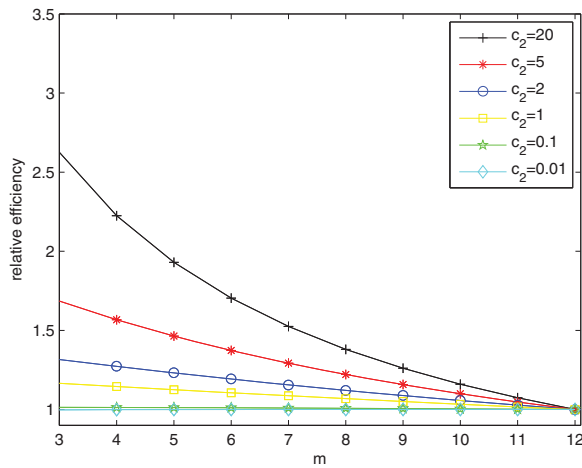
Figure 3. Relative efficiencies of ζ_m^* relative to ζ_{12}^* for the two-response linear mixed model (14) with costs $c_1 = 100$ and $c_2 \in \{0.01, 0.1, 1, 2, 5, 20\}$.

four measurements are available for each individual, and there is no replication at a certain time point. It is noted from Figure 3 that a large value of c_2 results in a small sample size, and we then consider the cost function (10) where $c_1 = 100$, $c_2 = 1.7$, for example. We find the D-optimal population design as follows:

$$\zeta^* = \left\{ \begin{matrix} (1, 2.809, 12) & (1, 2.162, 4.319, 12) \\ 0.2352 & 0.7648 \end{matrix} \right\},$$

which can be verified by Theorem 3.2. Therefore, according to Gagnon and Leonov (2005) and (11), the ratio of the frequencies is

$$\frac{N_1}{N_2} = \frac{w_1}{w_2} \cdot \frac{c_1 + 2c_2 m_2}{c_1 + 2c_2 m_1} = 0.3170,$$

which means that about 24.07% of the population are observed at $\{1, 2.8090, 12\}$, and about 75.93% should be measured at $\{1, 2.1623, 4.3190, 12\}$. Obviously, this population optimal design is more efficient than the single design ξ_{12}^* when considering the cost.

On the other hand, the result is a little different from Schmelter (2007a, 2007b), and each individual design involved in ζ^* is no longer D-optimal. When the value of c_2 varies, the individual designs remain the same, while the weight of each individual design changes, and the population design tends to single three or four points single-group design.

5. Conclusion

The problem of choosing the optimal population design is investigated for the multi-response linear mixed effects models with correlated errors. AR(1) structure is used to model the time series autocorrelation of the errors. A higher order AR may be worth consideration, however, it brings more unknown parameters into the model. Simulation study is used to show the effect of autocorrelation on designs and parameter estimations.

The equivalence theorems are provided to characterise the optimal population designs for the estimation of the fixed effects.

It should be noticed that the autocorrelation coefficient and the variance parameters in the random effect distribution are pre-specified in finding the D-optimal designs. In practice, the experimenter may get some prior information of the parameters, such as the range of parameter values. Then, a semi-Bayesian criterion is used in combination with the D-optimal criterion to find the optimal design robust to the autocorrelation coefficient. A topic for further research may be to construct optimal designs with respect to other optimality criteria, such as T-optimality criterion, or compound criterion. The result may be extended to polynomial mixed models with measurement error, and also to generalised linear mixed effects models.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Hongyan Jiang is a Ph.D. in Statistics, Shanghai Normal University, 2019, Associate Professor in Huaiyin Institute of Technology, Jiangsu, 2019–Present.

Rongxian Yue is a Ph.D. in Statistics, Hong Kong Baptist University, 1997. Postdoctoral Research Fellow in Statistics, East China Normal University, 1997–1999. Associate Professor (1999–2001) and Professor (2001–Present), Shanghai Normal University.

References

Atkinson, A. C., Donev, A. N., & Tobias, R. D. (2007). *Optimum experimental designs, with SAS*. Oxford University Press.

Azurduy, S. A. O. (2009). *Robust designs for longitudinal studies*. University of Maastricht.

Berger, M. P. F., Ouwens, M. J. N. M., & Tan, F. E. S. (2003). Robust designs for longitudinal mixed effects models. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New developments in psychometrics*. Springer.

Castañeda, L. M. E., & López-Rios, V. I. (2016). Optimal population designs for discrimination between two nested nonlinear mixed effects models. *Revista Ciencia En Desarrollo, 7*(1), 71–81. https://doi.org/10.19053/01217488.4233

Chi, E. M., & Reinsel, G. C. (1989). Models for longitudinal data with random effects and AR(1) errors. *Journal of the American Statistical Association*, *84*(406), 452–459. https://doi.org/10.1080/01621459.1989.10478790

Diggle, P. J. (1988). An approach to the analysis of repeated measurements. *Biometrics*, *44*(4), 959–971. https://doi.org/10.2307/2531727

Fedorov, V. V., Gagnon, R. C., & Leonov, S. L. (2002). Design of experiments with unknown parameters in variance. *Applied Stochastic Models in Business and Industry*, *18*(3), 207–218. https://doi.org/10.1002/(ISSN)1526-4025

Gagnon, R., & Leonov, S. (2005). Optimal population designs for PK models with serial sampling. *Journal of Biopharmaceutical Statistics*, *15*(1), 143–163. https://doi.org/10.1081/BIP-200040853

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *78*(1), 13–22. https://doi.org/10.1093/biomet/73.1.13

Liu, X., Yue, R. X., & Wong, W. K. (2019). D-optimal designs for multi-response linear mixed models. *Metrika*, *82*, 87–98. https://doi.org/10.1007/s00184-018-0679-7

Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data*. Springer Science+Business Media, Inc.

Nyberg, J., Höglund, R., Bergstrand, M., Karlsson, M. O., & Hooker, A. C. (2012). Serial correlation in optimal design for nonlinear mixed effects models. *Journal of Pharmacokinetics and Pharmacodynamics*, *39*(3), 239–249. https://doi.org/10.1007/s10928-012-9245-5

Ogungbenro, K., Gueorguieva, I., Majid, O., Graham, G., & Aarons, L. (2007). Optimal design for multiresponse pharmacokinetic-pharmacodynamic models – dealing with unbalanced designs. *Journal of Pharmacokinetics and Pharmacodynamic*, *34*(3), 313–331. https://doi.org/10.1007/s10928-006-9048-7

Roy, A. (2006). Estimating correlation coefficient between two variables with repeated observations using mixed effects model. *Biometrical Journal*, *48*(2), 286–301. https://doi.org/10.1002/(ISSN)1521-4036

Schmelter, T. (2007a). The optimality of single-group designs for certain mixed models. *Metrika*, *65*, 183–193. https://doi.org/10.1007/s00184-006-0068-5

Schmelter, T. (2007b). Considerations on group-wise identical designs for linear mixed models. *Journal of Statistical Planning and Inference*, *137*(12), 4003–4010. https://doi.org/10.1016/j.jspi.2007.04.017

Tan, F. E. S., & Berger, M. P. F. (1999). Optimal allocation of time points for the random effects model. *Communications in Statistics-Simulation and Computation*, *28*(2), 517–540. https://doi.org/10.1080/03610919908813563

Tekle, F. B., Tan, F. E. S., & Berger, M. P. F. (2008). D-optimal cohort designs for linear mixed-effects model. *Statistics in Medicine*, *27*(14), 2586–2600. https://doi.org/10.1002/sim.v27:14

Verbeke, G., Fieuws, S., Molenberghs, G., & Davidian, M. (2014). The analysis of multivariate longitudinal data: A review. *Statistical Methods in Medical Research*, *23*(1), 42–59. https://doi.org/10.1177/0962280212445834

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*, Springer Series in Statistics. Springer.

Wang, W. L., & Fan, T. H. (2010). ECM-based maximum likelihood inference for multivariate linear mixed models with autoregressive errors. *Computational Statistics and Data Analysis*, *54*(5), 1328–1341. https://doi.org/10.1016/j.csda.2009.11.021