Optimal Designs for Generalized Linear Mixed Models
using Penalized Quasi-likelihood Method

Yao Shi & John Stufken
Arizona State University, School of Mathematical and Statistical Sciences &
The University of North Carolina at Greensboro, Department of Mathematics and Statistics

School of Mathematical and Statistical Sciences
Arizona State University
PSA room 216, 901 S Palm Walk, Tempe, AZ, 85287
Email: yshi97@asu.edu

Introduction

In longitudinal studies, we may have many subjects, each with the characteristics measured repeatedly over time. For discrete responses, generalized linear models (GLMs) can be applied. Moreover, if the data is considered to be correlated, GLMMs could be applied. Especially, for binary longitudinal data, we could use a logistic link in GLMM.

In searching for a optimal design under GLMM with logistic link, the information matrix does not have a closed form and could be very slow to simulate. PQL method could be used to get an approximation of the information matrix, thus we study the performance by comparing the approximation with a numerical simulated exact information matrix and also MQL method.

Under different parameter settings, D- and A-optimal designs are found and the robustness of those designs are discussed with respect to mis-specified mean or variance-covariance matrix.

Numerical Simulation of the Information Matrix

Consider a GLMM with logistic link:

$$\log \frac{P(y_{ij}=1|b_i)}{1-P(y_{ij}=1|b_i)} = f(x_{ij})'b_i,$$

where y_{ij} is the jth observation of subject i with covariate value x_{ij}, $i = 1,\ldots,n$, $j = 1,\ldots,n_i$. Assuming b_i $i.i.d \sim N_p(\beta,\Sigma)$, the likelihood turns out to be

$$L(\beta) = \prod_{i=1}^{n}\int\prod_{j=1}^{n_i}\frac{\exp[f(x_{ij})'b_i]^{y_{ij}}}{1+\exp[f(x_{ij})'b_i]}\Phi(b_i;\beta,\Sigma)db_i = \prod_{i=1}^{n}L_i(\beta)$$

where Φ denotes the density of normal distribution.

The information matrix of β, say $I(\beta|X)$, is

$$-E_y\frac{\partial^2\log L}{\partial\beta\partial\beta'} = -\sum_{i=1}^{n}E_{y_i}\frac{\partial^2\log L_i}{\partial\beta\partial\beta'} = -nE_{y_i}\frac{\partial^2\log L_i}{\partial\beta\partial\beta'}.$$

The (h,q)th element of $E_{y_i}\frac{\partial^2\log L_i}{\partial\beta\partial\beta'}$ is

$$\sum_{y_i=z_i}\frac{\partial^2\log L_i(z_i)}{\partial\beta_h\beta_q}L_i(z_i) = -\sum_{y_i=z_i}\frac{1}{L_i(z_i)}\frac{\partial L_i(z_i)}{\partial\beta_h}\frac{\partial L_i(z_i)}{\partial\beta_q}.$$

This summation is over all possible outcomes of y_i, up to 2^{n_i}. It can take up to 800 seconds using an i5-8400 CPU for one design, which is too slow, but can serve as a good ruler of our approximation.

Approximation to the Information Matrix by PQL

Breslow and Clayton (1993) [1] discussed the penalized quasi-likelihood function to estimate β undel GLMMs. Based on PQL, for sequential design $\xi_i = \{x_{i1},\ldots,x_{ik_i}\}$, we can approximate the variance-covariance matrix of $\hat{\beta}$ given the true random effects by

$$cov(\hat{\beta}|b_i) \approx \left[\mathbf{F}_i'((\mathbf{V}_i)^{-1}+\mathbf{F}_i\Sigma\mathbf{F}_i')^{-1}\mathbf{F}_i\right]^{-1},$$

where $\mathbf{V}_i = diag(V_{i1},\ldots,V_{i,k_i})$ with V_{il} being the conditional variance function, which is $\frac{\exp[f(x_{il})'b_i]}{\{1+\exp[f(x_{il})'b_i]\}^2}$ here for logistic link, and $\mathbf{F}_i = (f(x_{i1}),\ldots,f(x_{in_i}))'$.

The information matrix for sequence i is then estimated by inverting $cov(\hat{\beta}|b_i)$:

$$\mathfrak{M}_i^{PQL}(\xi_i|\beta,\Sigma) = E_{b_i}(\mathbf{F}_i^T(\mathbf{V}_i^{-1}+\mathbf{F}_i\Sigma\mathbf{F}_i^T)^{-1}\mathbf{F}_i) \quad (1)$$

Suppose we have multiple sequences as $\xi = \{(\xi_i, w_i), i = 1,\ldots,n_s\}$, then the information matrix of ξ can be expressed as

$$\mathfrak{M}^{PQL}(\xi|\beta,\Sigma) = \sum_{i=1}^{n_s}w_i\mathfrak{M}_i^{PQL}(\xi_i|\beta,\Sigma) \quad (2)$$

The expectation in (1) can be obtained by a representative sample of b_i from the normal distribution $N(\beta,\Sigma)$, like a large enough random sample or other representative sets. We suggest using support points by Mak and Joseph (2018) [2] here.

Set $b_i = 0$, PQL approximation degenerates to MQL approximation, which is also mentioned in Breslow and Clayton (1993) [1], and can also be considered as an option.

Comparison

We then compare different approximations numerically under a 2-parameter logistic GLMM to ensure that we use a good proxy for the exact information matrix. β's are set to be $(1,-1)'$, $(3.5,-1)'$, and $(6,-1)'$, and variance structures are selected as $(\sigma_1,\sigma_2) = (1.7145, 1.05)$, $(6,.3)$ and $(.3,6)$, with $\rho = .5$. Only $\beta = (1,-1)'$ is shown here.

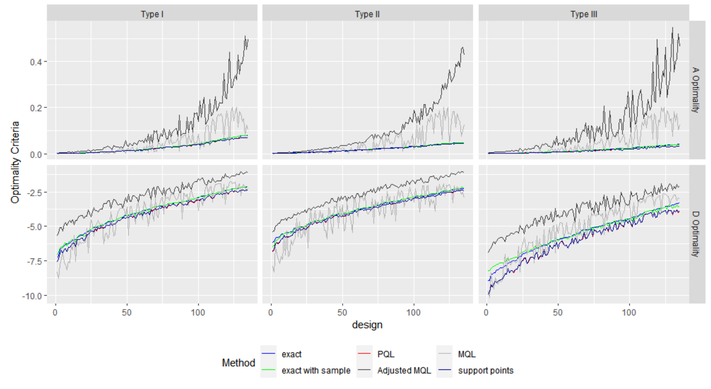

Figure 1: Comparing PQL, MQL and adjusted MQL approximations to the exact information matrix for $\beta = (1,-1)'$, exact information is evaluated on both numerical integration and random sample, and both random sample and support points are used in calculating (1)

- PQL indeed provides good enough approximation in keeping the order of the designs
- Support points can provide almost same results with a much smaller sample size than random sample (1000 v.s. 50, not shown in figure)

Optimal designs

Particle Swarm Optimization (PSO) is then applied to find locally A- and D-optimal designs for the mixed-effects logistic model, and we show one of the A-optimal results here. r, along the vertical axis, is a value multiplied with the variance matrix Σ to control the variation.

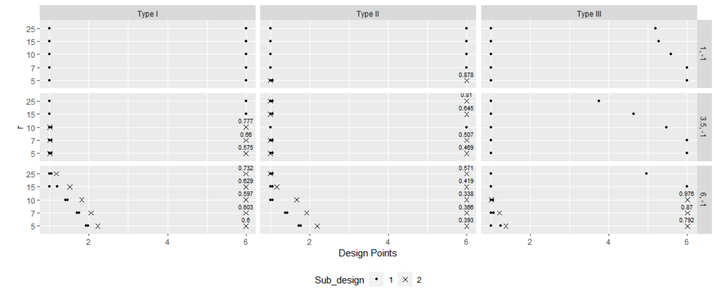

Figure 2: A-optimal designs with $n = 2$ observations per subject. The covariance type is shown in the top bar and the choice for β in the righthand bar. The value of r for the covariance matrix Σ is along the vertical axis, and the design region $[1,6]$ is shown along the horizontal axis.

- Two-sequence designs could be optimal

Robustness Study

Mis-specified Covariance Matrix

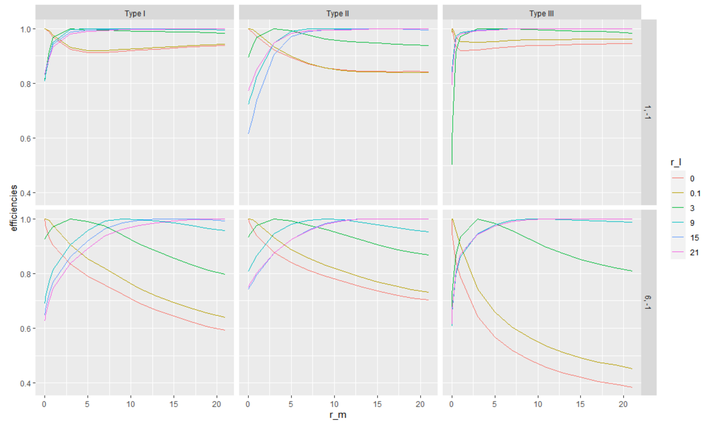

Figure 3: Robustness study: Efficiencies of D-optimal Designs, $n = 5$. The covariance type is shown in the top bar and the choice for β in the righthand bar. Value of r_m is along the horizontal axis, which corresponds to the true covariance matrix, and values of r_l are represented by the five different lines, each corresponds to the optimal design under such r_l. The efficiencies of these designs are shown along the vertical axis.

References

[1] N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. journal of the american statistical association. *Journal of the American Statistical Association*, 88(421):9–25, 1993.

[2] S. Mak and V. R. Joseph. Support points. *The Annals of Statistics*, 46:2562–2592, 2018.

[3] W. Yu and J. Stufken. Optimal designs for generalized linear mixed models using penalized quasi-likelihood method. *Unpublished*, 2019.