2. Generalized Linear Mixed Models

2.1. Introduction

- What is a GLMM?

- Some applications

Example 2.1. Fetal mortality in mouse litters

Source of data: Brooks et. al. (1997), Table 4: # of dead implants in 1328 litters of mice from untreated experimental animals.

Brooks et. al. (1997) used a Beta-binomial model

A mixed logistic model: Given random effects $\alpha_1, \ldots, \alpha_m$, y_{ij}, $1 \leq i \leq m$, $1 \leq j \leq n_i$ conditionally indep. binary,

$$\text{logit}\{\text{P}(y_{ij} = 1|\alpha)\} = \mu + \alpha_i.$$

Furthermore, α_i's indep. $\sim N(0, \sigma^2)$.

Example 2.2. The salamander mating data

McCullagh and Nelder, 1989, Ch. 14.5.

Three experiments: Summer 1986 (1), Fall 1986 (2).

The same group of 40 salamanders were used or the summer and first fall experiments. A new set of 40 animals was used in the second fall experiment.

McCullagh and Nelder (1989) propose a mixed logistic model with crossed random effects, a special case of GLMM.

For more applications, see, e. g., Breslow & Clayton (1993), Lee & Nelder (1996), Malec *et al.* (1997), Ghosh *et al.* (1998).

2.2. Definition of GLMM

a. Given $\alpha_1, \ldots, \alpha_m$, responses $y_1, \ldots, y_N$ are conditionally indep.;

b. the conditional density

$$f(y_i|\alpha) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c_i(y_i, \phi)\right\};$$

where ...

c. $\mu_i = E(y_i|\alpha)$,

$$g(\mu_i) = x_i'\beta + z_i'\alpha,$$

where β is a vector of unknown fixed effects.

d. It is often assumed that $\alpha \sim N(0, G)$, where $G = G(\theta)$ and θ is a vector of variance components.

2.3. Likelihood under GLMM

Example 2.3. Given $u_1, \ldots, u_{m_1}$ and $v_1, \ldots, v_{m_2}$, y_{ij}, $i = 1, \ldots, m_1$, $j = 1, \ldots, m_2$ cond. indep. binary with $p_{ij} = \mathrm{P}(y_{ij} = 1 | u, v)$ and $\mathrm{logit}(p_{ij}) = \mu + u_i + v_j$, where μ is an unknown parameter.

Furthermore, assume that the random effects are indep. with $u_i \sim N(0, \sigma_1^2)$, $v_j \sim N(0, \sigma_2^2)$.

The likelihood under this model can be expressed as

$$c - \frac{m_1}{2} \log(\sigma_1^2) - \frac{m_2}{2} \log(\sigma_2^2) + \mu y_{\cdot\cdot} + \log(\cdots),$$

where c is a constant, and

$$
\begin{aligned}
& \cdots \\
= & \int \cdots \int \left[\prod_{i=1}^{m_1} \prod_{j=1}^{m_2} \{1 + \exp(\mu + u_i + v_j)\}^{-1} \right] \\
& \times \exp \left(\sum_{i=1}^{m_1} u_i y_{i\cdot} + \sum_{j=1}^{m_2} v_j y_{\cdot j} \right. \\
& \left. -\frac{1}{2\sigma_1^2} \sum_{i=1}^{m_1} u_i^2 - \frac{1}{2\sigma_2^2} \sum_{j=1}^{m_2} v_j^2 \right) \\
& du_1 \cdots du_{m_1} dv_1 \cdots dv_{m_2}.
\end{aligned}
$$

2.4. Monte-Carlo EM (MCEM)

- EM algorithm

The "complete data", w, consists of the observed data, y, and unobserved data, ξ (e. g., the random effects).

E-step: Compute the conditional expectation

$$Q\{\theta|\theta^{(k)}\} \;=\; \mathrm{E}\left\{\log f(w|\theta)|y,\theta^{(k)}\right\}.$$

M-step: Maximizes $Q\{\theta|\theta^{(k)}\}$ with respect to θ to obtain the next step estimator $\theta^{(k+1)}$.

In GLMM, it is the E-step that causes problem.

• McCulloch (1994) used Gibbs sampler to approximate the E-step.

• McCulloch (1997) proposed three Monte-Carlo methods:

1. MCEM using Metropolis-Hastings algorithm for the E-step (e. g., Gelman *et al.* 2003);

2. A Monte-Carlo Newton-Raphson (MCNR) method with the help of Metropolis-Hastings algorithm;

3. A simulated maximum likelihood (SML) method using importance sampling.

Simulation results showed SML worked poorly compared to MCEM and MCNR.

• Booth & Hobert (1999) proposed two *automated* MCEM methods. Instead of using Markov chains, the latter methods used i. i. d. random sampling.

1. Importance sampling

The E-step is all about the calculation of

$$Q\{\psi|\psi^{(l)}\} = \mathrm{E}[\log\{f(y,\alpha|\psi)\}|y;\psi^{(l)}].$$

$$f(\alpha|y;\psi) \propto f(y|\alpha;\beta,\phi)f(\alpha|\theta),$$

where the normalizing constant is $f\{y|\psi^{(l)}\}$.

Let $\alpha_1^*, \ldots, \alpha_K^*$ be an i. i. d. sample generated from g, the importance sampling distribution. Then,

$$Q\{\psi|\psi^{(l)}\} \approx \frac{1}{K}\sum_{k=1}^{K} w_{kl} \log\{f(y, \alpha_k^*|\psi)\},$$

where

$$w_{kl} = \frac{f\{\alpha_k^*|y; \psi^{(l)}\}}{g(\alpha_k^*)}$$

is known as the *importance weights*.

Note: The unknown constant makes no difference in the M-step, because the maximization is over ψ (while the constant depends only on $\psi^{(l)}$).

Question: What g?

Booth and Hobert proposed to use a multivariate t-distribution whose mean and Var. match the Laplace approximations of the mean and Var. of $f(\alpha|y; \psi)$.

2. Rejection sampling

Write the conditional density as $f = cf_1f_2$, where c is the normalizing constant.

(i) First draw α from f_2 and, independently, u from the Uniform[0, 1] distribution.

(ii) If $u \leq f_1(\alpha)/\tau$, where $\tau = \sup_\alpha f_1(\alpha)$, accept α. Otherwise, return to (i).

• Advantages of i. i. d. sampling over Markov chains:

1. The assessment of the Monte-Carlo errors is straightforward. The latter is critical to the automated method (see below).

2. It is easier to establish central limit theorem (for normal approximation).

3. Faster: In a simulated example, the rejection and importance sampling methods are about 2.5 times and 30 times faster than the Metropolis-Hastings sampling method (McCulloch 1997).

2.5. Maximization by parts

Song *et al.* (2005) proposed a method which they called maximization by parts (MBP).

• Idea: Express the log-likelihood function as $l(\theta) = l_w(\theta) + l_e(\theta)$.

(i) The initial estimator, $\hat{\theta}_1$, is a solution to $\dot{l}_w(\theta) = 0$.

(ii) Then, use the equation $\dot{l}_w(\theta) = -\dot{l}_e(\hat{\theta}_1)$ to update in order to get $\hat{\theta}_2$.

(iii) Repeat (i) and (ii) until convergence.

• If BMP converges, the limit, $\hat{\theta}$, satisfies the likelihood equation $\dot{l}(\theta) = 0$.

• What $l_w(\theta)$?

A good choice of l_w is such that $\ddot{l}_e$ is smaller than $\ddot{l}_w$ in certain sense (Song *et al.* 2005).

Another condition for choosing l_w is that

$$\dot{l}_w(\theta) = 0$$

is an unbiased estimating equation.

Song *et al.* suggested that the hierarchical log-likelihood of Lee and Nelder (1996) could be used as $l_w(\theta)$ in the case of GLMM. However, the latter does not satisfy the above condition of unbiased estimating equation.

• How much does BMP help?

MBP has computational advantage in situations where $\ddot{l}$ is much more difficult to deal with (numerically or analytically) than $\dot{l}$. Example: The Gaussian copula model (e. g., Song 2000).

However, in GLMM, $\dot{l}$ is typically as difficult to evaluate as $\ddot{l}$. Still, there are many more $\ddot{l}$'s than $\dot{l}$'s.

Example 2.4. Suppose that, given α_i, $1 \le i \le m$, y_{ij} are cond. indep. with

$$\text{logit}\{\text{P}(y_{ij} = 1|u, v)\} = \beta_0 + \beta_1 x_{ij} + u_i + v_j,$$

and u_i's indep. $\sim N(0, \sigma^2)$.

For simplicity, assume that σ is known. Then, there are three different (one-dimensional) integrals in $\dot{l}$, and six different ones in $\ddot{l}$.

In general, if there are p unknown parameters, there may be as many as $p+1$ different integrals in $\dot{l}$, and as many as $(1/2)(p+1)(p+2)$ different integrals in $\ddot{l}$. If p is large, it is quite a saving in computation, provided that any single one of the integrals can be evaluated.

2.6. Bayesian inference

In addition to the GLMM assumptions, a prior for β and $G = \text{Var}(\alpha)$ is assumed.

- Posterior for (β, G)

$$f(\beta, G|y) = \frac{\int f(y|\beta,\alpha) f(\alpha|G) \pi(\beta, G) d\alpha}{\int f(y|\beta,\alpha) f(\alpha|G) \pi(\beta, G) d\alpha d\beta dG}$$

- Posterior for α

$$f(\alpha|y) = \frac{\int f(y|\beta,\alpha) f(\alpha|G) \pi(\beta, G) d\beta dG}{\int f(y|\beta,\alpha) f(\alpha|G) \pi(\beta, G) d\alpha d\beta dG}$$

• Computation: Gibbs sampler (e. g., Karim and Zeger 1992), etc.

• Advantage: Posterior rather than point estimates.

• Disadvantage: (i) computational intensive; (ii) imporper posterior (e. g., Hobert & Casella 1996).

2.7. Approximate inference

• Laplace approximation to integrals

Wish to approximate

$$\int \exp\{-q(x)\}dx,$$

where $q(\cdot)$ is "well-behaved" in that it achieves its minimum at $x = \tilde{x}$ with $q'(\tilde{x}) = 0$ and $q''(\tilde{x}) > 0$.

By Taylor expansion,

$$q(x) = q(\tilde{x}) + \frac{1}{2}q''(\tilde{x})(x - \tilde{x})^2 + \cdots$$

Thus, we have

$$\int \exp\{-q(x)\}dx \approx \sqrt{\frac{2\pi}{q''(\tilde{x})}} \exp\{-q(\tilde{x})\}.$$

• Using a multivariate version of the Laplace approximation, Breslow & Clayton (1993) developed a penalized quasi-likelihood method, or PQL. Similar methods were also proposed, e. g., by Schall (1991), Wolfinger & O'Connell (1993), McGilchrist (1994), and Lin & Breslow (1996).

• Lee & Nelder's (1996) hierarchical likelihood method is similar to PQL in spirit, but allows non-Gaussian distributions for the random effects.

• Advantage: (i) The Laplace-approximation-based methods are computationally attractive. (ii) The methods also provide estimates of the random effects.

• Disadvantage: Unfortunately, the methods are known to have some unsatisfactory properties. In particular, the resulting estimators are inconsistent under standard asymptotic assumptions (e. g., Jiang 1998). Furthermore, Lin and Breslow (1996) that PQL works well when the variances of the random effects are close to zero; otherwise, the bias can be substantial. Also see Kuk (1995).

2.8. Estimating equations

• Generalized estimating equations (GEE)

The method is well known in the analysis of longitudinal data (e. g., Liang and Zeger 1986, Prentice 1988).

However, the GEE does not apply to models with crossed random effects, such that in the salamander-mating example.

• Jiang (1998) proposed estimating equations that apply to GLMMs in general. The method leads to consistent estimators of the fixed effects and variance components. However, the estimators are inefficient.

• Jiang and Zhang (2001) proposed a two-step procedure to obtain more efficient estimators. Let S be a vector of base statistics. The 1st-step estimator of θ, $\tilde{\theta}$, is a solution to

$$B\{S - u(\theta)\} = 0,$$

where B is a known matrix, and $u(\theta) = \mathrm{E}_\theta(S)$.

• What B?

The optimal B is known to be $B^* = U'V^{-1}$, where $U = \partial u/\partial\theta'$ and $V = \mathrm{Var}_\theta(y)$.

Unfortunately, the best B depends on θ.

- The 2nd-step estimator, $\hat{\theta}$, is a solution to

$$\tilde{B}^*\{S - u(\theta)\} = 0,$$

where $\tilde{B}^* = B^*(\tilde{\theta})$.

- How does it work?

(i) The method only requires specification of the first two conditional moments of the data given the random effects, so it applies to a class of models wider than GLMMs.

(ii) Both 1st- and 2nd- estimators are consistent; 2nd-step estimators are efficient among a class of estimators.

(iii) Furthermore, Jiang and Zhang (2003) reported results from two simulated examples; in each case the 2nd-step estimator had about 40% reduction of the MSE compared to the 1st-step estimator.

2.9. Prediction of random effects

• Joint maximization of fixed and random effects provides estimates of random effects.

• Another look: Maximum a posterior (Jiang *et al.* 2001).

$$\max_{\alpha,\beta} L_{\text{J}}(\alpha,\beta) = \max_{\beta} \max_{\alpha} L_{\text{J}}(\alpha,\beta).$$

$$f(y,\alpha|\beta,\theta) = f(y|\beta,\theta) f(\alpha|y,\beta,\theta).$$

Note that the first factor on the right side does not involve α, while the second factor is the posterior of α (under a flat prior).

• Some computational issues

The joint estimates are typically obtained by solving

$$\frac{\partial l_J}{\partial \beta} = 0, \quad \frac{\partial l_J}{\partial \alpha} = 0.$$

However, in practice, the number of random effects may be quite large (e. g., McCullagh and Nelder 1989; Malec *et al.* 1997).

• A nonlinear Gauss-Seidel algorithm (Jiang 2000) - An Example:

$$\text{logit}\{\text{P}(y_{ij} = 1|u, v)\} = \mu + u_i + v_j,$$

$i = 1, \ldots, m$, $j = 1, \ldots, n$, where $u_i \sim N(0, \sigma^2)$, $v_j \sim N(0, \tau^2)$. The equations for joint maximization is equivalent to

$$\frac{u_i}{\sigma^2} + \sum_{j=1}^{n} \frac{\exp(\mu + u_i + v_j)}{1 + \exp(\mu + u_i + v_j)} = y_{i\cdot},$$

$1 \le i \le m$, and

$$\frac{v_j}{\tau^2} + \sum_{i=1}^{m} \frac{\exp(\mu + u_i + v_j)}{1 + \exp(\mu + u_i + v_j)} = y_{\cdot j},$$

$1 \le j \le n$, where $y_{i\cdot} = \sum_{j=1}^{n} y_{ij}$ and $y_{\cdot j} = \sum_{i=1}^{m} y_{ij}$.

• Empirical best prediction (EBP; Jiang and Lahiri 2001, Jiang 2003).

It is a two-step procedure. Let ζ denote a mixed effect.

Step I: Derive an expression for the best predictor $\tilde{\zeta} = \mathrm{E}(\zeta|y) = \psi(y, \theta)$.

Step II: replace θ by a consistent estimator, $\hat{\theta}$, to obtain the EBP $\hat{\zeta} = \psi(y, \hat{\theta})$.

• EBP with design-consistency

A feature of EBP is that it is model-based. If the assumed model fails, the predictor may perform poorly. Jiang and Lahiri (2005) developed a model-assisted EBP, which is design-consistent.

2.10. Future research and open problems

• Asymptotic behavior of the MLE for GLMM with crossed random effects (e. g., the salamander problem).

• Testing problems

e. g., Lin (1997), Lin & Carrol (1999), Song & Jiang (2000).

• Model diagnostics and model section