Generalized linear mixed models - computation of the likelihood function

Rasmus Waagepetersen
Department of Mathematics
Aalborg University
Denmark

November 9, 2023

Generalized linear mixed effects models

Consider stochastic vector $\mathbf{Y} = (Y_1, \ldots, Y_n)$ and vector of random effects $\mathbf{U} = (U_1, \ldots, U_m)$.

Two step formulation of GLMM:

- $\mathbf{U} \sim N(0, \Sigma)$.
- Given realization $\mathbf{u}$ of $\mathbf{U}$, Y_i independent and each follows density $f_i(y_i|\mathbf{u})$ with mean $\mu_i = g^{-1}(\eta_i)$ and linear predictor $\eta = X\beta + Z\mathbf{u}$.

I.e. conditional on $\mathbf{U}$, Y_i follows a generalized linear model.

NB: GLMM specified in terms of marginal density of $\mathbf{U}$ and conditional density of $\mathbf{Y}$ given $\mathbf{U}$. But the likelihood is the marginal density $f(\mathbf{y})$ which can be hard to evaluate !

We already saw one example: logistic regression with random effects.

Another common example: Poisson-log normal. Here

$$\mathbf{U} \sim N(0, \Sigma)$$
$$Y_i | \mathbf{U} = \mathbf{u} \sim \text{Pois}(\exp(\eta_i))$$

where $\eta_i = x_i \beta + z_i \mathbf{u}$

Likelihood for generalized linear mixed model

Likelihood for a generalized linear mixed model given by integral:

$$f(\mathbf{y}) = \int_{\mathbb{R}^m} f(\mathbf{y}, \mathbf{u})\mathrm{d}\mathbf{u} = \int_{\mathbb{R}^m} f(\mathbf{y}|\mathbf{u})f(\mathbf{u})\mathrm{d}\mathbf{u}$$

Difficult since $f(\mathbf{y}|\mathbf{u})f(\mathbf{u})$ is a very complex function.

Huge statistical literature on how to compute good approximations of the likelihood: Laplace approximation, numerical quadrature, Monte Carlo, Markov chain Monte Carlo,...

Example: logistic regression with random intercepts

$$
\begin{aligned}
U_j &\sim N(0,\tau^2),\ j=1,\dots,m \\
Y_j|U_j = u_j &\sim \text{binomial}(n_j,p_j) \\
\log(p_j/(1-p_j)) &= \eta_j = \beta + U_j \\
p_j &= \exp(\eta_j)/(1+\exp(\eta_j))
\end{aligned}
$$

Conditional density:

$$f(y|u;\beta) = \prod_j p_j^{y_j}(1-p_j)^{n_j-y_j} = \prod_j \frac{\exp(\beta+u_j)^{y_j}}{(1+\exp(\beta+u_j))^{n_j}}$$

Likelihood function ($u=(u_1,\dots,u_m)$)

$$\int_{\mathbb{R}^m} f(y|u;\beta) f(u;\tau^2)\mathrm{d}u = \prod_j \int_{\mathbb{R}} \frac{\exp(\beta+u_j)^{y_j}}{(1+\exp(\beta+u_j))^{n_j}} \frac{\exp(-u_j^2/(2\tau^2))}{\sqrt{2\pi\tau^2}}\mathrm{d}u_j$$

Integrals can not be evaluated in closed form.

Hierarchical model with independent random effects

Suppose $\mathbf{U} = (U_1, \dots, U_m)$ with the the U_i independent.

Moreover $\mathbf{Y} = (Y_{ij})_{ij}$, $i = 1, \dots, m$ and $j = 1, \dots, n_i$ where the conditional distribution of the $\mathbf{Y}_i = (Y_{ij})_j$ only depends on U_i.

Then we can factorize likelihood as

$$f(\mathbf{y}) = \prod_{i=1}^{m} \int_{\mathbb{R}} f(\mathbf{y}_i | u_i) f(u_i) \mathrm{d}u_i$$

That is, product of one-dimensional integrals.

Consider in the following computation of one-dimensional integral.

One-dimensional case

Compute

$$L(\theta) = \int_{\mathbb{R}} f(y|u;\beta) f(u;\tau^2) \mathrm{d}u$$

Some possibilities:

- Laplace approximation.
- Numerical integration/quadrature (e.g. Gaussian quadrature as in `glmer` `PROC NLMIXED` (SAS) or `GLLAM` (STATA)) (one level of random effects, dimensions one or two).

Laplace approximation

Let $g(u) = \log(f(y|u)f(u))$ and choose $\hat{u}$ so $g'(\hat{u}) = 0$ ($\hat{u} = \arg\max g(u)$).

Taylor expansion around $\hat{u}$:

$$g(u) \approx \tilde{g}(u) =$$
$$g(\hat{u}) + (u - \hat{u})g'(\hat{u}) + \frac{1}{2}(u - \hat{u})^2 g''(\hat{u}) = g(\hat{u}) - \frac{1}{2}(u - \hat{u})^2(-g''(\hat{u}))$$

I.e. $\exp(\tilde{g}(u))$ proportional to normal density $N(\mu_{LP}, \sigma^2_{LP})$, $\mu_{LP} = \hat{u}$ $\sigma^2_{LP} = -1/g''(\hat{u})$.

$$L(\theta) = \int_{\mathbb{R}} \exp(g(u))\mathrm{d}u \approx \int_{\mathbb{R}} \exp(\tilde{g}(u))\mathrm{d}u$$
$$= \exp(g(\hat{u})) \int_{\mathbb{R}} \exp\left(-\frac{1}{2\sigma^2_{LP}}(u - \mu_{LP})^2\right)\mathrm{d}u = \exp(g(\hat{u}))\sqrt{2\pi\sigma^2_{LP}}$$

Laplace approximation also works for for higher dimensions (multivariate Taylor expansion).

NB:

$$f(u|y) = f(y|u)f(u)/f(y) \propto \exp(g(u)) \approx const \exp\left(-\frac{1}{2\sigma^2_{LP}}(u-\mu_{LP})^2\right)$$

where $\mu_{LP} = \hat{u}\ \sigma^2_{LP} =, -1/g''(\hat{u})$.

Hence

$$U|Y = y \approx N\left(\mu_{LP}, \sigma^2_{LP}\right)$$

Note: μ_{LP} is mode of conditional distribution - used for prediction of random effects in `glmer` (`ranef()`).

Gaussian quadrature

Gauss-Hermite quadrature (numerical integration) is

$$\int_{\mathbb{R}} f(x)\phi(x)\mathrm{d}x \approx \sum_{i=1}^{n} w_i f(x_i)$$

where ϕ is the standard normal density and (x_i, w_i),$i = 1, \ldots, n$ are certain arguments and weights which can be looked up in a table.

We can replace $\approx$ with $=$ whenever f is a polynomial of degree $2n-1$ or less.

In other words (x_i, w_i), $i = 1, \ldots, n$ is the solution of the system of $2n$ equations

$$\int_{\mathbb{R}} x^k \phi(x)\mathrm{d}x = \sum_{i=1}^{n} w_i x_i^k, \quad k = 0, \ldots, 2n-1$$

where

$$\int_{\mathbb{R}} x^k \phi(x)\mathrm{d}x = 1[k \text{ even }](k-1)!! = 1[k \text{ even }](k-1)(k-1-2)(k-1-4)\ldots$$

Adaptive Gauss-Hermite quadrature

Naive application of Gauss-Hermite ($U \sim N(0, \tau^2)$):

$$\int f(y|u)f(u)\mathrm{d}u = \int f(y|\tau x)\phi(x)\mathrm{d}x$$

Now GH is applicable.

Adaptive GH:

$$\int f(y|u)f(u)\mathrm{d}u = \int \frac{f(y|u)f(u)}{\phi(u;\mu_{LP},\sigma^2_{LP})}\phi(u;\mu_{LP},\sigma^2_{LP})\mathrm{d}u =$$
$$\int \frac{f(y|\sigma_{LP}x+\mu_{LP})f(\sigma_{LP}x+\mu_{LP})}{\phi(x)}\sigma_{LP}\phi(x)\mathrm{d}x$$

(change of variable: $x = (u - \mu_{LP})/\sigma_{LP}$)

In my experience, adaptive GH is way more accurate than naive GH.

Advantage

$$\frac{f(y|u)f(u)}{\phi(u;\mu_{LP},\sigma^2_{LP})} = \frac{f(y|\sigma_{LP}x+\mu_{LP})f(\sigma_{LP}x+\mu_{LP})}{\phi(x)} \quad x = (u-\mu_{LP})/\sigma_{LP}$$

close to constant ($f(y)$) – hence adaptive G-H quadrature very accurate.

GH scheme with $n = 5$:

x	2.020	0.959	0.0000000	-0.959	-2.020
w	0.011	0.222	0.533	0.222	0.011

(x's are roots of Hermite polynomial computed using `ghq` in library `glmmML`).

(GH schemes for $n = 5$ and $n = 10$ available on web page)

Prediction of random effects for GLMM

Conditional mean

$$\mathbb{E}[U|Y = y] = \int uf(u|y)\mathrm{d}u$$

is minimum mean square error predictor, i.e.

$$\mathbb{E}(U - \hat{U})^2$$

is minimal with $\hat{U} = H(Y)$ where $H(y) = \mathbb{E}[U|Y = y]$

Difficult to analytically evaluate

$$\mathbb{E}[U|Y = y] = \int uf(y|u)f(u)/f(y)\mathrm{d}u$$

Computation of conditional expectations (prediction)

$$\mathbb{E}[U|Y=y] = \int u \frac{f(y|u)f(u)}{f(y)} \mathrm{d}u =$$

$$\frac{1}{f(y)} \int (\sigma_{LP} x + \mu_{LP}) \frac{f(y|\sigma_{LP} x + \mu_{LP}) f(\sigma_{LP} x + \mu_{LP})}{\phi(x)} \sigma_{LP} \phi(x) \mathrm{d}x$$

Note:

$$(\sigma_{LP} x + \mu_{LP}) \frac{f(y|\sigma_{LP} x + \mu_{LP}) f(\sigma_{LP} x + \mu_{LP})}{\phi(x)} \sigma_{LP}$$

behaves like a first order polynomial in x - hence GH still accurate.

Score function and Fisher information

Let

$$V_\theta(y,u) = \frac{\mathrm{d}}{\mathrm{d}\theta} \log f(y,u|\theta)$$

Then score and observed information are

$$u(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta} \log L(\theta) = \mathbb{E}_\theta[V_\theta(y,U)|Y=y] \tag{1}$$

and

$$\begin{aligned} j(\theta) &= -\frac{\mathrm{d}^2}{\mathrm{d}\theta^\mathsf{T}\mathrm{d}\theta} \log L(\theta) \\ &= -\big(\mathbb{E}_\theta[\mathrm{d}V_\theta(y,U)/\mathrm{d}\theta^\mathsf{T}|Y=y] + \mathbb{V}\mathrm{ar}_\theta[V_\theta(y,U)|Y=y]\big) \end{aligned} \tag{2}$$

Again the above expectations and variances can be evaluated using Laplace or adaptive GH.

Newton-Raphson iterations:

$$\theta_{l+1} = \theta_l + j(\theta_l)^{-1} u(\theta_l)$$

Difficult cases for numerical integration - dimension $m > 1$

- correlated random effects: multivariate density of **U** does not factorize
- crossed random effects: U_i and V_j independent $i = 1, \ldots, m$ $j = 1, \ldots, n$ but Y_{ij} depends on both U_i and V_j.

Not possible to factorize likelihood into low-dimensional integrals

Number of quadrature points $\approx k^m$ where k is number of quadrature points for $1D$ and m number of random effects – hence numerical quadrature may not be feasible.

Alternatives: Laplace-approximation or *Monte Carlo computation*.

Wheeze results with different values of nAGQ

Default nAGQ=1 means Laplace approximation:

```
> fiter=glmer(resp~age+smoke+(1|id),family=binomial,data=ol
> summary(fiter)
Random effects:
 Groups Name        Variance Std.Dev.
 id     (Intercept) 5.491    2.343
Number of obs: 2148, groups:  id, 537

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.37396    0.27496 -12.271   <2e-16 ***
age         -0.17677    0.06797  -2.601   0.0093 **
smoke        0.41478    0.28705   1.445   0.1485
```

5 quadrature points:

```
> fiter5=glmer(resp~age+smoke+(1|id),family=binomial,
                                   data=ohio,nAGQ=5)
Groups Name        Variance Std.Dev.
 id     (Intercept) 4.198    2.049
Number of obs: 2148, groups:  id, 537

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.02398    0.20353 -14.857  < 2e-16 ***
age         -0.17319    0.06718  -2.578  0.00994 **
smoke        0.39448    0.26305   1.500  0.13371
```

10 quadrature points:

```
> fiter10=glmer(resp~age+smoke+(1|id),family=binomial
                                        ,data=ohio,nAGQ=10)
Random effects:
 Groups Name        Variance Std.Dev.
 id     (Intercept) 4.614    2.148
Number of obs: 2148, groups:  id, 537

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.08959    0.21557 -14.332  < 2e-16 ***
age         -0.17533    0.06762  -2.593  0.00952 **
smoke        0.39799    0.27167   1.465  0.14293
```

Some sensivity regarding variance estimate. Fixed effects results quite stable.

Results with 20 quadrature points very similar to those with 10 quadrature points.

Laplace - mathematical details in one-dimension

(one dimension to avoid technicalities of multivariate Taylor)

Let

$$I_n = \int_{\mathbb{R}} \exp(nh(x))g(x)\mathrm{d}x$$

where $h(x)$ is three times differentiable and assume there exists $\hat{x}$ so that

1. $H = -h''(\hat{x}) > 0$ and $h'(\hat{x}) = 0$
2. for any $\Delta > 0$ there exists an $\epsilon > 0$ so that $h(\hat{x}) - h(x) > \epsilon$ for $|x - \hat{x}| > \Delta$
3. there exists a $\delta > 0$ so that $|h^{(3)}(x)| < K$ and $|g(x)| < C$ for $|x - \hat{x}| \leq \delta$
4. a) $\int_{\mathbb{R}} |g(x)|\mathrm{d}x < \infty$ or b) $\int_{\mathbb{R}} \exp(h(x))|g(x)|\mathrm{d}x < \infty$

Then

$$\frac{I_n}{\exp(nh(\hat{x}))g(\hat{x})\sqrt{2\pi n^{-1} H^{-1}}} \to 1 \tag{3}$$

as $n \to \infty$.

Relative error of approximation

Absolute error of approximation is

$$E_n = I_n - \exp(nh(\hat{x}))g(\hat{x})\sqrt{2\pi n^{-1} H^{-1}}$$

Previous result says that relative error

$$\frac{E_n}{I_n} \to 0$$

Strong result in case I_n is a small quantity (may not be enough that absolute error is "small")

We can say more:

$$\frac{I_n}{\exp(nh(\hat{x}))g(\hat{x})\sqrt{2\pi n^{-1} H^{-1}}} = 1 + O(n^{-1}).$$

That is, the relative error is of order n^{-1}.

Exercises

1. How does Laplace approximation look in the multivariate case ?
2. Show that adaptive GH with one quadrature point is equivalent to Laplace approximation.
3. Show the identities (1) and (2) (assuming differentiation and integration can be interchanged as needed).
4. Write down the likelihood in case of crossed random effects. What is the problem ?
5. Solve exercises on `exercises_lp_gh.pdf`
6. Carefully check the proof for Laplace approximation in Section 1 in note available on course webpage. If you like Taylor expansions you may also want to check Sections 2-3.
7. Consider the case of one Normal random effect $U \sim N(0, \tau^2)$ and observations $Y_1, \ldots, Y_n$ that are iid given U. Can you apply the formal result for the Laplace approximation to show convergence of the approximation of the likelihood ? Which problems do you face ?