Notes on How GLMMs are Fit to Data

William Ruth

In this document, we review how parameters are estimated in generalized linear mixed models (GLMMs). We start with an overview of the general problem and what parts are challenging. We then describe the method used by `glmer()` in the `lme4` package in `R`. I have been exploring some alternative estimators which may be more efficient, but in the interest of getting something written down as quickly as possible, I'm going to just focus on what I actually did.

1 Likelihood for GLMMs

The GLMM framework closely resembles that of generalized linear models (GLMs). We have a response variable, $Y \in \mathbb{R}$, whose mean is related to some covariates, $X \in \mathbb{R}^p$, via a link function, g, and a vector of regression coefficients, $\beta \in \mathbb{R}^p$. Formally, we write $g[\mathbb{E}(Y|X)] = X^T\beta$. The distribution of Y can be from any exponential family (in fact, this restriction can be weakened even further). In our analysis, the response variable is always binary, so $Y|X \sim \text{Bernoulli}(p)$, where $p = \mathbb{E}(Y|X)$. We call $X^T\beta$ the linear predictor, and denote it by η.

Extending from GLMs to GLMMs involves the addition of one or more covariates, $Z \in \mathbb{R}^Q$, and one or more regression coefficients, $U \in \mathbb{R}^Q$. The difference here is that U is random. Typically, we model $U \sim \text{N}(0, \Gamma(\theta))$, where Γ is parameterized by some low-dimensional vector, θ. We refer to β as the fixed-effects and U as the random-effects. The sets of covariates, X and Z, may or may not be disjoint.

Having introduced random effects to our model, we must now update the distribution of Y. We take the conditional distribution of Y given $U = u$ to be a GLM with linear predictor $\eta(u) = X^T\beta + Z^Tu$. The joint distribution of Y and U is thus

$$f_{Y,U}(y,u) = f_{Y|U}(y|U=u) \cdot f_U(u) \tag{1}$$

$$= \left[\left(\frac{e^{\eta(u)}}{1+e^{\eta(u)}}\right)^y \left(\frac{1}{1+e^{\eta(u)}}\right)^{1-y}\right] \cdot \left[(2\pi)^{-Q/2}|\Gamma|^{-1/2}\exp\left(-\frac{1}{2}u^T\Gamma^{-1}u\right)\right] \tag{2}$$

Note that U is unobserved, so the likelihood of our data comes from the marginal

distribution of Y, given by

$$f_Y(y) = \int_{\mathbb{R}^Q} f_{Y,U}(y,u)du \tag{3}$$

This integral has no closed form solution, but many methods exist to approximate it. Section 2 describes one such method in detail.

The likelihood of our observed data is the product of the marginal distribution given in (3) over all samples. Although we have suppressed it in our notation, the linear predictor, $\eta(u)$ depends on both X and Z, which differ across observations.

1.1 Clustered Data

In our problem, the data are divided into K disjoint groups, or clusters. We can reflect this clustering in the model equations by taking Z and Γ to be block-diagonal. More precisely, suppose that Y is obtained by concatenating responses across groups; i.e. $Y = (Y^{(1)}, \ldots, Y^{(K)})$. Similarly, we imagine having measured some small number of random-effects covariates, say q of them, separately across groups, $Z^{(1)}, \ldots, Z^{(K)}$. The random-effects covariate matrix given above is then the block-diagonal matrix with diagonal entries $Z^{(k)}$, and $Q = Kq$. Next, random effects are taken to be iid across groups, so Γ is a block-diagonal matrix with K copies of the same q-by-q covariance matrix, Σ, where Σ is the covariance matrix of a single group's random effects. Finally, we also write $u^{(k)}$ for the values of the random effects in group k.

While it's a bit cumbersome to articulate the equivalence between this group-wise formulation and the global model given earlier in this section, it is important to know that these two versions of the GLMM are equivalent. In particular, the two formulations are used almost interchangeably in the literature based on the author's taste.

I will mostly stick to the clustered data formulation of the model, but might jump back and forth sometimes. Hopefully, what I mean will be clear from context.

2 Maximizing the Likelihood: `lme4`

The `lme4` package in `R` includes a function called `glmer`, which fits GLMMs to data. This function uses a Laplace approximation to the integral in (3), which is then maximized. We follow the presentation in Section 7.7 of Demidenko (2004). First, maximize $f_{Y,U}$ over u to get the mode, u_*. Next, we use the multivariate Laplace approximation to the integral in (3). The general form of this approximation is

$$\int_{\mathbb{R}^q} e^{h(u)}du \approx (2\pi)^{Q/2} e^{h_*} \left| -\nabla_u^2 h(u)\big|_{u=u_*} \right|^{-1/2}, \text{or} \tag{4}$$

$$\log \int_{\mathbb{R}^q} e^{h(u)}du \approx \frac{Q}{2}\log(2\pi) + h_* - \frac{1}{2}\log\left| -\nabla_u^2 h(u)\big|_{u=u_*} \right| \tag{5}$$

where u_* is the maximizer of h, and $h_* = h(u_*)$ is the maximum value. Specializing to our problem, we write $\ell_{Y|U}(y, u)$ for the conditional log-likelihood of an observation y, given the relevant group's random effects, $U^{(k)} = u^{(k)}$. Then the function h to be maximized is

$$h(u) = h(u^{(1)}, \ldots, u^{(K)}) \tag{6}$$

$$= \sum_{k=1}^{K} \sum_{n=1}^{N_k} \ell_{Y|U}(y_i^{(k)}, u^{(k)}) - \frac{1}{2} \sum_{k=1}^{K} (u^{(k)})' \Sigma^{-1} u^{(k)} \tag{7}$$

where N_K is the number of observations in group K. Conveniently, h can be maximized analytically, giving the maximizer $u_* = (u_*^{(1)}, \ldots, u_*^{(K)})$ This isn't maximized analytically. I'm not sure how Demidenko proposes to find u_*. It's probably in his book somewhere. Substituting h and u_* into (5) and letting $\varphi_i^{(k)} = \exp[\beta' X_i^{(k)} + (u_*^{(k)})' Z_i^{(k)}]$, we get the following approximation to the marginal log-likelihood of Y:

$$\ell_Y(y) = \sum_{k=1}^{K} \sum_{i=1}^{N_k} y_i^{(k)} X_i^{(k)} - \sum_{k=1}^{K} \left[\sum_{i=1}^{N_k} \log(1 + \varphi_i^{(k)}) + \frac{1}{2} (u_*^{(k)})' \Sigma^{-1} (u_*^{(k)}) \right]$$

$$\frac{K}{2} \log \left| \Sigma^{-1} \right| - \frac{1}{2} \sum_{k=1}^{K} \log \left| \Sigma^{-1} + \sum_{i=1}^{N_k} \frac{\varphi_i^{(k)}}{\left(1 + \varphi_i^{(k)}\right)^2} Z_i^{(k)} \left[Z_i^{(k)} \right]' \right| \tag{8}$$

Demidenko gives some guidance on how to maximize this approximate likelihood in β and Σ.

3 The `glmer` Function in `lme4`

This section is based on an in-progress paper contained in the source files for the `lme4` package (Walker et al., 2024). The general idea is that we want to approximate the integral in (3) using Laplace's method. In order to do so, we must compute the maximizer for u in the joint likelihood $f_{Y,U}(y, u)$. Call this maximizer $\tilde{u}(\beta, \theta)$. We then plug $\tilde{u}$ into the Laplace approximation formula, (5), and maximize over β and θ. Much attention is paid to the problem of computing $\tilde{u}$, while little is said about how to optimize the Laplace approximation. While the former certainly requires care, the latter problem seems more challenging to me. In particular, the proposed method for computing $\tilde{u}$ is an iterative algorithm, so optimizing the resulting Laplace approximation requires nested iteration. Clearly, Walker et al. have a working implementation of this nested optimization; I would just prefer that they spend more time here in their exposition. Maybe the published version of the paper will spend more time here.

Anyway, on to what they actually did.

3.1 PIRLS

The method proposed by Walker et al. (2024) for computing $\tilde{u}$ is called the penalized iteratively reweighted least squares (PIRLS) algorithm. This method is closely related to the iteratively re-weighted least squares algorithm used to fit regular GLMs (see Section 2.5 of McCullagh and Nelder, 1989). Before describing the algorithm, we first remark that the random effects in our problem can be re-parameterized in terms of so-called "spherical" random effects. Write $\Sigma = \Lambda\Lambda'$ for the Cholesky factorization of our random effects' covariance matrix, and define $U = \Lambda V$, with $V \sim \mathrm{N}(0, I)$. Our linear predictor is then $\zeta(v) = \eta(\Lambda v) = X\beta + Z\Lambda v$. We get an equivalent joint distribution to (2) in terms of V by writing $f_{Y,V}(y, v)$ in terms of our new linear predictor, and taking $f_V(v)$ to be the standard multivariate normal density. For concreteness, we write

$$f_{Y,V}(y, v) = f_{Y|V}(y|V = v) \cdot f_V(v) \tag{9}$$

$$= \left[\left(\frac{e^{\zeta(v)}}{1 + e^{\zeta(v)}}\right)^{y} \left(\frac{1}{1 + e^{\zeta(v)}}\right)^{1-y}\right] \cdot \left[(2\pi)^{-Q/2} \exp\left(-\frac{1}{2} v^T v\right)\right] \tag{10}$$

and

$$f_Y(y) = \int_{\mathbb{R}^Q} f_{Y,V}(y, v) dv \tag{11}$$

An advantage of the new parameterization is that the joint likelihood can be written in a convenient form

$$f_{Y,V}(y, v) = (2\pi)^{-Q/2} \exp\left[-\frac{d(y, v) + \|v\|^2}{2}\right] \tag{12}$$

where $d(y, u)$ is called the deviance of the GLM defined for Y conditional on $V = v$, and $\|v\|$ is the Euclidean (ℓ^2) norm of v. This deviance is -2 times the log-likelihood. We observe now that maximizing the likelihood is equivalent to minimizing the so-called "penalided deviance": $d(y, v) + \|v\|^2$. This deviance is minimized by solving a sequence of penalized, weighted, non-linear least-squares problems:

$$v_{i+1} = \arg\min_{v} \left\| W_i^{1/2}[y - \mathbb{E}(Y|V = v_i)] \right\|^2 + \|v\|^2 \tag{13}$$

where W_i is a diagonal matrix with entries $\mathbb{V}(Y|V = v_i)$. The problem in (13) can further be re-cast in terms of the increments, $\delta_{i+1} = v_{i+1} - v_i$, by linearizing the inverse link function at $\mathbb{E}(Y|V = v_i)$ for iteration $i + 1$.

$$\delta_{i+1} = \arg\min_{\delta} \left\| \begin{bmatrix} W^{1/2}(y - \mathbb{E}(Y|V = v_i)) \\ v_i \end{bmatrix} - \begin{bmatrix} W^{1/2} M_i Z\Lambda \\ I \end{bmatrix} \delta \right\|^2 \tag{14}$$

where M_i is the gradient of the inverse link function evaluated at the linear predictor obtained from v_i. That is, $M_i = d\mathbb{E}(Y|V = v)/d\zeta|_{\zeta=\zeta(v_i)}$. Note that

the problem given by (14) is an ordinary quadratic program in δ, so the solution can be obtained analytically as the solution to a set of normal equations. The actual implementation in `lme4` uses the Cholesky factorization to solve the normal equations for δ_{i+1}. Call the corresponding Cholesky factor $L^{\ddagger}$.

This concludes our description of one iteration of the PIRLS algorithm. We obtain $\tilde{u}$ by repeating this iteration until some convergence criterion is met.

3.2 Laplace Approximation

Upon having obtained $\tilde{u}$ using the algorithm in Section 3.1, we equipped to compute the Laplace approximation to the marginal likelihood for Y. After some manipulation, this approximation reduces to $f(y, \tilde{u})|L|$, where L is the (sparse) Cholesky factor computed while solving the normal equations in the final iteration of PIRLS. See Section 2.2 of Walker et al. (2024) for a bit more detail.

Our final estimates for β and θ are obtained by maximizing the Laplace approximation to the marginal likelihood for Y. We call these maximizers, $\hat{\beta}$ and $\hat{\theta}$, the MLEs. It's not clear to me how to efficiently solve this optimization problem, since the function being optimized (the Laplace approximation) can only be evaluated by applying another iterative algorithm, PIRLS. Clearly, `lme4` is able to do this, it just seems like it would be worth also discussing.

3.3 Prediction of Random Effects

The random effects, U (or V), are unobserved random variables. As such, we don't estimate them, but predict them based on what we know about their conditional distribution given the observed data, Y. To this end, `lme4` uses the conditional mode of $U|Y = y$ as a predictor for U. Conveniently, this conditional mode can be obtained directly by applying the PIRLS algorithm (see Section 3.1) with $\beta = \hat{\beta}$ and $\theta = \hat{\theta}$, the MLEs.

4 Some Extensions

The problem of fitting GLMMs to data has been extensively studied. I highlight here a few other directions. I'm not advocating for their use right now, it's more important to get this project submitted. However, I do think that they're worth considering, if not exploring, in future work.

The main ideas I want to raise for now are quadrature as an alternative to the Laplace approximation, and restricted maximum likelihood.

[‡]Note that $LL' = P(\Lambda' Z' M W M Z \Lambda + I)P'$, where P is a permutation matrix which induces sparsity in L.

4.1 Quadrature

One of the hardest parts of estimation in GLMMs is the integral required to compute the marginal likelihood for Y. While the PIRLS algorithm is ingenious, another option is just to do numerical integration. In low-dimensional problems this is fine, but the accuracy of standard quadrature techniques is known to fall off quickly with the dimension of the problem. In fact, `lme4` includes an option for Gauss-Hermite quadrature, but only in problems with a single random effect. In our problem we have 2 or more, so trying to use this just gives an error.

Nevertheless, there are strategies for doing quadrature in multiple dimensions. In particular, there is a whole package called `mvQuad` (Weiser, 2023) which does multivariate quadrature. I haven't had time to explore this yet, but it does seem like a promising alternative to the Laplace approximation to the marginal likelihood.

4.2 Restricted Maximum Likelihood

The information in this section is based on Maestrini et al. (2024).

The conventional wisdom seems to be that maximum likelihood estimates of the random effects' covariance matrix tends to be biased toward zero in small samples. A suite of methods have been proposed to reduce this bias (at the expense of increasing the variance), collectively referred to as restricted maximum likelihood (REML). A motivation for this technique is the degrees-of-freedom correction used when estimating the residual variance in ordinary linear regression. We divide by $n - p$ instead of n, thereby making our variance estimator unbiased, but increasing its variance by a factor of $(1 - p/n)^{-2}$.

In linear mixed-effects models, the different flavours of REML are equivalent. However, in GLMMs they can give different results. See Maestrini et al. (2024) for a detailed review of the general REML strategy, as well as many different implementations.

References

Demidenko, E. (2004). *Mixed Models: Theory and Applications*. Wiley.

Maestrini, L., Hui, F. K. C., and Welsh, A. H. (2024). Restricted maximum likelihood estimation in generalized linear models. *arXiv*, (2402.12719).

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall/CRC, 2nd edition.

Walker, S. C., Christensen, R. H. B., Bates, D., Bolker, B. M., and Mächler, M. (2024). Fitting generalized linear mixed-effects models using `lme4`. `https://github.com/lme4/lme4/blob/master/misc/glmer_JSS/glmer.pdf`.

Weiser, C. (2023). *`mvQuad`: Methods for Multivariate Quadrature.* (R package version 1.0-8).