Bias induced by fitting GLMMs with dichotomous outcomes using penalized quasi-likelihood

Journal Title
XX(X):1–7

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Joshua Nugent[1], Bianca Doone[1] and Ken Kleinman[1]

Abstract

Generalized linear mixed models (GLMMs) are the most widely-used method for analyzing data from cluster-randomized trials (CRTs). Popular statistical software packages allow for different GLMM fitting algorithms, but one of these algorithms, penalized quasi-likelihood (PQL), has been shown to produce biased parameter estimates. We review the literature to assess how widely PQL may be used, and conduct a literature-informed simulation study to show the extent of the PQL bias in plausible CRT settings. We find that the algorithms employed are rarely reported in the literature, and that PQL bias is most extreme when the cluster size is small and the variability between clusters is large. Further, intraclass correlation coefficient (ICC) estimates from PQL-fitted models are also shown to vary by outcome prevalence and treatment effect. Alternatives to PQL estimation are demonstrated to be unbiased and feasible for most CRT data analysis needs. Analysts should not use PQL and should report fitting methods when reporting trial results.

Keywords

Cluster randomized trials, generalized linear mixed models, penalized quasi-likelihood, PQL

Background

Generalized linear mixed models (GLMMs) are a commonly used method for analyzing data from cluster randomized trials (CRTs). GLMMs extend generalized linear models (GLMs) by including an additional random-effects term in the linear predictor. This term captures variance between clusters - for example, the group-level differences between hospitals or classrooms. In settings where interventions are applied at the cluster level, GLMMs can disaggregate treatment effects from any preexisting underlying variance between clusters. In medical settings, CRTs with dichotomous outcomes are very common - for example, estimating the effect of a new infection control protocol on MRSA incidence, or the probability of a preterm birth for people enrolled in prenatal support groups - and GLMMs are a commonly-used tool for analysis.

The optimization problem of fitting a GLMM to data is a non-trivial task. Three common numerical methods for estimating the coefficients are *penalized quasi-likelihood* (PQL), *Gauss-Hermite quadrature* (GHQ), and *Laplace approximation*. Other methods, such as Newton quadrature, Monte Carlo integration, and Markov Chain Monte Carlo can be used as well [1], but since popular statistical software packages use PQL, GHQ, and Laplace approximation as their standard GLMM fitting algorithms, we will focus on those in this paper. The full mathematical details of these three main methods have been elaborated in other sources [2 3], and an overview of the technical aspects of GLMMs and the algorithms is given in the Supplemental Material.

Penalized quasi-likelihood was popularized by Breslow and Clayton [4], though similar methods were developed by others [5 6] around the same time. Though it is computationally efficient, especially for models with many random effects, PQL can induce bias in certain cases, in particular when the response variable distribution is far from normal [7 8 9 10]. Additionally, PQL produces Wald-type test statistics, not true likelihoods, making it unsuitable for use in the likelihood ratio test. Thus is cannot be used in nested model selection [1 3 11].

[1]University of Massachusetts, Amherst

Corresponding author:
Ken Kleinman, Department of Biostatistics and Epidemiology, School of Public Health and Health Sciences, University of Massachusetts, 715 North Pleasant Street, Amherst, MA 01003-9304, USA
Email: kkleinman@schoolph.umass.edu

Gauss-Hermite quadrature is a more computationally demanding method, but results in no discernible bias to coefficient estimates, and can produce true likelihood statistics for model comparison. The accuracy with which it computes the model fit is a function of how many *quadrature points* it uses in estimating the model. All else being equal, the computation time is roughly proportional to $(N_q)^u$, where N_q is the number of quadrature points and u is the number of random effects at all levels of the model[12 3]. For a model with 4 random effects and 5 quadrature points, $(N_q)^u = 5^4 = 625$. Doubling the number of points to 10 changes that result to $10,000$, a factor of 16. For data sets with large numbers of random effects, this can limit the utility of GHQ.

Luckily for the data analyst, many CRTs have only one random effect, so computation time will increase linearly with the number of quadrature points, rather than as a power function. Furthermore, if many models need to be compared, using one quadrature point can give preliminary results rapidly. Then, after that model selection process, the number of quadrature points can be increased to make the final estimates as accurate as computationally possible. Empirical results suggest that 7 or fewer quadrature points often give suitably accurate estimates[13].

Laplace approximation is equivalent to GHQ with a single quadrature point[14], and as such it produces true likelihoods for model comparison. While its accuracy is typically lower than GHQ with $N_q \geq 2$, under certain circumstances it performs quite well[14].

For reasons of computational efficiency, PQL was a useful method for fitting GLMMs when it was developed, but with the advent of more modern computers, less biased methods such as GHQ have become an attractive alternative. For CRTs with binary outcomes, where the bias in PQL is the most extreme[11 15], and where the presence of only one random effect is typical, using GHQ is the best option: fast enough, and, more importantly, unbiased.

Most modern statistical software packages have functions to fit GLMMs with dichotomous outcomes, such as PROC GLIMMIX in SAS, meglm in Stata, and glmer (from the lme4 package, as well as others) in R. However, the default fitting algorithm in each of those functions varies. In SAS PROC GLIMMIX, the default is PQL, with Laplace or GHQ available if specified. In R, the glmer function default is Laplace, with GHQ available if specified; PQL is only available in R via the glmmPQL function in the MASS package. In Stata, meglm defaults to GHQ with 7 quadrature points, with Laplace available if specified.

Given that many data analysts may be unfamiliar with the fitting options, a function's default settings are influential in the final results. Below, we investigate how often functions and algorithms are reported in the literature and use simulations to describe the bias induced by PQL in a literature-informed, plausible CRT scenario.

Methods

We started by conducting a literature review among recent CRTs with dichotomous outcomes to determine a) common values for cluster size and number of clusters and b) what software, functions, and fitting algorithms were used to analyze the data, if reported. The review, searching for the phrase "cluster randomized trial" in the title or abstract of the article, spanned two databases over two timeframes. First, we searched the The New England Journal of Medicine, The British Medical Journal, The Journal of the American Medical Association, and The Lancet from January 1, 2014 through August 31, 2018, using the Web of Science database. Second, we did a broader search of all articles in the PubMed database published between March 1st, 2018 and August 31, 2018. Full details of this review can be found in Doone et al. (2019).

Having identified candidate articles, we filtered to completed CRTs with dichotomous outcomes. The mean number of observations per cluster and number of clusters for each study was recorded, as well as the software and functions/algorithms the authors used, if available.

The second phase of our work was a simulation study to investigate the bias of different GLMM fitting algorithms for dichotomous outcomes. To maximize the utility of the results, our simulations used a range of plausible cluster counts and cluster sizes drawn from the literature review.

Our data-generating mechanism for the simulations was a simple logistic-link GLMM with one fixed intercept, one treatment effect, and one random intercept, defined as:

$$\text{logit}[\Pr(y_{ij}|b_j) = 1] = \beta_0 + \beta_1 x_{ij} + b_j \qquad (1)$$

with x_{ij} an indicator for treatment (1) or control (0) arm of the study for unit i in cluster j; $\Pr(y_{ij}|b_j)$ the probability of the outcome y for unit i in cluster j; e^{β_0} the baseline odds of the outcome across all clusters; e^{b_j} the odds ratio compared to baseline for the outcome, specific to units in cluster j relative to the mean cluster, with assumed distribution $b_j \sim N(0, \sigma^2)$; and β_1, our parameter of interest, the log odds ratio due to the treatment.

Software cited (specific procedures or functions)

	R	SAS	SPSS	Stata	Unreported
Top4 2014 - 2018 ($n = 42$)	11 (2)	13 (1)	2 (0)	18 (2)	4 (4)
PubMed Q2 2018 ($n = 43$)	3 (0)	11 (2)	8 (0)	14 (2)	10 (0)
Total ($n = 85$)	14 (2)	24 (3)	10 (0)	32 (4)	14 (0)

Table 1. Cited software used for data analysis. Numbers in parentheses indicate number of references to specific functions or fitting options within the software package. Some articles used multiple software packages, so the totals differ from n. While most articles report the software used, very few of them specifically reference the function or fitting algorithm.

	Median number of clusters (middle 50%)	Median observa cluster (middl
Top4 2014 - 2018 ($n = 42$)	46 (24 - 116)	110 (40 - 4
PubMed Q2 2018 ($n = 43$)	32 (15 - 70)	44 (14 - 2
Total ($n = 85$)	40 (16 - 90)	66 (19 - 3

Table 2. Common values for units/cluster and number of clusters from the literature review.

From that model, populations were generated with the following parameter values, informed by the literature review (see Table 2):

- Number of clusters $n \in \{10, 50, 100\}$
- Number of observations per cluster $p \in \{25, 100\}$
- β_0 values corresponding to a baseline prevalence of .01, .02, .03, .05, .1, and .2
- β_1 values corresponding to a treatment effect odds ratio of .5, .75, .9, 1.1, 1.33, 1.5, and 2
- σ^2 values of 1 (high cluster variability) and .1 (low cluster variability).

Using 3000 simulated datasets for each combination of parameters, logistic-link GLMMs were fit via PQL, GHQ, and Laplace using SAS/STAT software version 13.2 (SAS Institute Inc., Cary, NC).

The distribution of $\hat{\beta}_1$ estimates from each method was compared to the true value from the data-generating mechanism and absolute bias was measured as the difference between the two. The standard errors of estimates and the estimated cluster variance ($\hat{\sigma}^2$) were also collected from the fitted models, and model fitting CPU time was measured. Finally, the simulated populations were re-fit under the assumption that the outcome variable was normally distributed, without the logit link, and the intraclass correlation coefficient (ICC) estimated from those models was collected.

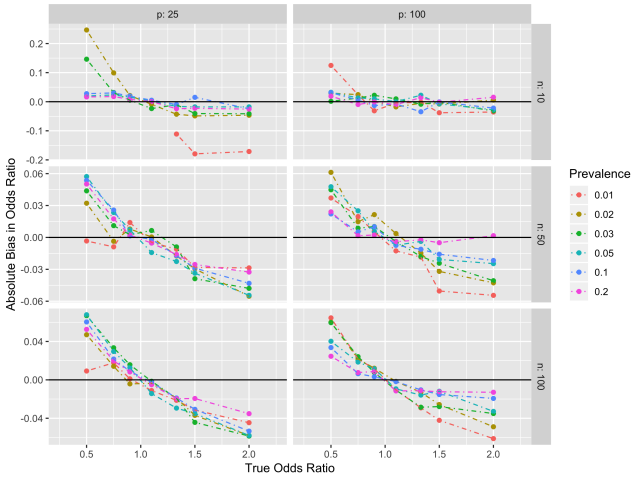

Figure 1. Odds ratio bias ($\exp\{\bar{\hat{\beta}}_1 - \beta_1\} - 1$) in PQL estimation, $\sigma^2 = 1$. Simulation runs with less than 80% convergence omitted. Note larger scale in the first row.

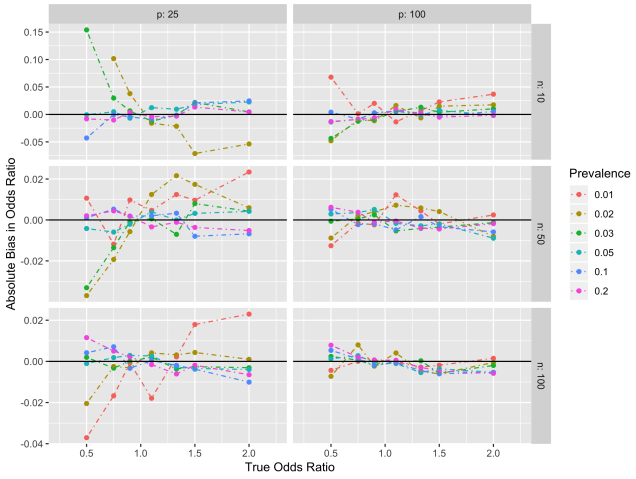

Figure 2. Odds ratio bias ($\exp\{\bar{\hat{\beta}}_1 - \beta_1\} - 1$) in PQL estimation, $\sigma^2 = 0.1$. Simulation runs with less than 80% convergence omitted.

Results

Results of the literature review are shown in Table 1. Of the 85 articles, only 9 identified the specific procedure (meglm or GLIMMIX, for example) used. Among the 24 articles that identified SAS as one of the software packages, only 3 specified which SAS procedure was used and only one identified the model fitting algorithm.

Parameters for the simulations were chosen after observing the common values in Table 2. The number of clusters was typically below 100, as was the number of observations per cluster, though the latter showed significantly more variability.

The results of the bias investigation in PQL estimation are shown in Figures 1-3. The results show bias towards the null: As the true odds ratio rises above 1, there is a negative bias, meaning the mean estimated odds ratio is closer to 1 than it should be. When the odds ratio is less than one, conversely, there a positive bias towards 1. Further, while the data are noisy, the bias is more pronounced for smaller cluster sizes and when the outcome's baseline prevalence

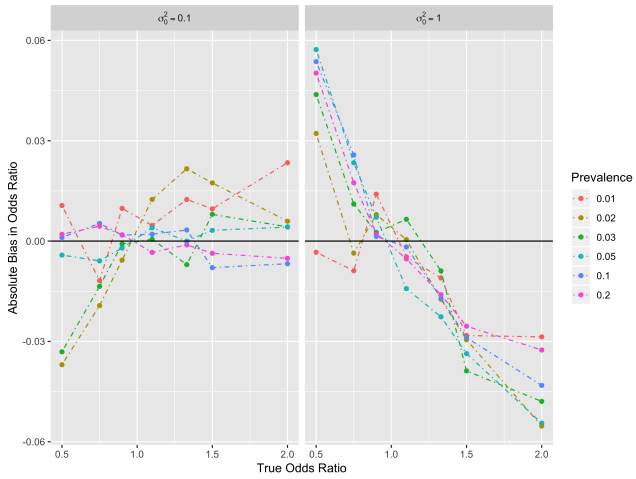

Figure 3. Odds ratio bias ($\exp\{\bar{\hat{\beta}}_1 - \beta_1\} - 1$) in PQL estimation for different σ^2, 25 units/cluster, 50 clusters.

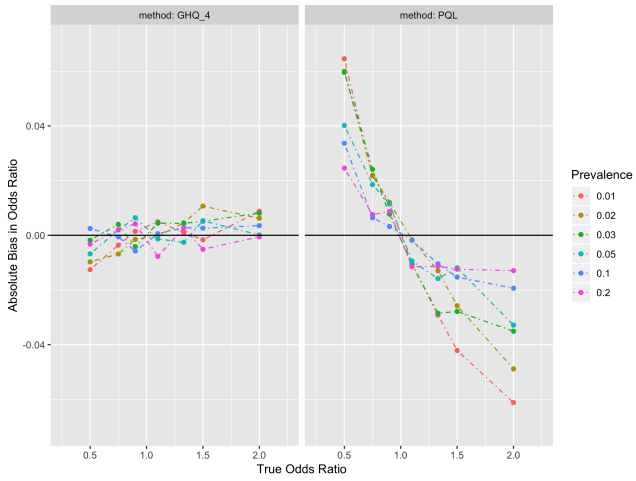

Figure 4. Odds ratio bias ($\exp\{\bar{\hat{\beta}}_1 - \beta_1\} - 1$) in GHQ (4 quadrature points) and PQL, $\sigma^2 = 1$, 100 units/cluster, 100 clusters.

is lowest. The bias occurs, to a differing extent, across all values of prevalence, treatment effect, cluster size, and number of clusters. Figure 3 demonstrates that the effect of high between-cluster variability is more pronounced bias.

The fitted models using GHQ (a representative example is given in Figure 4) and Laplace approximation (Figure 5) did not show a clear bias; in each, PQL is shown for reference. In a small number of simulation runs, particularly with a small number of clusters and rare outcomes, the GHQ and Laplace algorithms became unstable, leading to estimates that tended toward infinity, making the mean bias comparison unhelpful - this can be seen in the outlier in Figure 5 when the prevalence is .01 and the treatment effect odds ratio is 2. On the other hand, when the number of clusters is small, the PQL algorithm was less likely to converge, implying that limited data is a challenge for all of the algorithms.

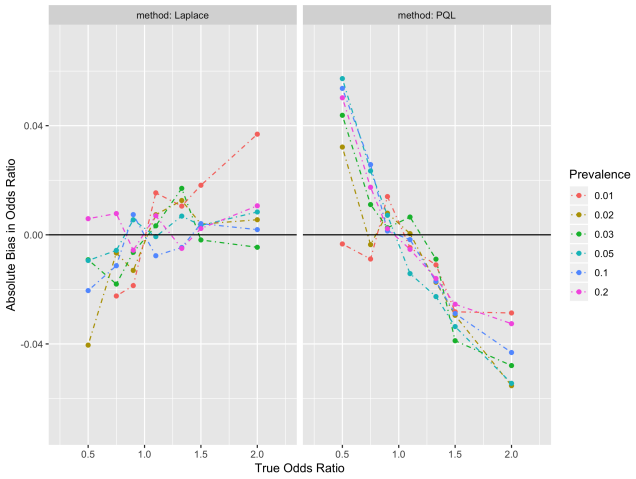

Figure 5. Odds ratio bias ($\exp\{\bar{\hat{\beta}}_1 - \beta_1\} - 1$) in Laplace and PQL, $\sigma^2 = 1$, 25 units/cluster, 50 clusters. Note that the lack on convergence in some of the Laplace models leads to the outlier at odds ratio 2.

Our simulations confirmed that the main advantage of using PQL over GHQ is speed. In addition, SAS's implementation of PQL outperforms the other methods. The mean CPU time for SAS to fit a single large data set (500 clusters, 500 observations per cluster) on a modern laptop is 3.7 seconds for PQL, 11 seconds for Laplace approximation, 18.4 seconds for GHQ with $N_q = 4$, and 27.1 seconds for GHQ with $N_q = 10$. Results for GHQ and the Laplace approximation are comparable for R's lme4 package, though the glmmPQL method from the MASS package, with a runtime of at 23.3 seconds, is significantly less efficient than in the SAS implementation.

For primarily historical reasons, analysts may be interested in the intraclass correlation coefficient (ICC), though it is not an actual parameter of the model. In normally distributed data, the ICC measures the proportion of total variance that is explained by the variance between groups. However, in the case of a non-normally distributed outcome variable, there has been considerable discussion about how to appropriately characterize and calculate the ICC[16,17]. We generated two ICC estimates using the PQL-fitted models from our simulation. First, a version that assumed a random intercept logistic model, implying an ICC of $\frac{\sigma^2}{\sigma^2 + \frac{\pi^2}{3}}$, based on the estimated between-cluster variance. Second, by fitting a linear mixed model that assumed the outcome variable was normally distributed, leading to the typical ANOVA-based calculation $\frac{\sigma^2}{\sigma^2 + \sigma_\epsilon^2}$, where σ_ϵ^2 represents the variance of the residuals. The results are shown in Figure 6 for the two values of σ^2 examined in out simulations. In both cases, the estimated ICC varies significantly by prevalence, treatment effect, and model class. Results when fitting the models with GHQ showed a similar pattern with the assumption of a

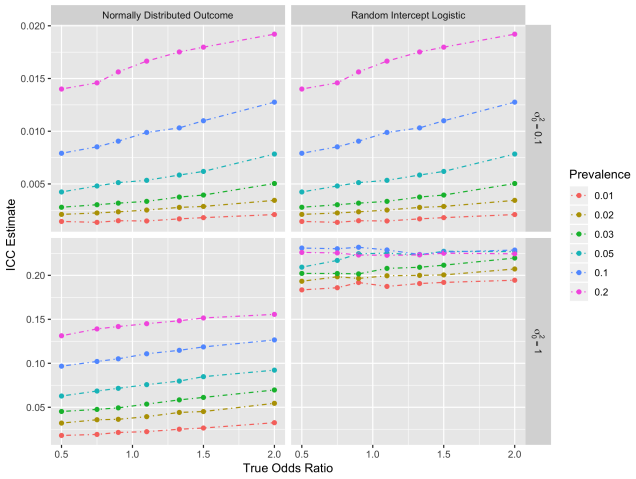

Figure 6. ICC estimates from models fit via PQL, $p = 100$, $n = 50$.

```
\begin{table}
\small\sf\centering
\caption{<Table caption.>}
\begin{tabular}{<table alignment>}
\toprule
<column headings>\\
\midrule
<table entries
(separated by & as usual)>\\
<table entries>\\
.
.
.\\
\bottomrule
\end{tabular}
\end{table}
```

Figure 7. Example table layout.

normal distribution for the outcome, but more consistency when using the $\frac{\sigma^2}{\sigma^2+\frac{\pi^2}{3}}$ formulation.

Discussion

For CRTs with sample sizes that occurr commonly in the literature, using PQL to estimate coefficients in random intercept logistic regression shows a noticeable bias towards the null, particularly if the true odds ratio is far from 1. GHQ, on the other hand, shows no noteiceable bias, so for the vast majority of cluster randomized trials with dichotomous outcomes, GHQ is superior to PQL when fitting models. To fit a single data set with a small number of random effects, and given than most CRTs have only one random effect, the speed of GHQ with 4-10 quadrature points is adequate and it produces no detectable bias. For data analysts who are experimenting with different nested models, using Laplace approximation during the model-selection stage can save time, and GHQ can be utilized for the fitting once the final model has been chosen. Laplace is also preferable to PQL in the model-selection process because PQL only provides quasi-likelihood, and hence it is not suited to nested model comparison with the likelihood ratio test.

The bias towards the null generated by PQL is more pronounced when clusters are small, between-cluster variance is high, and baseline incidence of an outcome is low. Given our simulations, we suspect that existing studies may have suffered this bias, though it is hard to be sure given that fitting methods are rarely reported in the literature. Statisticians should report methods/functions and the algorithm options in more detail. We should take care when selecting procedures for fitting GLMMs, particularly in SAS, where PQL is the default option.

Previous work investigating PQL estimation in scenarios with dichotomous outcomes has noted the bias [1,18], but not examined the full interactions between cluster size, number of clusters, baseline prevalence, true odds ratio, and cluster-level variance as have here. We hope this will make it easy for analysts to identify situations where bias could be present.

Given the bias it creates, why use PQL? As noted above, PQL fits models much more quickly than the Laplace or GHQ methods. However, our simulations showed that even for large data sets, none of the runtimes are prohibitively long, given a typical CRT model with one random intercept term. In a situation where many models need to be tested, the Laplace method could be used to compare models, and then for the final analysis, a more accurate GHQ fit could be made with a large number of quadrature points.

Finally, ICC estimates generated by these algorithms may vary substantially by the method used to calculate them and by the baseline prevalence of the outcome, and should be approached with a degree of skepticism.

TEMPLATE STUFF

The standard coding for a table is shown in Figure 7.

Funding

Grant Support: NIGMS: R01 GM121370

Supplemental material

GLMM Fitting

Mathematically, a GLMM can be modeled as

$$g(\mu_{it}) = \mathbf{x}_{it}^T\boldsymbol{\beta} + \mathbf{z}_{it}^T\mathbf{u}_i \quad (2)$$

with i a cluster indicator, t an observation indicator within cluster i, g the GLM link function, $\boldsymbol{\beta}$ a vector of coefficients for covariate values $\mathbf{x}_{it}$, and $\mathbf{z}_{it}^T$ a vector of coefficients for random effects $\mathbf{u}_i$, assumed to be distributed as multivariate normal with mean 0 and covariance matrix $\boldsymbol{\Sigma}$. When the outcomes are dichotomous, the link function g is typically the logit, and the mean μ_{it} is the probability of the outcome given the covariate values and cluster membership.

To fit a GLMM with a vector $\mathbf{x}_{it}$ and corresponding outcome vector $\mathbf{y}$, it is necessary to integrate the random effects $\mathbf{u}_i$ out of the likelihood function[8]. That likelihood function, the probability mass function of y as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$[7], is, in general,

$$\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}; \mathbf{y}) = f(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \int f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta}) f(\mathbf{u}; \boldsymbol{\Sigma}) d\mathbf{u}. \quad (3)$$

For many link functions of interest, including the logit link function for dichotomous outcomes considered in this paper and other situations where the response variable is discrete, the integral above does not have a closed-form solution, in part because it involves integrating the product of discrete and continuous densities[11]. Numerical methods are required to approximate the integral in these circumstances.

PQL iteratively fits a linear mixed model[10] to the data, essentially approximating the discrete density using a Gaussian density[11]. Further details of PQL have been discussed above.

Gauss-Hermite quadrature approximates the integral of a function $f(\cdot)$ multiplied by a normal density function; note that it is very similar to the likelihood function presented earlier where $f(\mathbf{u}; \boldsymbol{\Sigma})$ was a multivariate normal and $f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta})$ the conditional likelihood. For univariate cases,

$$\int_{-\infty}^{\infty} f(u) exp(-u^2) du \approx \sum_{k=1}^{q} c_k f(s_k) \quad (4)$$

where c_k are weights, sometimes from a table, and s_k are the each of the quadrature points used to approximate the normal density. More quadrature points results in a more accurate approximation of the integral, but is more computationally intensive, though various GHQ subvariants have been developed that increase efficiency and reduce the number of quadrature points needed[3]. With GHQ, inversion of the Fisher information matrix can provide standard errors for the maximum likelihood estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$.

The Laplace method approximates the likelihood using a second-order Taylor expansion[13] and is is equivalent to GHQ with a single quadrature point[14]. Simulation studies have found Laplace approximations to exhibit mild bias in coefficient estimates, and significant bias in estimation of the variance components[3].

References

1. Zhang H, Lu N, Feng C et al. On Fitting Generalized Linear Mixed-effects Models for Binary Responses using Different Statistical Packages. *Statistics in medicine* 2011; 30(20): 2562–2572. DOI:10.1002/sim.4265.
2. Wolfinger R and O'connell M. Generalized linear mixed models a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* 1993; 48(3-4): 233–243. DOI: 10.1080/00949659308811554.
3. Pinheiro JC and Chao EC. Efficient Laplacian and Adaptive Gaussian Quadrature Algorithms for Multilevel Generalized Linear Mixed Models. *Journal of Computational and Graphical Statistics* 2006; 15(1): 58–81.
4. Breslow NE and Clayton DG. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association* 1993; 88(421): 9–25.
5. Zeger SL, Liang KY and Albert PS. Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics* 1988; 44(4): 1049–1060. DOI:10.2307/2531734.
6. Engel B and Keen A. A simple approach for the analysis of generalizea linear mixed models. *Statistica Neerlandica* 1994; 48(1): 1–22. DOI:10.1111/j.1467-9574.1994.tb01428.x.
7. Agresti A. *Categorical Data Analysis*. Wiley Series in Probability and Statistics, Wiley, 2013. ISBN 978-0-470-46363-5.
8. Rodriguez G and Goldman N. An Assessment of Estimation Procedures for Multilevel Models with Binary Responses. *Journal of the Royal Statistical Society Series A (Statistics in Society)* 1995; 158(1): 73. DOI:10.2307/2983404.
9. Breslow NE and Lin X. Bias Correction in Generalised Linear Mixed Models with a Single Component of Dispersion. *Biometrika* 1995; 82(1): 81–91. DOI:10.2307/2337629.
10. Lin X and Breslow NE. Bias Correction in Generalized Linear Mixed Models With Multiple Components of Dispersion. *Journal of the American Statistical Association* 1996; 91(435): 1007–1016. DOI:10.2307/2291720.
11. Ng ES, Carpenter JR, Goldstein H et al. Estimation in generalised linear mixed models with binary outcomes by simulated maximum likelihood. *Statistical Modelling: An International Journal* 2006; 6(1): 23–42. DOI:10.1191/1471082X06st106oa.
12. StataCorp. Stata 15 Base Reference Manual, 2017.
13. Pinheiro JC and Bates DM. Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. *Journal of Computational and Graphical Statistics* 1995; 4(1): 12–35. DOI:10.2307/1390625.

14. Liu Q and Pierce DA. A Note on Gauss-Hermite Quadrature. *Biometrika* 1994; 81(3): 624–629. DOI:10.2307/2337136.
15. Lin X. Estimation using penalized quasilikelihood and quasi-pseudo-likelihood in Poisson mixed models. *Lifetime Data Analysis* 2007; 13(4): 533–544. DOI:10.1007/s10985-007-9071-z.
16. Wu S, Crespi CM and Wong WK. Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. *Contemporary Clinical Trials* 2012; 33(5): 869–880. DOI:10.1016/j.cct.2012.05.004.
17. Nakagawa Shinichi, Johnson Paul C D and Schielzeth Holger. The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface* 2017; 14(134): 20170213. DOI:10.1098/rsif.2017.0213.
18. Jang W and Lim J. A Numerical Study of PQL Estimation Biases in Generalized Linear Mixed Models Under Heterogeneity of Random Effects. *Communications in Statistics - Simulation and Computation* 2009; 38(4): 692–702. DOI:10.1080/03610910802627055. URL https://doi.org/10.1080/03610910802627055.