A NEW ESTIMATION METHOD FOR GENERALIZED LINEAR MIXED MODEL

O.Y. Halid and R.A. Adeleke
Department of Mathematical Sciences, Ekiti State University, Ado – Ekiti, Nigeria

ABSTRACT
There has been evolution of several estimation methods for generalized linear mixed models (GLMMs) under the likelihood and Bayesian frameworks with most plagued with certain disabilities. In this work, we proposed the new method which is pseudo-likelihood based as an alternative GLMM estimation method to overcome some of this disabilities. This new proposed methods are developed by approximating the intractable integrals encountered in the likelihood function of the GLMM by the Taylor’s approximation, with the three approximation steps in the analytic expression of the conventional pseudo-likelihood reduced to two steps. The performance of this new method is investigated by modeling the well-known Salamander Mating data that involves two 20-dimensional analytically intractable integrals using the binary logistic regression model, our GLMM of interest. The results of its implementation are observed to compare well with other commonly used methods in GLMM inference, in terms of convergence rate and accuracy.

KEYWORDS: Generalized linear mixed model, intractable integrals, Bayesian framework, Maximum likelihood, PL- pseudo-likelihood, PQL-penalized quasi likelihood, Gibbs Sampling, Moment method

I. INTRODUCTION
Mixed models extend classical linear regression models by including random or subject-specific effects (next to the (traditional) fixed effects) in the structure for the mean. For distributions from the exponential family, GLMMs extend GLMs by including random effects in the linear predictor. The random effects not only determine the correlation structure between observations on the same subject, they also take account of heterogeneity among subjects, due to unobserved characteristics and as such measure observations that are not independent.
In order words, GLMMs have their origin from a class of fixed effects regression models with several types of dependent variables (i.e. continuous, dichotomous, counts) namely Generalized linear models (GLMs). McCullagh and Nelder (1989) describe these in great detail and indicate that its root or origin is due to Nelder and Wedderburn (1972).
Generalized linear mixed models are also natural extension of linear mixed models and generalized linear models according .The conditional mean of the response given the random effect(s) is linked to the linear predictor. The linear predictor comprises of fixed and random effects.
In particular, the combination of random and fixed effects, along with the distributions in the exponential family makes it one of the most important classes of statistical models as it allows for modeling of binary, proportional, count and continuous data. As in linear mixed models, the GLMMs can accommodate correlated error structures such as found in longitudinal, time series and spatial statistics.
The model parameters can be estimated in closed form only for some very special situations, like normal-normal, beta-binomial or gamma-Poisson mixtures. But in other mixtures, especially when the number of nested random effects are more than two or the random effects are crossed, the closed form solutions are not possible or they are very complicated or tedious to obtain. In such situations, some approximation procedures are currently in practice. Penalized Quasi-Likelihood by Breslow and Clayton (1993), Expectation Maximization (EM) algorithm by Dempster, A.P. et al. (1977), Monte
Carlo EM, Simulated Maximum Likelihood and Monte Carlo Newton Raphson by McCulloch (1997) and Stochastic Approximation EM by Jank (2006) are mostly used estimation procedures. As pointed out earlier, the combination of a generalized linear model with normal random effects on the linear predictor scale, to give a ‘rich class of models’ have led to a wide variety of applications which are also found in Diggle et al (2002), Verbeke and Molenberghs (2000), Verbeke and Molenberghs (2005) and McCulloch et al (2008) to mention a few. Generalized linear mixed models (GLMMs) continue to grow in popularity due to their ability to directly acknowledge multiple levels of dependency and model different data types.

They are also popular due to their flexibility and effectiveness for analysis of grouped data and data with repeated observations. Its application for both Gaussian and non-Gaussian response variables makes it so versatile. Due to its versatility in accommodating random effects, they are widely used in the analysis of clustered data including longitudinal data or repeated measurements (Breslow and Clayton, 1993) and now they are the main vehicle for analysis of longitudinal data (Diggle, Heagerty, Liang, and Zeger, 2002, Fitzmaurice, Laird and Ware, 2004 to mention a few).

Consequently, GLMMs provide a useful approach for analyzing a wide variety of data structures which most practicing statisticians often encounter and there exist several methods for fitting these models. These methods are mostly classified as likelihood-based and Bayesian-based methods. There are also some other new methods for its estimation. As a matter of fact, there have been some literatures (such as Breslow and Clayton (1993), McCulloch (1997), Sinha (2004), Fong et al (2010) to mention a few) on the estimation procedures of the GLMMs.

These literatures and some other sources had earlier on presented some of these current methods such as the Penalized-Quasi likelihood (PQL), Marginal-Quasi likelihood (MQL), Restricted Maximum likelihood (REML), Adaptive Guass Quadrature (AGQ) and the Laplace Approximation procedures which are all under the likelihood framework. There have also been the Marcov Chain Monte Carlo (MCMC) methods with Metropolis-Hastings, Diffuse-Prior and Gibbs Sampling procedures under the Bayesian framework with proven usage validity.

Most of the methods for fitting GLMMs are iterative schemes which are plagued with a few disabilities such as cumbersomeness, slow convergence rates, biasedness, and lack of precision. Some of these methods of estimating them also exist as subroutines in some modern day computer statistical software packages such as SAS, R, STATA, open BUGS, lme4 to mention a few. The disabilities are there till date and it is believed that this work and future studies may gradually correct these disabilities and improve its versatility. For instance, one of the disabilities in these techniques is the inability of some likelihood-based methods to handle small sample sizes with a considerable level of accuracy as well as the challenges posed by cumbersome nature of the resulting integrals (in their likelihood) thereby making them computationally inefficient whereas the Bayesian–based methods which give more accurate variance estimate is very slow because of the choice of specifying the prior distributions. The latter method is also limited to only a few distributions.

Our focus therefore lies on creating alternative solution that overcome these intractable multidimensional integrals encountered in the implementation of likelihood-based methods on crossed designs.

Generally, the varieties of maximum likelihood estimation technique widely used under the likelihood-based methods requires integration techniques for calculation of the log-likelihood and to avoid or overcome the computational problems involving these cumbersome irreducibly high-dimensional integrals, several algorithms had been proposed (e.g. Sinha, 2004). These algorithms are capable of estimating model parameters by approximating the log-likelihood function. Although, these algorithms are useful for fitting these models but are efficient only under strict model assumptions. They can also be highly influenced by the presence of unusual data points (outliers) and less accurate for small sample sizes. A full maximum likelihood (ML) analysis based on the joint marginal likelihood of the responses can be used for estimating both fixed and random effects parameters in GLMM which requires numerical integration techniques for calculating the log-likelihood, score equations and information matrix. This is only limited to simple models but intractable for more complicated problems involving irreducibly high-dimensional integrals. Such is the case of the estimation problem of the Salamander mating data posed by the presence of two 20-dimensional intractable multidimensional integrals, which is our focus in this paper. We therefore proposed and developed the new method which is pseudo-likelihood based as an alternative method.
for estimating the binary logistic mixed model our GLMM of interest in this work using the Salamander mating data.
The problem of estimation via the likelihood approach was considered in section II and the proposed new method was spelt out in section III while section IV and V respectively contain application of the new method to the Salamander mating data and discussion of the results.

II. PROBLEM OF EVALUATION OF INTEGRALS OF GLMM FROM THE LIKELIHOOD APPROACH

To understand these computational difficulties we consider the following; suppose that, given the random effects \( u_i, \ldots, u_m \) and \( v_1, \ldots, v_m \), binary responses \( y_{ij} \), \( i = 1, \ldots, m \), \( j = 1, \ldots, m_2 \) are conditionally independent such that, with \( P(y_{ij} = 1 | u, v) \),

\[
\text{logit} \ (P_{ij}) = \mu + u_i + v_j
\]

where \( \mu \) is an unknown parameter, \( u = (u_i)_{i \in \mathcal{I}} \) and \( v = (v_j)_{j \in \mathcal{J}} \).

Furthermore, the random effects \( u_i, \ldots, u_m \) and \( v_1, \ldots, v_m \) are independent such that \( u_i \sim N(0, \sigma_1^2) \), \( v_j \sim N(0, \sigma_2^2) \) where the variances \( \sigma_1^2 \) and \( \sigma_2^2 \) are known.

Thus, the unknown parameters involved in this model are \( \psi = (\mu, \sigma_1^2, \sigma_2^2) \). It can be shown that the likelihood function under this model for estimating \( \psi \) can be expressed as;

\[
c - \frac{m_1}{2} \log(\sigma_1^2) - \frac{m_2}{2} \log(\sigma_2^2) + \mu y_{..} + \log \left[ \prod_{i=1}^{m} \prod_{j=1}^{m_2} (1 + \exp(\mu + u_i + v_j)) \right] \exp \left( \sum_{i=1}^{m} u_i y_{i..} + \sum_{j=1}^{m_2} v_j y_{..j} \right)
- \frac{1}{2\sigma_1^2} \sum_{i=1}^{m} u_i^2 - \frac{1}{2\sigma_2^2} \sum_{j=1}^{m_2} v_j^2
\]

\[
du_i \ldots du_m dv_1 \ldots dv_{m_2}
\]

(2)

Where \( c \) is a constant, \( y_{..} = \sum_{i=1}^{m} \sum_{j=1}^{m_2} y_{ij} \), \( y_{i..} = \sum_{j=1}^{m_2} y_{ij} \) and \( y_{..j} = \sum_{i=1}^{m} y_{ij} \).

The multidimensional integrals above have no closed-form expression and it cannot be further simplified. Furthermore, such an integral is difficult to evaluate even numerically.

We therefore consider solutions which avoid the problem of evaluation of the intractable multidimensional integrals (encountered in most likelihood approach) in closed forms through the development of a new set of pseudo-likelihood estimation methods.

These new methods are motivated and developed in this work through the reduction of the three approximation steps in the pseudo-likelihood (PL) estimation method of Wolfinger and O’Connell (1993) to two steps (namely one analytic and one probabilistic approximation).

III. THE NEW ESTIMATION METHOD

We attempt to solve this problem as follows for the GLMM, by considering a data vector of length \( n \) satisfying

\[
Y = \mu + \varepsilon
\]

and a differentiable monotonic link function \( g(\cdot) \) such that

\[
g(\mu) = X\beta + Z\alpha
\]

Where \( \alpha \) is a vector of unknown random effects with known model matrix \( Z \), \( \beta \) is a vector of unknown fixed effects with known model matrix \( X \) of rank \( p \). The mean of a GLMM is given by

\[
E(Y | \alpha, \beta) = g^{-1}[X'\beta + Z\alpha] = g^{-1}(\eta) = \mu
\]

(4)
Recall that $\alpha \sim N(0, G)$ and $\text{var}(Y | \alpha) = A^{1/2} R(\rho) A^{1/2}$ and we assume that $G(\alpha)$ is diagonal and positive definite. More so, the covariance matrix $G$ and correlation matrix $R$ depends on parameter vector $\alpha$ and $\rho$ respectively. $g^{-1}(\cdot)$ is the inverse of (3)

3.1. Analytic Approximation

Following Wolfinger and O’Connell (1993), we seek the first Taylor’s series expansion of $\mu$ around $\tilde{\beta}$ and $\tilde{\alpha}$, in (4) yielding

$$g^{-1}(\eta) = g^{-1}(\tilde{\eta}) + \Delta X (\beta - \tilde{\beta}) + \Delta (\alpha - \tilde{\alpha})$$

where

$$\Delta = \left( \frac{\partial g^{-1}(\eta)}{\partial \eta} \right)_{\hat{\beta}, \hat{\alpha}}$$

is a diagonal matrix of derivatives of conditional mean evaluated at $\tilde{\beta}$ and $\tilde{\alpha}$ on the expansion locus. Upon re-arranging terms in (5) yields,

$$y = X \beta + Z \alpha + \epsilon$$

which is a linear mixed model with pseudo-response $y$, fixed effects $\beta$ and random effects $\alpha$ and

$$\text{var}(\epsilon) = \Delta^{-1} A^{1/2} \sigma^2 R(\rho) A^{1/2} \Delta^{-1} = \text{var}(y | \alpha)$$

such that $y | \alpha, \beta \sim N[X \beta + Z \alpha, \Delta^{-1} A^{1/2} \sigma^2 R(\rho) A^{1/2} \Delta^{-1}]$ where $\theta = (\sigma^2, v', \rho')'$ is the parameter vector containing all unknowns in $G$ and $R$, with a further assumption that the distribution of $y$ is known.

3.2 Probabilistic Approximation

The formulation of the pseudo-data above in (6) can be viewed as a transformation that creates data that are approximately Gaussian. As a matter of fact the new response variable is analogous to the modified dependent variable used in the iteratively-weighted least squares algorithm of Nelder and Wedderburn (1972) giving rise to the Gaussian likelihood corresponding to the linear mixed model (normal mixed model) of (6) for $I_G$

$$l_G = -\frac{1}{2} \log|\mathbf{V}| - \frac{1}{2} \phi^{-1}(y - X \beta)^T V^{-1}(y - X \beta) - \frac{n}{2} \log 2\pi$$

$$= \log|V(\theta)| - \frac{1}{2} (y - X \beta)^T V(\theta)^{-1} (y - X \beta) - \frac{n}{2} \log 2\pi$$

An additional scale/dispersion parameter $\phi$ has been introduced and $l_G$ reparameterized in terms of ratios of $\phi$ so that

$$V(\theta) = ZGZ^T + \Delta^{-1} A^{1/2} \sigma^2 RA^{1/2} \Delta^{-1}$$

$l_G$ above is maximized for $\beta$ and $\phi$ yielding the maximum pseudo log likelihood $l_M$ and the restricted pseudo log likelihood $l_R$ as follows ;

$$l_M (\theta, y) = -\frac{1}{2} \log|V(\theta)| - \frac{1}{2} r^T V(\theta)^{-1} r - \frac{n}{2} \log 2\pi$$

$$l_R (\theta, y) = -\frac{1}{2} \log|V(\theta)| - \frac{1}{2} r^T V(\theta)^{-1} r - \frac{1}{2} \log|X'V(\theta)^{-1}X| - \frac{n-p}{2} \log 2\pi$$

Define minus two times the pseudo-log-likelihoods in both cases and represent the constants as $c_1$ and $c_2$ so that the objective functions become

$$-2l_M (\theta, \bar{y}) = l_R = \log|V(\theta)| + r^T V(\theta)^{-1} r + c_1$$
$-2l(\theta; \tilde{y}) = l_r = \log |V(\theta)| + rV(\theta)^{-1}r + \log |X'V(\theta)^{-1}X| + c_2 \tag{11}$

where $r = \tilde{y} - X\left(X'V(\theta)^{-1}X\right)^{-1}X'V(\theta)^{-1}\tilde{y}$, $c_1$ and $c_2$ are constants that do not depend on $\theta$ or $\beta$. Equation (10) & (11) are the respective objective functions for the new pseudo maximum likelihood estimator and the new pseudo restricted maximum likelihood estimator which are profiled to get

$$l_{sl}(\psi, \tilde{y}) = \log |Q(\psi)| + n\log \left(q'Q(\psi)^{-1}q\right) + q + c_1 \tag{12}$$

$$l_r(\psi, \tilde{y}) = \log |Q(\psi)| + (n - p)\log \left(q'Q(\psi)^{-1}q\right) + \log |X'Q(\psi)^{-1}X| + c_2 \tag{13}$$

The two pseudo profiled log likelihood equations above will be maximized to obtain the parameter estimates of $\beta$ and $\alpha$.

Once we have $\hat{\theta}$, we use generalized least squares (equal to maximum likelihood for (10) to estimate $\beta$ and $\alpha$ with

$$\hat{\beta} = (X'V(\hat{\theta})^{-1}X)^{-1}X'V(\hat{\theta})^{-1}\tilde{y} \tag{14}$$

which also equals

$$\hat{\beta} = (X'Q(\hat{\theta})^{-1}X)^{-1}X'Q(\hat{\theta})^{-1}\tilde{y}$$

when minimizing the profiled pseudo-log-likelihoods (12) or (13) to obtain $\psi$, we have

$$\hat{\alpha} = GZ'V(\hat{\theta})^{-1}r \tag{15}$$

The analytic part involves creation of a pseudo model (4) upon rearrangement of terms in the first Taylor’s expansion of the inverse link function of GLMM in (3) while the probabilistic approximation assumes that the response variable of the pseudo model follows the Gaussian distribution giving rise to the new restricted pseudo-log-likelihood and the new maximum pseudo log-likelihood functions which are maximized appropriately to obtain our parameters of interest.

The new estimation method will emerge by maximizing (12) and (13) for the random and fixed effect parameters along the expansion loci.

IV. APPLICATION OF THE NEW METHOD TO THE SALAMANDER MATING DATA

These new methods are applied for fitting the binary logistic regression model, our GLMM of interest to the famous salamander mating data (from the summer experiment carried out in 1986 by Arnold and Vernell, published in McCullagh and Nelder 1989). This data was adopted because they are commonly used for GLMM inferences. This is evident in fairly recent literatures namely Chan et al (2005) as well as Eid and Al-Eid (2013) which have illustrated their estimation methods by analyzing this data.

The data consist of mating of two geographically isolated population of 40 salamanders namely Rough Butt (RB) and Whiteside (WS) with 20 females and 20 males. Each female salamander was mated to three male salamanders from its own population and another three from the other population using a crossed design giving rise to 40 correlated observations. The main scientific question of interest is whether or not the mating between population is as successful as that within population. Another question is to see if heterogeneity in mating between individual animals (males and females) exists and if so, whether it is greater for males or females.

The model fits (and analysis by the new method) were carried out such that these questions are answered and estimates of random effects, fixed effects and mating success probabilities result. These results will be compared in terms of accuracy and convergence rate, with the result of some methods within and outside the likelihood and Bayesian frameworks, which are standard methods in existing literatures such as moment method of McCullagh & Nelder (1989), Gibbs Sampling method of Karim & Zeger (1992), Penalized Quasi Likelihood method of Breslow & Clayton (1993), Pseudo-likelihood
method of Wolfinger and O’Connell (1993), Modified Laplace approximation technique of Shun(1997) as well as Gibbs-MC approach of Chan et al (2005) to mention a few. There exist some computer statistical software packages such as R, SAS, and STATA for fitting this model but the latest version of SAS (9.4) will be particularly used due to its recency and versatility.

4.1 The Binary Logistic Mixed Model

Let \( Y_{ij} \) be the variable and \( y_{ij} \) the actual observation, for the pairing of the \( i^{th} \) female with the \( j^{th} \) male, \( i, j = 1,...,N \) where \( N = 20 \) is the number of males and of females and let \( I, J \) be \( R \) (Rough Butt) or \( W \) (White Side), according to the population to which the \( i^{th} \) female and the \( j^{th} \) male belong.

We assume that the responses are independent conditionally on the subjects and on the crossing effect and that the subject and crossing effects are additive on the transformed scale of the response.

With the logit link, the model is

\[
\text{logit} \ P(Y_{ij} = 1 | \alpha, \beta) = X' \beta + \alpha^F_i + \alpha^M_j, \quad i, j \in \Omega
\]  

(16)

Which is also called the binary logistic mixed model.

Where \( X_i \) is a vector of covariates, \( \alpha = (\alpha^F, \alpha^M) \) is the vector of male and female random effects respectively, that is, they are independent random variables representing female and male random effects (20 each and 10 from each population).

The fixed effects \( \beta = (\beta_{RR}, \beta_{RW}, \beta_{WR}, \beta_{WW}) \) represents the vector of crossing effects (the order is female-male). It also denotes the average logit of mating between females of population \( i \) and males of population \( j \).

\( \Omega \) is the set of all pairings or observations in the experiment (\( \Omega \) has 40 elements);

\( \alpha^F_i \) and \( \alpha^M_j \) are individual random effects (subject effects) for which it is assumed that

\[
\alpha^F_i \sim N(0, \sigma^2_F) \quad \text{and} \quad \alpha^M_j \sim N(0, \sigma^2_M)
\]

where \( \alpha^F_i \) and \( \alpha^M_j \) are independent.

Note that the variances of the random effects depend on the sex, but not on the population membership of the animals.

Following the theory of GLM (as stated in chapter three), we denote the R.H.S of (16) by \( \eta_{ij} \), the probability in the L.H.S by \( \pi_{ij} \) and the logistic transform by \( g^{-1} \) with this notation, (16) becomes:

\[
\pi_{ij} = g(\eta_{ij}) = X_i' \beta + \alpha^F_i + \alpha^M_j \quad \text{for pairs} \ (i, j)
\]  

(17)

The crossing effect \( \beta \) is of main interest, as it determines the unconditional (population-level) probability of a successful mating for each crossing \( \pi_{ij} \)

\[
\pi_{ij} = E\left[ g\left( X_i' \beta + \alpha^F_i + \alpha^M_j \right) \right]
\]  

(18)

where the expectation is taken with respect to the distribution of the random male and female effects.

Obviously, the calculation of (18) involves also the variances \( \sigma^2_F \) and \( \sigma^2_M \). The random effects \( \alpha^F_i \) and \( \alpha^M_j \) of the individual animals are not of particular interest here. The parameter of interest consist of the fixed effects, \( \beta \) and the variances of the random effects, \( \sigma^2_F \) and \( \sigma^2_M \).

Formally, we put \( \delta = (\sigma^2_F, \sigma^2_M) \) and \( \theta = (\beta, \delta) \)

The conditional likelihood is

\[
L(\beta, \alpha | y) = P(y | \alpha, \beta) = \prod_{(i,j) \in \Omega} \left\{ g(\eta_{ij}) \right\}^{y_{ij}} \left\{ 1 - g(\eta_{ij}) \right\}^{1-y_{ij}} = \prod_{(i,j) \in \Omega} \frac{\exp(y_{ij} \eta_{ij})}{1 + \exp(\eta_{ij})}
\]
\[
\exp\left(\sum \frac{y_{ij} \beta_{ij} + \sum y_j \alpha_{ij}^F + \sum y_i \alpha_{ij}^M}{\prod (1 + \exp(\beta_{ij} + \alpha_{ij}^F + \alpha_{ij}^M))}\right)
\]  \hspace{1cm} (19)

where \( y_{ij} = \sum y_{ij} \) and the sum extends over all observations from the crossing \((I, J)\), \( y_i \) and \( y_j \) are the total number of successful mating for the \( i^{th} \) female and \( j^{th} \) male respectively and \( i, j \) extend over all females and males respectively in the experiment.

The likelihood is obtained by integrating out the random effects

\[
L(\theta | y) = E[L(\beta, \alpha | y) | \theta, y] = \int P(y | \beta, \alpha) P(\alpha | \delta) d\alpha
\]  \hspace{1cm} (20)

where \( P(\alpha | \delta) \) is the probability density of the random effects given by (17). Due to the design of the experiment, the above 40 dimensional integral may be decomposed in a product of two 20-dimensional integrals corresponding to the two independent tables (in Table 4.1) which cannot be reduced any further.

**TABLE 4.1** : The Salamander Mating Data for the Summer Experiment

This integrand in (20) above will be approximated and evaluated by the new pseudo-likelihood approach in the succeeding section so as to obtain the required parameter estimates (crossing effects and subject effect)

### 4.2 Parameter Estimation Algorithm

Suppose that we have initial values \( \hat{\beta}^{[0]} \) and \( \hat{\alpha}^{[0]} \) we have defined our random intercept model with a likelihood function and mixed effects covariance matrix \( G \), then the following iterative parameter estimation algorithm follows from Wolfinger and O’Connell (1993).

- **Step 1.** Form pseudo-data \( \hat{y}^{[m+1]} \)
- **Step 2.** Estimate \( \hat{\psi}^{[m+1]} \) using \( \hat{\alpha}^{[m+1]} \) in (12) or (13) minimizing for \( \psi \) and then estimate \( \sigma^2 \) using (15). This requires iterative minimization methods and we will call this step 2 optimization.
- **Step 3.** Estimate \( \hat{\beta} \) using (14) and \( \hat{\alpha} \) using (15) and then set \( \hat{\beta}^{[m+1]} = \hat{\beta} \) and \( \hat{\alpha}^{[m+1]} = \hat{\alpha} \)

This process of updating \( \hat{\beta}^{[m]} \) and \( \hat{\alpha}^{[m]} \) is called the step 3 optimization

- **Step 4.** Set \( m = m + 1 \) and go to step 1 of convergence criteria are not satisfied. Each iteration through \( m \) is called an outer iteration.
step 3 optimization is closely related to iterative weighted least squares as used by McCullagh and Nelder (1989) for estimating parameters of GLMM when assuming that all of the data are independent.

When there are large data sets with temporal autocorrelation Ver Hoef et al (2010) noted that the step 2 optimization can be very slow because of the minimization of (12) or (13) requires matrix inverses and determinants whose dimensions are equal to the number of observations. This does also apply to our case as it is a case of crossed design.

4.3 Implementation and Results of the New Method

The response variable ‘mating’ follows a binary distribution conditional on the random effects (according to the binary logistic mixed model), hence the mean of the data is an event probability and the logit of the probability is linearly related to the linear predictor of the model.

There are two female and two male populations. The variable have 10 levels each. Consequently, we have 20 females and 20 males with four mating types. In each of the PL methods, there are two covariance parameters in the optimization leading to two variance components and the initial optimization computes pseudo-data based on the response values in the data set rather than from estimates of a generalized linear model fit.

We fitted this binary logistic mixed model in (16) to the salamander mating data via the new pseudo-likelihood algorithm highlighted in section 4.2 to obtain the fixed effect parameter $\beta$ and random effect $\alpha$ as well as the probability of mating success of the different species of the animals.

The implementation were carried out for new maximum pseudo-log-likelihood and new restricted pseudo-log-likelihood using the profiled maximum and restricted pseudo-log-likelihood in (12) and (13) respectively. Each of these implementations were carried out on the basis of Taylor’s series expansion around $\hat{\beta}$ and $E(\hat{\alpha})$ as well as $\hat{\beta}$ and $\hat{\alpha}$ giving rise to the new RSPL, MSPL, RMPL and MMPL methods.

The new RSPL is a restricted PL method based on Taylor’s series expansion of $\hat{\beta}$ and $\hat{\alpha}$, the new MSPL is a maximum PL method based on Taylor’s series expansion of fixed effect $\beta$ and $\alpha$, the new RMPL is a restricted PL method based on Taylor’s series expansion around $\hat{\beta}$ and $E(\hat{\alpha})$, while the new MMPL is a maximum likelihood PL method based on Taylor’s series expansion around $\hat{\beta}$ and $E(\hat{\alpha})$.

In addition to this, we applied the Laplace approximation and Quad methods which are likelihood-based methods. The Laplace approximation method performs maximum likelihood estimation based on Laplace approximation of the marginal log likelihood of the GLMM whereas the quad method approximated the marginal log likelihood with an adaptive Gauss-Hermite Quadrature rule.

The results of the implementations of Gibbs-EM, Bayes (Gibbs sampling), PQL, Laplace approximation, quad method as well as the MCEM on the data (by past authors such as Karim and Zeger (1992), Booth and Hobert (1999), Schall (1991), McCullagh and Nelder (1989) etc) are compared with these new PL methods.

The covariance parameter (subject / random effects) estimates, fixed effect (crossing effects) estimates and mating probability (marginal population-level probabilities) estimates / logit estimates obtained from the implementation of the new PL and likelihood-based methods as well as other existing standard methods are presented as follows:
V. DISCUSSION OF RESULTS

In this work, we developed and used the new pseudo-likelihood methods (RSPL, MSPL, RMPL, MMPL) which had roots from Wolfinger and O’Connell (1993) to calculate the pseudo-likelihood estimates for the parameters involved in model (16). We also applied the existing ML methods (Laplace approximation and Adaptive Gaussian quadrature methods AGQ). The parameters are the crossing effects (fixed effects) and the subject effects (random effects). The marginal or population-level probabilities of mating success were also obtained by each of these methods. Since most of the previous researchers have analyzed separately the data from the experiment, we included for comparison their results based on this experiment. In particular, some Bayesian-based methods, the moment and the PQL methods were compared with the new PL and ML methods.

From the numerical results of these implementations, It is clear that the new MMPL and new RMPL methods gave the same (identical) numerical results for the fixed effects and the mating probabilities. Also, it is obvious that the new MSPL, RMPL and MMPL methods gave closely related variance components estimates but the RMPL converged faster than the MSPL and MMPL. It also converged faster than any other method since it has the least number of iterations. The Laplace and the AGQ procedures which are ML methods gave exactly the same numerical estimates of the fixed effects, random effects and mating probabilities. The Bayesian-based methods (Bayes’ Gibbs sampling, Gibbs-EM and MCEM) also gave exactly the same mating probabilities but different fixed and random effects estimates. Moreso, the fixed effects estimates of both Gibbs-EM and Bayes’ (Gibbs sampling) are somewhat close.

This evidently shows that methods with similar computational techniques or similar framework often times, give very closely related results or sometimes exactly the same results. The same can be asserted about the new PL methods MMPL, MSPL and RMPL as well the ML methods (namely Laplace and Quad) which gave the same estimates for variance component, fixed effects and mating probabilities.

In the nutshell, all the results have similar patterns in accordance to the category of their method of estimation. It is noteworthy that all these methods differ also in their ability to estimate the various relevant quantities.

The new RSPL method stand out amongst the new PL and ML methods considered in this work since it gave precise variance component estimates, $\sigma^2_r = 1.4099, \sigma^2_m = 0.08963$ which coincided with PQL and REPL methods with estimates $\sigma^2_r = 1.41$ and $\sigma^2_m = 0.09$ by Breslow and Clayton (1993) as
well as Drum and McCullagh (1993) respectively. The variance component estimate of the new RSPL is also in reasonable agreement with the moment method of McCullagh and Nelder (1989) with the exception of male variance. Although the fixed effects and the mating probability estimates are not reported in these literatures.

The new RSPL result also compares well with the estimate of regression coefficients and variance components for the Salamander data (using model in 16) obtained through the REML approach by Schall (1991).

The ML methods gave fixed effects estimates PL and ML methods that are very close to that of the MCMC methods.

As regards the mating probability estimates, the (RSPL, MMPL, RMPL, MSPL, Quad and Laplace) gave closely related results.

From the result of the RSPL implementation, we also observe that after pairing of male rough butt salamander, with a whiteside female instead of a female from its own population, the logit drops from 1.1629 to -1.4119 with corresponding estimated probabilities of mating success \( \hat{\pi}_{RR} = 0.7619 \) and \( \hat{\pi}_{WR} = 0.1959 \). If the same comparisons are made in pairings with whiteside males rather than its own population, the logit drops from 1.0151 to 0.7839 whereas the corresponding estimated mating probability are \( \hat{\pi}_{WW} = 0.7340 \) and \( \hat{\pi}_{RW} = 0.6865 \). It follows that the whiteside females appear to be successful with their own population and whiteside males appear to succeed equally well with their female patners of the two populations.

We also observe that the other three new PL methods (and even the ML methods) result in estimates of the mating probabilities and logits which are very close and are identical to the new RSPL since pairing of salamander of the same population yielded higher mating success and lower mating success if otherwise.

VI. CONCLUSION AND RECOMMENDATION

The new RSPL stands out (for variance component estimation) among the new PL methods developed since its variance component estimates coincided with the REML, REPL and PQL results by Drum and McCullagh (1993), Schall (1991) and Breslow and Clayton (1993) respectively. It also provide a good means of estimating the variance component since its result is in reasonable agreement with the moment method of McCullagh and Nelder (1989) with the exception of the male variance.

We also conclude that any of the new PL methods or ML methods can be used because most of them gave a good account of the fixed effects estimates as well as the mating probabilities with the exception of the variance component estimates.

It is also clear that the choice of our computational method or technique depends on the quantity or parameter of interest. We recommend that, for a binary correlated data (crossed design), the new RSPL method is the most ideal for the computation of the variance components(random effects) while the other three PL methods are recommended for computation of the fixed effects and the mating probabilities.

All the four new PL methods in this work are also alternatives to overcoming computational difficulties encountered with the evaluation of likelihood functions of correlated binary data since the estimates of their fixed estimates and mating probabilities compare well with the proven techniques such as moment, Bayes (Gibbs sampling), MCEM methods developed by McCullagh and Nelder (1989), Karim and Zeger (1992) and Booth and Hobert (1997) respectively.

The PQL method on the other hand produces severely biased estimates for the variance components when modeling correlated binary data as later pointed out by Booth and Hobert (1999), the new RSPL in this work gives a less biased random effect estimate and converges reasonably faster.

The numerical results showed that the new PL methods work reasonably well when calculating the new pseudo-likelihood estimator of the parameters in the GLMMs for modeling correlated binary data since each of these methods gave high mating probability for the animals of the same species and low mating probability for animals of different species with the exception of the whiteside female which mates indiscriminately. This is evident from the significant interaction that exist between the male and female populations irrespective of the methods used. Infact, on a general note, the percentages of successful mating between the male and female animals from the two populations are in the
neighborhood of 70 percent for WS-WS, 23 percent for WS-RB, 67 percent for RB-WS and 73 percent for RB-RB. Thus, the percentage for WS-RB is much lower than all three other cases, which have similar percentages. As a result, the female WS salamander was highly significant whereas its male counterpart was insignificant for all methods. These results are also consistent with result of Jiang and Zhang (2001) and all the previous result earlier mentioned in this work, which found the interaction term significant.

REFERENCES


AUTHORS BIOGRAPHY

R. A. Adeleke was born on the 2nd of March, 1960 in Osan Ekiti, Ekiti State, Nigeria. He bagged his Bachelor of Science Degree B.Sc (Hon.) in Mathematics at the Ondo State University (now Ekiti State University) in 1986. He also bagged a Master of Science Degree (M.Sc) in Statistics at the University of Ibadan, Ibadan, Oyo State, Nigeria, in 1991. He bagged his Ph.D in Statistics at the Department of Mathematical Sciences, University of Ado-Ekiti (now Ekiti State University), Ekiti State, Nigeria. He has being a lecturer/researcher of Statistics to date since his first appointment as a Graduate Assistant in 1987 at the Department of Mathematical Sciences, Ekiti State University, Ekiti State, Nigeria. He is a member of some professional bodies which include Mathematical Association of Nigeria (MAN) and National Statistical Association (NSA), Nigeria. He is also the incumbent president of National Statistical Association, Ekiti State Chapter, Ekiti State, Nigeria. He is currently an associate professor with numerous scholarly publications. He is married and blessed with four children.

O.Y. Halid was born on the 15th of September 1982 in Ibadan Oyo State, Nigeria. He bagged a Bachelor of Science Degree in Statistics at the Obafemi Awolowo University, Ile-Ife, Osun State, Nigeria in 2004. He also bagged a Masters Degree in Statistics at the University of Ado-Ekiti (now Ekiti State University, Ado-Ekiti, Ekiti State, Nigeria in 2009. He has a few publications in Statistics and is a member of a few professional bodies which include National Statistical Association (NSA), Nigeria, Royal Statistical Society (RSS), United Kingdom and America Statistical Association (ASA), USA. He has being a lecturer of Statistics at the Department of Mathematical Sciences, Ekiti State University, Ado-Ekiti, Ekiti State, Nigeria since 2009 and his area of research in Statistics is Computational Econometrics. He is currently awaiting his final oral defence of his Ph.D in Statistics. Mr. Halid is married and blessed with a son.