Likelihood-based and Bayesian Methods for Tweedie Compound Poisson Linear Mixed Models

Yanwei Zhang
actuary_zhang@hotmail.com

Abstract

The Tweedie compound Poisson distribution is a subclass of the exponential dispersion family with a power variance function, in which the value of the power index lies in the interval $(1, 2)$. It is well known that the Tweedie compound Poisson density function is not analytically tractable, and numerical procedures that allow the density to be accurately and fast evaluated did not appear until fairly recently. Unsurprisingly, there has been little statistical literature devoted to full maximum likelihood inference for Tweedie compound Poisson mixed models. To date, the focus has been on estimation methods in the quasi-likelihood framework. Further, Tweedie compound Poisson mixed models involve an unknown variance function, which has a significant impact on hypothesis tests and predictive uncertainty measures. The estimation of the unknown variance function is thus of independent interest in many applications. However, quasi-likelihood-based methods are not well suited to this task. This paper presents several likelihood-based inferential methods for the Tweedie compound Poisson mixed model that enable estimation of the variance function from the data. These algorithms include the likelihood approximation method, in which both the integral over the random effects and the compound Poisson density function are evaluated numerically; and the latent variable approach, in which maximum likelihood estimation is carried out via the Monte Carlo EM algorithm, without the need for approximating the density function. In addition, we derive the corresponding Markov Chain Monte Carlo algorithm for a Bayesian formulation of the mixed model. We demonstrate the use of the various methods through a numerical example, and conduct an array of simulation studies to evaluate the statistical properties of the proposed estimators.

Keywords: Adaptive Gauss-Hermite quadrature, Extended quasi-likelihood, Laplace approximation, Monte Carlo EM, Maximum likelihood estimation, Mixed models, Penalized quasi-likelihood, Tweedie compound Poisson distribution.

1 Introduction

The exponential dispersion model (Jørgensen 1987) plays an important role in modern applied data analysis, as it is the underlying response distribution in many commonly used statistical models. A two-parameter representation of the exponential dispersion model is

$$p(y|\theta,\phi) = a(y,\phi)\exp\left(\frac{y\theta-\kappa(\theta)}{\phi}\right), \tag{1.1}$$

where a and κ are known functions, θ is the natural parameter and $\phi > 0$ is the dispersion parameter. For this family of distributions, we have the well-known relationships $E(y) = \mu = \kappa'(\theta)$ and $Var(y) = \phi\kappa''(\theta)$ (e.g., see McCullagh and Nelder 1989). Since the mapping from θ to μ is one-to-one (Barndorff-Nielsen 1978), $\kappa''(\theta)$ can also be represented as a function of μ, denoted by $V(\mu)$. This is generally known as the variance function, which uniquely defines an exponential dispersion model (Jørgensen 1987).

The paper focuses on the exponential dispersion model with a power variance function $V(\mu) = \mu^p$, in which the value of the index parameter p lies in the interval $(1, 2)$. This particular distribution is generated by a compound Poisson-Gamma distribution and has a probability mass at the origin accompanied by a skewed continuous distribution on the positive real line. Henceforth we refer to it as the Tweedie compound Poisson distribution, or simply the compound Poisson distribution. Extensive applications of this distribution, mainly in the context of generalized linear models [GLM], have been found in a wide range of fields where continuous data with exact zeros regularly arise. Dunn and Smyth (2005) give an excellent account of the various areas where the Tweedie compound Poisson distribution has been applied.

Nevertheless, it is well known that the normalizing quantity $a(y, \phi, p)$ (we express the normalizing quantity as $a(y, \phi, p)$ to emphasize its dependence on the index parameter p) for the compound Poisson distribution is not analytically tractable, and the density of the compound Poisson distribution cannot be expressed in a closed form. Numerical methods that enable accurate and fast evaluation of the density function did not appear until fairly recently (Dunn and Smyth 2005, 2008). Not surprisingly, implementations of the compound Poisson mixed models, to date, have been limited to the penalized quasi-likelihood [PQL] method (Breslow and Clayton 1993), in which the inferential scheme only requires knowledge of the first two moments. Nevertheless, before the PQL method can be applied, the variance function, i.e., the index parameter, must be determined. In many applications, the value of the index parameter is specified beforehand based on expert judgment. This is only acceptable when the goal of the study is to estimate regression coefficients, upon which the variance function has negligible effects (e.g., see Davidian and Carroll 1987, Dunn and Smyth 2005). Problems arise when statistical inference goes beyond mere regression coefficient estimation. For example, in their paper on variance function estimation, Davidian and Carroll (1987) state, "how well one models and estimates the variances will have substantial impact on prediction and calibration based on the estimated mean response" and "far from being only a nuisance parameter, the structural variance parameter can be an important part of a statistical analysis". Indeed, the study of Peters *et. al* (2009) shows that estimation of the index parameter in the Tweedie compound Poisson GLM is of special interest to the insurance industry, as it has a material impact on the uncertainty measures of the predicted outstanding liability, a critical element used in strategic planning, accounting and risk management by insurance companies.

Methods that enable estimation of the index parameter have appeared for generalized linear models, in which the index parameter is determined by maximizing the profile extended quasi-likelihood (Cox and Reid 1987, Nelder and Pregibon 1987), or the profile likelihood evaluated using the density approximation methods (Dunn and Smyth 2005, 2008). However, two issues complicate implementing these in the PQL-based compound Poisson mixed models. First, an inherent difficulty with the extended quasi-likelihood approach is that it cannot handle exact zero values and requires the response variable to be adjusted away from zero by adding an small positive number (see Nelder and Pregibon 1987). However, Dunn and Smyth (2005) point out that this adjustment is inappropriate for modeling the compound Poisson data, as the parameter estimates are extremely sensitive to the choice of the adjustment. We experience a similar problem when applying the penalized extended quasi-likelihood method to the compound Poisson mixed model. Second, the underlying objective function optimized in PQL is not truly an approximation of the likelihood function (Pinheiro and Chao 2006), which precludes its use in the profile likelihood to make inference of the index parameter.

The intractable density function and the unknown variance function have presented distinct challenges to statistical inference problems involving compound Poisson mixed models. In this paper, we derive several likelihood-based methods for compound Poisson mixed models that enable estimation of the variance function. Given the recent development in numerical approximation of the compound Poisson density, existing inference methods for mixed models using Laplace approximation (Tierney and Kadane 1986) and adaptive Gauss-Hermite quadrature [AGQ] (Liu and Pierce 1994) can be modified and implemented. These methods evaluate the likelihood by approximating the integral over the distribution of the random effects. The approximated likelihood is then optimized numerically to produce parameter estimates. In many existing mixed models, the normalizing quantity in (1.1) does not enter the likelihood to be maximized either because it is simply a constant involving no parameters (e.g., Bernoulli, Poisson and Binomial), or because it is a function of only the dispersion parameter which can be profiled out of the likelihood (e.g., Normal). In contrast, the Laplace and quadrature approximations of the compound Poisson mixed model must take into account the normalizing quantity, as it depends, in a complicated way, on both the dispersion parameter and the index parameter, whose values are to be estimated from the numerical optimization.

The dependence of the above likelihood approximation methods on the compound Poisson density implies that they are subject to the degree of accuracy of the density approximation methods, and further, they will fail when the underlying compound Poisson density evaluation methods encounter numerical difficulties. For example, the series evaluation method of Dunn and Smyth

(2005) may fail to work in certain regions of parameter space where the number of terms required to approximate the density to a given level of accuracy is prohibitively large. We consider an alternative method that treats the unobserved Poisson variable implicit in the compound Poisson distribution, as well as the model's random effects, as latent variables (Laird and Ware 1982). The corresponding maximum likelihood estimation can be carried out via the Monte Carlo EM [MCEM] algorithm (Wei and Tanner 1990, McCulloch 1997). This latent variable approach avoids the need to approximate the compound Poisson density numerically. Rather, a Monte Carlo expectation step is implemented in which samples of the latent variables are drawn from the conditional distribution of the latent variables on the observed data.

In addition, we derive the Markov Chain Monte Carlo [MCMC] (Gelman *et. al* 2003) method, providing a Bayesian formulation of the compound Poisson mixed model (Zeger and Karim 1991). Corresponding to the above frequentist methods, there are two ways of implementing the MCMC algorithm: one relying on direct compound Poisson density approximation and the other exploiting latent variables. The MCEM and MCMC methods are more general than their likelihood approximation method counterparts: they can accommodate various random effect structures or random effects from non-Normal distributions. However, by their Monte Carlo nature, they are more computationally demanding and subject to Monte Carlo errors.

Details of these algorithms will be presented in section 3 after a brief review of the compound Poisson distribution and the density approximation method in section 2. Section 4 will apply the proposed algorithms to a data set analyzed in Dunn and Smyth (2005). In section 5, we conduct a series of simulation studies to investigate the statistical properties of the proposed algorithms. Section 6 will provide concluding comments.

2 The compound Poisson distribution

In this section, we briefly review the compound Poisson distribution considered in the paper. In particular, we describe the series evaluation method (Dunn and Smyth 2005) to approximate the intractable density function, which is required by certain likelihood-based estimation routines presented in the next section.

2.1 The compound Poisson distribution as an exponential dispersion model

It can be shown (e.g., see Jørgensen 1987) that the exponential dispersion model, with $V(\mu) = \mu^p$ for some known value of $p \in (1, 2)$, collapses to a compound Poisson-Gamma random variable generated in the following way:

$$Y = \sum_{i=1}^{T} X_i, \; T \sim Pois(\lambda), \; X_i \overset{\text{iid}}{\sim} Ga(\alpha, \gamma), \; T \perp X_i, \tag{2.1}$$

where $Pois(\lambda)$ denotes a Poisson random variable with mean λ, and $Ga(\alpha, \gamma)$ denotes a Gamma random variable with mean and variance equal to $\alpha\gamma$ and $\alpha\gamma^2$, respectively. Implicit in this definition is that if $T = 0$ then $Y = 0$, thereby allowing the distribution to have a probability mass at the origin. When $T > 0$, the response variable Y is the sum of T i.i.d Gamma random variables, implying that $Y|T \sim Ga(T\alpha, \gamma)$. As a result, the compound Poisson distribution has a probability mass at zero accompanied by a skewed continuous distribution on the positive real line. This distinctive feature makes it ideally suited for modeling continuous data with exact zeros that frequently arise in many applied fields. Indeed, for certain applications, there is some intuitive appeal to exploit this distribution, where the underlying data can be considered as generated by a compound process. For example,

- in actuarial science, Y is the aggregate claim amount for a covered risk, T the number of reported claims and X_i the insurance payment for the i_{th} claim;
- in rainfall modeling, Y is the total amount of precipitation in a given period, T the number of rain events and X_i the intensity of the precipitation for the i_{th} rain event;
- in ecological studies of stock abundance, Y is the total biomass in a certain area, T the number of patches of organisms and X_i the measured biomass for the i_{th} patch.

By means of deriving and equating the cumulant generating functions for (1.1) and (2.1), we can work out the relationship between the two sets of parameters in the two representations as:

$$\begin{aligned} \mu &= \lambda\alpha\gamma, & \lambda &= \frac{\mu^{2-p}}{\phi(2-p)}, \\ p &= \frac{\alpha+2}{\alpha+1}, & \alpha &= \frac{2-p}{p-1}, \\ \phi &= \frac{\lambda^{1-p} \cdot (\alpha\gamma)^{2-p}}{2-p}, & \gamma &= \phi(p-1)\mu^{p-1}. \end{aligned} \tag{2.2}$$

2.2 Numerical approximation to the density function

From (2.1), the joint distribution of the compound Poisson and the Poisson variables can be derived as

$$p(y, t|\lambda, \alpha, \gamma) = p(y|t, \alpha, \gamma)p(t|\lambda) = \begin{cases} \exp(-\lambda) & (0, 0) \\ \frac{y^{t\alpha-1}\exp(-y/\gamma)}{\Gamma(t\alpha)\gamma^{t\alpha}} \cdot \frac{\lambda^t \exp(-\lambda)}{t!} & \mathbb{R}^+ \times \mathbb{Z}^+. \end{cases} \tag{2.3}$$

To recover the marginal distribution of Y, we integrate out T in (2.3), that is, $p(y|\lambda, \alpha, \gamma) = \sum_{t=0}^{\infty} p(y, t|\lambda, \alpha, \gamma)$. Equating this with (1.1) and canceling out common terms, we obtain the normalizing quantity in (1.1) as

$$a(y, \phi, p) = \frac{1}{y} \sum_{t=1}^{\infty} \frac{y^t}{(p-1)^{t\alpha} \phi^{t(1+\alpha)} (2-p)^t t! \Gamma(t\alpha)} = \frac{1}{y} \sum_{t=1}^{\infty} W_t. \tag{2.4}$$

It is well known (e.g., see Dunn and Smyth 2005) that the above formula for the normalizing quantity does not have a closed-form expression. However, methods have been proposed that can approximate it reasonably well. Dunn and Smyth (2005) show that as a function of t, (2.4) is strictly convex, and that W_t decays faster than geometrically on either side of the mode of W_t, denoted $t_{\max}$. As a result, we can replace the infinite sum with finite sum over important terms. That is, we can find the limits $t_L < t_{\max}$ and $t_U > t_{\max}$ such that W_{t_L} and W_{t_U} are less than $\epsilon W_{t_{\max}}$ for some given threshold ϵ, and evaluate the normalizing quantity as

$$\hat{a}(y, \phi, p) = \frac{1}{y} \sum_{t=t_L}^{t_U} W_t. \tag{2.5}$$

Using the Stirling's formula to replace the Gamma function in (2.4), and taking the derivative with respect to t, we can find the approximate mode $t_{\max}$ as

$$t_{\max} = \frac{y^{2-p}}{(2-p)\phi}. \tag{2.6}$$

We evaluate the formula for W_t in (2.4) at $t_{\max}$ to obtain $W_{t_{\max}}$, which is then used to determine the two limits t_L and t_U.

The above series evaluation method is fairly straightforward to implement, however, Dunn and Smyth (2005) note that the number of terms required to approximate the density to a given accuracy could increase without bound in certain regions. An alternative approach (Dunn and Smyth 2008) evaluates the density by Fourier inversion of the characteristic function, and has better performance in these situations. Nevertheless, the more complex Fourier inversion method does not dominate the series evaluation approach universally, as it has been found to be less accurate in other certain parameter domains. In general, these two methods are regarded as complementary.

3 Compound Poisson linear mixed models

In this section, we consider statistical inference of the compound Poisson linear mixed model. Suppose that there are N observations. The mixed model assumes that the $N \times 1$ mean response vector $\boldsymbol{\mu}$ is stipulated by some linear predictors through a monotonic link function η as

$$\eta(\boldsymbol{\mu}) = \boldsymbol{X\beta} + \boldsymbol{Zb}, \tag{3.1}$$

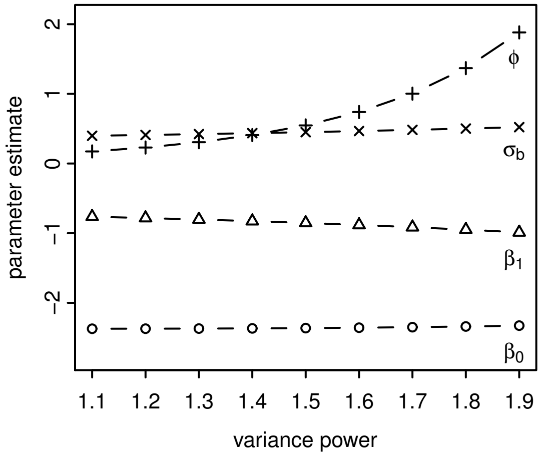

Figure 1: Estimates of the two fixed-effects (β_0 and β_1), the variance component (σ_b) and the dispersion parameter (ϕ) using the penalized quasi-likelihood method with a given variance power $p = 1.1, 1.2, \cdots, 1.9$.

where $\boldsymbol{\beta}$ is a $J \times 1$ vector of fixed effects, $\boldsymbol{b}$ a $K \times 1$ vector of random effects, and $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the associated design matrices. Moreover, the random effects have their own distribution

$$\boldsymbol{b} \sim N(\mathbf{0}, \boldsymbol{\Sigma}). \tag{3.2}$$

This is the typical setup in a mixed model. We have seen in section 2.1 that for a given p, the compound Poisson distribution is a member of the exponential dispersion family. In such a case, existing inferential procedures developed for the exponential dispersion family (e.g., McCulloch and Searle 2001) can be readily applied. However, the index parameter p is generally unknown beforehand, in which the compound Poisson distribution can no longer be expressed in the form of the exponential dispersion family. Methods that enable estimation of the index parameter along with other parameters of interest must be employed.

Existing theory (e.g., Smyth 1996) suggests that for the compound Poisson model, the mean $\boldsymbol{\mu}$ is orthogonal to p and ϕ, meaning that the mean, and thus the regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{b}$, vary slowly as p and ϕ change. For this reason, it is perhaps reasonable to specify the index parameter based on expert judgment when the goal of the study is solely to estimate regression coefficients. However, problems arise when statistical inference goes beyond mere regression coefficient estimation, as the index parameter significantly impacts the estimation of the dispersion parameter, which, in turn, has a substantial influence on the estimation of asymptotic standard errors of the regression coefficients, statistical hypothesis tests, and predictive uncertainty measures.

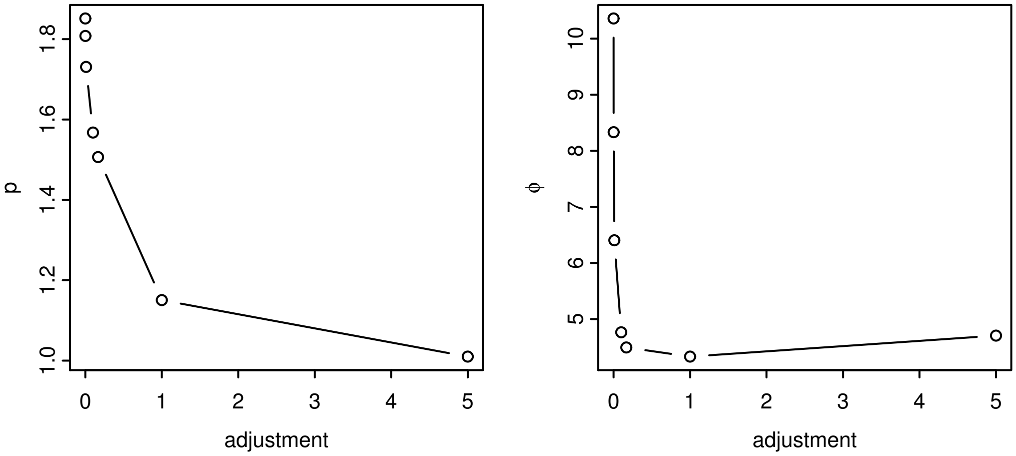

Figure 2: Plot of the estimates of the index parameter (p) and the dispersion parameter (ϕ) corresponding to seven different adjustments adopted in the penalized extended quasi-likelihood method.

The impact of the index p on the estimation of the other parameters in a mixed model is illustrated in Figure 1, in which we fit nine compound Poisson mixed models to a real data set using the penalized quasi-likelihood method [PQL] for each value of $p \in \{1.1, 1.2, \cdots, 1.9\}$. The parameter estimates of the two fixed effects (β_0 and β_1), the variance component (σ_b) and the dispersion parameter (ϕ) are then plotted against the value of the index parameter used. From Figure 1, it is noteworthy that the value of the index parameter has a significant impact on the estimation of the dispersion parameter, but only a slight effect on the estimation of the mean parameters and the variance component.

The penalized quasi-likelihood method, however, is not equipped to estimate the variance function. A natural modification, the extended quasi-likelihood (Nelder and Pregibon 1987), can be exploited, in which the PQL is used to estimate the fixed and random effects, the variance component and the marginal deviance for a given value of p, and the resulting profiled extended quasi-likelihood is maximized to produce the estimate of p. Unfortunately, the extended quasi-likelihood involves a term $\log(V(y))$, which will become infinite if $y = 0$. To overcome this issue, the observed zeros are adjusted away from the origin by adding a small positive constant c. However, this adjustment is inappropriate for the compound Poisson distribution because the resulting parameter estimates are extremely sensitive to the choice of the adjustment. For example, Figure 2 shows the estimates of the parameter p and ϕ corresponding to six different values of c used in the penalized extended quasi-likelihood method [PEQL].

We see that the penalized extended quasi-likelihood method suffers from the dependency on the ad hoc adjustment and is not well suited for estimating the variance function in the compound Poisson mixed model. In the following, we derive several likelihood-based methods that enable the variance function to be estimated from the data. Further, in contrast to the PQL method, these methods are approximations to the true likelihood, thus permitting the construction of likelihood ratio tests for comparing nested models.

3.1 Laplace approximation

For convenience in deriving the conditional mode of the random effects needed in the following, we further express the variance component in (3.2) in terms of the relative covariance factor $\boldsymbol{\Lambda}$ such that

$$\boldsymbol{\Sigma} = \phi \boldsymbol{\Lambda} \boldsymbol{\Lambda}'. \tag{3.3}$$

As a result, the specification of the mean vector in (3.1) can be further expressed as

$$\eta(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{u} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}^*\boldsymbol{u}, \tag{3.4}$$

where $\boldsymbol{u} \sim N(\boldsymbol{0}, \phi \boldsymbol{I})$.

Given this formulation, we have two sets of random variables, the observation $\boldsymbol{Y}$ and the unobserved (scaled) random effects $\boldsymbol{u}$, and we are interested at estimating the set of parameters $\boldsymbol{\Theta} = (\boldsymbol{\beta}, \phi, p, \boldsymbol{\Lambda})$. These parameters are estimated by maximizing the observed likelihood, that is, the marginal likelihood where the random effects are integrated out from the joint likelihood of $(\boldsymbol{Y}, \boldsymbol{u})$:

$$p(\boldsymbol{y}|\boldsymbol{\Theta}) = \int p(\boldsymbol{y}, \boldsymbol{u}|\boldsymbol{\Theta}) d\boldsymbol{u} = \int p(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\Theta}) p(\boldsymbol{u}|\phi) d\boldsymbol{u}. \tag{3.5}$$

Since the integral in (3.5) is often intractable, we evaluate it via the Laplace approximation (Tierney and Kadane 1986), in which the integrand in (3.5) is replaced by its second-order Taylor series at the conditional mode of $\boldsymbol{u}$ given the current value of $\boldsymbol{\Theta}$. The resulting marginal likelihood is a function of only $\boldsymbol{\Theta}$, which can be maximized using numerical optimization procedures to find the maximum likelihood estimate of $\boldsymbol{\Theta}$.

The conditional mode of $\boldsymbol{u}$ can be located via the widely used Fisher's scoring algorithm, often referred to as the penalized iteratively re-weighted least squares in the context of mixed models (e.g., Bates *et. al* 2012). To derive the conditional mode, we first note that the joint loglikelihood of $(\boldsymbol{Y}, \boldsymbol{u})$ is:

$$\ell(\boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{u}) = \log p(\boldsymbol{y}, \boldsymbol{u}|\boldsymbol{\Theta}) = \sum_{i=1}^{N} \log p(y_i|\boldsymbol{u}, \boldsymbol{\Theta}) - \frac{K}{2}\log\phi - \frac{\boldsymbol{u}'\boldsymbol{u}}{2\phi}, \tag{3.6}$$

where $\log p(y_i|\boldsymbol{u}, \boldsymbol{\Theta})$ is the conditional loglikelihood of the data given the random effects. Since the conditional distribution of $\boldsymbol{u}$ is proportional to the joint distribution of $(\boldsymbol{Y}, \boldsymbol{u})$, the conditional mode of $\boldsymbol{u}$ can be located by maximizing the joint loglikelihood in (3.6) with respect to $\boldsymbol{u}$. We denote $g(\mu) = \partial\eta/\partial\mu$. From (3.6), we can compute the derivative of the joint loglikelihood with respect to u_r, the r_{th} element of $\boldsymbol{u}$, as

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{u})}{\partial u_r} &= \sum_{i=1}^{N} \frac{\partial \log p(y_i|\boldsymbol{u}, \boldsymbol{\Theta})}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial u_r} - \frac{u_r}{\phi} \\ &= \frac{1}{\phi} \left(\sum_{i=1}^{N} \frac{1}{g(\mu_i)^2 V(\mu_i)} (y_i - \mu_i) g(\mu_i) Z^*_{ir} - u_r \right). \end{aligned}$$

Therefore, the gradient can be expressed, in matrix notation, as

$$\frac{\partial \ell(\boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{u})}{\partial \boldsymbol{u}} = \frac{1}{\phi} \left(\boldsymbol{Z}^{*'} \boldsymbol{W} \boldsymbol{g}(\boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu}) - \boldsymbol{u} \right), \tag{3.7}$$

where $\boldsymbol{g}(\boldsymbol{\mu})$ is an $N \times N$ diagonal matrix whose i_{th} diagonal element is $g(\mu_i)$ and $\boldsymbol{W}$ is an $N \times N$ diagonal matrix whose i_{th} diagonal element w_{ii} satisfies $w_{ii}^{-1} = g(\mu_i)^2 V(\mu_i)$. Differentiating (3.7) again with respect to $\boldsymbol{u'}$ and taking expectation with respect to the distribution of $\boldsymbol{Y}$, we can approximate the Hessian matrix as

$$\frac{\partial \ell^2(\boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{u})}{\partial \boldsymbol{u} \partial \boldsymbol{u'}} \approx -\frac{1}{\phi} \left(\boldsymbol{Z}^{*'} \boldsymbol{W} \boldsymbol{Z}^* + \boldsymbol{I} \right) = -\frac{1}{\phi} \boldsymbol{L}\boldsymbol{L}', \tag{3.8}$$

where $\boldsymbol{L}$ is the Cholesky factor of $\boldsymbol{Z}^{*'} \boldsymbol{W} \boldsymbol{Z}^* + \boldsymbol{I}$. With (3.7) and (3.8), we can define an iterative scheme that eventually leads to the conditional mode of $\boldsymbol{u}$, where the update of $\boldsymbol{u}$ in each iteration is given by

$$\boldsymbol{u}_{new} = \boldsymbol{u}_{old} + \left(\boldsymbol{Z}^{*'} \boldsymbol{W} \boldsymbol{Z}^* + \boldsymbol{I} \right)^{-1} \left(\boldsymbol{Z}^{*'} \boldsymbol{W} \boldsymbol{g}(\boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu}) - \boldsymbol{u}_{old} \right). \tag{3.9}$$

In the above, the weight matrix $\boldsymbol{W}$ and the mean response $\boldsymbol{\mu}$ are updated in each iteration based on the value of $\boldsymbol{u}_{old}$. As a result of the definition in (3.3), the dispersion parameter from (3.7) and (3.8) cancels out and does not enter the scoring algorithm. This is similar to the scoring algorithm in generalized linear models (McCullagh and Nelder 1989). Convergence of the above scoring algorithm is declared when the relative change in the linear predictors falls below a threshold.

Denoting the found conditional mode as $\hat{\boldsymbol{u}}$, we estimate the variance of $\hat{\boldsymbol{u}}$ by inverting the negative approximated Hessian matrix as

$$Var(\hat{\boldsymbol{u}}) \approx \phi \left(\boldsymbol{L}\boldsymbol{L}' \right)^{-1}. \tag{3.10}$$

With the conditional mode and the approximate Hessian matrix, we compute the Laplace approximation of the marginal loglikelihood as

$$\begin{aligned}\ell(\boldsymbol{\Theta};\boldsymbol{y}) &\approx \ell(\boldsymbol{\Theta};\boldsymbol{y},\hat{\boldsymbol{u}}) + \frac{1}{2}\log|Var(\hat{\boldsymbol{u}})| \\ &= \sum_{i=1}^{N}\log p(y_i|\boldsymbol{\Theta},\hat{\boldsymbol{u}}) + \frac{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}}{2\phi} - \log|\boldsymbol{L}|. \end{aligned} \tag{3.11}$$

In many existing mixed models, the conditional loglikelihood $\log p(y_i|\boldsymbol{\Theta},\hat{\boldsymbol{u}})$ is often replaced by half of the deviance residuals (McCullagh and Nelder 1989), and the normalizing quantity in (1.1) is left out either because it is simply a constant with no parameters in it (e.g., Bernoulli, Poisson and Binomial), or because it involves only the dispersion parameter which can be profiled out of the marginal likelihood (e.g., Normal). The normalizing quantity in the compound Poisson distribution, however, depends on both the dispersion parameter and the index parameter in a complicated way, and it must be computed numerically and included in the conditional likelihood. The marginal loglikelihood in (3.11) is then maximized numerically to produce the estimate of $\boldsymbol{\Theta}$, subject to the constraints: $\phi > 0$, $p \in (1,2)$ and $diag(\boldsymbol{\Lambda}) > 0$.

3.2 Adaptive Gauss-Hermite quadrature

When there are no crossed random effects, the integral in (3.5) can also be approximated using the more accurate adaptive Gauss-Hermite quadrature [AGQ] (Liu and Pierce 1994), which replaces the integral by a weighted sum of the integrand at specified knots. These knots are zeros of the Hermite polynomial, but are transformed so that the integrand is sampled in an appropriate region. When there is only one knot, the AGQ method collapses to the Laplace approximation.

Specifically, an L-knot AGQ approximation to the integral of a general function $g(t)$ is

$$\int g(t)dt \approx \sqrt{2}\hat{\sigma}\sum_{l=1}^{L} w_l \exp(x_l^2) g(\hat{\mu} + \sqrt{2}\hat{\sigma}x_l), \tag{3.12}$$

where x_l and w_l are the pre-determined zeros and the corresponding weights of the Hermite polynomial, and $\hat{\mu}$ and $\hat{\sigma}$ are the mode of $g(t)$ and the square root of the variation of $g(t)$ at its mode, respectively.

For simplicity in the formulation, we suppose there is a single grouping factor with K levels so that there is only one random effect u_k for each level. Then the data can be partitioned into K groups so that each group is conditionally independent to each other given the random effects. We denote $p(\boldsymbol{y}_k, u_k|\boldsymbol{\Theta})$ as the joint density of the observations and the random effect for group k, so that the joint likelihood can be written as $p(\boldsymbol{y},\boldsymbol{u}|\boldsymbol{\Theta}) = \prod_{k=1}^{K} p(\boldsymbol{y}_k, u_k|\boldsymbol{\Theta})$. Therefore, the AGQ approximation to (3.5) is

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\Theta}) &= \prod_{k=1}^{K} \int p(\boldsymbol{y}_k, u_k|\boldsymbol{\Theta}) du_k \\ &= \prod_{k=1}^{K} \left(\sqrt{2}\hat{\sigma}_{u_k} \sum_{l=1}^{L} w_l \exp(x_l^2) p(\boldsymbol{y}_k, \hat{u}_k + \sqrt{2}\hat{\sigma}_{u_k} x_l|\boldsymbol{\Theta}) \right), \end{aligned} \tag{3.13}$$

where $\hat{u}_k$ is the conditional mode and $\hat{\sigma}_{u_k} = \sqrt{\phi}/L_{kk}$ is the standard deviation of $\hat{u}_k$, as found in (3.9) and (3.10).

3.3 Monte Carlo EM

The above two methods approximate the integral in (3.5) explicitly and in computing the marginal likelihood, they rely on the capability to directly evaluate the density function of the compound Poisson distribution. The Monte Carlo EM [MCEM] algorithm (Wei and Tanner 1990) to be presented here is fundamentally distinct in that latent variables are employed and direct compound Poisson density evaluation is avoided. Therefore, it is independent of the degree of accuracy in the density approximation. Moreover, it is a general-purpose optimization procedure that can handle flexible random effect specifications. For example, the distribution of the random effects are not necessarily limited to be Normal. For this reason, we specify the distribution of the random effects as $p(\boldsymbol{b}|\boldsymbol{\Sigma})$, depending on some unknown parameter $\boldsymbol{\Sigma}$, and re-define $\boldsymbol{\Theta} = (\boldsymbol{\beta}, \phi, p, \boldsymbol{\Sigma})$. Throughout the next two subsections, we use the unscaled random effects $\boldsymbol{b}$ instead of the scaled version $\boldsymbol{u}$ used in the previous two subsections.

In the MCEM algorithm, both the unobserved Poisson variable $\boldsymbol{T}$ implicit in the compound Poisson distribution and the random effects $\boldsymbol{b}$ are treated as latent variables. Thus, the complete data is $(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{b})$, and the corresponding joint loglikelihood can be derived as

$$\ell(\boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{t}, \boldsymbol{b}) = \sum_{i=1}^{N} \log p(y_i, t_i|\boldsymbol{\Theta}, \boldsymbol{b}) + \log p(\boldsymbol{b}|\boldsymbol{\Sigma}). \tag{3.14}$$

Using the joint distribution of (Y, T) in (2.3), and plugging the expression for (μ_i, ϕ, p) in (2.2), the first term in (3.14) can be further expressed as

$$\begin{aligned} \sum_{i=1}^{N} \log p(y_i, t_i|\boldsymbol{\Theta}, \boldsymbol{b}) &= \sum_{y_i=0} \log p(y_i, t_i|\boldsymbol{\Theta}, \boldsymbol{b}) + \sum_{y_i>0} \log p(y_i, t_i|\boldsymbol{\Theta}, \boldsymbol{b}) \\ &= -\frac{1}{\phi} \sum_{i=1}^{N} \frac{\mu_i^{2-p}}{2-p} - \sum_{y_i>0} \left(\frac{y_i}{\phi(p-1)\mu_i^{p-1}} + \log\Gamma(t_i \frac{2-p}{p-1}) + \log t_i! \right) \end{aligned}$$

$$+\sum_{y_i>0} t_i \left(\frac{2-p}{p-1} \log \frac{y_i}{p-1} - \frac{\log \phi}{p-1} - \log(2-p) \right). \tag{3.15}$$

In the above, the constant term involving only y's has been dropped out.

In the E-step of the EM algorithm, given the current value of the parameters, $\mathbf{\Theta}_{old}$, we take the expectation of (3.14) with respect to the conditional distribution of the latent variables $(\boldsymbol{T}, \boldsymbol{b})$ given the observed data (we refer to this as the posterior distribution in the following discussion):

$$Q(\mathbf{\Theta}; \mathbf{\Theta}_{old}) = E_{(\boldsymbol{T},\boldsymbol{b})|\boldsymbol{Y}}[\ell(\mathbf{\Theta}; \boldsymbol{y}, \boldsymbol{t}, \boldsymbol{b})]. \tag{3.16}$$

However, the expectation with respect to the posterior distribution of $(\boldsymbol{T}, \boldsymbol{b})$ is intractable. We can resort to a Monte Carlo approach to evaluate it via simulated values from the posterior distribution. Suppose that we have M simulated samples from the posterior distribution of $(\boldsymbol{T}, \boldsymbol{b})$, then (3.16) can be approximated by

$$\hat{Q}(\mathbf{\Theta}; \mathbf{\Theta}_{old}) = \frac{1}{M} \sum_{m=1}^{M} \ell \left(\mathbf{\Theta}; \boldsymbol{y}, \boldsymbol{t}^{(m)}, \boldsymbol{b}^{(m)} \right), \tag{3.17}$$

where $\boldsymbol{t}^{(m)}$ and $\boldsymbol{b}^{(m)}$ are the m_{th} simulated sample. In particular, we note that t_i and μ_i are additive in (3.15), implying that the posterior distributions of the two latent variables $\boldsymbol{T}$ and $\boldsymbol{b}$ are independent. Therefore, simulations of $\boldsymbol{T}$ and $\boldsymbol{b}$ can be performed separately. The simulation of $\boldsymbol{T}$ from the posterior is relatively straightforward by noting that it factorizes over i and each t_i can be simulated using rejection sampling (e.g., see Robert and Casella 2004). The posterior distribution for $\boldsymbol{b}$ is more involved, and depending on the form of $p(\boldsymbol{b}|\mathbf{\Sigma})$, either rejection sampling (Booth and Hobert 1999) or Markov Chain Monte Carlo methods (McCulloch 1997) can be used.

In the subsequent M-step, a new estimate of $\mathbf{\Theta}$ is found by maximizing the Monte Carlo estimate $\hat{Q}(\mathbf{\Theta}; \mathbf{\Theta}_{old})$. The conditional maximization approach in Meng and Rubin (1993) is used so that we can update each component of $\mathbf{\Theta}$ sequentially. To update $\boldsymbol{\beta}$, we use the iterative Newton-Raphson approach from McCulloch (1997):

$$\boldsymbol{\beta}_{new} = \boldsymbol{\beta}_{old} + \hat{E}_{\boldsymbol{b}|\boldsymbol{Y}}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{E}_{\boldsymbol{b}|\boldsymbol{Y}}[\boldsymbol{W}\boldsymbol{g}(\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})], \tag{3.18}$$

where $\boldsymbol{W}$ and $\boldsymbol{g}(\boldsymbol{\mu})$ are as defined in (3.7) and computed using the value of $\mathbf{\Theta}_{old}$. The notation $\hat{E}$ represents the Monte Carlo estimate of the expectation. The update of ϕ also has a closed-form solution:

$$\phi_{new} = \frac{\frac{p-1}{2-p}\sum_{i=1}^{N} \hat{E}_{\boldsymbol{u}|\boldsymbol{Y}}(\mu_i^{2-p}) + \sum_{y_i>0} y_i \hat{E}_{\boldsymbol{u}|\boldsymbol{Y}}(\mu_i^{1-p})}{\sum_{y_i>0} \hat{E}_{\boldsymbol{t}|\boldsymbol{Y}}(t_i)}. \tag{3.19}$$

The maximization over $\mathbf{\Sigma}$ is often simple and the solution depends on the form of $p(\boldsymbol{b}|\mathbf{\Sigma})$ and the structure of $\mathbf{\Sigma}$. For example, if $\boldsymbol{b}$ is Normally distributed and $\mathbf{\Sigma} = \sigma^2\boldsymbol{I}$, then $\sigma^2_{new} = \frac{1}{K}\hat{E}_{\boldsymbol{b}|\boldsymbol{Y}}(\boldsymbol{b}'\boldsymbol{b})$.

Maximization of p is achieved through a constrained optimization procedure with the restriction $p \in (1, 2)$.

The E-step and M-step as described in the above will be implemented iteratively and the algorithm will converge to a local maximum (e.g., see Wei and Tanner 1990).

3.4 Markov Chain Monte Carlo

Markov Chain Monte Carlo [MCMC] is another popular approach to handling mixed models, in which the model is formulated in a Bayesian setting and inference is made based on simulated samples from the posterior distribution of the parameters. The MCMC approach is also a general algorithm that can accommodate various random effect structures or random effects from non-Normal distributions. To make inference for the set of parameters $\boldsymbol{\Theta} = (\boldsymbol{\beta}, \phi, p, \boldsymbol{\Sigma})$, we draw simulated samples from the posterior distribution of $\boldsymbol{\Theta}$ given the data. The simulation is implemented through a Gibbs sampler (Gelman *et. al* 2003), which sequentially samples parameters from their lower-dimensional full conditional distributions over many iterations.

There are two ways in formulating the model based on whether the latent Poisson variable is employed. The first approach does not utilize the latent Poisson variable and similar to the likelihood approximation methods, it must make direct evaluation of the conditional distribution of the data on the random effects. That is, we estimate the posterior distribution of $(\boldsymbol{\Theta}, \boldsymbol{b})$ as:

$$p(\boldsymbol{\Theta}, \boldsymbol{b}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\beta}, \phi, p, \boldsymbol{b})p(\boldsymbol{b}|\boldsymbol{\Sigma})p(\boldsymbol{\Theta}), \tag{3.20}$$

where $p(\boldsymbol{\Theta})$ is the prior distribution for $\boldsymbol{\Theta}$. An alternative approach, similar to the MCEM algorithm above, makes use of the latent Poisson variable and computes the joint posterior distribution of $(\boldsymbol{\Theta}, \boldsymbol{T}, \boldsymbol{b})$ as

$$p(\boldsymbol{\Theta}, \boldsymbol{t}, \boldsymbol{b}|\boldsymbol{y}) \propto p(\boldsymbol{y}, \boldsymbol{t}|\boldsymbol{\beta}, \phi, p, \boldsymbol{b})p(\boldsymbol{b}|\boldsymbol{\Sigma})p(\boldsymbol{\Theta}). \tag{3.21}$$

In contrast to the first approach, the latter avoids the need for approximating the compound Poisson density function by simulating the latent variables from the full conditionals. In both cases, we draw simulations from the joint posterior, maintain the simulated values of $\boldsymbol{\Theta}$, and make inference about $\boldsymbol{\Theta}$ based on these simulated samples.

Indeed, the MCEM algorithm has a close relationship with the latter Gibbs sampler. By setting $M = 1$ in the E-step, we obtain one simulated sample of $\boldsymbol{T}$ and $\boldsymbol{b}$ from the posterior. Then, we replace each step in the M-step with one simulated sample from the full conditionals. This constitutes one iteration in the Gibbs sampling.

			Inner zone		Outer zone		All	
Plant	Stock	Spacing	# Samples	Mean RLD	# Samples	Mean RLD	# Samples	Mean RLD
1	Mark	5x3	40	0.080	51	0.019	91	0.046
2	Mark	5x3	35	0.061	51	0.011	86	0.032
3	Mark	4x2	26	0.061	18	0.006	44	0.039
4	Mark	4x2	26	0.112	15	0.029	41	0.082
5	MM106	5x3	36	0.123	49	0.053	85	0.083
6	MM106	5x3	34	0.106	47	0.099	81	0.102
7	M26	4x2	24	0.100	17	0.098	41	0.099
8	M26	4x2	24	0.146	18	0.061	42	0.110

Table 1: The number of samples and the mean of the root length density [RLD] by plant, root stock, plant spacing and root zone.

4 Illustrative examples

In this section, we demonstrate the use of the compound Poisson mixed model using the fine root data set from the study conducted by de Silva *et. al* (1999). The study examines possible factors that may affect the length density of fine roots, the main component of the root system through which vascular plants absorb water and nutrients. Specifically, the aim of their study is to investigate how the distribution of fine root length density [RLD] is affected by the geometry of the structural root system and the type of the root stock. Data are collected on eight apple trees, which are grafted onto one of three different root stocks (Mark, MM106 and M26) and planted at two different between-row × within-row spacings (4 × 2 meters and 5 × 3 meters). For each of the apple trees, a number of soil core sampling units are taken from which the fine roots are extracted. The total length of the fine roots in each core sample is measured, and the RLD is calculated by dividing the total length (cm) of the fine roots by the volume of the core sample (cm^3). Each of these samples is further classified as belonging to an inner or outer zone relative to each plant.

An exploratory data analysis is presented in Table 1, in which the number of samples in the data and the sample mean of the RLD are shown by tree, root stock, plant spacing and root zone. It is apparent that this study is not a full factorial design: the Mark root stock is tested at both plant spacings but the MM106 stock only at the 5 × 3 spacing and the M26 stock only at the 4 × 2 spacing. Further, for each of the plants, the observed sample mean RLD is much larger in the inner zone than that in the outer zone, this difference being greater for the Mark stock than the other two.

A distinct feature of this data is that the variable of primary interest, the RLD, has a mixed distribution: 37.8% of the RLD data are exact zeros while the rest take positive continuous values. The zeros are corresponding to the soil core samples that contain no fine roots, and accurately modeling the probability of zero RLD is considered an important part of the study. Noting that no transformation to normality is likely to be successful for this data due to the large proportion of zeros, Dunn and Smyth (2005) exploit the compound Poisson generalized linear model in their analyses. They specify a model including the stock and zone effects, the factors of primary interest in the study, as well as their interactions. After these effects are accounted for, the contribution from plant and spacing is not significant.

Dunn and Smyth (2005) find that the above model provides reasonable estimates of the stock and zone effects and is helpful to capture the observed pattern of zeros, but further comment that "a more complete treatment of this data might include fitting a generalized linear mixed model with plant as a random effect". Indeed, plant is a blocking factor in the randomized block design of de Silva *et. al* (1999), and thus, engenders a source of variation that must be accounted for when conclusions are drawn about the population, rather than about these particular plants themselves. Further, including plant as a random effect will account for the correlation induced due to multiple samples from the same plant.

Laplace and adaptive Gauss-Hermite quadrature. The above mixed model is most conveniently handled using the `cpglmm` function available in the R package `cplm`, which provides the Laplace and adaptive Gauss-Hermite quadrature methods to estimate the compound Poisson mixed model. For example, the following estimates the model using a 15-knot adaptive Gauss-Hermite quadrature:

```
fit <- cpglmm(RLD ~ Stock * Zone + (1 | Plant), data = FineRoot, nAGQ = 15)
```

The Laplace method is invoked when the number of quadrature knots is set to one (`nAGQ = 1`).

Monte Carlo EM. In the MCEM algorithm, we draw simulations of both the latent Poisson variable $\boldsymbol{T}$ and the random effects $\boldsymbol{b}$ via rejection sampling. A zero-truncated Poisson proposal distribution is exploited when producing samples of the latent Poisson variable $\boldsymbol{T}$, while a multivariate t-distribution with 6 degrees of freedom is used for sampling the random effects $\boldsymbol{b}$. The mean and variance of the multivariate t-distribution are chosen to match the mode and curvature of the posterior $p(\boldsymbol{b}|\boldsymbol{y})$, respectively. The sample size is set to 100 for the first 40 iterations, and increased to $3,000$ thereafter. To ensure that the algorithm is not stopped prematurely due to Monte Carlo errors, we perform three separate runs, each starting at a distinct set of initial values. The convergence history of each parameter is monitored and shown in Figure 3 (for the mean parameters, only the plots for β_0 - β_2 are included). It can be seen that within the first 40 iterations, all parameters have reached the neighborhood of the corresponding maximum likelihood estimates, which are from the 15-knot quadrature estimation and represented by the horizontal dashed gray lines. At convergence, all three runs result in parameter estimates with three-decimal accuracy, indicating that Monte Carlo errors are unlikely to be influential in the resulting statistical inference. To achieve higher accuracy, however, the number of Monte Carlo samples needed may be prohibitively large. Alternatively, one could construct an automated algorithm that increases the number of the simulated samples as the algorithm progresses by gauging the approximate Monte Carlo error in each iteration (Booth and Hobert 1999).

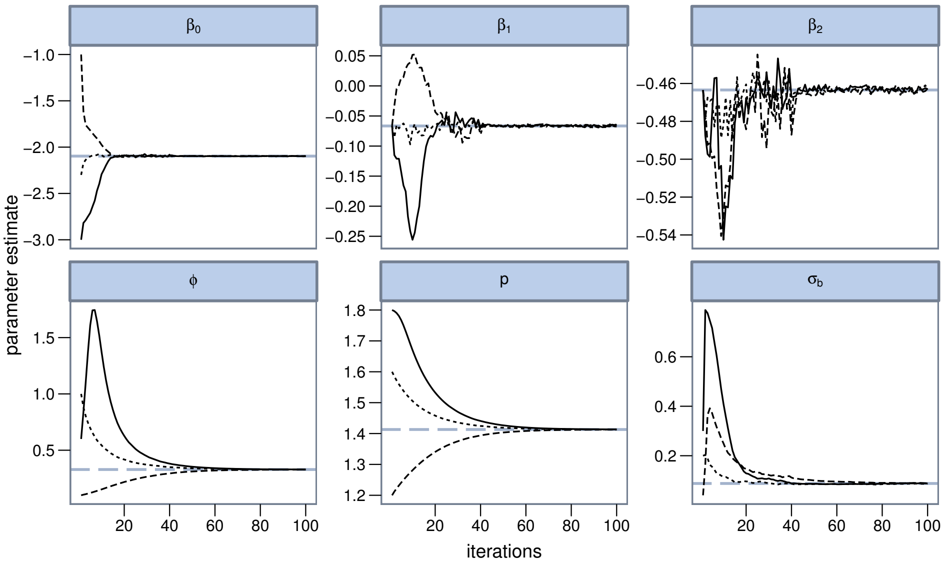

Figure 3: The iteration history for three runs of the MCEM algorithm with distinct starting values.

Markov Chain Monte Carlo. To implement the MCMC procedure, we formulate the model in a Bayesian setting. In particular, prior distributions for all parameters must be specified. In order to allow the parameters to be estimated largely from the data, non-informative priors are exploited: $\beta_j \sim N(0, 100^2)$, $j = 0, \cdots, 5$, $\phi \sim U(0, 100)$, $p \sim U(1, 2)$, and $\sigma_b^{-2} \sim Ga(0.001, 1000)$. Because of the conjugate Gamma prior, the posterior conditional distribution of σ_b^{-2} is still Gamma, for which samples could be directly generated. We implement the MCMC algorithm with direct density approximation in (3.20), and use the random walk Metropolis-Hastings algorithm (Gelman *et al.* 2003) for simulations of β_j, $j = 0, \cdots, 5$, b_k, $k = 1, \cdots, 8$, ϕ and p, in which a preliminary run with 5,000 iterations is used to tune the variances of the Normal proposal distributions so that the acceptance rate for each parameter is about 50%. The samples from this preliminary run are discarded and not used in the analysis. We then run 25,000 iterations in three parallel chains, discarding the burn-in period of the first 5,000 iterations at which point the approximate convergence is achieved (the potential scale reduction factors of Gelman and Rubin 1992, are below 1.1 for all parameters). To reduce autocorrelation, we use every 20th iteration of each chain. This results in 1,000 simulation draws per chain and 3,000 simulated samples in total. Inference is then made using these simulated sample values.

The corresponding parameter estimates from these algorithms are exhibited in Table 2, where the numbers in parentheses are the estimated standard errors. Estimates from the penalized extended quasi-likelihood approach are also included for comparison, in which a constant $c = 0.001$ is added to the zeros when making inference of the index parameter. In all these models, the inner zone and the root stock M26 serve as the reference level and their estimates are fixed as zero. We see

	PEQL	Laplace	AGQ	MCEM	MCMC
β_0	-2.096(0.18)	-2.098(0.17)	-2.097(0.17)	-2.097(0.17)	-2.102(0.22)
β_1	-0.068(0.24)	-0.066(0.22)	-0.067(0.22)	-0.066(0.23)	-0.066(0.30)
β_2	-0.462(0.22)	-0.463(0.20)	-0.463(0.20)	-0.464(0.20)	-0.462(0.27)
β_3	-0.447(0.29)	-0.447(0.26)	-0.447(0.26)	-0.447(0.27)	-0.443(0.26)
β_4	0.028(0.36)	0.026(0.31)	0.026(0.31)	0.025(0.35)	0.013(0.32)
β_5	-1.168(0.36)	-1.166(0.32)	-1.166(0.32)	-1.166(0.34)	-1.168(0.32)
ϕ	0.599	0.329	0.329	0.329	0.338
p	1.554	1.413	1.413	1.413	1.418
σ_b	0.063	0.088	0.088	0.088	0.158

Table 2: Parameter estimates (β_0 - Intercept, β_1- MM106, β_2 - Mark, β_3 - Outer, β_4 - Outer:MM106, β_5 - Outer:Mark) for the root length density data using different estimation methods. Approximate standard errors for the fixed effects are reported in parentheses.

that the estimates across the three likelihood-based algorithms, i.e., Laplace, AGQ and MCEM, are highly consistent, with negligible difference within the first three decimal places. By comparison, the penalized extended quasi-likelihood approach produces noticeably different estimates except for the fixed effects. The discrepancy for the dispersion ϕ and the index p is expected given that their estimates depend on the value of the constant c. In addition, the MCMC estimates are comparable to the other estimates expect for the variance component, for which the estimate is almost twice as large. This difference is likely because the likelihood-based methods tend to underestimate the variance component when the number of groups is small (8 plants in the example), as is shown in the simulation study in the next section. An advantage of the Bayesian approach is that posterior distributions of all quantities are readily available. For example, Figure 4 shows the posterior distribution of the index parameter from the MCMC simulations. We see that the value of the index parameter mainly lies in the interval $(1.35, 1.5)$. Moreover, the parameter estimates suggest that the estimated RLD is much smaller in the outer zone (β_3) than that in the inner zone - the average RLD in the outer zone is about $\exp(-0.447) = 64\%$ of that in the inner zone, which is consistent with the exploratory results in Table 1. However, this effect is only marginally significant (outside one standard deviation but within two standard deviations from the origin) in the presence of the interaction terms. For the root stock effect, roots with Mark stock (β_2) tend to have significantly lower RLD, averaged at about 30% of that from the other two root stocks, while root stock MM106 (β_1) and M26 are not statistically distinguishable.

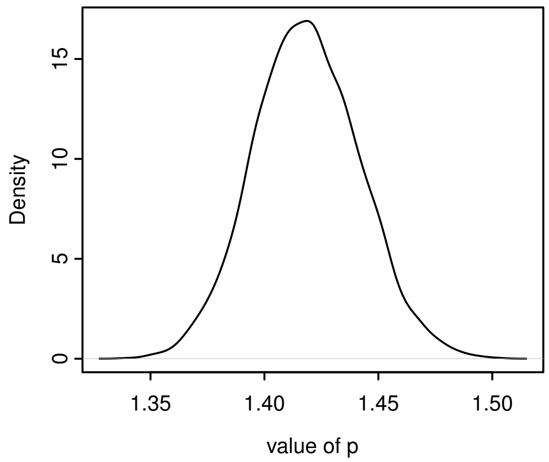

Figure 4: Posterior distribution of the index parameter from the MCMC approach.

5 Simulation studies

In this section, we conduct two sets of simulation studies to evaluate the statistical properties of the proposed algorithms for the compound Poisson mixed model. In the first study, we consider the Laplace and adaptive Gauss-Hermite quadrature methods, and compare them with the widely used penalized quasi-likelihood algorithm. The simulation is based on simulated covariates, and the aim is to investigate the performance of the proposed likelihood approximation estimators in different settings. The second study uses the fine root data set analyzed in the last section and includes the latent variable approach and the MCMC algorithm.

5.1 Simulation study I

It is understood that the performance of mixed model estimation procedures can depend on such factors as the magnitude of the variance component and the number of groups/clusters corresponding to the random effects. For this reason, we design and implement two pairs of simulations, each pair allowing only one of the variance component (σ_b) and the number of groups (g) to take contrasted values while controling all the other factors. This gives rise to four simulation scenarios in total as below:

1-a: $g = 20$, $\sigma_b = 0.1$, $p = 1.5$;

1-b: $g = 20$, $\sigma_b = 1$, $p = 1.5$;

2-a: $g = 5$, $\sigma_b = 1$, $p = 1.5$;

2-b: $g = 100$, $\sigma_b = 1$, $p = 1.5$;

The aim is to investigate not only the general performance of the likelihood approximation estimators, but also how and to what extent the performance may be influenced by the above factors. For example, comparing the results from 2-a and 2-b enables us to examine the effect of the number of groups on the variance component estimation.

In all simulations, the sample size of each simulated data set is set to $N = 500$, and for each simulation study, a covariate x simulated from a standard Normal distribution is included as the sole predictor in addition to the intercept. The true parameter values associated with the intercept and the covariate are set to $\boldsymbol{\beta} = (\beta_0, \beta_1)' = (-1, 1)'$, and the dispersion is set to $\phi = 1$. They are held constant across all simulations. To create a simulation data set, we simulate the random effects from a Normal distribution with standard deviation σ_b given in each of the above scenarios, compute the expected values of the response variable using (3.1) with a logarithmic link function, based on the simulated random effects and the given values of $\boldsymbol{\beta}$, and generate the response variable from the Tweedie compound Poisson distribution according to (2.1) and (2.2). For each simulation scenario, we create $S = 200$ data sets as above. For each data set, we fit a compound Poisson mixed model with x as the predictor using PQL, Laplace and AGQ, respectively. For the PQL approach, the true value of the index parameter in each scenario is used. The PEQL approach is not adopted mainly because its dependence upon the adjustment of the zeros could severely contaminate the comparison. For the AGQ algorithm, seven knots are specified, which proves to provide sufficient accuracy for most simulated data sets in the study. Running the simulation in each scenario results in $S = 200$ estimates of each parameter (say θ). We summarize the simulation result by the average value ($\hat{\theta} = \sum_{i=1}^{S} \theta_i / S$), the relative estimation bias ($(\hat{\theta} - \theta)/\theta$) and the mean square error ($\sum_{i=1}^{S} (\theta_i - \theta)^2 / S$). These statistics are reported in Table 3.

Several conclusions emerge from these reported statistics. First, all algorithms produce relatively little or no bias in the estimation of the fixed effects $\boldsymbol{\beta}$, where the bias is generally below 6%. In addition, the estimates of the dispersion ϕ and the index p are unbiased under the Laplace and AGQ methods, the biases across all scenarios being smaller than 2%. The conclusion for the variance component, however, appears to depend on several factors. Most notably, the result tends to strongly relate to the magnitude of the variance component: for scenario 1-a with a small variance component ($\sigma_b = 0.1$), the bias is as high as 20% for the Laplace and AGQ algorithms while this bias essentially disappears for $\sigma_b = 1$ in scenario 1-b ($\sigma_b = 1$). In addition, when there is a small number of groups ($g = 5$ in 2-a), the bias could be considerably large, more than 16% in most cases. On the other hand, increasing the number of groups to 20 (1-b) or 100 (2-b) leads to almost unbiased estimates.

Model	β_0	β_1	ϕ	p	σ_b	β_0	β_1	ϕ	p	σ_b
	Average estimate									
	1-a: $\sigma_b = 0.1$					1-b: $\sigma_b = 1$				
True	-1.000	1.000	1.000	1.500	0.100	-1.000	1.000	1.000	1.500	1.000
PQL	-1.003	1.002	0.985		0.080	-1.014	1.001	0.932		0.988
Laplace	-1.005	1.002	0.992	1.498	0.079	-1.027	1.001	0.992	1.498	0.993
AGQ7	-1.003	1.002	0.992	1.498	0.079	-1.012	1.001	0.993	1.498	0.987
	2-a: $g = 5$					2-b: $g = 100$				
True	-1.000	1.000	1.000	1.500	1.000	-1.000	1.000	1.000	1.500	1.000
PQL	-0.960	1.002	0.980		0.838	-1.026	0.999	0.788		1.034
Laplace	-0.963	1.002	0.997	1.498	0.839	-1.058	1.000	0.989	1.498	1.002
AGQ7	-0.960	1.002	0.997	1.498	0.838	-1.002	0.999	0.990	1.499	0.998
	Relative estimation bias (%)									
	1-a: $\sigma_b = 0.1$					1-b: $\sigma_b = 1$				
PQL	0.333	0.187	-1.529		-19.837	1.354	0.134	-6.758		-1.166
Laplace	0.514	0.201	-0.784	-0.133	-20.987	2.717	0.118	-0.767	-0.164	-0.724
AGQ7	0.315	0.199	-0.786	-0.133	-20.848	1.194	0.118	-0.732	-0.160	-1.298
	2-a: $g = 5$					2-b: $g = 100$				
PQL	-4.038	0.168	-2.035		-16.228	2.567	-0.057	-21.204		3.351
Laplace	-3.657	0.179	-0.329	-0.133	-16.096	5.830	-0.045	-1.070	-0.146	0.185
AGQ7	-4.045	0.179	-0.326	-0.132	-16.245	0.176	-0.139	-0.969	-0.096	-0.223
	Mean square error									
	1-a: $\sigma_b = 0.1$					1-b: $\sigma_b = 1$				
PQL	0.431	0.348	0.690		0.733	5.922	0.305	1.019		2.791
Laplace	0.434	0.347	0.442	0.041	0.706	6.022	0.306	0.562	0.046	2.822
AGQ7	0.428	0.347	0.442	0.041	0.708	5.891	0.306	0.562	0.046	2.778
	2-a: $g = 5$					2-b: $g = 100$				
PQL	19.446	0.259	0.818		12.945	1.722	0.375	4.831		1.995
Laplace	19.484	0.258	0.614	0.042	12.944	2.004	0.374	0.725	0.045	1.056
AGQ7	19.441	0.258	0.614	0.042	12.945	1.598	0.373	0.718	0.044	0.995

Table 3: Average values of parameter estimates, relative estimation biases (in 100s) and mean square errors (in 100s) under different estimation algorithms in each of the four simulation scenarios.

We also note that using the true value of p, the PQL method produces average estimates that are consistent with those from the likelihood-based methods. However, the estimates of the variance component and the dispersion in PQL are worse than those from the likelihood-based methods, as measured by the mean square errors. In some scenario (e.g., 2-b), the mean square errors of PQL are twice as large as those of the likelihood-based methods. The difference would have been even larger if we had allowed the index to be estimated from the data. As for the fixed effects, PQL is inferior to the quadrature method while slightly more accurate than the Laplace approximation. Comparison of the two likelihood-based methods reveals that the AGQ approach outperforms the Laplace approximation in estimating the fixed effects, reducing both the bias and the mean square error of the Laplace approximation almost across all scenarios. Such improvement is most noticeable when the number of groups and the variance component are large (2-b), in which the 7-knot AGQ reduces about 25% of the mean square error of the Laplace approximation. Nevertheless, it is not guaranteed that the AGQ method yields more accurate random component estimation. The two likelihood-based methods tend to have comparable performance in most scenarios, with the Laplace approximation producing somewhat smaller mean square errors of the variance component in two scenarios of the simulation study. In addition, the two methods perform equally well in estimating the dispersion and the index parameters, resulting in almost identical mean square errors in all scenarios. Lastly, we observe that for such application as 2-a where precise estimation of parameters is difficult due to insufficient between-group information, all three methods tend to have similar performance.

5.2 Simulation study II

In the second simulation study, we take the design matrix from the fine root length density example, and investigate the performance of computationally demanding algorithms such as the MCEM method, which is based on latent variables, and the MCMC method, which is formulated in the Bayesian setting. We set the true values of the fixed effects ($\boldsymbol{\beta}$) to be the maximum likelihood estimates from Table 2, and $(\phi, p, \sigma_b)' = (0.3, 1.5, 1)'$. We simulate 100 data sets, and in each data set the response variable is generated in the same fashion as described in the first simulation study. For each simulated data set, we fit the compound Poisson mixed model using three algorithms: AGQ, MCEM and MCMC. We use the quadrature method as a reference model. The MCEM algorithm is run similarly as in the example section: we use 100 samples for the first 40 iterations and 3,000 samples thereafter. For the MCMC method, we use a preliminary run of 2,000 iterations to tune the proposal variances in the random walk Metropolis algorithm, after which we run another 11,000 iterations in one chain, burin in the first 1,000 iterations and draw one sample every 10th iteration. The posterior sample medians are then used to make inference of the parameters.

We summarize the simulation result by the average value, the relative estimation bias, and the mean square error as before, and report these statistics in Table 4. We first notice that there are large relative biases for some of the fixed effects. However, this is mainly because the true values of

Model	β_0	β_1	β_2	β_3	β_4	β_5	ϕ	p	σ_b
	Average estimate								
True	-2.100	-0.070	-0.460	-0.450	0.030	-1.170	0.300	1.500	1.000
AGQ7	-2.160	-0.058	-0.368	-0.403	-0.020	-1.209	0.295	1.495	0.745
MCEM	-2.165	-0.056	-0.368	-0.403	-0.020	-1.209	0.295	1.495	0.746
MCMC	-2.130	-0.116	-0.429	-0.408	-0.014	-1.204	0.302	1.499	1.051
	Relative estimation bias (%)								
AGQ7	2.875	-16.962	-20.014	-10.425	-165.596	3.315	-1.559	-0.301	-25.469
MCEM	3.101	-20.440	-19.894	-10.425	-165.627	3.319	-1.566	-0.302	-25.439
MCMC	1.422	65.891	-6.825	-9.325	-148.087	2.905	0.689	-0.061	5.142
	Mean square error								
AGQ7	45.648	98.131	54.648	4.714	6.737	6.790	0.074	0.037	14.658
MCEM	45.809	98.470	54.912	4.715	6.740	6.793	0.074	0.037	14.641
MCMC	67.822	161.714	73.568	4.760	6.817	6.843	0.078	0.036	19.867

Table 4: Average values of parameter estimates, relative estimation biases (in 100s) and mean square errors (in 100s) under different estimation algorithms in the second simulation study.

these fixed effects are fairly small. As a second measure, we also compute the absolute biases for the fixed effects, the maximum of which is indeed less than 0.1. Additionally, the maximum likelihood estimates of the variance component are downwardly biased (25%) because there are only a small number of groups corresponding to the random effects. In contrast, the MCMC method appears not to suffer much from this and generates a markedly accurate estimate of the variance component, the bias being about 5%. This result is in line with that of Browne and Draper (2006), which finds that the Bayesian estimate of the variance component is approximately unbiased in Normal hierarchical models. Indeed, all parameter estimates from the MCMC method are less biased than those from the likelihood-based methods, except for the fixed effect β_1. Nevertheless, judged by the mean square errors, the Bayesian estimates are not necessarily better. This is perhaps due to the Monte Carlo errors inherent in the MCMC procedure. For example, Figure 5 shows the distribution of the estimates of the parameter σ_b from the quadrature and MCMC methods. We see that a large portion of the AGQ estimates are below the true value. In contrast, the distribution of the MCMC estimates is centered around the true value, but more spread out and thick-tailed.

6 Discussion and conclusion

The intractable density function and the unknown variance function have presented considerable challenges to statistical inference problems involving compound Poisson mixed models. To date, the focus has been on estimation methods within the quasi-likelihood framework. While fast and easy to

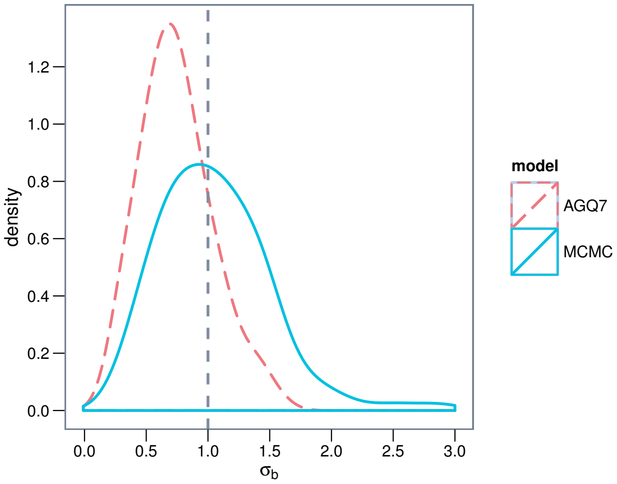

Figure 5: The distribution of the estimated variance component under the AGQ and MCMC algorithms in the simulation study.

implement, the penalized quasi-likelihood method is not equipped to estimate the variance function. This can have a considerable impact on hypothesis tests and parameter uncertainty measures in the compound Poisson mixed models. The natural modification utilizing the extended quasi-likelihood, however, cannot yield accurate and robust estimation of the variance function owing to the ad hoc adjustment of observed zeros in order to make the method feasible.

In contrast with the quasi-likelihood-based methods, this paper has presented several likelihood-based inferential algorithms that enable estimation of the variance function. These methods can be further categorized into two groups: the likelihood approximation method and the latent variable approach. Implementing the likelihood approximation methods, such as the Laplace approximation and the adaptive Gauss-Hermite quadrature, relies on the capability to numerically evaluate the compound Poisson density. In contrast, the latent variable approach avoids direct density evaluation. Rather, maximum likelihood estimation is carried out via the Monte Carlo EM algorithm. However, since the E-step in the EM algorithm involves simulating latent variables for Monte Carlo approximations, the algorithm can be computationally intensive. Implementation of importance sampling can reduce this computational cost by retaining old samples for use in later iterations. A related point is the Monte Carlo error inherent in the simulation-based method. Failing to account for this could cause the algorithm to stop prematurely and has the risk of making incorrect inference. This risk can be controlled by gauging the Monte Carlo error as in Booth and Hobert (1999) and Levine and Casella (2001) and increasing the number of Monte Carlo samples as the algorithm converges. Further, the Monte Carlo EM algorithm has a close relationship with Markov Chain Monte Carlo methods, and can be turned into a Gibbs sampler with minor modifications.

Compared to the likelihood approximation methods, the Monte Carlo EM and Markov Chain Monte Carlo methods can accommodate more general random effect structures or random effects from non-Normal distributions. Bayesian formulation and estimation of the mixed models also have the advantage to account for all sources of variability and produce posterior predictive distributions that are of considerable utility for formal decision making problems. However, these computationally demanding algorithms are less suited to large-scaled data problems, compared to the likelihood approximation methods. The Laplace approximation is markedly fast, reasonably accurate and widely applicable for a range of model structures. The quadrature method, on the other hand, relies heavily on the compound Poisson density evaluation and is considerably slower. It will become exceedingly slow when there are multiple random effects and lots of quadrature knots, because of the need for a large number of density evaluations.

We have also demonstrated the use of these algorithms through a numerical example, and conducted an array of simulation studies to evaluate their performance. We have found that the likelihood approximation methods are essentially unbiased for the fixed effects, the dispersion parameter and the index parameter, and perform substantially better than the penalized quasi-likelihood method in estimating the variance component. However, they could still produce downwardly biased estimation of the variance component when there is a small number of groups or the variance component is small. In these situations, the MCMC estimates based on posterior medians are generally less biased.

In situations where there is potentially large bias in the maximum likelihood estimation of the variance component, it is necessary to modify the estimation procedure by including bias-correction adjustments (e.g., Liao and Lipsitz 2002). For normal linear mixed models, such bias can be effectively corrected using the restricted maximum likelihood [REML] (Patterson and Thompson 1971). Extensions of the REML-type estimation to non-normal models generally fall in the framework of bias correction for profile estimating equations (McCullagh and Tibshirani 1990, Jørgenson and Knudsen 2004). Notably, Liao and Lipsitz (2002) derive the bias-corrected profile score equation for the variance component in generalized linear mixed models. Their bias correction procedure also relies on the MCEM algorithm, and when applied to the compound Poisson mixed models, it amounts to adjusting the expectation in (3.17) by a Monte Carlo estimate of the bias-correction term.

A final note is that the likelihood approximation methods presented here, namely, the Laplace approximation and the adaptive Guass-Hermite quadrature methods, can be readily modified to work for the exponential dispersion model $V(\mu) = \mu^p$ with $p > 2$. In this case, we evaluate the conditional likelihood $p(\boldsymbol{y}|\boldsymbol{u})$ using the density approximation methods derived for $p > 2$ in Dunn and Smyth (2005, 2008) and set the constraint in the numerical optimization to be $p > 2$.

7 References

1. Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory.* Chichester: Wiley.
2. Bates D., Mächler M. and Bolker B. (2012). Fitting linear mixed-effects models using lme4, *Journal of Statistical Software*, *forthcoming.*
3. Booth, J. G., and Hobert, J. P. (1999). Maximizing Generalized Linear Mixed Model Likelihoods with an Automated Monte Carlo EM Algorithm. *Journal of the Royal Statistical Society Series B*, 61, 265-285.
4. Breslow, N.E. and Clayton, D.G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 88 (421): 9-25.
5. Browne, W. J., and Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1(3), 473-514.
6. Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society Series B*, 49, 1-39.
7. Davidian, M. and Carroll, R. J. (1987). Variance function estimation. *Journal of the American Statistical Association*, 82, 1079-1091.
8. de Silva, H. N., Hall, A. J., Tustin, D. S. and Gandar, P. W. (1999). Analysis of distribution of root length density of apple trees on different dwarfing rootstocks. *Annals of Botany*, 83: 335-345.
9. Dunn, P.K. and Smyth, G.K. (2005). Series evaluation of Tweedie exponential dispersion models densities. *Statistics and Computing*, 15, 267-280.
10. Dunn, P.K. and Smyth, G.K. (2008). Evaluation of Tweedie exponential dispersion model densities by Fourier inversion. *Statistics and Computing*, 18, 73-86.
11. Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* 7, 457-511.
12. Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis.* London: Chapman and Hall.
13. Jørgensen, B. (1987). Exponential dispersion models (with discussion). *Journal of the Royal Statistical Society Series B*, 49: 127-162.
14. Jørgensen, B. and Knudsen S. J. (2004). Parameter Orthogonality and Bias Adjustment for Estimating Functions. *Scandinavian Journal of Statistics*, 31, 93-114.
15. Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963-974.
16. Levine R. A. and Casella G. (2001). Implementations of the Monte Carlo EM Algorithm. *Journal of Computational and Graphical Statistics*, 10:3, 422-439.
17. Liao J. G. and Lipsitz S. R. (2002). A type of restricted maximum likelihood estimator of variance components in generalised linear mixed models. *Biometrika*, 89, 2, 401-409.

18. Liu Q. and Pierce D.A. (1994) A Note on Gauss-Hermite Quadrature, *Biometrika*, 81, 3, 624-629.
19. McCullagh, P. and Nelder, J.(1989). *Generalized Linear Models*, Boca Raton: Chapman and Hall.
20. McCullagh, P. and Tibshirani, R. (1990). A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society Series B*, 52, 325-344
21. McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92, 162-170.
22. McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear and Mixed Models.* John Wiley & Sons.
23. Meng, X-L. and Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80 (2): 267-278.
24. Nelder, J.A. and Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, 74: 221-232.
25. Patterson, H. D. and Thompson R. N. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58, 545-554.
26. Peters G. W., Shevchenko P. V. and Wuthrich M. V. (2009). Model Uncertainty in Claims Reserving within Tweedie's Compound Poisson Models. *ASTIN Bulletin*, 39(1), 1-33.
27. Pinheiro, J. C. and Chao, E. C. (2006). Efficient Laplacian and Adaptive Gaussian Quadrature Algorithms for Multilevel Generalized Linear Mixed Models. *Journal of Computational and Graphical Statistics*, 15, 1, 58-81.
28. Robert, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods.* New York: Springer-Verlag.
29. Smyth, G.K. (1996). Regression analysis of quantity data with exact zeros. Proceedings of the Second Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management. Technology Management Centre, University of Queensland, pp. 572-580.
30. Tierney, L. and Kadane J. B. (1986), Accurate Approximation for Posterior Moments and Marginal Densities, *Journal of the American Statistical Association*, 81, 82-86.
31. Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, 61, 439-47.
32. Wei, G. C. G., and Tanner, M. A. (1990). A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85, 699-704.
33. Zeger, S. L. and Karim R. M. (1991). Generalized Linear Models With Random Effects; A Gibbs Sampling Approach. *Journal of the American Statistical Association*, 86, 413, 79-86.