Statistical Modelling 2005; **5**: 21–37

Two-component mixtures of generalized linear mixed effects models for cluster correlated data

Daniel B Hall and Lihua Wang
Department of Statistics, University of Georgia, Athens, Georgia, USA

Abstract: Finite mixtures of generalized linear mixed effect models are presented to handle situations where within-cluster correlation and heterogeneity (subpopulations) exist simultaneously. For this class of model, we consider maximum likelihood (ML) as our main approach to estimation. Owing to the complexity of the marginal loglikelihood of this model, the EM algorithm is employed to facilitate computation. The major obstacle in this procedure is to integrate over the random effects' distribution to evaluate the expectation in the E step. When assuming normally distributed random effects, we consider adaptive Gaussian quadrature to perform this integration numerically. We also discuss nonparametric ML estimation under a relaxation of the normality assumption on the random effects. Two real data sets are analysed to compare our proposed model with other existing models and illustrate our estimation methods.

Key words: adaptive Gaussian quadrature; EM algorithm; finite mixture of distributions; random effects

Data and software link available from: http://stat.uibk.ac.at/SMIJ
Received March 2004; **revised** October 2004; **accepted** November 2004

1 Introduction

Finite mixture models with regression structure have a long and extensive literature and have been commonly used in fields such as epidemiology, medicine, genetics, economics, engineering, marketing and in the physical and social sciences. Much of this work has focused on mixtures of normal distributions (e.g., McLachlan and Basford, 1988), but non-normal mixtures have received attention as well in the recent literature. Some of this work has included regression structure in the linear predictor only (Deb and Trivedi, 1997; Dietz and Böhning, 1997; Jansen, 1993), whereas other authors have considered covariates in both the linear predictor and the mixing probability (Thompson *et al.*, 1998; Wang and Puterman, 1998). Of course, models without covariates occur as a special case, and such models have been considered by Titterington *et al.* (1985) and Lindsay (1995), among others. A special case of the two component mixture occurs when one component is a degenerate distribution with point mass of one at zero. Such models are known as zero inflated regression models and include zero inflated Poisson (ZIP); (Lambert, 1992), zero

Address for correspondence: Daniel B Hall, Department of Statistics, University of Georgia, Athens, Georgia, 30602-1952, USA. E-mail: dhall@stat.uga.edu

 10.1191/1471082X05st090oa

inflated negative binomial, zero inflated binomial (ZIB); (Hall, 2000) and others (reviewed in Ridout *et al.*, 1998).

Recently, many researchers have incorporated random effects into a wide variety of regression models to account for correlated response and multiple sources of variance. Generalized linear models (GLMs) with fixed and random (mixed) effects and normal theory nonlinear mixed effects models are two model classes that have attracted an enormous amount of attention in recent years (Davidian and Giltinan, 1995; 2003; McCulloch and Searle, 2001; e.g., for recent reviews). A recent example falling slightly outside the class of generalized linear mixed models (GLMMs) was provided by Booth *et al.* (2003) who considered loglinear mixed models for counts based on the negative binomial distribution. In a mixture model context, van Duijn and Bockenholt (1995) presented a latent class Poisson model for analysing overdispersed repeated count data. Hall (2000) added random effects to ZIP and ZIB models. Zero inflated regression models with mixed effects for clustered continuous data have been considered by Olsen and Schafer (2001) and Berk and Lachenbruch (2002).

In this paper, we formulate a class of regression models based on the two component mixture of generalized linear mixed effect models (two-component GLMMs). This class can be viewed as an extension of finite mixtures of GLMs (Jansen, 1993) in which cluster specific random effects are included to account for within cluster correlation. Alternatively, it can be viewed as an extension of GLMMs in which a second component is added. We envision that finite mixtures of GLMMs will have application primarily in problems where there is some readily identified heterogeneity in the population so that the data represent a small number of subpopulations that cannot be directly identified except through the value of the response. We focus on the case in which it is reasonable to hypothesize two latent subpopulations underlying the data. For example, disease counts from epidemic and non epidemic years, weekly epileptic seizure counts from patients who have 'good weeks' and 'bad weeks', arrhythmia counts from a sample of clinically normal patients that is contaminated with abnormal patients, counts from honest and dishonest self-reports pertaining to some stigmatized act, counts from adherers and non adherers to some study protocol, and so on. GLMs, finite mixtures of GLMs, ZIP, ZIB and many other models are special cases of this broad class.

The difficulty of parameter estimation in mixture models is well known. A major advance came with the publication of the seminal paper of Dempster *et al.* (1977) on the EM algorithm. With the EM algorithm, latent variables or 'missing data' are introduced, which allows finite mixture models to be fit by iteratively fitting weighted versions of the component models. So, for example, a K component finite mixture of GLMs can be fit via maximum likelihood (ML) by fitting K weighted GLMs, updating the weights and iterating to convergence. Mixture models with random effects pose an additional challenge to ML estimation as the marginal likelihood involves an integral that cannot be evaluated in closed form. This challenge is similar to that found with ordinary (nonmixture) GLMMs and other nonlinear mixed models.

The paper is organized as follows: we formulate the two component mixture of GLMMs in Section 2. In Section 3, we outline the EM algorithm and consider various methods of handling the required integration with respect to the missing data. The model class and estimation methods are illustrated with two real data examples in Section 4. Finally, we give a brief discussion in Section 5.

2 Two-component mixture of GLMMs

Suppose, we observe an N-dimensional response vector $\mathbf{y}$ containing data from C independent clusters, so that $\mathbf{y} = (\mathbf{y}_1^{\mathrm{T}}, \ldots, \mathbf{y}_C^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^{\mathrm{T}}$. We assume that, conditional on a q-dimensional vector of random effects $\mathbf{b}_i$, the random variable Y_{ij} associated with observation y_{ij} follows a two-component mixture distribution

$$Y_{ij}|\mathbf{b}_i \sim \begin{cases} F_1(y_{ij}|\mathbf{b}_i; \zeta_{1ij}, \sigma_1), & \text{with probability } p_{ij} \\ F_2(y_{ij}|\mathbf{b}_i; \zeta_{2ij}, \sigma_2), & \text{with probability } 1 - p_{ij} \end{cases}$$

Here, F_1 and F_2 are assumed to be exponential dispersion family distributions, with densities $f_k(y_{ij}|\mathbf{b}_i; \zeta_{kij}, \sigma_k) = h_k(y_{ij}, \sigma_k) \exp[\{\zeta_{kij} y_{ij} - \kappa_k(\zeta_{kij})\} w_{ij}/\sigma_k]$, $k = 1, 2$, respectively, where the w_{ij}'s are known constants (e.g., binomial denominators). The functions κ_1 and κ_2 are cumulant generating functions, so F_1 and F_2 have (conditional) means $\mu_{1ij} = \kappa_1'(\zeta_{1ij})$ and $\mu_{2ij} = \kappa_2'(\zeta_{2ij})$.

We assume the canonical parameters $\boldsymbol{\zeta}_{ki} = (\zeta_{ki1}, \ldots, \zeta_{kin_i})^{\mathrm{T}}$, $k = 1, 2$, are related to covariates and cluster specific random effects through GLM type specifications. Specifically, for canonical link functions, we assume

$$\begin{aligned} \boldsymbol{\zeta}_{1i}(\boldsymbol{\mu}_{1i}) = \boldsymbol{\eta}_{1i} &= \mathbf{X}_i \boldsymbol{\alpha} + \mathbf{U}_{1i} \mathbf{D}_1^{\mathrm{T}/2} \mathbf{b}_{1i} \quad \text{or} \quad \boldsymbol{\mu}_{1i} = \boldsymbol{\zeta}_{1i}^{-1} = (\boldsymbol{\eta}_{1i}) \\ \boldsymbol{\zeta}_{2i}(\boldsymbol{\mu}_{2i}) = \boldsymbol{\eta}_{2i} &= \mathbf{Z}_i \boldsymbol{\beta} + \mathbf{U}_{2i} \mathbf{D}_2^{\mathrm{T}/2} \mathbf{b}_{2i} + \mathbf{U}_{3i} \mathbf{D}_3^{\mathrm{T}/2} \mathbf{b}_{1i} \quad \text{or} \quad \boldsymbol{\mu}_{2i} = \boldsymbol{\zeta}_{2i}^{-1}(\boldsymbol{\eta}_{2i}) \end{aligned} \tag{2.1}$$

Here, $\mathbf{X}_i$ and $\mathbf{Z}_i$ are $n_i \times r_1$ and $n_i \times r_2$ design matrices, respectively, for fixed effects parameters $\boldsymbol{\alpha}$ and β; $\mathbf{b}_i = (\mathbf{b}_{1i}^{\mathrm{T}}, \mathbf{b}_{2i}^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{b}_{1i}$ and $\mathbf{b}_{2i}$ are of dimension q_1 and q_2, respectively, where $q = q_1 + q_2$; and $\mathbf{U}_{ki}$, $k = 1, 2, 3$, are random effects design matrices. In some applications, we may drop the term $\mathbf{U}_{3i} \mathbf{D}_3^{\mathrm{T}/2} \mathbf{b}_{1i}$ from the model, but in general it is present to allow covariance between the two linear predictors. We assume $\mathbf{b}_1, \ldots, \mathbf{b}_C$ are independent, identically distributed $N_q(\mathbf{0}, \mathbf{I}_q)$ vectors, and model covariance through the shared random effects $\mathbf{b}_{1i}$ in the two linear predictors and through the lower triangular scale matrices $\mathbf{D}_k^{\mathrm{T}/2}$, $k = 1, 2, 3$. That is, model (2.1) implies

$$\begin{aligned} \mathrm{var}(\boldsymbol{\eta}_{1i}) &= \mathbf{U}_{1i} \mathbf{D}_1 \mathbf{U}_{1i}^{\mathrm{T}} \\ \mathrm{var}(\boldsymbol{\eta}_{2i}) &= \mathbf{U}_{2i} \mathbf{D}_2 \mathbf{U}_{2i}^{\mathrm{T}} + \mathbf{U}_{3i} \mathbf{D}_3 \mathbf{U}_{3i}^{\mathrm{T}} \\ \mathrm{cov}(\boldsymbol{\eta}_{1i}, \boldsymbol{\eta}_{2i}^{\mathrm{T}}) &= \mathbf{U}_{1i} \mathbf{D}_1^{\mathrm{T}/2} \mathbf{D}_3^{1/2} \mathbf{U}_{3i}^{\mathrm{T}} \end{aligned}$$

Here, $\mathbf{D}_k$ contains variance components along the diagonal and covariance components on the off diagonal. We assume that each $\mathbf{D}_k$ is parameterized by a vector $\boldsymbol{\theta}_k$, where we adopt an unconstrained Cholesky parameterization (Pinheiro and Bates, 1996) in which the elements of $\boldsymbol{\theta}_k$ are the nonzero entries in the upper triangular Cholesky factor $\mathbf{D}_k^{1/2}$. That is, $\boldsymbol{\theta}_k = \mathrm{vech}(\mathbf{D}_k^{1/2})$, $k = 1, 2, 3$, where vech stacks the columns of its matrix argument including only those elements on and above the diagonal. (Note that our definition of vech differs from the usual usage in which the elements on and below the diagonal are stacked.)

The model form given by Equation (2.1) is quite flexible, allowing a wide variety of random effects specifications. For example, a model with random cluster specific intercepts might assume $\eta_{1ij} = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\alpha} + \theta_1 b_{1i}$ and $\eta_{2ij} = \mathbf{z}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \theta_2 b_{2i}$. This implies independent cluster effects in the two components. Correlated components can be induced by assuming $\eta_{1ij} = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\alpha} + \theta_1 b_{1i}$ and $\eta_{2ij} = \mathbf{z}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \theta_2 b_{2i} + \theta_3 b_{1i}$, which leads to $\mathrm{corr}(\eta_{1ij}, \eta_{2iij}) = \theta_3/\sqrt{\theta_2^2 + \theta_3^2}$. The form given by Equation (2.1) simply generalizes these cases to higher dimension, allowing random slope and intercept models and other more general random effects structures. An alternative approach would have been to allow correlated random effects $\mathbf{b}_{1i}, \mathbf{b}_{2i}$, say, where $\mathbf{b}_{ki}$ appears only in the kth linear predictor and $\mathrm{cov}(\mathbf{b}_{1i}, \mathbf{b}_{2i}) \neq \mathbf{0}$. However, this more straight forward approach, which is essentially a reparametrization of the model we focus on, is not as conducive to estimation via the EM algorithm because it leads to a complete data likelihood, which does not factor cleanly into terms for each component in the mixture. [In particular, the second and third terms of formula (3.3) defined subsequently, which is the expected complete data loglikelihood used in the EM algorithm, would share parameters pertaining to $\mathrm{corr}(\tilde{\mathbf{b}}_{1i}, \tilde{\mathbf{b}}_{2i})$.]

Note that in Equation (2.1) we have assumed canonical links, but this is not necessary. In general, we allow known links g_1 and g_2 so that $\mu_{1ij} = g_1^{-1}(\eta_{1ij})$ and $\mu_{2ij} = g_2^{-1}(\eta_{2ij})$. Furthermore, we assume that the mixing mechanisms for each observation are independent, with probabilities $\mathbf{p}_i = (p_{i1}, \ldots, p_{in_i})^{\mathrm{T}}$, $i = 1, \ldots, C$, each following a regression model of the form $g_p(\mathbf{p}_i) = \mathbf{W}_i\boldsymbol{\gamma}$, involving a known link function g_p, unknown regression parameter $\boldsymbol{\gamma}$ and $n_i \times s$ design matrix $\mathbf{W}_i$. Typically, g_p will be taken to be the logit link, but the probit, complementary log–log, or other link function can be chosen here.

Let $\tilde{\boldsymbol{\alpha}} = (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\theta}_1^{\mathrm{T}})^{\mathrm{T}}$ and $\tilde{\boldsymbol{\beta}} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\theta}_2^{\mathrm{T}}, \boldsymbol{\theta}_3^{\mathrm{T}})^{\mathrm{T}}$, and denote the combined vector of model parameters as $\boldsymbol{\delta} = (\tilde{\boldsymbol{\alpha}}^{\mathrm{T}}, \tilde{\boldsymbol{\beta}}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}}, \sigma_1, \sigma_2)^{\mathrm{T}}$. The loglikelihood for $\boldsymbol{\delta}$ based on $\mathbf{y}$ is given by

$$\ell(\boldsymbol{\delta}; \mathbf{y}) = \sum_{i=1}^{C} \log\left\{ \int \prod_{j=1}^{n_i} f(y_{ij}|\mathbf{b}_i; \boldsymbol{\delta}) \phi_q(\mathbf{b}_i) \mathrm{d}\mathbf{b}_i \right\}$$

where $f(y_{ij}|\mathbf{b}_i; \boldsymbol{\delta}) = \{p_{ij}(\boldsymbol{\gamma})\} f_1(y_{ij}|\mathbf{b}_i; \tilde{\boldsymbol{\alpha}}, \sigma_1) - \{1 - p_{ij}(\boldsymbol{\gamma})\} f_2(y_{ij}|\mathbf{b}_i; \tilde{\boldsymbol{\beta}}, \sigma_2)$, $\phi_q(\cdot)$ denotes the q-dimensional standard normal density function, and the integral is q-dimensional.

3 Fitting the two-component mixture model via the EM algorithm

The complications of parameter estimation in mixture models are simplified considerably by applying the EM algorithm. Let u_{ij}, $i = 1, \ldots, C$, $j = 1, \ldots, n_i$ denote the component membership; u_{ij} equals one if Y_{ij} is drawn from distribution F_1 and equals zero if Y_{ij} is drawn from F_2. Then the 'complete' data for the EM algorithm are $(\mathbf{y}, \mathbf{u}, \mathbf{b})$. Here, $(\mathbf{u}, \mathbf{b})$ play the role of missing data, where $\mathbf{u} = (u_{11}, \cdots, u_{Cn_C})^{\mathrm{T}}$. On the basis of

the complete data ($\mathbf{y}$, $\mathbf{u}$, $\underset{\sim}{\mathbf{b}}$), the loglikelihood is given by $\log f(\mathbf{b}) + \log f(\mathbf{u}|\mathbf{b}; \gamma) + \log f(\mathbf{y}|\mathbf{u}, \mathbf{b}; \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \sigma_1, \sigma_2)$, which has kernel

$$\sum_{i=1}^{C}\sum_{j=1}^{n_i}[u_{ij}\log p_{ij}(\gamma) + (1-u_{ij})\log\{1-p_{ij}(\gamma)\}] + \sum_{i=1}^{C}\sum_{j=1}^{n_i} u_{ij}\log f_1(y_{ij}|\mathbf{b}_i; \zeta_{1ij}, \sigma_1)$$
$$+ \sum_{i=1}^{C}\sum_{j=1}^{n_i}(1-u_{ij})\log f_2(y_{ij}|\mathbf{b}_i; \zeta_{2ij}, \sigma_2) \equiv \sum_{i=1}^{C}\sum_{j=1}^{n_i}\ell^c(\boldsymbol{\delta}; y_{ij}, u_{ij}|\mathbf{b}_i)$$

where $\mathbf{b} = (\mathbf{b}_1, \ldots, \mathbf{b}_C)^{\mathrm{T}}$. On the basis of this complete data loglikelihood, the EM algorithm is applied both to ML estimation in Section 3.1 and to nonparametric ML estimation (NPML) in Section 3.2.

3.1 ML estimation for normal random effects

Given a starting value for the parameter vector $\boldsymbol{\delta}$, the EM algorithm yields ML estimates by alternating between an expectation step and a maximization step. At convergence, computation of an observed information matrix for parameter standard errors and Wald-type statistical inference is based upon Oakes formula (Oakes, 1999). Refer to Wang (2004) for details.

3.1.1 E step

In the $(h+1)$th iteration of EM algorithm, we compute $Q(\boldsymbol{\delta}|\boldsymbol{\delta}^{(h)}) = E\{\log f(\mathbf{y}, \mathbf{u}|\mathbf{b}; \boldsymbol{\delta})|\mathbf{y}, \boldsymbol{\delta}^{(h)}\}$ in the E step, where the expectation is with respect to the joint distribution of $\mathbf{u}$, $\mathbf{b}$ given $\mathbf{y}$ and $\boldsymbol{\delta}^{(h)}$. This conditional expectation can be taken in two stages where the inner expectation is with respect to $\mathbf{u}$ only. As $\log f(\mathbf{y}, \mathbf{u}|\mathbf{b}; \boldsymbol{\delta})$ is linear with respect to $\mathbf{u}$, this inner expectation can be taken simply by substituting $\mathbf{u}^{(h)} = E(\mathbf{u}|\mathbf{y}, \mathbf{b}, \boldsymbol{\delta}^{(h)})$ for $\mathbf{u}$. The vector $\mathbf{u}^{(h)}$ is easily computed, with elements

$$u_{ij}^{(h)}(\mathbf{b}_i) = \left[1 + \frac{1-p_{ij}(\gamma^{(h)})}{p_{ij}(\gamma^{(h)})}\frac{f_2\{y_{ij}|\mathbf{b}_i; \tilde{\boldsymbol{\beta}}^{(h)}, \sigma_2^{(h)}\}}{f_1\{y_{ij}|\mathbf{b}_i; \tilde{\boldsymbol{\alpha}}^{(h)}, \sigma_1^{(h)}\}}\right]^{-1} \tag{3.1}$$

Here, the superscript (h) indicates evaluation at the value obtained in the hth step of the algorithm. Note that $\mathbf{u}^{(h)}$ is a function of $\mathbf{b}_i$, so we have indicated that dependence in the notation $u_{ij}^{(h)}(\mathbf{b}_i)$. Taking the outer expectation and dropping terms not involving $\boldsymbol{\delta}$, we obtain

$$Q(\boldsymbol{\delta}|\boldsymbol{\delta}^{(h)}) = \sum_{i=1}^{C}\frac{\sum_{j=1}^{n_i}\int \ell^c(\boldsymbol{\delta}; y_{ij}, u_{ij}^{(h)}(\mathbf{b}_i)|\mathbf{b}_i) f(\mathbf{y}_i|\mathbf{b}_i; \boldsymbol{\delta}^{(h)})\phi_q(\mathbf{b}_i)\mathrm{d}\mathbf{b}_i}{\int f(\mathbf{y}_i|\mathbf{b}_i; \boldsymbol{\delta}^{(h)})\phi_q(\mathbf{b}_i)\mathrm{d}\mathbf{b}_i} \tag{3.2}$$

The integrals in Equation (3.2) are now with respect to the random effects b only, which must be 'integrated out' of Q. We propose performing this integration via

adaptive Gaussian quadrature (AGQ) (Liu and Pierce, 1994; Pinheiro and Bates, 1995). Let $\hat{\mathbf{b}}_i^1$ and $\hat{\mathbf{b}}_i^2$ denote the modes of the integrands in the numerator and denominator, respectively, of Equation (3.2), and let $g_1(\mathbf{b}_i) \equiv \sum_{j=1}^{n_i} \ell^c(\boldsymbol{\delta}; y_{ij}, u_{ij}^{(h)}(\mathbf{b}_i)|\mathbf{b}_i) f(\mathbf{y}_i|\mathbf{b}_i; \boldsymbol{\delta}^{(h)}) \phi_q(\mathbf{b}_i)$ and $g_2(\mathbf{b}_i) \equiv f(\mathbf{y}_i|\mathbf{b}_i; \boldsymbol{\delta}^{(h)}) \phi_q(\mathbf{b}_i)$ from equation (3.2). In addition, let $\hat{\boldsymbol{\Gamma}}_{1i}$ and $\hat{\boldsymbol{\Gamma}}_{2i}$ be the Hessian matrices of $\log\, g_1(\mathbf{b}_i)$ and $\log\, g_2(\mathbf{b}_i)$ evaluated at $\hat{\mathbf{b}}_i^1$ and $\hat{\mathbf{b}}_i^2$, and let $\boldsymbol{\pi}_{\ell_1,\ldots,\ell_q} = (\pi_{\ell_1}, \ldots, \pi_{\ell_q})^{\mathrm{T}}$ and $\mathbf{z}_{\ell_1,\ldots,\ell_q} = (z_{\ell_1}, \ldots, z_{\ell_q})^{\mathrm{T}}$, where $\pi_1, \ldots, \pi_m$ and $z_1, \ldots, z_m$ are m-point ordinary Gaussian quadrature (OGQ) weights and abscissas, respectively. Then the quadrature points under AGQ are shifted and rescaled versions of $\mathbf{z}_{\ell_1,\ldots,\ell_q}$ as follows: $\mathbf{b}_{i\ell_1,\ldots,\ell_q}^{1*} = (b_{i\ell_1}^{1*}, \ldots, b_{i\ell_q}^{1*})^{\mathrm{T}} = \hat{\mathbf{b}}_i^1 + 2^{q/2}\hat{\boldsymbol{\Gamma}}_{1i}^{-1/2}\mathbf{z}_{\ell_1,\ldots,\ell_q}$ and $\mathbf{b}_{i\ell_1,\ldots,\ell_q}^{2*} = (b_{i\ell_1}^{2*}, \ldots, b_{i\ell_q}^{2*})^{\mathrm{T}} = \hat{\mathbf{b}}_i^2 + 2^{q/2}\hat{\boldsymbol{\Gamma}}_{2i}^{-1/2}\mathbf{z}_{\ell_1,\ldots,\ell_q}$ for $g_1(\mathbf{b}_i)$ and $g_2(\mathbf{b}_i)$, respectively. The corresponding AGQ weights are $(\pi_{\ell_1}^*, \ldots, \pi_{\ell_q}^*)^{\mathrm{T}}$, where $\pi_i^* = \pi_i \exp(z_i^2)$.

Hence, at the E step, $Q(\boldsymbol{\delta}|\boldsymbol{\delta}^{(h)})$ is approximated by

$$\begin{aligned}\sum_{i,j}\Bigg(&\sum_{\ell_1,\ldots,\ell_q}^{m} w_{i\ell_1,\ldots,\ell_q}^{(h)} [u_{ij}^{(h)}(\mathbf{b}_{i\ell_1,\ldots,\ell_q}^{1*}) \log\, p_{ij}(\boldsymbol{\gamma}) + \{1 - u_{ij}^{(h)}(\mathbf{b}_{i\ell_1,\ldots,\ell_q}^{1*})\} \log\{1 - p_{ij}(\boldsymbol{\gamma})\}] \\ &+ \sum_{\ell_1,\ldots,\ell_q}^{m} w_{i\ell_1,\ldots,\ell_q}^{(h)} u_{ij}^{(h)}(\mathbf{b}_{i\ell_1,\ldots,\ell_q}^{1*}) \log\left\{f_1(y_{ij}|\mathbf{b}_{i\ell_1,\ldots,\ell_q}^{1*}; \tilde{\boldsymbol{\alpha}}, \sigma_1)\right\} \\ &+ \sum_{\ell_1,\ldots,\ell_q}^{m} w_{i\ell_1,\ldots,\ell_q}^{(h)} \{1 - u_{ij}^{(h)}(\mathbf{b}_{i\ell_1,\ldots,\ell_q}^{1*})\} \log\left\{f_2(y_{ij}|\mathbf{b}_{i\ell_1,\ldots,\ell_q}^{1*}; \tilde{\boldsymbol{\beta}}, \sigma_2)\right\}\Bigg) \end{aligned} \tag{3.3}$$

where

$$w_{i,\ell_1,\ldots,\ell_q}^{(h)} = \frac{|\hat{\boldsymbol{\Gamma}}_{1i}|^{-1/2} f(\mathbf{y}_i|\mathbf{b}_{i\ell_1,\ldots,\ell_q}^{1*}; \boldsymbol{\delta}^{(h)}) \phi_q(\mathbf{b}_{i\ell_1,\ldots,\ell_q}^{1*}) \prod_{n=1}^{q} \pi_{\ell_n}^*}{|\hat{\boldsymbol{\Gamma}}_{2i}|^{-1/2} \sum_{\ell_1,\ldots,\ell_q}^{m} \left[f(\mathbf{y}_i|\mathbf{b}_{i\ell_1,\ldots,\ell_q}^{2*}; \boldsymbol{\delta}^{(h)}) \phi_q(\mathbf{b}_{i\ell_1,\ldots,\ell_q}^{2*}) \prod_{n=1}^{q} \pi_{\ell_n}^*\right]}$$

are weights that do not involve $\boldsymbol{\delta}$.

3.1.2 M step

In the $(h+1)$th iteration of the algorithm, the M step maximizes the approximation to $Q(\boldsymbol{\delta}|\boldsymbol{\delta}^{(h)})$ given by Equation (3.3) with respect to $\boldsymbol{\delta}$. Notice that $Q(\boldsymbol{\delta}|\boldsymbol{\delta}^{(h)})$ has a relatively simple form that allows it to be maximized in a straightforward way. From Equation (3.3), the approximation can be seen to be a sum of three terms: first, a weighted binomial loglikelihood involving $\boldsymbol{\gamma}$ only; secondly, a weighted exponential dispersion family loglikelihood involving only $\boldsymbol{\alpha}$, $\boldsymbol{\theta}_1$ and σ_1; and thirdly, a weighted exponential dispersion family loglikelihood involving only $\boldsymbol{\beta}$, $\boldsymbol{\theta}_2$, $\boldsymbol{\theta}_3$ and σ_2. Therefore, the M step for $\boldsymbol{\delta}$ can be done by separately maximizing the three terms in $Q(\boldsymbol{\delta}|\boldsymbol{\delta}^{(h)})$. For each term, this can be done by fitting a weighted version of a standard GLM.

M Step for $\boldsymbol{\gamma}$. Maximization of $Q(\boldsymbol{\delta}|\boldsymbol{\delta}^{(h)})$ with respect to $\boldsymbol{\gamma}$ can be accomplished by fitting a weighted binomial regression of the $u_{ij}^{(h)}(\mathbf{b}_{i\ell_1,\ldots,\ell_q}^{1*})$'s on $\mathbf{W}_i \otimes \mathbf{1}_{m^q}$ with weights $w_{i\ell_1,\ldots,\ell_q}^{(h)}$. Here $\mathbf{1}_k$ is the $k \times 1$ vector of ones. For instance, for g_p taken to be the logit

link, we would perform a weighted logistic regression with a $Nm^q \times 1$ response vector formed by stacking the $u_{ij}^{(h)}(\mathbf{b}^{1*}_{i\ell_1,\ldots,\ell_q})$'s in such a way so that the indices $i, j, \ell_1, \ldots, \ell_q$ cycle through their values most quickly from right to left. The design matrix for this regression is formed by repeating each row of $\mathbf{W} = (\mathbf{W}_1^{\mathrm{T}}, \ldots, \mathbf{W}_C^{\mathrm{T}})^{\mathrm{T}}$ m^q times, and the weight for the $(i, j, \ell_1, \ldots, \ell_q)$th response is given by $w^{(h)}_{i\ell_1,\ldots,\ell_q}$ (constant over j).

M Step for $\tilde{\boldsymbol{\alpha}}, \sigma_1$. Maximization of $Q(\boldsymbol{\delta}|\boldsymbol{\delta}^{(h)})$ with respect to $\tilde{\boldsymbol{\alpha}}$ and σ_1 can be done by again fitting a weighted GLM. Let $\mathbf{X}^* = [(\mathbf{X} \otimes \mathbf{1}_{m^{q_1}}), \mathbf{u}_1^*]$ where $\mathbf{u}_1^*$ is the $Nm^{q_1} \times q_1(q_1+1)/2$ matrix with $(i, j, \ell_1, \ldots, \ell_{q_1})$th row equal to $\{\mathrm{vech}(\mathbf{b}^{1*}_{i\ell_1,\ldots,\ell_{q_1}}\mathbf{u}_{1ij})\}^{\mathrm{T}}$, where $\mathbf{u}_{1ij}$ is the jth row of the random effects' design matrix $\mathbf{u}_{1i}$. Then maximization with respect to $\tilde{\boldsymbol{\alpha}}$ and σ_1 can be accomplished by fitting a weighted GLM with mean $g_1^{-1}(\mathbf{X}^*\tilde{\boldsymbol{\alpha}})$, response vector $\mathbf{y} \otimes \mathbf{1}_{m^{q_1}}$ and weight $w^{(h)}_{i\ell_1,\ldots,\ell_{q_1}} u^{(h)}_{ij}(\mathbf{b}^{1*}_{i\ell_1,\ldots,\ell_{q_1}})$ corresponding to the $(i, j, \ell_1, \ldots, \ell_{q_1})$th element of the response vector.

M Step for $\tilde{\boldsymbol{\beta}}, \sigma_2$. Maximization with respect to $\tilde{\boldsymbol{\beta}}$ and σ_2 can be done by maximizing the third term of $Q(\boldsymbol{\delta}|\boldsymbol{\delta}^{(h)})$. This step proceeds in a similar manner as the M step for $\tilde{\boldsymbol{\alpha}}$ and σ_1. Again, we fit a weighted GLM based on the expanded data set. The design matrix in this regression is $\mathbf{Z}^* = [(\mathbf{Z} \otimes \mathbf{1}_{m^q}), \mathbf{u}_2^*, \mathbf{u}_3^*]$, where $\mathbf{u}_2^*$ is the $Nm^{q_2} \times q_2(q_2+1)/2$ matrix with $(i, j, \ell_{q_1+1}, \ldots, \ell_q)$th row equal to $\{\mathrm{vech}(\mathbf{b}^{1*}_{i\ell_{q_1+1},\ldots,\ell_q}\mathbf{u}_{2ij})\}^{\mathrm{T}}$, $\mathbf{u}_3^*$ is the $Nm^{q_1} \times q_1(q_1+1)/2$ matrix with $(i, j, \ell_1, \ldots, \ell_{q_1})^{th}$ row equal to $\{\mathrm{vech}(\mathbf{b}^{1*}_{i\ell_1,\ldots,\ell_{q_1}}\mathbf{u}_{3ij})\}^{\mathrm{T}}$ and $\mathbf{u}_{kij}$ is the jth row of the random effects' design matrix $\mathbf{u}_{ki}$, $k = 2, 3$. The mean function is $g_2^{-1}(\mathbf{Z}^*\tilde{\boldsymbol{\beta}})$, the response vector is $\mathbf{y} \otimes \mathbf{1}_{m^q}$ and the weight associated with the $(i, j, \ell_1, \ldots, \ell_q)$th response is $w^{(h)}_{i\ell_1,\ldots,\ell_q}\{1 - u^{(h)}_{ij}(\mathbf{b}^{1*}_{i\ell_1,\ldots,\ell_q})\}$.

3.2 NPML estimation

One limitation of the modeling approach described earlier is the normality assumption on the random effects. Although effects of mis-specification of the random effects distribution have been found to be mild in simpler contexts (Heagerty and Kurland, 2001; Neuhaus *et al.*, 1992), it still may desirable to estimate the random effects' distribution nonparametrically when little is known about the mixing distribution or if it is believed to be highly skewed or otherwise non-normal. This approach, known as NPML, has been developed by many authors (e.g., Aitkin, 1999; Follmann and Lambert, 1989; Hinde and Wood, 1987; Laird, 1978; Lindsay, 1983) in simpler contexts; we follow Aitkin (1999) and adapt his methods to the two component GLMM setting.

Aitkin's (1999) approach to NPML estimation can be seen as a modification of Gaussian quadrature in which the weights (i.e., mass points) and abscissas are estimated from the data rather than taken as fixed constants. This can be done as part of the EM algorithm as outlined in Section 3.1 by incorporating the abscissas and masses as parameters of the complete data loglikelihood. The procedure is most easily described for one dimensional random effects, so for the moment assume a random intercept model with $q_1 = q_2 = 1$ dimensional random effects in each component's linear predictor. That is, for now suppose the two GLMMs have linear predictors $\eta_{1ij} = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\alpha} + \theta_1 b_{1i}$ and $\eta_{2ij} = \mathbf{z}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \theta_2 b_{2i}$, respectively. In the parametric setup, there is mixing over the continuous distribution of $\theta_k b_{ki}$, $k = 1, 2$, in each component. In NPML, we replace these continuous distributions with discrete ones with masses at the unknown values $\mathbf{b}_k^* = (b_{k1}^*, b_{k2}^*, \ldots, b_{km}^*)^{\mathrm{T}}$, $k = 1, 2$. Thus for each observation i, j, we

obtain m linear predictors in each component: $\eta^*_{1ij\ell} = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\alpha} + b^*_{1\ell}$, $\ell = 1, \ldots, m$, and $\eta^*_{2ij\ell} = \mathbf{z}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + b^*_{2\ell}$, $\ell = 1, \ldots, m$, with unknown masses $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)^{\mathrm{T}}$. The parameters $\mathbf{b}^*_1$, $\mathbf{b}^*_2$, and $\boldsymbol{\pi}$ describing the mixing distribution are regarded as nuisance parameters, with interest centered on the regression parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

To describe the EM algorithm for NPML, redefine $\boldsymbol{\delta} \equiv (\boldsymbol{\alpha}, \sigma_1, \mathbf{b}^*_1, \boldsymbol{\beta}, \sigma_2, \mathbf{b}^*_2, \boldsymbol{\gamma})^{\mathrm{T}}$. Then the E step yields $Q(\boldsymbol{\delta}, \boldsymbol{\pi}|\boldsymbol{\delta}^{(h)}, \boldsymbol{\pi}^{(h)})$, given by [cf. Equation (3.3)]

$$\sum_{i,j}\Bigg(\sum_{\ell=1}^{m} w_{i\ell}^{(h)}[u_{ij}^{(h)}(\mathbf{b}_1^{*(h)}, \mathbf{b}_2^{*(h)})\log p_{ij}(\boldsymbol{\gamma}) + \{1 - u_{ij}^{(h)}(\mathbf{b}_1^{*(h)}, \mathbf{b}_2^{*(h)})\}\log\{1 - p_{ij}(\boldsymbol{\gamma})\}]$$
$$+ \sum_{\ell=1}^{m} w_{i\ell}^{(h)} u_{ij}^{(h)}(\mathbf{b}_1^{*(h)}, \mathbf{b}_2^{*(h)})\log f_1(y_{ij}|\mathbf{b}_i^*; \tilde{\boldsymbol{\alpha}}, \sigma_1)$$
$$+ \sum_{\ell=1}^{m} w_{i\ell}^{(h)}\{1 - u_{ij}^{(h)}(\mathbf{b}_1^{*(h)}, \mathbf{b}_2^{*(h)})\}\log f_2(y_{ij}|\mathbf{b}_i^*; \tilde{\boldsymbol{\beta}}, \sigma_2)\Bigg) + \sum_{i=1}^{C}\sum_{\ell=1}^{m} w_{i\ell}^{(h)}\log(\pi_\ell) \qquad (3.4)$$

where $w_{i\ell}^{(h)} = f(\mathbf{y}_i|b_{1\ell}^{*(h)}, b_{2\ell}^{*(h)}; \boldsymbol{\delta}^{(h)})\pi_\ell^{(h)} / \sum_{\ell'=1}^{m} f(\mathbf{y}_i|b_{1\ell'}^{*(h)}, b_{2\ell'}^{*(h)}; \boldsymbol{\delta}^{(h)})\pi_{\ell'}^{(h)}$. Comparing the earlier expression for $Q(\boldsymbol{\delta}, \boldsymbol{\pi}|\boldsymbol{\delta}^{(h)}, \boldsymbol{\pi}^{(h)})$ to Equation (3.3), we see that we have almost the same form, with just an extra term for $\boldsymbol{\pi}$ in Equation (3.4). Therefore, the M step proceeds along the same lines as described previously.

M Step for $\boldsymbol{\gamma}$. This can be done by fitting a weighted binomial regression of the $u_{ij}^{(h)}(\mathbf{b}_1^{*(h)}, \mathbf{b}_2^{*(h)})$'s on $\mathbf{W}_i \otimes \mathbf{1}_m$ with weights $w_{i\ell}^{(h)}$.

M Step for $\boldsymbol{\alpha}, \sigma_1, \mathbf{b}^*_1$. Let $\mathbf{X}^* = [(\mathbf{X} \otimes \mathbf{1}_m), \mathbf{I}_n \otimes \mathbf{1}_N]$. Then maximization with respect to $\boldsymbol{\alpha}$, σ_1 and $\mathbf{b}^*_1$ consists of fitting a weighted GLM with mean $g_1^{-1}\{\mathbf{X}^*(\boldsymbol{\alpha}^{\mathrm{T}}, \mathbf{b}_1^{*T})^{\mathrm{T}}\}$, response vector $\mathbf{y} \otimes \mathbf{1}_m$ and weight $w_{i\ell}^{(h)} u_{ij}^{(h)}(b^*_{1\ell}, b^*_{2\ell})$ corresponding to the (i, j, ℓ)th element of the response vector.

M Step for $\boldsymbol{\beta}, \sigma_2, \mathbf{b}^*_2$. Again, we fit a weighted GLM based on the expanded data set. The design matrix in this regression is $\mathbf{Z}^* = [(\mathbf{Z} \otimes \mathbf{1}_m), \mathbf{I}_n \otimes \mathbf{1}_N]$, the mean function is $g_2^{-1}(\mathbf{Z}^*[\boldsymbol{\beta}^{\mathrm{T}}, \mathbf{b}_2^{*\mathrm{T}}]^{\mathrm{T}})$, the response vector is $\mathbf{y} \otimes \mathbf{1}_m$ and the weight associated with the (i, j, ℓ)th response is $w_{i\ell}^{(h)}\{1 - u_{ij}^{(h)}(b^*_{1\ell}, b^*_{2\ell})\}$.

M Step for $\boldsymbol{\pi}$. Maximization with respect to $\boldsymbol{\pi}$ can be done by maximizing the fourth term of Equation (3.4). This maximization yields the closed form solution $\pi_\ell^{(h+1)} = \sum_{i=1}^{C} w_{i\ell}^{(h)} / C$, $\ell = 1, \ldots, m$.

Extension to more than one dimensional random effects is straightforward. In that case, the linear predictors for the two components take the form $\eta_{1ij} = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\alpha} + \mathbf{U}_{1ij}^{\mathrm{T}}\mathbf{D}_1^{\mathrm{T}/2}\mathbf{b}_{1i}$ and $\eta_{2ij} = \mathbf{z}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{U}_{2ij}^{\mathrm{T}}\mathbf{D}_2^{\mathrm{T}/2}\mathbf{b}_{2i}$, respectively. Note that we have dropped the $\mathbf{U}_{3ij}^{\mathrm{T}}\mathbf{D}_3^{\mathrm{T}/2}\mathbf{b}_{1i}$ term from η_{2ij} here [cf. model (2.1)]. However, in the NPML approach no assumption (such as independence) concerning the joint distribution of $\mathbf{D}_1^{\mathrm{T}/2}\mathbf{b}_{1i}$ and $\mathbf{D}_2^{\mathrm{T}/2}\mathbf{b}_{2i}$ is imposed, so correlation between the two linear predictors is permitted automatically and the term $\mathbf{U}_{3ij}^{\mathrm{T}}\mathbf{D}_3^{\mathrm{T}/2}\mathbf{b}_{1i}$ becomes unnecessary. Again, NPML estimation results in m linear predictors per component with masses $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)^{\mathrm{T}}$: $\eta_{1ij} = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\alpha} + \mathbf{U}_{ij}^{\mathrm{T}}\mathbf{b}^*_{1\ell}$, $\eta_{2ij} = \mathbf{z}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{U}_{ij}^{\mathrm{T}}\mathbf{b}^*_{2\ell}$, $\ell = 1, \ldots, m$, where the $\mathbf{b}^*_{k\ell}$s are unknown q_k-dimensional parameters to be estimated. Note that although it may be necessary to choose a larger value of m to capture the joint distribution of $\mathbf{b}_{ki}$ when $q_k > 1$, the computational effort when using NPML increases

linearly with q. In contrast, quadrature methods involve the computation of m^q term sums for each observation in the data set.

In the NPML context, we adapt the method of Friedl and Kauermann (2000) to our context to obtain an expected information matrix at convergence. Again, for details see Wang (2004).

4 Examples

4.1 Measles data

As an illustration of our methodology, we analyse annual measles data that were collected for each of 15 counties in the United States between 1985 and 1991. For each county, the annual number of preschoolers with measles was recorded as well as two variables related to measles incidence: immunization rate and number of preschoolers per county. These data are presented and analysed by Sherman and le Cessie (1997). They employed a bootstrap method for dependent data to get bootstrap replicates from 15 counties. For each bootstrap resample, the parameters of a loglinear regression of number of cases on immunization rate were estimated by maximizing the Poisson GLM likelihood under independence, using the natural logarithm of the number of children as an offset.

An interesting result of their analysis is the histogram of the 1000 bootstrap slope estimates they obtained (Sherman and le Cessie, 1997: 914, Figure 1). This histogram shows a clear bimodal shape, suggesting a possible mixture structure underlying the data set. Such a structure is also apparent in a simple plot of the data by county (Figure 1). In Figure 1, there appears to be a mix of high and low incidences which may reflect the epidemiology of this disease. Intuitively, it seems reasonable that there may be epidemic and nonepidemic years. Because the categorization of any given year as an epidemic year is determined by the magnitude of the response, it is necessary to infer the latent class structure through a mixture model rather than to directly model it in the linear predictor of a GLM. Because the data are also clustered by county, we feel that a two component GLMM is a natural model to consider here. In addition, such a model will allow us to separately quantify covariate effects in the two components. For example, the effect of vaccination rate on the incidence of measles may be quite different in epidemic and nonepidemic years.

Let y_{ij} be the number of cases in county i, $(i = 1, \ldots, 15)$ in year j, $(j = 1, \ldots, 7)$, and let b_i be a one-dimensional random county effect for county i. We considered several models for these data all of which are special cases of a two-component GLMM of the form $Y_{ij}|b_i \sim p_{ij}\text{Poisson}(\lambda_{1ij}|b_i) + (1 - p_{ij})\text{Poisson}(\lambda_{2ij}|b_2)$, where

$$\begin{aligned}
\log(\lambda_{1ij}) &= \alpha_0 + \alpha_1 \text{rate}_{ij} + \theta_1 b_{1i} + \log(n_{ij}) \\
\log(\lambda_{2ij}) &= \beta_0 + \beta_1 \text{rate}_{ij} + \theta_2 b_{2i} + \theta_3 b_{1i} + \log(n_{ij}) \\
\text{logit}(p_{ij}) &= \gamma_0 + \gamma_1 \text{rate}_{ij}
\end{aligned} \tag{4.1}$$

and where α_0, β_0, γ_0 are fixed intercepts and α_1, β_1, γ_1 are fixed effects of immunization rate for the two Poisson means λ_{1ij}, λ_{2ij} and the mixing probability p_{ij}, respectively.

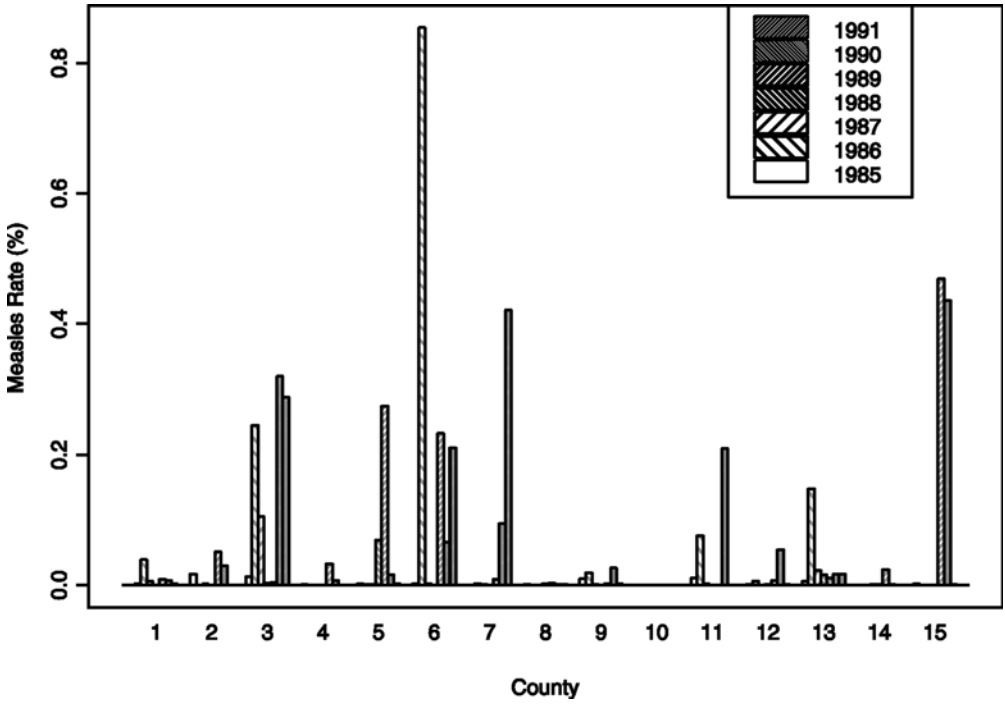

Figure 1 Measles data. Years are grouped together for each county 1985–1991

In addition, $\log(n_{ij})$ represents an offset corresponding to the natural logarithm of the number of children in the ith county during the jth year.

Table 1 lists eight models of this general form and the corresponding values of -2 times the maximized loglikelihood (-2ℓ) and the Akaike information criterion (AIC). All of the models in Table 1 except models 2 and 8 were fit with ML. Of these, model 1 is a (nonmixture) GLMM, model 3 is a mixture model without random effects and models 4–6 are two component GLMMs with identical fixed effects, but different assumptions regarding the random effects structure. Among these models, it is clear that models 1 and 3 fit much worse than the rest, suggesting that a two component structure and county specific random effects are necessary here. Models 5 and 6 have very similar values of -2ℓ, with AIC preferring the more parsimonious model 6. To investigate whether the mixing probability varied with immunization rate, we refit model 6 with immunization rate included as a covariate in the linear predictor for $\text{logit}(p_{ij})$. A comparison of this model, listed as model 7 in Table 1, with model 6 via either a likelihood ratio test or a AIC suggests that $\gamma_1 = 0$. Thus, among the models fit with ML, we prefer model 6 for these data.

To illustrate the NPML approach, we refit model 6 dropping the assumption of normality on the random effects. This model is listed as model 8 in Table 1. For model 8, we followed the strategy described by Friedl and Kauermann (2000) and started the fitting procedure from a large value of m ($m = 12$), and then reduced m systematically

Table 1 Comparison of different models for the measles data

Model number	Model type	Random effects	Fitting method	$\mathbf{W}_{ij}^{\mathrm{T}}\gamma$	− 2 Loglik	AIC
1	GLMM	$\theta_1 b_{1i}$	ML	N/A	10174.0	10180.0
2	GLMM	$\theta_1 b_{1i}$ [a]	NPML	N/A	2677.7	2699.7
3	2-GLM	N/A	ML	γ_0	2712.3	2722.3
4	2-GLMM	$\theta_1 b_{1i},\ \theta_2 b_{2i} + \theta_3 b_{1i}$	ML	γ_0	1959.7	1975.7
5	2-GLMM	$\theta_1 b_{1i},\ \theta_3 b_{1i}$	ML	γ_0	2140.1	2154.1
6	2-GLMM	$\theta_1 b_{1i},\ \theta_2 b_{2i}$	ML	γ_0	1960.3	1974.3
7	2-GLMM	$\theta_1 b_{1i},\ \theta_2 b_{2i}$	ML	$\gamma_0 + \gamma_1$ rate	1959.4	1975.4
8	2-GLMM	$\theta_1 b_{1i},\ \theta_2 b_{2i}$ [a]	NPML	γ_0	1914.1	1962.1

[a]NPML model is most similar to this parametric description, having random intercept(s) in each linear predictor. However, the normality assumption is dropped, and the θs are absorbed into the mass points.

until all quadrature points are different and no quadrature weights are very small (less than 0.01). In this case, we stopped the fitting procedure at $m = 7$. According to the AIC criterion, model 8 offers a slight improvement in fit over model 6.

To further investigate the suitability of the models, we fit in this example, we follow the approach of Vieira *et al.* (2000) who suggested the use of half normal plots as goodness-of-fit tools. Half normal plots for the GLMM (model 1), two-component GLM (model 3) and the two component GLMM with independent normal random effects (model 6) appear in Figure 2(a–c). The plots display the absolute values of the Pearson residuals versus half normal scores, with simulated envelopes based on the assumed model evaluated at the estimated parameter values. A suitable model is indicated by the observed values falling within the simulated envelope. The Pearson residuals are defined as $[y_{ij} - \widehat{E(Y_{ij})}]/\sqrt{\widehat{\mathrm{var}(Y_{ij})}}$, where $E(Y_{ij}) = E\{E(Y_{ij}|b_i)\}$, $\mathrm{var}(Y_{ij}) = E(Y_{ij}^2) - \{E(Y_{ij})\}^2 = E\{E(Y_{ij}^2|b_i)\} - \{E(Y_{ij})\}^2$ for the mixed models. The marginal expectations here were evaluated using 20-point OGQ, and the hats indicate evaluation at the final parameter estimates, which were obtained using 11-point AGQ. For the two-

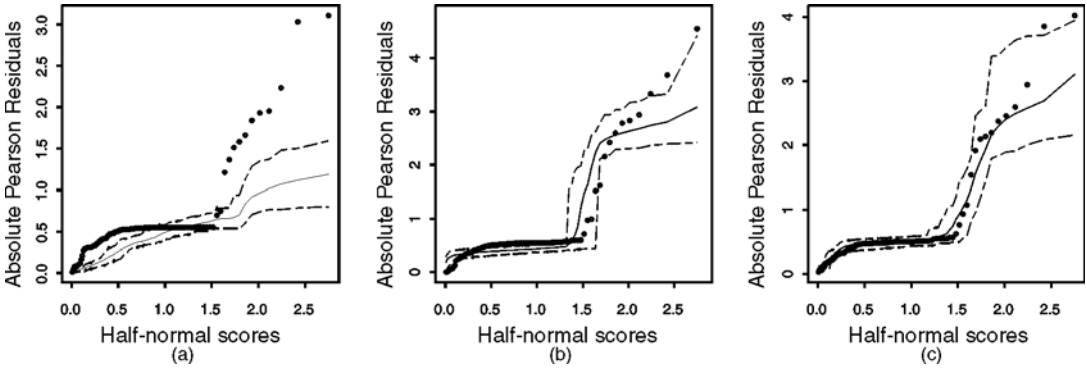

Figure 2 Half normal plot for assessing goodness-of-fit of models 1 (a), 3 (b) and 6 (c). These three models are GLMM, two-component GLM, and two-component GLMM, respectively

component GLM, $E(Y_{ij}) = p_{ij}\lambda_{1ij} + (1 - p_{ij})\lambda_{2ij}$, $\text{var}(Y_{ij}) = p_{ij}(\lambda_{1ij} + \lambda_{1ij}^2) + (1 - p_{ij})(\lambda_{2ij} + \lambda_{2ij}^2) - \{E(Y_{ij})\}^2$, where λ_{1ij} and λ_{2ij} are means for each Poisson component.

Figure 2(a) clearly indicates that the one-component GLMM model is inadequate for the measles data because almost all points fall outside of the simulated envelope. Figure 2(b) shows that the two-component GLM improves the fit, but in the left half of the plot there are still many points outside the envelope. In Figure 2(c), nearly all points are along the simulated means, confirming that a two-component GLMM fits these data best.

The fitting results mentioned earlier are consistent with expectations based on a preliminary examination of the data and some consideration of the epidemiology of this disease. Because of the mixture of large and small incidences of measles and the epidemic phenomenon, we expected that two components would be necessary to model these data. This is borne out by the vast improvement in fit from a one to a two-component GLM. In their analysis of these data, Sherman and le Cessie (1997) found that county number 6 was largely responsible for the bimodal structure of the bootstrap distribution of the estimated rate effect. However, we found no evidence that the necessity of a second component was driven by this or by any other single county's data. For model 6, the posterior component membership probabilities (3.1) were computed at convergence. According to these posterior probabilities, all but one county (county 10) had at least one observation (year) classified in each component of the mixture. As the data are also clustered, the within county correlation must also be accounted for. We have chosen to account for this correlation through a random county effect, and this approach improves the fit compared with the fixed effect two-component GLM.

The parameter estimates from models 6 and 8 are summarized in Table 2. Writing somewhat informally, the ML results imply

$$\log\begin{pmatrix}\text{predicted}\\ \text{measles}\\ \text{rate}\end{pmatrix} = \begin{cases} 1.97 - 0.133(\text{imm. rate}) & \\ \quad + 0.669(\text{county effect}), & \text{in outbreak years} \\ -5.53 - 0.0767(\text{imm. rate}) & \\ \quad + 1.06(\text{county effect}), & \text{in nonoutbreak years} \end{cases}$$

Table 2 Parameter estimates (standard errors) from models 6 and 8 fit to measles data

Parameter	AGQ	NPML
α_0	1.97 (1.58)	–
α_1	−0.133 (0.0232)	−0.143 (0.0137)
β_0	−5.53 (2.43)	–
β_1	−0.0767 (0.0354)	−0.0962 (0.00811)
γ_0	−0.992 (0.230)	−0.931 (0.204)
θ_1	0.669 (0.138)	–
θ_2	1.06 (0.253)	–

Note: The two components have been exchanged in these two models.

If we suppose an immunization rate of 70%, say and set the county effect to zero, the fitted model predicts

$$\text{annual measles rate} = \begin{cases} 0.000649, & \text{in outbreak years} \\ 0.0000185, & \text{in nonoutbreak years} \end{cases}$$

with an estimated probability of an outbreak of $\text{logit}^{-1}\{-0.992\} = 0.271$. The NPML results are similar. The parameters α_0, β_0, θ_1, and θ_2 have been absorbed into the quadrature points and weights in the NPML approach. However, the immunization rate effects for outbreak and nonoutbreak years, -0.143 and -0.0962, respectively, are comparable, though somewhat larger, than those for ML. The estimated probability of a measles outbreak, $\text{logit}^{-1}\{-0.931\} = 0.283$, is similar to that for ML.

An associate editor has inquired how the two component mixture fit with NPML differs from a one-component model with NPML. The concern here is that the latter model is fit as a finite mixture of GLMs and the former is fit as a two-component mixture of finite mixtures of GLMs. Thus, both models are finite mixtures of GLMs, and it is not completely obvious that the two-component model would result in a substantially different fit or that it would even be identifiable. The difference between the two types of models has to do with the level at which mixing occurs. In both cases, the NPML approach assumes that cluster specific random effects are drawn from a discrete probability distribution with finite support. That is, discrete mixing is assumed at the cluster level. In the two component case, there is additional finite mixing at the individual observation level. To highlight this difference, we fit the one-component GLMM with NPML. This model is listed as model 2 in Table 1 and is based on $m = 5$ mass points. Although this model fit substantially better than model 1 (fit with ML), it is not competitive with the two-component GLMMs. In the context of this example, it appears that it is useful to account for heterogeneity both between counties and from year to year within counties. In addition, the superior fit of models 4–8 relative to model 2 undermines the notion that the two-component model fits better only because of an inappropriate distributional assumption for the random effects in the one-component model. Because the NPML models fit better than the corresponding ML models, it does appear that the normality assumption on the random effects may be violated. However, a two-component structure is still necessary for modeling these data adequately.

4.2 Whitefly data

Our second example involves data from a horticulture experiment to investigate the efficacy of several different means of applying pesticide to control whiteflies on greenhouse raised poinsettia plants. The data arise from a randomized complete block design with repeated measures taken over 12 weeks. Eighteen experimental units were formed from 54 plants, with units consisting of three plants each. These units were randomized to six treatments in three blocks. The response variable of interest here is the number of surviving whiteflies out of the total number placed on the plant two weeks previously. These data are discussed in more detail in van Iersel *et al.* (2000). In that paper, ZIB regression models were used to analyse these data, with random effects at the plant level

to account for correlation among the repeated measures on a given plant. We return to this problem to investigate whether a two-component mixture of GLMMs can improve upon the fit of a ZIB mixed model for these data. That is, the question is whether an improvement can be achieved by allowing the second component to be a binomial with relatively small success probability rather than a zero success probability.

Let $y_{ijk\ell}$ be the number of live adult whiteflies on plant k ($k = 1, \ldots, 54$) in treatment i ($i = 1, \ldots, 6$) in block j ($j = 1, \ldots, 3$) measured at time $\ell(\ell = 1, \ldots, 12)$. Let $n_{ijk\ell}$ be the total number of whiteflies placed on the leaf of plant k in treatment i in block j measured at time ℓ. Further, let α_i be the ith treatment effect, β_j be the jth block effect, τ_ℓ be the ℓth week effect and b_{1k} and b_{2k} each be one-dimensional random plant effects for plant k. For simplicity, we consider models containing only main effects (treatment, block and week) and plant specific random intercepts. Specifically, the results we present here all pertain to special cases of the model that assumes $Y_{ijk\ell}|b_k \sim p_{ijk\ell}\text{Binomial}(n_{ijk\ell}, \pi_{1ijk\ell}|b_k) + (1 - p_{ijk\ell})\text{Binomial}(n_{ijk\ell}, \pi_{2ijk\ell}|b_k)$, where

$$\begin{aligned}
\text{logit}(\pi_{1ijk\ell}) &= \mu_1 + \alpha_{1i}\text{treatment}_i + \beta_{1j}\text{block}_j + \tau_{1\ell}\text{week}_\ell + \theta_1 b_{1k} \\
\text{logit}(\pi_{2ijk\ell}) &= \mu_2 + \alpha_{2i}\text{treatment}_i + \beta_{2j}\text{block}_j + \tau_{2\ell}\text{week}_\ell + \theta_2 b_{2k} + \theta_3 b_{1k} \\
\text{logit}(p_{ijk\ell}) &= \mu_3 + \alpha_{3i}\text{treatment } t_i + \beta_{3j}\text{block}_j + \tau_{3\ell}\text{week}_\ell
\end{aligned} \tag{4.2}$$

Fit statistics for model (4.2) and some simpler models appear in Table 3. Models 1, 4, 6, 7 and 8 in this table were fit with ML using five-point AGQ. Among these models, the two-component GLMM with proportional plant specific random effects in each linear predictor, model 7, yielded the smallest AIC. A similar two component GLMM with plant effects in each component was also fit using NPML (model 9). In fitting model 9, we followed the same procedure as described in Section 4.1, which resulted in $m = 5$ non-negligible mass points and a slightly worse fit according to the AIC statistic. For purposes of comparison, we also fit a one-component GLMM with both ML and NPML ($m = 7$), a ZIB model, a ZIB mixed model and a two-component GLM, all of which were special cases of Equation 4.2. That is, they had the same linear predictors for the first component (without random effects for models 3 and 5) and mixing probability as in Equation (4.2).

Table 3 Comparison of different models for whitefly data

Model number	Model type	Random effects	Fitting method	− 2 Loglik	AIC
1	GLMM	$\theta_1 b_{1k}$	ML	2409.0	2449.0
2	GLMM	$\theta_1 b_{1k}$[a]	NPML	2392.2	2456.2
3	ZIB	N/A	ML	1928.7	2004.7
4	ZIB-Mixed	$\theta_1 b_{1k}$	ML	1883.3	1961.3
5	2-GLM	N/A	ML	1628.2	1742.2
6	2-GLMM	$\theta_1 b_{1k},\ \theta_2 b_{2k} + \theta_3 b_{1k}$	ML	1606.5	1726.5
7	2-GLMM	$\theta_1 b_{1k},\ \theta_3 b_{1k}$	ML	1607.0	1725.0
8	2-GLMM	$\theta_1 b_{1k},\ \theta_2 b_{2k}$	ML	1642.9	1760.9
9	2-GLMM	$\theta_1 b_{1k},\ \theta_2 b_{2k}$[a]	NPML	1596.4	1736.4

[a]Again, NPML model is most similar to this parametric description.

From Table 3, we find that the two component models are better than the corresponding one component models. In addition, models with random plant effects are better than the corresponding models without random effects. From these results, it is clear that both random effects and second component are necessary here. In addition, a nondegenerate (nonzero) second component also improves the fit over a ZIB model. That is, two-component GLMMs fit best here, with proportional random effects slightly preferred over the other random effects structures we considered.

5 Discussion

In this paper, we have formulated a class of two-component mixtures of GLMMs for clustered data and described how the EM algorithm, combined with quadrature methods, can be used to fit these models using ML estimation. Extension of this model class to more than two components is possible and is, in principle, straightforward. However, the complexity of the model, its notation, and its fitting algorithm will grow rapidly with the number of components, and it is not clear that the practical value of such models justifies consideration of cases beyond two or three components.

Our model class allows for correlation due to clustering to be accounted for through the inclusion of cluster specific random effects. Clearly, there are other valid approaches for accounting for within cluster correlation. Alternatives include marginal models (Rosen *et al.*, 2000) and transition models (Park and Basawa, 2002). As in the nonmixture case, which approach is most appropriate for accounting for the correlation will depend upon the application.

Extension to multilevel models (multiple nested levels of clustering), crossed random effects and other more general random effects structures is an important area of future research. One attractive approach for this extension is to use a Monte Carlo EM algorithm (McCulloch, 1997) in place of our EM algorithm with quadrature. The main challenge to implement the Monte Carlo EM in this context is sampling from $f(\mathbf{b}|\mathbf{y};\boldsymbol{\delta})$, the conditional distribution of the random effects given the observed data. We have had some success with this approach, but have found that the computing time is prohibitively long for practical use. Another possibility raised by an anonymous referee is to use simulated ML (refer, for example, McCulloch and Searle, 2001, Section 10.3.e). However, this approach is well known to be very inefficient for nonoptimal importance samplers, and, as mentioned previously, the optimal choice, $f(\mathbf{b}|\mathbf{y};\boldsymbol{\delta})$, can be quite difficult to sample from. Furthermore, simulation based (McCulloch, 1997) and analytical (Jank and Booth, 2003) investigations have found simulated ML estimation to perform poorly relative to Monte Carlo EM, which we plan to investigate further in future work.

Acknowledgements

The authors wish to thank two anonymous referees and an associate editor whose comments led to significant improvements in the paper.

References

Aitkin M (1999) A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55**, 117–28.

Berk KN and Lachenbruch PA (2002) Repeated measures with zeros. *Statistical Methods in Medical Research* **11**, 303–16.

Booth JG, Casella G, Friedl H and Hobert JP (2003) Negative binomial loglinear mixed models. *Statistical Modelling: An International Journal* **3**, 179–91.

Davidian M and Giltinan DM (1995) *Nonlinear models for repeated measurement data*. New York: Chapman and Hall.

Davidian M and Giltinan DM (2003) Nonlinear models for repeated measurement data: an overview and update. *Joural of Agricultural, Biological, and Environmental Statistics* **8**, 387–19.

Deb P and Trivedi PK (1997) Demand for medical care by the elderly: a finite mixture approach. *Journal of Applied Econometrics* **12**, 313–36.

Dempster AP, Laird NM and Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* **39**, 1–38, (with discussion).

Dietz E and Böhning, D (1997) The use of two-component mixture models with one completely or partly known component. *Computational Statistics* **12**, 219–34.

van Duijn MAJ and Bockenholt U (1995) Mixture models for the analysis of repeated count data. *Applied Statistics* **44**, 473–85.

Follmann D and Lambert D (1989) Generalizing logistic regression non-parametrically. *Journal of the American Statistical Association* **84**, 295–300.

Friedl H and Kauermann G (2000) Standard errors for EM estimates in generalized linear models with random effects. *Biometrics* **56**, 761–67.

Hall DB (2000) Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics* **56**, 1030–39.

Heagerty PJ and Kurland BF (2001) Misspecified maximum likelihood estimates and generalized linear mixed models. *Biometrika* **88**, 973–85.

Hinde JP and Wood ATA (1987) Binomial variance component models with a non-parametric assumption concerning random effects. In Crouchley, R, editor *Longitudinal data analysis*, Aldershot, Hants: Avebury.

van Iersel M, Oetting R and Hall DB (2000) Imidacloprid applications by subirrigation for control of silverleaf whitefly (Homoptera: Aleyrodidae) on poinsettia. *Journal of Economic Entomology* **93**, 813–19.

Jank W and Booth J (2003) Efficiency of Monte Carlo EM and simulated maximum likelihood in two-stage hierarchical models. *Journal of Computational and Graphical Statistics* **12**, 214–29.

Jansen RC (1993) Maximum likelihood in a generalized linear finite mixture model by using the EM algorithm. *Biometrics* **49**, 227–31.

Laird NM (1978) Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, **73**, 805–11.

Lambert D (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**, 1–14.

Lindsay BG (1983) The geometry of likelihoods: a general theory. *Annals of Statistics* **11**, 86–94.

Lindsay BG (1995) *Mixture models: theory, geometry and applications*. The Institute of Mathematical Statistics and the American Statistical Association.

Liu Q and Pierce DA (1994) A note on Gaussian–Hermite quadrature. *Biometrika* **81**, 624–29.

McCulloch CE (1997) Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* **92**, 162–70.

McCulloch CE and Searle SR (2001) *Generalized, Linear, and Mixed Models*. New York: Wiley.

McLachlan GJ and Basford KE (1988) *Mixture models: inference and applications to clustering*. New York: Marcel Dekker.

Neuhaus JM, Hauck WW and Kalbfleisch JD (1992) The effects of mixture distribution misspecification when fitting mixed effects logistic models. *Biometrika* **79**, 755–62.

Oakes D (1999) Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society: Series B* **61**, 479–82.

Olsen MK and Schafer JL (2001) A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association* **96**, 730–45.

Park JG and Basawa IV (2002) Estimation for mixtures of Markov processes. *Statistics and Probability Letters* **59**, 235–44.

Pinheiro JC and Bates DM (1995) Approximations to the loglikelihood function in the nonlinear mixed effects model. *Journal of Computational and Graphical Statistics* **4**, 12–35.

Pinheiro JC and Bates DM (1996) Unconstrained parameterizations for variance–covariance matrices. *Statistics and Computing* **6**, 289–96.

Ridout M, Demétrio CGB and Hinde J (1998) Models for count data with many zeros. Invited Paper, The XIXth International Biometric Conference, Cape Town, South Africa, 179–92.

Rosen O, Jiang WX and Tanner MA (2000) Mixtures of marginal models. *Biometrika* **87**, 391–404.

Sherman M and le Cessie S (1997) A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models. *Communications in Statistics–Simulation* **26**, 901–25.

Thompson TJ, Smith PJ and Boyle JP (1998) Finite mixture models with concomitant information: assessing diagnostic criteria for diabetes. *Applied Statistics* **47**, 393–404.

Titterington DM, Smith AFM and Makov UE (1985) *Statistical analysis of finite mixture distributions*. New York: Wiley.

Vieira AMC, Hinde JP, and Demétrio CGB (2000) Zero-inflated proportion data models applied to a biological control assay. *Journal of Applied Statistics* **27**, 373–89.

Wang L (2004) *Parameter Estimation for Mixtures of Generalized Linear Mixed-effects Models*. Unpublished PhD dissertation, Department of Statistics. University of Georgia, can be accessed via authors website: http:/www.stat.uga.edu/~dhall/pub/lihuasdissert.pdf.

Wang P and Puterman ML (1998) Mixed logistic regression models. *Journal of Agricultural, Biological, and Environmental Statistics* **3**, 175–200.