Negative Binomial Loglinear Mixed Models

James G. Booth[1*], George Casella[1], Herwig Friedl[2] and James P. Hobert[1]

[1]Department of Statistics
University of Florida
Gainesville, FL 32611, USA

and

[2]Institute of Statistics
Technical University Graz
Steyrergasse 17
A-8010 Graz, Austria

August 2001

*Booth and Hobert's research supported by NSF Grant DMS-00-72827. Casella's research supported by NSF Grant DMS-99-71586.

Abstract

The Poisson loglinear model is a common choice for explaining variability in counts. However, in many practical circumstances the restriction that the mean and variance are equal is not realistic. Overdispersion with respect to the Poisson distribution can be modeled explicitly by integrating with respect to a mixture distribution, and use of the conjugate gamma mixing distribution leads to a negative binomial loglinear model. This paper extends the negative binomial loglinear model to the case of dependent counts, where dependence among the counts is handled by including linear combinations of random effects in the linear predictor. If we assume that the vector of random effects is multivariate normal, then arbitrary forms of dependence can be modeled by appropriate specification of the covariance structure. Although the likelihood function for the resulting model is not tractable, maximum likelihood estimates (and standard errors) can be found using the NLMIXED procedure in SAS or, in more complicated examples, using a Monte Carlo EM algorithm. An alternate approach is to leave the random effects completely unspecified and attempt to estimate them using nonparametric maximum likelihood. The methodologies are illustrated with several examples.

Key words and phrases: Monte Carlo EM; NLMIXED procedure; Nonparametric maximum likelihood; Overdispersion; Random effects;

1 Introduction

In settings where it is of interest to describe the way in which counts vary as a function of explanatory variables, there are often many factors that are not or cannot be measured. Omission of such factors from the model typically translates into overdispersion with respect to the Poisson distribution; that is, observed variability much greater than the mean (Cox, 1983). This additional variability can be modeled explicitly by integrating with respect to a mixture distribution, and use of the conjugate gamma mixing distribution leads to a negative binomial model for the counts whose index parameter accounts for potential overdispersion with respect to Poisson variation.

In Section 2 we describe the negative binomial model, and give details on the treatment of the random factors. Section 3 details methods for maximizing the resulting likelihood. In particular, we find that SAS Proc NLMIXED will often work, but when it does not (as in our example in Section 4.2) a Monte Carlo EM algorithm can be implemented in a straightforward way. We also show that the nonparametric version of our model, where the random effects distribution is unspecified, results in a closed-form expression of the expected complete-data likelihood and, hence, a simple EM algorithm. Section 4 analyzes two data sets, and illustrates that the negative binomial model can lead to more meaningful (less biased) parameter estimates, and that the computations are quite tractable. Section 5 contains a final discussion, and there are five appendices with technical details and data.

2 A Negative Binomial Loglinear Mixed Model

We begin by describing the negative binomial loglinear model. Let $V_1, \ldots, V_n$ denote an independent and identically distributed (iid) sample of unit mean gamma random variables with shape parameter α; that is, $f(v_1) \propto v_1^{\alpha-1} \exp\{-\alpha v_1\} I(v_1 > 0)$. Suppose that, conditional on v_i, the ith count Y_i has a Poisson distribution with mean $v_i \mu_i$, so $Y_i | v_i \sim \text{Poisson}(v_i \mu_i)$. The counts are then marginally independent negative binomial random variables with mass functions given by

$$\Pr(Y_i = y; \alpha, \mu_i) = \frac{\Gamma(y+\alpha)}{\Gamma(\alpha) y!} \left(\frac{\alpha}{\mu_i + \alpha}\right)^{\alpha} \left(\frac{\mu_i}{\mu_i + \alpha}\right)^{y} \tag{1}$$

where $y \in \{0, 1, 2, \ldots\}$. We write $Y_i \sim \text{nb}(\alpha, \mu_i)$ and denote the right-hand side of (1) by $\text{nb}(y; \alpha, \mu_i)$. Using iterated expectation and variance, it can be shown that the marginal mean and variance of Y_i are μ_i and $\mu_i + \mu_i^2/\alpha$. Hence, the parameter α quantifies the amount of overdispersion with $\alpha = \infty$ corresponding to no overdispersion with respect to the Poisson distribution.

Suppose that $Y_i \sim \text{nb}(\alpha, \mu_i)$ and that μ_i is related to a vector of explanatory variables, $\mathbf{x}_i$, through a loglinear model, $\log \mu_i = \mathbf{x}_i' \boldsymbol{\beta}$. This alternative to the Poisson regression model has become fairly standard (see e.g. Land, McCall and Nagin, 1996). In fact, maximum likelihood fitting can now be accomplished using the GENMOD procedure in SAS (SAS Institute Inc., 1997), and methods for maximizing the resulting likelihood function are described in Appendix A. (Note that this model does not quite fit into the *quasi-likelihood* framework introduced by Breslow (1984). In that model, the conditional mean of Y_i is assumed to have the form $\exp\{\mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i\}$ where "...the ϵ_i are random error terms having means 0 and a constant unknown variance σ^2.")

One obvious fully parametric approach is to assume the random errors are iid

$N(0,\sigma^2)$, but this does not lead to a tractable likelihood. Instead, if we assume that the random errors are logarithms of gamma variables, we obtain a closed form negative binomial likelihood with components of the form given in (1). In our model the conditional mean of Y_i is $\exp\{\mathbf{x}_i'\boldsymbol{\beta} + \log(v_i)\}$, where $\mathrm{E}\log(V_i) < \log \mathrm{E}(V_i) = 0$. If one attempts to create mean-zero errors by adding $-\mathrm{E}\log(V_i)$ to the linear predictor, this leads to a model in which the regression parameter and the error variance both depend on α, and are thus not functionally independent. This does not seem consistent with the model of Breslow (1984).

Multiple counts observed at the same site or on the same subject at different times are typically correlated. A standard approach to modeling such correlation is to introduce random effects into the linear predictors. Specifically, suppose we have count data Y_{ij}, $i = 1, \ldots, r$ and $j = 1, \ldots, n_i$, where Y_{ij} denotes the jth measurement on the ith subject (or cluster). Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$ and suppose that $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are known vectors of covariates associated with Y_{ij}. We assume that, conditional on a (q-dimensional) vector of *cluster specific random effects*, $\mathbf{u}_i$, the elements of $\mathbf{Y}_i$ are independent negative binomial random variables

$$Y_{ij}|\mathbf{u}_i \sim \mathrm{nb}(\alpha, \mu_{ij}) \text{ with } \mu_{ij} = \mathrm{E}(y_{ij}|\mathbf{u}_i) = \exp\{\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{u}_i\} \tag{2}$$

where $\boldsymbol{\beta}$ is an unknown $p \times 1$ vector of regression coefficients. We note here that the case $r = 1$ may be of interest in some settings; for example, if subject levels are crossed.

The normal distribution is a convenient choice for the random effects because it allows for a flexible specification of the correlation structure. Hence, we assume that $\mathbf{u}_1, \ldots, \mathbf{u}_r$ are iid $\mathrm{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$ and that $\boldsymbol{\Sigma}$ is known up to a vector of variance components, $\boldsymbol{\theta}$. We denote the entire vector of unknown parameters as $\boldsymbol{\psi} = (\alpha, \boldsymbol{\beta}, \boldsymbol{\theta})$.

We also consider the situation where the random effect distribution is unknown and left totally unspecified. The nonparametric maximum likelihood (NPML) estimate is then an appealing choice for this setting. The NPML technique was used in Aitkin (1996) to account for overdispersion in generalized linear Poisson and binomial models, by allowing for response-specific random intercepts. The model's extension to clustered observations, assumed to be correlated, is discussed in Aitkin (1999) and also in Alfò and Aitkin (2000). As shown in Kiefer and Wolfowitz (1956) the random effect distribution can be consistently estimated as $r \to \infty$. From the results in Laird (1978) and Lindsay (1983) it is known that this estimator is a discrete distribution defined on a finite number, say K, of q-dimensional mass points $\boldsymbol{\zeta}_k$ with associated masses π_k, $k = 1, \ldots, K$. So here, instead of estimating the variance components, we have to estimate K unknown pairs $(\boldsymbol{\zeta}_k, \pi_k)_{k=1}^{K}$, which is often done conditionally on a given number K.

3 Maximizing the likelihood

3.1 Parametric maximum likelihood

Let $\phi(\cdot; \boldsymbol{\Sigma})$ denote the q-variate normal density with mean zero and covariance matrix $\boldsymbol{\Sigma}$. Under the assumption that the $\mathbf{u}_i$'s are iid $\mathrm{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$, the joint density of $\mathbf{u}$ and $\mathbf{y}$ is given by

$$f(\mathbf{y}, \mathbf{u}; \boldsymbol{\psi}) = \prod_{i=1}^{r} f(\mathbf{y}_i, \mathbf{u}_i; \boldsymbol{\psi}) = \prod_{i=1}^{r} f(\mathbf{y}_i | \mathbf{u}_i; \alpha, \boldsymbol{\beta})\, \phi(\mathbf{u}_i; \boldsymbol{\Sigma})$$

where $f(\mathbf{y}_i | \mathbf{u}_i; \alpha, \boldsymbol{\beta})$ is given by (1). Since the random effects are not observable, the likelihood function is given by

$$L(\boldsymbol{\psi}; \mathbf{y}) = \int f(\mathbf{y}, \mathbf{u}; \boldsymbol{\psi})\, d\mathbf{u} = \prod_{i=1}^{r} \int f(\mathbf{y}_i | \mathbf{u}_i; \alpha, \boldsymbol{\beta})\, \phi(\mathbf{u}_i; \boldsymbol{\Sigma})\, d\mathbf{u}_i \,.$$

Recall that $\log \mu_{ij}$ is the linear predictor and hence μ_{ij} is a nonlinear function of $\mathbf{u}_i$. Consequently, L involves r intractable integrals of dimension q making direct maximization with respect to the parameters infeasible in general.

The intractable integrals in the two examples discussed in Section 4 are 1-dimensional and 2-dimensional respectively. In such cases numerical integration, for example the adaptive Gaussian quadrature method used in the SAS procedure NLMIXED, can be used to maximize the likelihood function. However, when the dimension approaches 5 or 10, Monte Carlo methods are typically preferable to numerical integration.

3.2 Implementation of Monte Carlo EM

To understand the application of a Monte Carlo EM (MCEM) algorithm, we first consider application of the (deterministic) EM algorithm (Dempster, Laird and Rubin, 1977) where $\mathbf{u}$ is viewed as the missing data. Let $\boldsymbol{\psi}^{(t)}$ denote the value of the estimate at the tth iteration. Then $\boldsymbol{\psi}^{(t+1)}$ is calculated by maximizing

$$Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(t)}) = \mathrm{E}[\log f(\mathbf{y}, \mathbf{u}; \boldsymbol{\psi}) | \mathbf{y}; \boldsymbol{\psi}^{(t)}]$$

with respect to $\boldsymbol{\psi}$. As the notation suggests, the expectation is with respect to $f(\mathbf{u}|\mathbf{y}; \boldsymbol{\psi}^{(t)})$. In order to calculate $f(\mathbf{u}|\mathbf{y}; \boldsymbol{\psi}^{(t)})$, we need $f(\mathbf{y}, \mathbf{u}; \boldsymbol{\psi}^{(t)})$ and $f(\mathbf{y}; \boldsymbol{\psi}^{(t)})$. While we have the former in closed form, the latter is, of course, simply $L(\boldsymbol{\psi}^{(t)}; \mathbf{y})$, which we do not have. Hence, there is no hope for analytical evaluation of Q.

The MCEM algorithm (Wei and Tanner, 1990; Booth and Hobert, 1999) allows us to circumvent the intractable expectation in the E-step via Monte Carlo. Given an iid sample $\mathbf{u}_{t,1}, \ldots, \mathbf{u}_{t,m}$ from $f(\mathbf{u}|\mathbf{y}; \boldsymbol{\psi}^{(t)})$ (see Appendix B for details on drawing such a sample) we maximize

$$\hat{Q}(\boldsymbol{\psi}; \boldsymbol{\psi}^{(t)}) = \frac{1}{m} \sum_{l=1}^{m} \log f(\mathbf{y}, \mathbf{u}_{t,l}; \boldsymbol{\psi}) \tag{3}$$

with respect to $\boldsymbol{\psi}$ to get $\boldsymbol{\psi}^{(t+1)}$. Note that $\hat{Q}$ separates into terms involving different subsets of parameters;

$$\hat{Q}(\boldsymbol{\psi};\boldsymbol{\psi}^{(t)}) \;=\; \frac{1}{m}\sum_{l=1}^{m}\sum_{i=1}^{r}\sum_{j=1}^{n_i}\log f(y_{ij}|\mathbf{u}_{t,l,i};\alpha,\boldsymbol{\beta}) + \frac{1}{m}\sum_{l=1}^{m}\sum_{i=1}^{r}\log\phi(\mathbf{u}_{t,l,i};\boldsymbol{\Sigma})\;.$$

where the first term only involves the parameters $(\alpha,\boldsymbol{\beta})$ and the second term only involves $\boldsymbol{\Sigma}$, so maximization with respect to these parameters can be done separately. Since the second piece is proportional to the log-likelihood associated with a random sample of size rm from the $\mathrm{N}_q(\mathbf{0},\boldsymbol{\Sigma})$ distribution, the maximizing value of $\boldsymbol{\Sigma}$ is simply

$$\boldsymbol{\Sigma}^{(t+1)} = \frac{\sum_{l=1}^{m}\sum_{i=1}^{r}\mathbf{u}_{t,l,i}\mathbf{u}'_{t,l,i}}{mr}\;.$$

As for the maximization with respect to $(\alpha,\boldsymbol{\beta})$, since $f(y_{ij}|\mathbf{u}_{t,l,i};\alpha,\boldsymbol{\beta})$ is the mass function of a $\mathrm{nb}\big(y_{ij};\alpha,\exp\{\mathbf{x}'_{ij}\boldsymbol{\beta}+\mathbf{z}'_{ij}\mathbf{u}_{t,l,i}\}\big)$, this is the log-likelihood corresponding to a negative binomial loglinear model with known *offsets* $(\mathbf{z}'_{ij}\mathbf{u}_{t,l,i})$ in the linear predictor. Thus, the techniques described in Appendix A, to maximize the *fixed effects model*, can be used to find $(\alpha^{(t+1)},\boldsymbol{\beta}^{(t+1)})$.

Booth and Hobert (1999) provide an automated Monte Carlo EM algorithm in which the value of m increases with the iteration number and convergence is diagnosed by approximating the Monte Carlo error. There are several parameters that must be chosen by the user in order to implement this algorithm. For the applications in this paper, we used (in Booth and Hobert's notation) $\delta_1 = 0.001$, $\delta_2 = 0.003$, $k = 5$, $\alpha = 0.15$ and we started the algorithm with $m = 20$. Finally, we note that modifications of the MCEM algorithm we have described involving importance sampling (Booth and Hobert, 1999) or even Markov chain Monte Carlo (McCulloch, 1997) can be used in cases in which rejection sampling becomes too inefficient.

3.3 Nonparametric maximum likelihood

In this section we relax the assumption of Gaussian random effects, and we instead assume that the $\mathbf{u}_i$'s are iid with (unknown) density g. The likelihood is therefore given by

$$L(\alpha, \boldsymbol{\beta}, g; \mathbf{y}) = \prod_{i=1}^{r} \int f(\mathbf{y}_i | \mathbf{u}_i; \alpha, \boldsymbol{\beta})\, g(\mathbf{u}_i)\, d\mathbf{u}_i \, . \tag{4}$$

For K a fixed positive integer, the results of Laird (1978) and Lindsay (1983) imply that we can take g to be an unknown discrete distribution with K mass points, $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1', \ldots, \boldsymbol{\zeta}_k')'$, and corresponding masses $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)'$. The vector of unknowns therefore becomes $\boldsymbol{\psi} = (\alpha, \boldsymbol{\beta}', \boldsymbol{\zeta}', \boldsymbol{\pi}')'$.

Consider again application of the (deterministic) EM algorithm with $\mathbf{u}$ playing the role of the missing data. Suppose that the result of the tth iteration of EM is $\alpha^{(t)}$, $\beta^{(t)}$ and $g^{(t)} = (\boldsymbol{\zeta}_k^{(t)}; \pi_k^{(t)})_{k=1}^{K}$. Because $g^{(t)}$ is a K-point distribution, the conditional distribution of $\mathbf{u}_i$ given $\mathbf{y}_i$ is also a K-point distribution with the same exact mass points as $g^{(t)}$. The probabilities, however, are given by $w_{ik}^{(t)}$, where

$$w_{ik}^{(t)} = \frac{\left[\prod_{j=1}^{n_i} f(y_{ij} | \mathbf{u}_i; \alpha^{(t)}, \mu_{ijk}^{(t)})\right] \pi_k^{(t)}}{\sum_{l=1}^{K} \left[\prod_{j=1}^{n_i} f(y_{ij} | \mathbf{u}_i; \alpha^{(t)}, \mu_{ijl}^{(t)})\right] \pi_l^{(t)}}$$

with $f(y_{ij} | \mathbf{u}_i; \alpha^{(t)}, \mu_{ijk}^{(t)})$ the mass function of a $\mathrm{nb}\Big(y_{ij}; \alpha^{(t)}, \mu_{ijk}^{(t)}\Big)$ with $\log \mu_{ijk}^{(t)} = \mathbf{x}_{ij}' \boldsymbol{\beta}^{(t)} + \mathbf{z}_{ij}' \boldsymbol{\zeta}_k^{(t)}$. Thus, unlike the parametric case, here we can write Q in closed form

$$Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(t)}) = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \sum_{k=1}^{K} w_{ik}^{(t)} \log \left[f(y_{ij} | \mathbf{u}_i; \alpha, \mu_{ijk}) \right] + \sum_{i=1}^{r} \sum_{k=1}^{K} w_{ik}^{(t)} \log \pi_k \, , \tag{5}$$

where $\log \mu_{ijk} = \mathbf{x}_{ij}' \boldsymbol{\beta} + \mathbf{z}_{ij}' \boldsymbol{\zeta}_k$.

In the subsequent M-step, $Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(t)})$ is maximized with respect to $\boldsymbol{\psi}$ to obtain $\boldsymbol{\psi}^{(t+1)}$. Since $\boldsymbol{\zeta}_k$ appears only in the linear predictor, its role in the maximization is essentially the same as that of $\boldsymbol{\beta}$ in the parametric case. The maximization with respect to $(\alpha, \boldsymbol{\beta}, \boldsymbol{\zeta})$ can be done independently of the maximization with respect to $\boldsymbol{\pi}$. This latter maximization must be done under the constraint that $\sum_{k=1}^{K} \pi_k = 1$. Using a Lagrange multiplier, λ, we need to solve the equations

$$\frac{\partial}{\partial \pi_k}\left[\sum_{i=1}^{r}\sum_{k=1}^{K} w_{ik}^{(t)} \log \pi_k - \lambda\left(\sum_{l=1}^{K}\pi_l - 1\right)\right] = \frac{1}{\pi_k}\sum_{i=1}^{r} w_{ik}^{(t)} - \lambda = 0 \ .$$

Summing these K equations leads to the result $\lambda = r$ which in turn leads to

$$\pi_k^{(t+1)} = \frac{1}{r}\sum_{i=1}^{r} w_{ik}^{(t)} \ .$$

The maximization with respect to α, $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$ is similar to the M-step in the MCEM setting. Maximizing (5) with respect to the index α is equivalent to solving the equation

$$N\left\{\log\alpha + 1 - \psi(\alpha)\right\} + \sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K} w_{ik}^{(t)}\left\{\psi(y_{ij}+\alpha) - \log(\mu_{ijk}+\alpha) - \frac{y_{ij}+\alpha}{\mu_{ijk}+\alpha}\right\} = 0 \ , \quad (6)$$

where $N = \sum_{i=1}^{r} n_i$ and $\psi(x) = \frac{\partial}{\partial x}\log\Gamma(x)$. As for $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$, we have to solve

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K} w_{ik}^{(t)}\left[\frac{y_{ij}-\mu_{ijk}}{1+\mu_{ijk}/\alpha}\right]\mathbf{x}_{ij} = \mathbf{0} \quad (7)$$

and the system

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i} w_{ik}^{(t)}\left[\frac{y_{ij}-\mu_{ijk}}{1+\mu_{ijk}/\alpha}\right]\mathbf{z}_{ij} = \mathbf{0} \ , \qquad k = 1,\ldots,K \ , \quad (8)$$

respectively. Equations (6), (7) and (8) can be solved easily using a weighted version of the (negative binomial) maximum likelihood estimators as discussed in Appendix A. The difference is that here we have to base the algorithm on an enlarged data structure, where K replicates of the original data are included, and where, in addition, each design row is extended by a K level dummy factor interacting with all elements in $\mathbf{z}_{ij}$.

One of the advantages of the NPML procedure is that it enables us to approximate the marginal likelihood (4) in a simple discrete way. Replacing g by $\hat{g}$ yields the nonparametric version of the log-likelihood function

$$l_K(\hat{\alpha}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\pi}}; \mathbf{y}) = \sum_{i=1}^{r} \log \left\{ \sum_{k=1}^{K} \Big[\prod_{j=1}^{n_i} f(y_{ij}|\mathbf{u}_i; \hat{\alpha}, \hat{\mu}_{ijk}) \Big] \hat{\pi}_k \right\}, \tag{9}$$

where $\hat{\boldsymbol{\mu}}$ comprises the parameter estimates $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}})$. This likelihood can be easily calculated by using the denominators of the weights w_{ik} at convergence. As an example of how (9) can be used, choosing $K = 1$ results in

$$\begin{aligned} l_1(\alpha, \boldsymbol{\mu}, \boldsymbol{\pi}; \mathbf{y}) &= \sum_{i=1}^{r} \log \Big[\prod_{j=1}^{n_i} f(y_{ij}|\mathbf{u}_i; \alpha, \mu_{ij}) \Big] = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \log[f(y_{ij}|\mathbf{u}_i; \alpha, \mu_{ij})] \\ &=: l_1(\alpha, \boldsymbol{\mu}; \mathbf{y}) \ , \end{aligned}$$

which is the log-likelihood of N independent negative binomial responses without any randomness in the linear predictor.

Another interesting example is the saturated version of (9), where every observed response y_{ij} is exactly modeled by its conditional means μ_{ijk}. That is, we allow for a total of N parameters in the considered model resulting in the log-likelihood

$$\begin{aligned} l_K(\alpha, \mathbf{y}, \boldsymbol{\pi}; \mathbf{y}) &= \sum_{i=1}^{r} \log \left\{ \sum_{k=1}^{K} \Big[\prod_{j=1}^{n_i} f(y_{ij}|\mathbf{u}_i; \alpha, y_{ij}) \Big] \pi_k \right\} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \log[f(y_{ij}|\mathbf{u}_i; \alpha, y_{ij})] \\ &= l_1(\alpha, \mathbf{y}; \mathbf{y}) \end{aligned}$$

for arbitrary $\boldsymbol{\pi}$'s satisfying $\sum_k \pi_k = 1$. Hence, $l_1(\alpha, \mathbf{y}; \mathbf{y}) = l_K(\alpha, \mathbf{y}, \boldsymbol{\pi}; \mathbf{y})$ for different values of K and this saturated likelihood is a monotone increasing function in α (which does not hold for the likelihood under the model (2)). We will use the maximum achievable value (at $\alpha = \infty$, the independent Poisson model) to define the respective deviance measure as minus twice the difference between maximum log-likelihood achieved and the maximum log-likelihood achievable under the model; that is,

$$\begin{aligned}\mathrm{dev}_K &= -2[l_K(\hat{\alpha}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\pi}}; \mathbf{y}) - l_1(\infty, \mathbf{y}; \mathbf{y})] \\ &= -2\left[l_K(\hat{\alpha}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\pi}}; \mathbf{y}) - \sum_{i=1}^{r}\sum_{j=1}^{n_i}\Big(y_{ij}\log y_{ij} - \log(y_{ij}!) - y_{ij}\Big)\right] . \qquad (10)\end{aligned}$$

The above definition of the deviance also enables us to chose an appropriate value K. In the examples we always start with $K = 1$ and increase K as long as there is evidence that dev_K gets smaller. Note that increasing K to $K+1$ introduces another $q+1$ parameters in our model.

In order to compare the NPML results to the parametric results, we calculate the first two moments of the estimated random effect distribution, that is, the moments with respect to $\hat{g}$ which are

$$\widehat{\mathrm{E}}(\mathbf{u}) = \sum_{k=1}^{K}\hat{\pi}_k\hat{\boldsymbol{\zeta}}_k \quad \text{and} \quad \widehat{\mathrm{Var}}(\mathbf{u}) = \sum_{k=1}^{K}\hat{\pi}_k(\hat{\boldsymbol{\zeta}}_k - \widehat{\mathrm{E}}(\mathbf{u}))(\hat{\boldsymbol{\zeta}}_k - \widehat{\mathrm{E}}(\mathbf{u}))' .$$

Standard errors of the fixed effects are then computed by replacing the conditional moments in Louis's (1982) variance formula by their nonparametric versions (see Appendix D for details).

4 Examples

4.1 Application to anchovy larvae data

Casella, Robson, Schwager and Youngs (1985) discuss data on the entrainment of Bay Anchovy larvae at the Bowline Point hydroelectric plant in New York State. The data, given in Appendix E, consists of two replicate larvae counts and associated water volumes measured in cubic meters. A preliminary analysis of the data indicates an obvious negative trend in the counts over time. The replicate measurements are also highly correlated.

Let (y_{i1}, y_{i2}) denote the ith replicate pair of counts, $i = 1, \ldots, 49$. We consider the following loglinear model for the expected counts:

$$\log \mu_{ij} = \beta_0 + \beta_1 \mathrm{Vol} + \beta_2 \mathrm{Day} + u_i \ , \tag{11}$$

where Vol is the logarithm of volume and Day is the date in August when the pair of counts were collected. Since it is natural to expect the means to be proportional to volume we expect β_1 to be close to 1. Fitting this model with the replicate effects, u_i, fixed indicates substantial overdispersion with respect to Poisson variability (Deviance = 445, df = 48). Thus, we consider the model in which the counts are conditionally independent $\mathrm{nb}(\alpha, \mu_{ij})$ variables given the replicate effects, which we assume are iid $N(0, \sigma^2)$. The parameter estimates and their standard errors are given in Table 1. Also given are the corresponding Poisson GLMM and NPML fits. The Poisson and negative binomial model fits were obtained using the SAS NLMIXED procedure, and the SAS code for the negative binomial model is given in Appendix D. It is interesting to note that the Poisson GLMM fit produces an estimate of β_1 far below the anticipated value of 1, although its standard error is large enough for the difference to be statistically insignificant.

In the NPML analysis of the negative binomial mixed model with K mass points $-2l_K$ is reduced only from 922.7 ($K = 1$, independence model) to 913.5 ($K = 2$). Since the saturated (independent) Poisson log-likelihood is $-2l_1(\infty, \mathbf{y}; \mathbf{y}) = 558.6$ we get $\mathrm{dev}_2 = 354.9$ for the deviance defined in (10). Compared to $\mathrm{dev}_1 = 364.1$, the gain of 9.2 deviance units is due to the fact that the model now allows for two more parameters. The mass points (with their associated masses) are $\hat{\zeta}_1 = 3.25$ ($\hat{\pi}_1 = 0.36$) and $\hat{\zeta}_2 = 2.58$ ($\hat{\pi}_2 = 0.64$) reflecting a mean and standard deviation of $\widehat{\mathrm{E}}(u) = 2.82$ and $\widehat{\mathrm{Var}}(u)^{1/2} = 0.32$, respectively. But we have to submit both moments to a careful

Table 1: *Different model fits for the anchovy data.*

	Fit			
	Poisson		Negative Binomial	
Variable	parametric	NPML(K=5)	parametric	NPML(K=2)
Fixed effects				
Constant	4.23 (1.87)	3.77	1.65 (2.42)	2.82
Vol	0.31 (0.46)	0.37 (0.14)	0.96 (0.60)	0.67 (0.69)
Day	-0.09 (0.01)	-0.08 (.004)	-0.10 (0.01)	-0.10 (0.01)
Random effects				
Intercept ($\sqrt{\sigma^2}$)	0.22 (0.05)	0.47	0.14 (0.05)	0.32
Index (α)			7.27 (1.77)	5.99 (1.24)

examination because they are based on a two-point distribution. A value of $K > 2$ would need more parameters but would only slightly change the log-likelihood (e.g. $\text{dev}_3 = 350.2$). It is interesting to note that the index parameter increases from $\hat{\alpha} = 3.74$ ($K = 1$) to $\hat{\alpha} = 5.99$ ($K = 2$). To recognize the impact of the index parameter we also want to point out the Poisson GLMM deviance of $\text{dev}_5 = 656.0$, which is for a model with 11 parameters.

4.2 Application to epilepsy data

Thall and Vail (1990) describe an experiment in which 59 epileptics were randomly assigned to one of two treatment groups. The number of seizures experienced by each patient over four consecutive two-week periods following treatment were recorded. A baseline count of the number of seizures during the eight week period prior to treatment is also available. Let y_{ij} denote the number of seizures experienced by the ith patient over the jth two-week period, where $i = 1, \ldots, 59$ and $j = 1, 2, 3, 4$. Thall and Vail consider a multiplicative model for the mean number of seizures with the following covariates.

Base:	the logarithm of baseline/4
Age:	the logarithm of the patients age in years
Trt:	the 0/1 indicator of treatment group
V4:	the 0/1 indicator for the fourth two week period

Their model also includes the Base$\times$Trt interaction term. In their analysis, dependence between the four counts from the same subject was handled using semi-parametric GEE methodology.

Breslow and Clayton (1993) consider several models for the same data. Their Model I is a Poisson loglinear model for the mean with the same linear predictor suggested by Thall and Vail; Model II includes a subject level random effect to account for dependence between the repeated counts; and Model III includes an additional unit level random effect to account for overdispersion with respect to the Poisson distribution. In a fourth model, the V4 indicator is replaced by a trend variable, Visit, defined as $(j-2.5)/5$ for $j=1,2,3,4$, and a bivariate subject level random intercept and slope. Breslow and Clayton fit these models using PQL, which can be viewed as an approximate ML method with normal random effects.

In Model IV of Breslow and Clayton, conditional on the random effects the counts are implicitly assumed to be independent Poisson variables. However, the fixed effects version of their model, allowing subject specific intercepts and slopes, clearly indicates overdispersion with respect to Poisson variability with a deviance statistic of 280.7 with 118 degrees of freedom. In our analysis of this data we account for this overdispersion by assuming the counts are conditionally independent negative binomial variables.

Our full model is as follows. Conditional upon subject level bivariate normal random effect $\mathbf{u}_i$, $i = 1, \ldots, 59$, the counts, y_{ij}, are independent $\text{nb}(\alpha, \mu_{ij})$ variables, with means satisfying the loglinear model:

$$\log \mu_{ij} = \beta_0 + \beta_1 \text{Base} + \beta_2 \text{Trt} + \beta_3 \text{Base} \times \text{Trt} + \beta_4 \text{Age} + \beta_5 \text{Visit} + u_{i0} + u_{i1} \text{Visit} \,, \quad (12)$$

where the $\mathbf{u}_i$'s are iid $N_2(\mathbf{0}, \mathbf{\Sigma})$

In Table 2, note the difference between the variance component estimates obtained using PQL and the ML estimates for the corresponding Poisson model obtained using NLMIXED. Note also that the random slope is significantly greater than zero.

The fit of the full negative binomial model using NLMIXED was very unstable. Different starting values led to different estimates and very different standard errors. Application of the MCEM algorithm in this problem suggests that the random slope variance is zero. (The EM algorithm essentially grinds to a halt when one of the variance components is zero. The MCEM algorithm was run for a large number of iterations, with all of the estimates, except for slope variance and the covariance, converging quickly. These latter two estimates appeared to be slowly converging towards zero.)

As in the example before, we apply the NPML method again and decrease $-2l_K$ from 1295.0 ($K = 1$), 1254.1 ($K = 2$), etc., to 1234.7 ($K = 6$) by increasing K. A larger value of K yields only minor log-likelihood changes (e.g. $-2l_7 = 1234.5$), which is often the result of estimated mass points that are no longer distinguishable (e.g. for $K \geq 8$). The corresponding saturated log-likelihood independent Poisson model gives $-2l_1(\infty, \mathbf{y}; \mathbf{y}) = 766.0$, and we therefore select the 22 parameter ($K = 6$) model with $\text{dev}_6 = 468.7$, which is essentially smaller than $\text{dev}_1 = 529.0$, the deviance of the corresponding fixed effects negative binomial model. This is comparable to the deviance value 261.8 from an overparameterized negative binomial model that allows for both fixed patient-specific intercepts and slopes, and hence includes 119 parameters. Since we used six mass points, which are $\hat{\boldsymbol{\zeta}}_1 = (-0.82, 0.52)'$, $\hat{\boldsymbol{\zeta}}_2 = (-0.97, -3.83)'$, $\hat{\boldsymbol{\zeta}}_3 = (-1.60, -0.24)'$, $\hat{\boldsymbol{\zeta}}_4 = (-2.03, -0.50)'$, $\hat{\boldsymbol{\zeta}}_5 = (-2.82, -1.80)'$, and $\hat{\boldsymbol{\zeta}}_6 = (-2.93, 1.55)'$, with $\hat{\boldsymbol{\pi}} = (0.07, 0.02, 0.49, 0.30, 0.05, 0.07)'$, we should be able to interpret the estimated moments of the bivariate random effects distribution. We also found that an increase of K also results in increasing variances of the random effects and in larger estimates of α.

5 Discussion

The Poisson model, although often suggested by the underlying process, suffers from its lack of flexibility in modeling variances. The resulting overdispersion from the model can result in biased estimates of the other parameters and, among other problems, difficulties in interpretations. The negative binomial alternative, while maintaining much of the interpretation of the Poisson model, allows a more flexible variance structure which reduces the bias in parameter estimates. This last point is illustrated in the two examples presented, where the negative binomial model leads to more meaningful parameter estimates and inferences.

In addition to the modeling benefits, we have seen that the negative binomial model can be made computationally feasible. If the overdispersion random effect, v, is modeled as a log-Gamma, it can be integrated out analytically in the likelihood. In contrast, if v is assumed normal (log-normal) then analytical integration is not possible. The resulting likelihood maximization must either be done numerically, or v must be treated as missing data in an EM algorithm, greatly increasing the ratio of missing to observed data and hence adversely effecting the performance of EM.

For our negative binomial model, the SAS procedure NLMIXED can be used.

However, for analyses that go beyond the current capabilities of this procedure, an MCEM algorithm is straightforward to implement and performs quite satisfactorily.

We have also seen that a fully nonparametric approach, where no form for the random effects distribution is specified, is an attractive and computationally feasible alternative model.

A Algorithms for fitting the fixed effects model

Suppose that $Y_1, \ldots, Y_n$ are independent with $Y_i \sim \mathrm{nb}(\alpha, \mu_i)$ and $\log \mu_i = \mathbf{x}_i' \boldsymbol{\beta}$. We consider only maximum likelihood (ML) estimation of $(\alpha, \boldsymbol{\beta})$. (See Lawless (1987) for details concerning an alternative estimation method based on the method-of-moments.) Using (1), the log-likelihood associated with the data $\mathbf{y} = (y_1, \ldots, y_n)$ is given by,

$$\begin{aligned} l(\alpha, \boldsymbol{\beta}) &= n \left\{ \alpha \log \alpha - \log \Gamma(\alpha) \right\} \\ &\quad + \sum_{i=1}^{n} \left\{ \log \Gamma(y_i + \alpha) - \log y_i! - \alpha \log(\mu_i + \alpha) + y_i \log \left(\frac{\mu_i}{\mu_i + \alpha} \right) \right\} \quad (13) \end{aligned}$$

The ML estimates $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ jointly maximize (13) with respect to the parameters. First, define

$$l_\alpha(\alpha, \boldsymbol{\beta}) = \frac{\partial l(\alpha, \boldsymbol{\beta})}{\partial \alpha} \text{ and } l_{\boldsymbol{\beta}}(\alpha, \boldsymbol{\beta}) = \frac{\partial l(\alpha, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} ,$$

and, extending this notation in an obvious way, let $l_{\alpha\alpha}$, $l_{\alpha\boldsymbol{\beta}}$ and $l_{\boldsymbol{\beta}\boldsymbol{\beta}}$ denote the second derivatives. If the log-likelihood function is sufficiently regular, the ML estimates can be obtained directly by solving the likelihood equations,

$$l_\alpha(\alpha, \boldsymbol{\beta}) = n \left\{ \log \alpha + 1 - \psi(\alpha) \right\} + \sum_{i=1}^{n} \left\{ \psi(y_i + \alpha) - \log(\mu_i + \alpha) - \frac{y_i + \alpha}{\mu_i + \alpha} \right\} = 0 \,, \quad (14)$$

where ψ denotes the digamma function; that is, $\psi(z) = \frac{\partial}{\partial z} \log \Gamma(z)$, and

$$l_{\boldsymbol{\beta}}(\alpha, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ y_i - \frac{y_i + \alpha}{\mu_i + \alpha} \mu_i \right\} \mathbf{x}_i = \sum_{i=1}^{n} \frac{(y_i - \mu_i)\, \mathbf{x}_i}{1 + \mu_i / \alpha} = \mathbf{0}\,. \tag{15}$$

Note that putting $\alpha = \infty$ in (15) yields the scoring equation for the corresponding Poisson generalized linear model (GLM).

Defining $\psi'(z) = \frac{\partial^2}{\partial z^2} \log \Gamma(z)$, the second derivatives are

$$\begin{aligned}
l_{\alpha\alpha}(\alpha, \boldsymbol{\beta}) &= n \left\{ \frac{1}{\alpha} - \psi'(\alpha) \right\} + \sum_{i=1}^{n} \left\{ \psi'(y_i + \alpha) - \frac{1}{\mu_i + \alpha} - \frac{\mu_i - y_i}{(\mu_i + \alpha)^2} \right\} \\
l_{\alpha\boldsymbol{\beta}}(\alpha, \boldsymbol{\beta}) &= \sum_{i=1}^{n} \left\{ \frac{\mu_i (y_i - \mu_i)}{(\mu_i + \alpha)^2} \right\} \mathbf{x}_i \\
l_{\boldsymbol{\beta}\boldsymbol{\beta}}(\alpha, \boldsymbol{\beta}) &= -\alpha \sum_{i=1}^{n} \left\{ \frac{\mu_i (y_i + \alpha)}{(\mu_i + \alpha)^2} \right\} \mathbf{x}_i \mathbf{x}_i' = -\mathbf{X}'\, \mathbf{D}(\alpha, \boldsymbol{\beta})\, \mathbf{X}
\end{aligned}$$

where $\mathbf{D}(\alpha, \boldsymbol{\beta})$ is an $n \times n$ diagonal matrix with ith diagonal element given by

$$d_{ii} = [\alpha \mu_i (y_i + \alpha)] / (\mu_i + \alpha)^2.$$

Finally, let

$$I_{\boldsymbol{\beta}\boldsymbol{\beta}}(\alpha, \boldsymbol{\beta}) = -\mathrm{E}[l_{\boldsymbol{\beta}\boldsymbol{\beta}}(\alpha, \boldsymbol{\beta})] = \mathbf{X}'\, \mathrm{E}[\mathbf{D}(\alpha, \boldsymbol{\beta})]\, \mathbf{X} = \mathbf{X}'\, \mathbf{W}(\alpha, \boldsymbol{\beta})\, \mathbf{X}\ ,$$

where $\mathbf{W}(\alpha, \boldsymbol{\beta})$ is an $n \times n$ diagonal matrix of iterative weights with ith diagonal element given by $w_{ii} = \mu_i / (1 + \mu_i / \alpha)$.

We now describe three different iterative methods of solving the likelihood equations. Let t and $t + 1$ denote the *current* and *next* iterations, respectively.

M METHOD: A standard approach to solving the likelihood equations, (14) and (15), is direct application of the Newton-Raphson algorithm:

$$\begin{pmatrix} \alpha^{(t+1)} \\ \boldsymbol{\beta}^{(t+1)} \end{pmatrix} = \begin{pmatrix} \alpha^{(t)} \\ \boldsymbol{\beta}^{(t)} \end{pmatrix} - \begin{pmatrix} l_{\alpha\alpha}(\alpha^{(t)}, \boldsymbol{\beta}^{(t)}) & l_{\alpha\boldsymbol{\beta}}(\alpha^{(t)}, \boldsymbol{\beta}^{(t)}) \\ l'_{\alpha\boldsymbol{\beta}}(\alpha^{(t)}, \boldsymbol{\beta}^{(t)}) & l_{\boldsymbol{\beta}\boldsymbol{\beta}}(\alpha^{(t)}, \boldsymbol{\beta}^{(t)}) \end{pmatrix}^{-1} \begin{pmatrix} l_{\alpha}(\alpha^{(t)}, \boldsymbol{\beta}^{(t)}) \\ l_{\boldsymbol{\beta}}(\alpha^{(t)}, \boldsymbol{\beta}^{(t)}) \end{pmatrix} . \tag{16}$$

One disadvantage of this algorithm is the lack of an obvious starting value.

CM METHOD: A popular alternative to direct Newton-Raphson is to iterate between solving for α and $\boldsymbol{\beta}$; that is, solve for the next value of α conditionally given the current $\boldsymbol{\beta}$ value and vice-versa.

For fixed α, the likelihood equations for $\boldsymbol{\beta}$ are those of a GLM with log link and variance function, $V(\mu) = \mu(1 + \mu/\alpha)$. Thus, these equations can be solved by iteratively reweighted least squares (IRLS). Specifically, let $\mathbf{z}(\boldsymbol{\beta})$ be the $n \times 1$ working response vector whose ith element is $\log \mu_i + (y_i - \mu_i)/\mu_i$. Then given $(\alpha^{(t)}, \boldsymbol{\beta}^{(t)})$, we update $\boldsymbol{\beta}$ using

$$\boldsymbol{\beta}^{(t+1)} = (\mathbf{X}' \mathbf{W}(\alpha^{(t)}, \boldsymbol{\beta}^{(t)}) \mathbf{X})^{-1} \mathbf{X}' \mathbf{W}(\alpha^{(t)}, \boldsymbol{\beta}^{(t)}) \mathbf{z}(\boldsymbol{\beta}^{(t)}) \,. \tag{17}$$

For a fixed value of $\boldsymbol{\beta}$, the likelihood equation for α involves only one unknown and can easily be solved for numerically using a univariate Newton-Raphson algorithm:

$$\alpha^{(t+1)} = \alpha^{(t)} - \frac{l_\alpha(\alpha^{(t)}, \boldsymbol{\beta}^{(t+1)})}{l_{\alpha\alpha}(\alpha^{(t)}, \boldsymbol{\beta}^{(t+1)})} \,. \tag{18}$$

We now consider starting values for this algorithm. Since the Poisson loglinear model is the limiting case as $\alpha \to \infty$, it makes sense to use the values $\alpha = \infty$ and $\mu_i = y_i + \delta$, for some small $\delta > 0$, in (17) to initiate the IRLS procedure. However, $\alpha^{(0)} = \infty$ is not a valid starting value for use in (18). A starting value for α can be obtained by the method of moments: Assuming that the counts are iid with mean μ and setting the sample mean and variance equal to the corresponding population versions yields the equations $\bar{y} = \mu$ and $s^2 = \mu(1 + \mu/\alpha)$. This yields the starting value $\alpha^{(0)} = \bar{y}^2/(s^2 - \bar{y})$.

HCM METHOD: Note that after the initial step, (17) is equivalent to the Fisher scoring algorithm

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + I^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\alpha^{(t)}, \boldsymbol{\beta}^{(t)}) l_{\boldsymbol{\beta}}(\alpha^{(t)}, \boldsymbol{\beta}^{(t)})\,. \tag{19}$$

This suggests a hybrid maximization algorithm in which α and $\boldsymbol{\beta}$ are alternately updated using (18) and (19). That is, a single step of Fisher scoring is followed by a single Newton-Raphson update using (18) and this is followed by another single step of Fisher scoring, etc.

An alternative method of finding the MLE's is to consider $v_1, \ldots, v_n$ as missing data and apply the EM algorithm (Dempster et al., 1977). While EM and its variants (eg. ECM) are effective in this situation, they are much slower than the Newton-Raphson-type algorithms described above.

B Maximizing $\hat{Q}$ with respect to $\boldsymbol{\psi}$

As in Section 3.2, the likelihood breaks into two parts and we can write

$$\hat{Q}(\boldsymbol{\psi}; \boldsymbol{\psi}^{(t)}) \quad = \quad \frac{1}{m}\sum_{l=1}^{m}\sum_{i=1}^{r}\sum_{j=1}^{n_i} \log f(y_{ij}|\mathbf{u}_{i,t,l}; \alpha, \boldsymbol{\beta}) + \frac{1}{m}\sum_{l=1}^{m}\sum_{i=1}^{r} \log f(\mathbf{u}_{i,t,l}; \boldsymbol{\theta}),$$

so the maximization with respect to $(\alpha, \boldsymbol{\beta})$ can be done independently of that with respect to $\boldsymbol{\theta}$. The first piece can be maximized via Newton-Raphson. Here are the derivatives:

$$m\frac{\partial}{\partial\alpha} Q_m(\boldsymbol{\psi}; \boldsymbol{\psi}^{(t)}) = mN(1 + \log\alpha - \psi(\alpha)) + m\left[\sum_{i=1}^{r}\sum_{j=1}^{n_i} \psi(y_{ij} + \alpha)\right]$$
$$- \sum_{l=1}^{m}\sum_{i=1}^{r}\sum_{j=1}^{n_i}\left[\frac{\alpha + y_{ij}}{\alpha + \exp\{\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_{i,t,l}\}} + \log(\alpha + \exp\{\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_{i,t,l}\})\right]$$

where $N = \sum_{i=1}^{r} n_i$.

$$m\frac{\partial^2}{\partial\alpha^2}Q_m(\boldsymbol{\psi};\boldsymbol{\psi}^{(t)}) = mN\left(\frac{1}{\alpha} - \psi'(\alpha)\right) + m\left[\sum_{i=1}^{r}\sum_{j=1}^{n_i}\psi'(y_{ij}+\alpha)\right]$$
$$-\sum_{l=1}^{m}\sum_{i=1}^{r}\sum_{j=1}^{n_i}\left[\frac{\exp\{\mathbf{x}_{ij}'\boldsymbol{\beta}+\mathbf{z}_{ij}'\mathbf{u}_{i,t,l}\} - y_{ij}}{\left(\alpha+\exp\{\mathbf{x}_{ij}'\boldsymbol{\beta}+\mathbf{z}_{ij}'\mathbf{u}_{i,t,l}\}\right)^2} + \frac{1}{\alpha+\exp\{\mathbf{x}_{ij}'\boldsymbol{\beta}+\mathbf{z}_{ij}'\mathbf{u}_{i,t,l}\}}\right]$$

$$m\frac{\partial}{\partial\boldsymbol{\beta}}Q_m(\boldsymbol{\psi};\boldsymbol{\psi}^{(t)}) = \sum_{l=1}^{m}\sum_{i=1}^{r}\sum_{j=1}^{n_i}\left[\frac{\alpha(y_{ij}-\exp\{\mathbf{x}_{ij}'\boldsymbol{\beta}+\mathbf{z}_{ij}'\mathbf{u}_{i,t,l}\})}{\alpha+\exp\{\mathbf{x}_{ij}'\boldsymbol{\beta}+\mathbf{z}_{ij}'\mathbf{u}_{i,t,l}\}}\right]\mathbf{x}_{ij}$$

$$m\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}Q_m(\boldsymbol{\psi};\boldsymbol{\psi}^{(t)}) = -\sum_{l=1}^{m}\sum_{i=1}^{r}\sum_{j=1}^{n_i}\mathbf{x}_{ij}\left[\frac{\alpha(\alpha+y_{ij})\exp\{\mathbf{x}_{ij}'\boldsymbol{\beta}+\mathbf{z}_{ij}'\mathbf{u}_{i,t,l}\}}{\left(\alpha+\exp\{\mathbf{x}_{ij}'\boldsymbol{\beta}+\mathbf{z}_{ij}'\mathbf{u}_{i,t,l}\}\right)^2}\right]\mathbf{x}_{ij}'$$

$$m\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\alpha}Q_m(\boldsymbol{\psi};\boldsymbol{\psi}^{(t)}) = \sum_{l=1}^{m}\sum_{i=1}^{r}\sum_{j=1}^{n_i}\left[\frac{\exp\{\mathbf{x}_{ij}'\boldsymbol{\beta}+\mathbf{z}_{ij}'\mathbf{u}_{i,t,l}\}(y_{ij}-\exp\{\mathbf{x}_{ij}'\boldsymbol{\beta}+\mathbf{z}_{ij}'\mathbf{u}_{i,t,l}\})}{\left(\alpha+\exp\{\mathbf{x}_{ij}'\boldsymbol{\beta}+\mathbf{z}_{ij}'\mathbf{u}_{i,t,l}\}\right)^2}\right]\mathbf{x}_{ij}$$

The second piece is just a normal likelihood so the maximizing value of $\boldsymbol{\Sigma}$ is

$$\frac{\sum_{l=1}^{m}\sum_{i=1}^{r}\mathbf{u}_{i,t,l}\mathbf{u}_{i,t,l}'}{mr}$$

C Rejection Sampling in MCEM

To sample from $f(\mathbf{u}|\mathbf{y};\boldsymbol{\psi}^{(t)})$, the factorization

$$f(\mathbf{u}|\mathbf{y};\boldsymbol{\psi}) = \prod_{i=1}^{r} f(\mathbf{u}_i|\mathbf{y}_i;\boldsymbol{\psi})$$

allows us to restrict attention to sampling from

$$f(\mathbf{u}_i|\mathbf{y}_i;\boldsymbol{\psi}) = \frac{f(\mathbf{y}_i|\mathbf{u}_i;\alpha,\boldsymbol{\beta})\phi(\mathbf{u}_i;\boldsymbol{\Sigma})}{f(\mathbf{y}_i;\boldsymbol{\psi})} = c\phi(\mathbf{u}_i;\boldsymbol{\Sigma})\prod_{j=1}^{n_i}\left(\frac{\alpha}{\mu_{ij}+\alpha}\right)^{\alpha}\left(\frac{\mu_{ij}}{\mu_{ij}+\alpha}\right)^{y_{ij}} \quad (20)$$

where c does not depend on $\mathbf{u}_i$. (Because we are describing how to sample from $f(\mathbf{u}|\mathbf{y};\boldsymbol{\psi}^{(t)})$, we should be writing all of the components of $\boldsymbol{\psi}$ with a (t) superscript. However, these superscripts are notationally cumbersome and are therefore not used.)

Equation (20) shows that we can use a rejection sampler (Robert and Casella, 1999) with candidate $\phi(\cdot;\boldsymbol{\Sigma})$ if we can calculate

$$\sup_{\mathbf{u}_i}\prod_{j=1}^{n_i}\left(\frac{\alpha}{\mu_{ij}+\alpha}\right)^{\alpha}\left(\frac{\mu_{ij}}{\mu_{ij}+\alpha}\right)^{y_{ij}} .$$

Since $\mu_{ij}=\exp\{\mathbf{x}_{ij}'\boldsymbol{\beta}+\mathbf{z}_{ij}'\mathbf{u}_i\}$, an equivalent problem is to find the $\mathbf{u}_i$ that maximizes

$$\alpha n_i\log\alpha-\sum_{j=1}^{n_i}(\alpha+y_{ij})\log(\mu_{ij}+\alpha)+\sum_{j=1}^{n_i}y_{ij}\log\mu_{ij} \tag{21}$$

and this can be accomplished using Newton-Raphson. The first derivative of (21) with respect to $\mathbf{u}_i$ is

$$\alpha\sum_{j=1}^{n_i}\mathbf{z}_{ij}\left[\frac{y_{ij}-\mu_{ij}}{\mu_{ij}+\alpha}\right]$$

and the second derivative is

$$-\sum_{j=1}^{n_i}\mathbf{z}_{ij}\left[\frac{\alpha\mu_{ij}(\alpha+y_{ij})}{(\mu_{ij}+\alpha)^2}\right]\mathbf{z}_{ij}' .$$

Thus, for $l=1,\ldots,m$, we are able to simulate $\mathbf{u}_{t,l}$ by simulating r independent q-vectors; that is, $\mathbf{u}_{t,l}=(\mathbf{u}_{t,l,1}',\ldots,\mathbf{u}_{t,l,r}')'$.

D NPML Standard Errors

We will make use of the observed information matrix and calculate it by applying Louis' formula $F_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\psi};\mathbf{y})=\mathrm{E}(-l_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\psi};\mathbf{y},\mathbf{u})|\mathbf{y};\boldsymbol{\psi})-\mathrm{Var}(l_{\boldsymbol{\psi}}(\boldsymbol{\psi};\mathbf{y},\mathbf{u})|\mathbf{y};\boldsymbol{\psi})$ in order to get standard errors of the estimates of $\boldsymbol{\psi}=(\alpha,\boldsymbol{\beta})$ (see Louis, 1982). Generally, the variance-covariance matrix of an EM estimate $\hat{\boldsymbol{\psi}}$ is then estimated by the inverse of

$$F_{\boldsymbol{\psi}\boldsymbol{\psi}}(\hat{\boldsymbol{\psi}};\mathbf{y})=\mathrm{E}(-l_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\psi};\mathbf{y},\mathbf{u})|\mathbf{y};\hat{\boldsymbol{\psi}})-\mathrm{Var}(l_{\boldsymbol{\psi}}(\boldsymbol{\psi};\mathbf{y},\mathbf{u})|\mathbf{y};\hat{\boldsymbol{\psi}}) . \tag{22}$$

Because the complete log-likelihood function factors under our model as

$$l(\boldsymbol{\psi};\mathbf{y},\mathbf{u})=\log f(\mathbf{y},\mathbf{u};\boldsymbol{\psi})=\sum_{i=1}^{r}\log f(\mathbf{u}_i)+\sum_{i=1}^{r}\sum_{j=1}^{n_i}\log\mathrm{nb}(y_{ij}|\mathbf{u}_i;\boldsymbol{\psi}) ,$$

where $f(\mathbf{u})$ denotes an arbitrary random effect density, the complete $\boldsymbol{\psi}$-score is based on the conditionally assumed negative binomial model. Let $l_{\boldsymbol{\psi}}(\boldsymbol{\psi}; y_{ij}, \mathbf{u}_i) = \partial \log \mathrm{nb}(y_{ij}|\mathbf{u}_i; \boldsymbol{\psi})/\partial \boldsymbol{\psi}$ denote such a score contribution of the jth observation in the ith cluster. Note that each $l_{\boldsymbol{\psi}}(\boldsymbol{\psi}; y_{ij}, \mathbf{u}_i)$, $j = 1, \ldots, n_i$, depends on the same $\mathbf{u}_i$ under the model (2). This allows us to use the representation $l_{\boldsymbol{\psi}}(\boldsymbol{\psi}; \mathbf{y}_i, \mathbf{u}_i) = \sum_{j=1}^{n_i} l_{\boldsymbol{\psi}}(\boldsymbol{\psi}; y_{ij}, \mathbf{u}_i)$ for the score of the ith cluster.

Since $\mathrm{E}(l_{\boldsymbol{\psi}}(\boldsymbol{\psi}; \mathbf{y}, \mathbf{u})|\mathbf{y}; \hat{\boldsymbol{\psi}}) = \mathbf{0}$ just defines the EM estimate, the relevant part of the so-called missing information reduces to

$$\mathrm{Var}(l_{\boldsymbol{\psi}}(\boldsymbol{\psi}; \mathbf{y}, \mathbf{u})|\mathbf{y}; \hat{\boldsymbol{\psi}}) = \mathrm{E}(l_{\boldsymbol{\psi}}(\boldsymbol{\psi}; \mathbf{y}, \mathbf{u}) l'_{\boldsymbol{\psi}}(\boldsymbol{\psi}; \mathbf{y}, \mathbf{u})|\mathbf{y}; \hat{\boldsymbol{\psi}}) ,$$

and we have

$$\begin{aligned}\mathrm{E}\big(l_{\boldsymbol{\psi}}(\boldsymbol{\psi}; \mathbf{y}, \mathbf{u}) l'_{\boldsymbol{\psi}}(\boldsymbol{\psi}; \mathbf{y}, \mathbf{u})\big|\mathbf{y}; \boldsymbol{\psi}\big) &= \sum_{i=1}^{r} \mathrm{E}\big(l_{\boldsymbol{\psi}}(\boldsymbol{\psi}; \mathbf{y}_i, \mathbf{u}_i) l'_{\boldsymbol{\psi}}(\boldsymbol{\psi}; \mathbf{y}_i, \mathbf{u}_i)\big|\mathbf{y}_i; \boldsymbol{\psi}\big) \\ &\quad + \sum_{i=1}^{r} \sum_{i' \neq i} \mathrm{E}\big(l_{\boldsymbol{\psi}}(\boldsymbol{\psi}; \mathbf{y}_i, \mathbf{u}_i)\big|\mathbf{y}_i; \boldsymbol{\psi}\big) \mathrm{E}\big(l'_{\boldsymbol{\psi}}(\boldsymbol{\psi}; \mathbf{y}_{i'}, \mathbf{u}_{i'})\big|\mathbf{y}_{i'}; \boldsymbol{\psi}\big)\end{aligned}$$

by applying the NPML technique. The K point approximation to the conditional means results in

$$\begin{aligned}\mathrm{E}\big(l_{\boldsymbol{\psi}}(\boldsymbol{\psi}; \mathbf{y}, \mathbf{u}) l'_{\boldsymbol{\psi}}(\boldsymbol{\psi}; \mathbf{y}, \mathbf{u})\big|\mathbf{y}; \boldsymbol{\psi}\big) &\approx \sum_{i=1}^{r} \sum_{k=1}^{K} l_{\boldsymbol{\psi}}(\boldsymbol{\psi}; \mathbf{y}_i, \boldsymbol{\zeta}_k) l'_{\boldsymbol{\psi}}(\boldsymbol{\psi}; \mathbf{y}_i, \boldsymbol{\zeta}_k) w_{ik} \\ &+ \sum_{i=1}^{r} \sum_{i' \neq i} \sum_{k=1}^{K} \sum_{k'=1}^{K} l_{\boldsymbol{\psi}}(\boldsymbol{\psi}; \mathbf{y}_i, \boldsymbol{\zeta}_k) w_{ik} l'_{\boldsymbol{\psi}}(\boldsymbol{\psi}; \mathbf{y}_{i'}, \boldsymbol{\zeta}_{k'}) w_{i'k'} ,\end{aligned}$$

with $l_{\boldsymbol{\psi}}(\boldsymbol{\psi}; \mathbf{y}_i, \boldsymbol{\zeta}_k) = \sum_{j=1}^{n_i} l_{\boldsymbol{\psi}}(\boldsymbol{\psi}; y_{ij}, \boldsymbol{\zeta}_k)$, where these scores are defined on the left sides of formulas (6), (7), and (8). Hence,

$$\begin{aligned}\mathrm{Var}\big(l_{\boldsymbol{\psi}}(\boldsymbol{\psi}; \mathbf{y}, \mathbf{u})\big|\mathbf{y}; \hat{\boldsymbol{\psi}}\big) &\approx \sum_{i=1}^{r} \sum_{k=1}^{K} l_{\boldsymbol{\psi}}(\boldsymbol{\psi}; \mathbf{y}_i, \hat{\boldsymbol{\zeta}}_k) l'_{\boldsymbol{\psi}}(\boldsymbol{\psi}; \mathbf{y}_i, \hat{\boldsymbol{\zeta}}_k) \hat{w}_{ik} \\ &- \sum_{i=1}^{r} \sum_{k=1}^{K} \sum_{k'=1}^{K} l_{\boldsymbol{\psi}}(\boldsymbol{\psi}; \mathbf{y}_i, \hat{\boldsymbol{\zeta}}_k) \hat{w}_{ik} l'_{\boldsymbol{\psi}}(\boldsymbol{\psi}; \mathbf{y}_i, \hat{\boldsymbol{\zeta}}_{k'}) \hat{w}_{ik'} . \quad (23)\end{aligned}$$

As already shown in the derivations of Appendix B, the estimated complete information matrix $\mathrm{E}(-l_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\psi};\mathbf{y},\mathbf{u})|\mathbf{y};\boldsymbol{\psi})$ consists of

$$\begin{aligned}
\mathrm{E}(l_{\alpha\alpha}(\boldsymbol{\psi};\mathbf{y},\mathbf{u})|\mathbf{y};\hat{\boldsymbol{\psi}}) &= N\left\{\frac{1}{\hat{\alpha}} - \psi'(\hat{\alpha})\right\} \\
&+ \sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K}\left\{\psi'(y_{ij}+\hat{\alpha}) - \frac{1}{\hat{\mu}_{ijk}+\hat{\alpha}} - \frac{\hat{\mu}_{ijk}-y_{ij}}{(\hat{\mu}_{ijk}+\hat{\alpha})^2}\right\}\hat{w}_{ik} \qquad (24)\\
\mathrm{E}(l_{\alpha\boldsymbol{\beta}}(\boldsymbol{\psi};\mathbf{y},\mathbf{u})|\mathbf{y};\hat{\boldsymbol{\psi}}) &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K}\mathbf{x}_{ij}\left\{\frac{\hat{\mu}_{ijk}(y_{ij}-\hat{\mu}_{ijk})}{(\hat{\mu}_{ijk}+\hat{\alpha})^2}\right\}\hat{w}_{ik} \qquad (25)\\
\mathrm{E}(l_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\psi};\mathbf{y},\mathbf{u})|\mathbf{y};\hat{\boldsymbol{\psi}}) &= -\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K}\mathbf{x}_{ij}\left\{\frac{\hat{\alpha}\hat{\mu}_{ijk}(y_{ij}+\hat{\alpha})}{(\hat{\mu}_{ijk}+\hat{\alpha})^2}\right\}\mathbf{x}'_{ij}\hat{w}_{ik}\ , \qquad (26)
\end{aligned}$$

and, for the derivatives with respect to $\boldsymbol{\zeta}$, we get expressions similar to (25) or (26).

Usually, the complete information is automatically provided as a by-product of the EM algorithm at convergence. However, some software products (like the `glm.nb` macro in S-PLUS) does not make use of the mixed derivative (25) and inverts (24) and (26) separately, or they use the expected information and not the observed one.

Together with (23), we can now calculate (22). Although derived in a different context, the same variance-covariance estimator was also obtained in Friedl and Kauermann (2000) for generalized linear (canonical link) random intercept models with responses from the exponential family.

E Anchovy data and NLMIXED code

```
data anchovy;
input day vol1 y1 vol2 y2 @@;
pair=_n_;
cards;
 8 62.7 158 63.7 147   8 45.8 143 46.9  72   9 58.0 106 57.4  71
 9 59.8  57 58.9  74   9 60.0  97 62.6 141   9 59.3  80 63.3  85
 9 67.4 109 69.4 119   9 64.8 109 58.8  96  10 62.8 350 63.8 142
10 63.1 224 65.2 175  10 63.2 109 64.7 137  10 62.1 214 61.2  80
```

```
11 58.6  84 55.5  45  12 77.6  37 71.4   9  12 54.1  67 51.6  24
12 53.9  46 54.4  24  13 76.4  44 77.8  38  15 55.6 118 56.9 142
15 52.6  30 51.5  48  15 77.5 116 77.9  93  16 64.0 137 66.1  59
16 67.9  61 71.6  50  16 66.2  49 67.1  76  17 64.3  37 61.3   9
17 57.2  91 59.8  68  18 76.2  55 76.5  99  18 75.5  91 77.2  63
18 61.5  50 58.8  38  19 66.9  63 63.3  69  19 64.1  28 63.4  34
19 63.1  41 69.3  38  19 73.1  44 71.1  26  22 69.5  53 64.8  54
22 62.2  46 66.0  13  23 80.1  72 79.8 100  23 60.0  24 59.5  36
24 56.9  19 60.5   8  24 69.2  32 70.2  25  24 70.7  29 73.8  34
25 69.6  13 68.2  31  25 66.2   8 66.3  20  25 77.0  12 70.5  43
26 55.0  16 52.7  24  26 71.8  20 66.9  17  26 61.0  14 64.6  13
26 75.5  17 76.9  10  27 68.3  22 67.2  28  27 75.5  18 74.1  15
27 72.4  23 72.9  60
;
data set1; set anchovy; y=y1; lv=log(vol1); keep pair day lv y;
data set2; set anchovy; y=y2; lv=log(vol2); keep pair day lv y;
data new; set set1 set2;
proc sort; by pair;
proc nlmixed;
parms b0=.5 b1=1 b2=-.1 s2u=.1 a=4;
eta=b0+b1*lv+b2*day+u;
m=exp(eta);
p=m/(m+a);
ll=lgamma(y+a)-lgamma(y+1)-lgamma(a);
ll=ll+a*log(1-p)+y*log(p);
model y~general(ll);
random u~normal(0,s2u)
subject=pair;
run;
```

References

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models, *Statistics and Computing* **6**: 251–262.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models, *Biometrics* **55**: 117–128.

Alfò, M. and Aitkin, M. (2000). Random coefficient models for binary longitudinal responses with attrition, *Statistics and Computing* **10**: 279–288.

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.

Breslow, N. E. (1984). Extra-Poisson variation in log-linear models, *Applied Statistics* **33**: 38–44.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* **88**: 9–25.

Casella, G., Robson, D. S., Schwager, S. J. and Youngs, W. D. (1985). Evaluation of entrainment abundance sampling designs, Prepared under contract with Consolidated Edison Company of New York, Inc.

Cox, D. R. (1983). Some remarks on overdispersion, *Biometrika* **70**: 269–274.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society,* Series B **39**: 1–38.

Friedl, H. and Kauermann, G. (2000). Standard errors for EM estimates for generalized linear models with random effects, *Biometrics* **56**: 761–767.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters, *The Annals of Mathematical Statistics* **27**: 887–906.

Laird, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution, *Journal of the American Statistical Association* **73**: 805–811.

Land, K. C., McCall, P. L. and Nagin, D. S. (1996). A comparison of Poisson, negative binomial, and semiparametric mixed Poisson regression models, *Sociological Methods and Research* **24**: 387–442.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression, *The Canadian Journal of Statistics* **15**: 209–225.

Lindsay, B. G. (1983). The geometry of mixture likelihoods: A general theory, *The Annals of Statistics* **11**: 86–94.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm, *Journal of the Royal Statistical Society,* Series B **44**: 226–233.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models, *Journal of the American Statistical Association* **92**: 162–170.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*, Springer, New York.

SAS Institute Inc. (1997). *SAS/STAT Software: Changes and Enhancements Through Release 6.12*, Cary, NC: SAS Institute Inc.

Thall, P. F. and Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion, *Biometrics* **46**: 657–671.

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.

Table 2: *Different model fits for the epilepsy data.*

	Fit				
	B&C (PQL)	Poisson		Negative Binomial	
Variable	Model IV	para.	NPML(K=7)	para.*	NPML(K=6)
Fixed effects					
Constant	-1.27	-1.37	-1.51	-1.34	-1.82
	(1.2)	(1.20)		(1.18)	
Base	0.87	0.89	0.94	0.89	0.91
	(0.14)	(0.13)	(0.07)	(0.13)	(0.09)
Trt	-0.91	-0.93	-0.64	-0.93	-1.06
	(0.41)	(0.40)	(0.27)	(0.40)	(0.34)
Base$\times$Trt	0.33	0.34	0.29	0.34	0.46
	(0.21)	(0.20)	(0.11)	(0.20)	(0.14)
Age	0.46	0.48	0.45	0.48	0.58
	(0.36)	(0.35)	(0.18)	(0.35)	(0.25)
Visit	-0.26	-0.27	-0.29	-0.27	-0.27
	(0.16)	(0.16)		(0.17)	
Random effects					
Intercept ($\sqrt{\sigma_1^2}$)	0.52	0.25	0.53	0.22	0.51
	(0.06)	(0.06)		(0.06)	
Slope ($\sqrt{\sigma_2^2}$)	0.74	0.53	0.91	0	0.80
	(0.16)	(0.23)			
Covariance (σ_{12})	-0.01	0.003	0.01	0	-0.05
	(0.03)	(0.09)			
Index (α)	-	-	-	7.46	10.15
				(1.76)	(2.61)

* The estimates shown were obtained using NLMIXED for the model with a random intercept only.