Statistical Modelling 2003; **3**: 179–191

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, USA
[2]Institute of Statistics, Technical University Graz, Graz, Austria

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 complex forms of dependence can be modelled 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: Monte Carlo EM; NLMIXED procedure; nonparametric maximum likelihood; overdispersion; random effects

Data and software link available from: http://stat.uibk.ac.at/SMIJ
Received February 2003; **revised** May 2003; **accepted** June 2003

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 modelled 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.

More specifically, 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

Address for correspondence: Department of Statistics, PO Box 118545, University of Florida, Gainesville, FL 32611-8545, USA.

10.1191/1471082X03st058oa

marginally independent negative binomial random variables with mass functions given by

$$\Pr(Y_i = y; \alpha, \mu_i) = \frac{\Gamma(y+\alpha)}{\Gamma(\alpha)y!}\left(\frac{\alpha}{\mu_i+\alpha}\right)^{\alpha}\left(\frac{\mu_i}{\mu_i+\alpha}\right)^{y} \tag{1.1}$$

where $y \in \{0, 1, 2, \dots\}$. We write $Y_i \sim \text{nb}(\alpha, \mu_i)$ and denote the right-hand side of (1.1) 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, for example, Land *et al.*, 1996). For example, maximum likelihood fitting can now be accomplished using the GENMOD procedure in SAS (SAS Institute Inc., 1997), and the macro 'glm.nb' from the MASS library in S-Plus (Venables and Ripley, 1997). In particular, the glm.nb macro takes advantage of the fact that, for fixed values of the shape parameter, the $\text{nb}(y, \alpha, \mu)$ mass function is a one-parameter exponential family. It follows that estimation of the regression parameters in the corresponding loglinear model can be accomplished using iteratively reweighted least squares (McCullagh and Nelder, 1989, subsection 2.5).

We obtained the $\text{nb}(y; \alpha, \mu)$ mass function by assuming a count Y to be conditionally Poisson with mean m, where m has a gamma distribution with shape α and scale α/μ. An alternative, negative binomial model is obtained by assuming m has a gamma distribution with shape $\alpha\mu$ and scale α (McCullagh and Nelder, 1989, subsection 6.2). For this model the marginal mean and variance are $E(Y) = \mu$ and $\text{var}(Y) = \mu(1 + 1/\alpha)$, respectively. However, in this case the marginal distribution of Y is not a one-parameter exponential family for fixed α. Thus, this parameterization is not as amenable to efficient model fitting procedures as the previous model, and for this reason is seldom used in practice.

If multiple counts are observed at the same site or on the same subject at different times, they 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}, $i = 1, \dots, r$ and $j = 1, \dots, n_i$, where Y_{ij} denotes the jth measurement on the ith subject (or cluster). Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{in_i})'$ and suppose that $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are known vectors of covariates associated with Y_{ij}. Note that $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ may or may not have common components. 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

$$Y_{ij}|\mathbf{u}_i \sim \text{nb}(\alpha, \mu_{ij}) \quad \text{with } \mu_{ij} = \mathrm{E}(y_{ij}|\mathbf{u}_i) = \exp\{\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{u}_i\} \tag{1.2}$$

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 $\mathrm{N}_q(0, \boldsymbol{\Sigma})$ and that $\boldsymbol{\Sigma}$ is known up to a vector of variance components, $\boldsymbol{\theta}$. We denote the entire vector of unknown parameters as $\boldsymbol{\psi} = (\alpha, \boldsymbol{\beta}, \boldsymbol{\theta})$.

EXAMPLE 1: Casella *et al.* (1985) discuss data on the entrainment of Bay Anchovy larvae at the Bowline Point hydroelectric plant in New York State. The data consist of two replicate larvae counts and associated water volumes measured in cubic metres. 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, $i = 1, \ldots, 49$. In subsection 3.1 we consider the negative binomial loglinear mixed model in which, conditional on iid random effects, $u_i \sim \mathrm{N}(0, \sigma^2)$, the expected counts satisfy

$$\log \mu_{ij} = \beta_0 + \beta_1 \mathrm{Vol} + \beta_2 \mathrm{Day} + u_i \tag{1.3}$$

where Vol is the logarithm of volume and Day is the date in August when the pair of counts were collected.

EXAMPLE 2: 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. 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 patient's 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 × 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 in which the V4 indicator is replaced by a trend variable, Visit, defined as $(j - 2.5)/5$ for $j = 1, 2, 3, 4$. They consider a Poisson loglinear mixed model with a bivariate subject level random intercept and slope (associated with Visit). In Subsection 3.2 we show that this model fails to adequately account for overdispersion and consider the corresponding negative binomial model. Breslow and Clayton fit these models using penalized quasi-likelihood (PQL), which can be viewed as an approximate maximum likelihood (ML) method with normal random effects.

EXAMPLE 3: Waller and Zelterman (1997) analyse a longitudinal study of 121 senior citizens enrolled in a health plan in Minnesota. The data consist of the number of times each person visited and called the medical clinic in each of four consecutive six-month periods. Let y_{ijk} denote the count for subject i, event (visit or call) j, and period k,

$i = 1, \ldots, 121$, $j = 1, 2$, $k = 1, 2, 3, 4$. Waller and Zelterman consider the loglinear model

$$\log \mu_{ijk} = \mu + u_i + \beta_j + \gamma_k + v_{ij} + w_{ik} \tag{1.4}$$

in which u_i is the (fixed) effect of subject i, β_j is an event (visit or call) effect, γ_k is the effect of period k, and v_{ij} and w_{ik} are fixed subject × event and subject × period interactions. Variability about the mean and dependence among the eight counts for each subject were modelled by assuming a negative multinomial distribution. In contrast, our approach is to treat subject, and subject × event and subject × period interactions, as a random effects. We model these effects as independent zero mean normal variables with variance components σ^2_u, σ^2_v and σ^2_w, respectively. Then, conditionally on the random effects, the counts are independent negative binomial variables, with $y_{ijk} \sim \mathrm{nb}(\alpha, \mu_{ijk})$.

In the next section we outline methods for maximizing the likelihood function for a negative binomial loglinear mixed model. In particular, we find that the SAS procedure NLMIXED can be used in simple cases. When it cannot (as in our Example 3) a Monte Carlo EM algorithm can be implemented in a straightforward way. In subsection 2.3 we describe a semiparametric version of our model in which the random effects distribution is unspecified. We show that this results in a closed-form expression of the expected complete-data likelihood and, hence, a simple EM algorithm. This method is applied in Example 1 and gives similar results to our parametric approach. The results of our model fits for Examples 1–3 are described in section 3. These illustrate that the negative binomial model can lead to more meaningful (less biased) parameter estimates, and that the computations are quite tractable. Section 4 contains a final discussion. Some technical details concerning the implementation of MCEM are given in the Appendix.

2 Maximum likelihood estimation

2.1 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 $\mathrm{N}_q(0, \boldsymbol{\Sigma})$, 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 $f(\mathbf{y}_i|\mathbf{u}_i; \alpha, \boldsymbol{\beta}) = \prod_j \mathrm{nb}(y_{ij}; \alpha, \mu_{ij})$ with $\mu_{ij} = \exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_i)$. Since the random effects are not observable, the likelihood function is given by

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

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 Examples 1 and 2 are one-dimensional and two-dimensional, respectively. In such cases numerical integration, for example, the adaptive Gaussian quadrature method used in the SAS procedure NLMIXED, can often be used to maximize the likelihood function. However, for higher dimensions, such as in Example 3 (see subsection 3.3), Monte Carlo methods may be required.

2.2 Implementation of Monte Carlo EM

Consider first a (deterministic) EM algorithm (Dempster *et al.*, 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; McCulloch, 1997; Booth and Hobert, 1999; Caffo *et al.*, 2002) 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 the Appendix for details on drawing such a sample using an accept/reject algorithm) we maximize

$$\hat{Q}(\boldsymbol{\psi};\boldsymbol{\psi}^{(t)}) = \frac{1}{m}\sum_{l=1}^{m}\log f(\mathbf{y},\mathbf{u}_{t,l};\boldsymbol{\psi}) \tag{2.1}$$

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 $\boldsymbol{\theta}$ (through $\boldsymbol{\Sigma}$), and so maximization with respect to these parameters can be done separately. Maximization with respect to $\boldsymbol{\theta}$ is typically straightforward, since the second piece is proportional to the log-likelihood associated with a random sample of size rm from the $\mathrm{N}_q(0,\boldsymbol{\Sigma}(\boldsymbol{\theta}))$ distribution. In particular, if the form of $\boldsymbol{\Sigma}$ is unspecified

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

As for the maximization with respect to $(\alpha, \boldsymbol{\beta})$, since $f = \text{nb}(y_{ij}; \alpha, \exp\{\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_{t,l,i}\})$, the first piece 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.

Modifications of this basic MCEM algorithm involving importance sampling (Booth and Hobert, 1999) or even Markov chain Monte Carlo (McCulloch, 1997; Caffo *et al.*, 2002) can be used in cases where rejection sampling is prohibitively inefficient. In addition, Booth and Hobert (1999) discuss an algorithm in which the value of m increases with the iteration number and convergence is diagnosed by quantifying the Monte Carlo error (see also Caffo *et al.*, 2002).

2.3 Nonparametric maximum likelihood

Suppose 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

$$L(\alpha, \boldsymbol{\beta}, g; \mathbf{y}) = \prod_{i=1}^{r} \int f(\mathbf{y}_i|\mathbf{u}_i; \alpha, \boldsymbol{\beta}) g(\mathbf{u}_i)\, d\mathbf{u}_i \tag{2.2}$$

Kiefer and Wolfowitz (1956) show that the unspecified random effects 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 typically done conditionally on a fixed value of K. The vector of unknowns therefore becomes $\boldsymbol{\psi} = (\alpha, \boldsymbol{\beta}', \boldsymbol{\zeta}', \boldsymbol{\pi}')'$, where $\boldsymbol{\zeta} = (\boldsymbol{\zeta}'_1, \ldots, \boldsymbol{\zeta}'_K)'$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$ denote the vectors of mass points and their corresponding masses, respectively.

Maximizing (2.2) with respect to $\boldsymbol{\psi}$ yields the nonparametric maximum likelihood (NPML) estimate of the random effects distribution g. 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 extension to clustered observations, assumed to be correlated, is discussed in Aitkin (1999) and also in Alfò and Aitkin (2000).

The maximizer of (2.2) is most easily obtained via 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 $\alpha^{(t)}$, $\boldsymbol{\beta}^{(t)}$ 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 $f = \mathrm{nb}(y_{ij}; \alpha^{(t)}, \mu_{ijk}^{(t)})$ 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

$$Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(t)}) = \sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K} w_{ik}^{(t)} \log[f(y_{ij}|\mathbf{u}_i; \alpha, \mu_{ijk})] + \sum_{i=1}^{r}\sum_{k=1}^{K} w_{ik}^{(t)} \log \pi_k \qquad (2.3)$$

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 argument 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, involving a weighted version of the negative binomial loglinear model likelihood. Another difference is that the algorithm involves 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 (2.2) in a simple discrete way. Replacing g by $\hat{g}$ yields an approximation to the nonparametric version of the log-likelihood function

$$l_K(\hat{\alpha}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\pi}}; \mathbf{y}) = \sum_{i=1}^{r} \log\left\{\sum_{k=1}^{K}\left[\prod_{j=1}^{n_i} f(y_{ij}|\mathbf{u}_i; \hat{\alpha}, \hat{\mu}_{ijk})\right]\hat{\pi}_k\right\} \qquad (2.4)$$

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. In particular, we note that choosing $K = 1$ results in the log-likelihood of $N = \sum_i n_i$ independent negative binomial responses without any randomness in the linear predictor.

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

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

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

Table 1 Different model fits for the Anchovy data

	Fit			
	Poisson		Negative binomial	
Variable	Parametric	NPMLI ($K = 5$)	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 (0.004)	−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)

3 Examples

3.1 Application to Anchovy larvae data

Consider the model (1.3) for the Bay Anchovy larvae data discussed in the Introduction. 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 nb(α, μ_{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. The data and SAS code for the negative binomial loglinear mixed model is available from the journal website (http://www.statmod.com/). Since it is natural to expect the means to be proportional to volume we expect β_1 to be close to 1. It is interesting to note that the Poisson GLMM fit produces an estimate of β_1 far below this value, although its standard error is large enough for the difference to be statistically insignificant.

In the NPML analysis of the negative binomial mixed model we obtained maximum $-2l_K$ values of 922.7, 913.5 and 908.8 for $K = 1, 2, 3$, respectively. Thus, based on the decrease of 9.2 at the cost of two additional parameters, a random effects distribution with two mass points appears to be significantly better than a single point mass. 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 $\hat{E}(u) = 2.82$ and $\widehat{\mathrm{Var}}(u)^{1/2} = 0.32$, respectively. Increasing K to 3 decreases $-2l_K$ by a further 4.7. However, this is at the expense of two additional parameters.

3.2 Application to epilepsy data

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 $j = 1, 2, 3, 4$. In Model IV of Breslow and Clayton (1993) the counts are implicitly assumed to be independent Poisson variables, conditional on the random effects. 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 nb(α, μ_{ij}) variables, with means satisfying the loglinear model:

$$\log \mu_{ij} = \beta_0 + \beta_1 \text{Base} + \beta_2 \text{Trt} + \beta_3 \text{Base} \times \text{Trt} + \beta_4 \text{Age} + \beta_5 \text{Visit} + u_{i0} + u_{i1} \text{Visit} \qquad (3.1)$$

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

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.)

Thus, the effect of accounting for overdispersion is to eliminate the variance component allowing for subject-specific variability in the slope parameter associated with Visit. As in the Anchovy data example, the Poisson model attempts to accom-

Table 2 Different model fits for the epilepsy data

	Poisson		
Variable	PQL	ML	Negative binomial[a]
Fixed effects			
Constant	−1.27 (1.20)	−1.37 (1.20)	−1.34 (1.18)
Base	0.87 (0.14)	0.89 (0.13)	0.89 (0.13)
Trt	−0.91 (0.41)	−0.93 (0.40)	−0.93 (0.40)
Base × Trt	0.33 (0.21)	0.34 (0.20)	0.34 (0.20)
Age	0.46 (0.36)	0.48 (0.35)	0.48 (0.35)
Visit	−0.26 (0.16)	−0.27 (0.16)	−0.27 (0.17)
Random effects			
Intercept $\left(\sqrt{\sigma_1^2}\right)$	0.52 (0.06)	0.25 (0.06)	0.22 (0.06)
Slope $\left(\sqrt{\sigma_2^2}\right)$	0.74 (0.16)	0.53 (0.23)	0
Covariance (σ_{12})	−0.01 (0.03)	0.00 (0.09)	0
Index (α)	—	—	7.46 (1.76)

[a] The estimates shown were obtained using NLMIXED for the model with a random intercept only. The data and SAS code for this model are available from the journal website.

modate the additional variability leading to biased estimation of one or more parameters.

3.3 Minnesota clinic data

Waller and Zelterman (1997) fit model (1.4) to the Minnesota clinic data with all effects fixed. Their model accounts for dependence by assuming that the eight counts per subject follow a negative multinomial distribution. However, they point out that the index parameter (corresponding to α in our negative binomial model) is not significantly different from ∞. In fact the Pearson chi-squared goodness-of-fit statistic is 400.1 with 363 degrees of freedom for the Poisson model with fixed effects. This implies that there is little or no overdispersion with respect to the Poisson distribution.

We feel that a more natural approach to dealing with the dependence between repeated counts on the same subject is to treat the subject effects as random. Thus, as indicated in section 1 we initially attempted to fit a negative binomial loglinear mixed model with iid seven-dimensional random effects

$$(u_i, v_{i1}, v_{i2}, w_{i1}, w_{i2}, w_{i3}, w_{i4})$$

where $i = 1, \ldots, 121$ associated with the subjects. The random effects vectors are independent zero-mean normals with variance components σ_u^2, σ_v^2 and σ_w^2. Because of the high dimension and the crossing of subject and fixed factors it is not possible to fit this model using numerical quadrature methods. Thus, we used the MCEM methods outlined in subsection 2.2. However, as expected, the estimate of the index parameter α increased with each EM iteration suggesting that the limiting Poisson loglinear mixed model is reasonable for this data. As in Example 2, the convergence of EM is very slow when the ML estimate of a dispersion parameter ($1/\alpha$ in this case) is close to zero. The performance of MCEM is much better for the Poisson loglinear mixed model.

The final model fit is summarized in Table 3. We note that the fixed effects estimates have very small standard errors. However, these effects are dominated by random subject-to-subject variability. All three variance components are significantly greater than zero, which is consistent with the corresponding fixed effects analysis.

Table 4 is a Monte Carlo approximation of the correlation between the eight repeated measurements on a given subject based on 100 000 simulated response vectors from the fitted model. Not surprisingly, the table indicates a high degree of positive correlation between the eight counts.

Table 3 Poisson GLMM model fit to the Minnesota clinic data using the MCEM algorithm

	Fixed effects					Variance components		
	μ	β_2	γ_2	γ_3	γ_4	σ_u	σ_v	σ_w
Estimate	1.64	−0.12	0.35	0.23	0.17	1.04	0.60	0.60
Std. error	0.001	0.001	0.001	0.001	0.001	0.091	0.053	0.036

Table 4 Approximate correlation matrix for eight within-subject counts based on 100 000 simulations ($v1$ = number of visits during the first six-month period, $c1$ = number of calls, and so on.)

*v*1	*c*1	*v*2	*c*2	*v*3	*c*3	*v*4	*c*4
1.00	0.65	0.65	0.42	0.68	0.41	0.62	0.39
	1.00	0.43	0.65	0.41	0.64	0.39	0.63
		1.00	0.68	0.63	0.41	0.63	0.41
			1.00	0.40	0.63	0.41	0.66
				1.00	0.64	0.59	0.38
					1.00	0.38	0.63
						1.00	0.63
							1.00

4 Discussion

The Poisson model, although often suggested by the underlying process, suffers from its lack of flexibility in modelling 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 that reduces the bias in parameter estimates. This last point is illustrated in Examples 1 and 2, 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 modelled 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 affecting the performance of EM.

When the structure of the random effects is simple the negative binomial mixed model can be fit using existing software, such as the SAS procedure NLMIXED. However, for analyses that go beyond the current capabilities of this procedure, an MCEM algorithm is straightforward to implement.

We have also discussed a fully nonparametric approach, where no form for the random effects distribution is specified. This may be an attractive and computationally feasible alternative model, particularly in the simple random intercept case.

Acknowledgements

Booth and Hobert's research was supported by NSF Grant DMS-00-72827. Casella's research was supported by NSF Grant DMS-99-71586. The authors are grateful to John Hinde and to a pair of anonymous referees for helpful comments and suggestions that led to improvements in the paper. The authors also wish to thank Ludwig Heigenhauser for allowing them to use his S-plus code for NPML.

References

Aitkin M (1996) A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, **6**, 251–62.

Aitkin M (l999) A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, **55**, 117–28.

Alfò M, Aitkin M (2000) Random coefficient models for binary longitudinal responses with attrition. *Statistics and Computing*, **10**, 279–88.

Booth, JG, Hobert, JP (1999) Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society*, Series B, **61**, 265–85.

Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Caffo BS, Jank W, Jones GL (2002) Ascent-based Monte Carlo EM. Technical report, Johns Hopkins University.

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

Cox DR (1983) Some remarks on overdispersion. *Biometrika*, **70**, 269–74.

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

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

Land KC, McCall PL, Nagin DS (1996) A comparison of Poisson, negative binomial, and semiparametric mixed Poisson regression models. *Sociological Methods and Research*, **24**, 387–442.

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

Louis TA (l982) Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society*, Series B, **44**, 226–33.

McCullagh P, Nelder JA (1989) *Generalized linear models*, 2nd edn. London: Chapman & Hall.

McCulloch CE (l997) Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, **92**, 162–70.

Robert CP, Casella G (1999) *Monte Carlo statistical methods.* New York: Springer.

SAS Institute Inc. (1997) SAS/STAT software: Changes and enhancements through release 6.12. Cary, NC: SAS Institute Inc.

Thall PF, Vail SC (1990) Some covariance models for longitudinal count data with overdispersion. *Biometrics*, **46**, 657–71.

Venables WN, Ripley BD (1997) *Modern applied statistics with S-Plus*, 2nd edn. New York: Springer-Verlag.

Waller LA, Zelterman D (1997) Loglinear modeling with the negative multinomial distribution. *Biometrics*, **53**, 971–82.

Wei GCG, Tanner MA (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.

Appendix: Rejection sampling in MCEM

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

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

allows us to restrict attention to sampling from

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

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 (A.1) shows that we can use a rejection sampler (Robert and Casella, 1999) with candidate $\phi(\cdot;\, \boldsymbol{\Sigma})$ if we can calculate

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

An equivalent problem is to find the $\mathbf{u}_i$ that maximizes

$$\alpha n_i \log \alpha - \sum_{j=1}^{n_i} (\alpha + y_{ij}) \log(\mu_{ij}+\alpha) + \sum_{j=1}^{n_i} y_{ij} \log \mu_{ij} \qquad \text{(A.2)}$$

and this can be accomplished using Newton–Raphson. The first derivative of (A.2) 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, $\mathbf{u}_{t,l} = (\mathbf{u}'_{t,l,1}, \ldots, \mathbf{u}'_{t,l,r})'$.