Standard Errors for EM Estimates in Generalized Linear Models with Random Effects

Herwig Friedl* Göran Kauermann†

February 1, 2000

Abstract

A procedure is derived for computing standard errors of EM estimates in generalized linear models with random effects. Quadrature formulae are used to approximate the integrals in the EM algorithm, where two different approaches are pursued: Gauß-Hermite quadrature, in case of Gaussian random effects, and nonparametric maximum likelihood estimation for an unspecified random effect distribution. An approximation of the expected Fisher information matrix is derived from an expansion of the EM estimating equations. This allows for inferential arguments based on EM estimates, as demonstrated by an example and simulations.

Keywords: EM algorithm, Estimating equations, Gauß-Hermite quadrature, Mixture model, Nonparametric maximum likelihood estimation, Random effect model.

*Institute of Statistics; Technical University Graz; Steyrergasse 17; 8010 Graz, Austria

†Institute of Statistics; Ludwig-Maximilians-University Munich; Akademiestr. 1; 80796 Munich, Germany

1 Introduction

We consider generalized linear models with random effects. Let $y_i = (y_{i1}, \ldots, y_{in_i})^t$ be the response variables taken in the ith cluster, $i = 1, \ldots, n$. The clusters are assumed to be independent and for each cluster we have the $n_i \times p$ design matrix $x_i = (x_{i1}, \ldots, x_{in_i})^t$ of explanatory quantities for the measurements. All observations within the i-th cluster share the same unobserved random effect z_i. The mean response for given x_i and z_i is modelled by the conditional generalized linear model

$$E(y_{ij}|z_i) = \mu_{ij} = h(x_{ij}^t\beta + z_i) \tag{1}$$

where $h(\cdot)$ is the inverse link function and β denotes the associated p dimensional parameter vector of interest.

We make use of model (1) to investigate data describing the outdoor bacteria concentration in the air. Over a period of one year, every two weeks the bacteria content in the air was measured at two different locations in Graz, Austria. One of the two locations was near a neighborhood composting system while the other location served as reference. The focus of interest is on inference about the location effect. As additional covariate among others the temperature was recorded. The bacteria content was determined by a particle fractionating viable (microbiological) air sampler, which simulates the human respiratory system. Air is drawn through the sampler and airborne microorganisms are collected on six stages (filters) with increasing fineness. On the first three stages particles are collected which are usually filtered out in the nose, pharynx and trachea. On stages 4-6 small-sized microorganisms are collected which can be dangerous for humans since they are small enough to be inhalled into the lung and hence can reach bronchi and alveoli. The stages itself consists of Petri dishes with agar medium and the number of colonies forming units (*cfu*) monitored on the agar surface is taken as response variable. We concentrate only on the smallest bacteria. This means for each measurement we consider the three dimensional vector $y_i = (y_{i4}, y_{i5}, y_{i6})^t$, with y_{ij} as the number of *cfu* formed from bacteria filtered on the j-th stage. We model the correlation among these counts by a random effect z_i. As discussed later the random effect therefore not only covers the correlation but also copes for overdispersion found in the data where Poisson variability is assumed. The overdispersion can be explained by missing meteorological covariates which are not accounted for in our model. We investigate how the bacteria content differs at the two locations.

Though the random effects z_i are unobserved, they are assumed to be independently and identically distributed with density $f(z)$. Conditional on z_i, the n_i components of y_i are independent, i.e. $f(y_i|z_i;\beta) = \prod_j f(y_{ij}|z_i;\beta)$, with $f(y|z)$ from the exponential family. For $n_i > 1$, i.e. if more than one observation per cluster is observed, (1) allows to model dependent observations, see e.g. Diggle et al. (1994, chapter 9). If $n_i = 1$, (1) is known as random effect model which provides a general and convenient way for modeling overdispersion, see e.g. Aitkin (1996). For a general discussion on generalized linear models with random effects and their estimation we refer to Breslow and Clayton (1993), Aitkin (1999) and Böhning (1999) and references therein.

The observed (marginal) log likelihood is $l(\beta) = \sum_i \log \int f(y_i|z_i;\beta) f(z_i) dz_i$. The EM algorithm (Dempster et al., 1977) provides a flexible tool to maximize $l(\beta)$. Treating z_i as missing observations provides the Q function

$$Q(\beta|\tilde{\beta}) = \sum_{i=1}^{n} \int \log f(y_i, z_i;\beta) f(z_i|y_i;\tilde{\beta}) dz_i, \tag{2}$$

where $f(y, z;\beta)$ is the joint density of y and z. The calculation of $Q(\cdot|\tilde{\beta})$ for given $\tilde{\beta}$ represents the E step, maximization of $Q(\beta|\tilde{\beta})$ for β leads to the M step and iteration then yields the EM algorithm. Considering (2), one finds that integration is required in the E step. In general, its analytic calculation is often not easily possible. An exception occurs for conjugate distributions, e.g. for normal responses and normal random effects (see also Lee and Nelder, 1996). To overcome the problem of integration we approximate the integral by a quadrature formula. This means we substitute the integral by a finite weighted sum, where z is replaced by some masspoints ζ_k with corresponding masses π_k for $k = 1, \ldots, K$. It is a traditional and also convenient approach to assume a normal distribution for the random effect. This allows us to apply a Gauß-Hermite (GH) quadrature, where the masses and masspoints are known and available from tables (see e.g. Abramowitz and Stegun, 1964). More generality is however achieved if we leave the distribution of the random effect unspecified. This implies that both, π_k and ζ_k are treated as unknown and have to be estimated from the data. The approach is also known as nonparametric maximum likelihood (NPML) estimation as discussed by Simar (1976), Laird (1978) or Jewell (1982) and developed to random effect models by Aitkin (1996, 1999), see also Böhning (1999) for further references.

A disadvantage of the EM algorithm is that it does not automatically provide estimates for the variance-covariance matrix of $\hat{\beta}$. As pointed out in McLachlan and Krishnan (1997) for a variety of examples this point is closely related to the problem of slow convergence. In this paper we derive the Fisher matrix for the EM estimates based on quadrature formulae. This is achieved by embedding the EM algorithm for random effect models in an estimating equation framework. A Taylor expansion is then applied which allows to derive the expected second derivative of (2). The approach is given for both, normally distributed random effects using a GH quadrature as well as for NPML estimates leaving the random effect distribution unspecified. We show how the Fisher matrix can be calculated in practice and demonstrate its use in our data example. A simulation study is given to show the performance of our routine and to compare our approach with the technique suggested in Dietz and Böhning (1995).

For a general discussion of the EM algorithm we refer to Meng and van Dyk (1997) and references given there. Several authors have contributed to the derivation of variance estimates for EM estimates. Louis (1982) provided a very general derivation of the observed information matrix and showed that this matrix can be rewritten as a difference of the complete and the missing information. Meilijson (1989) used the empirical Fisher information matrix to estimate the observed Fisher information for the *i.i.d.* case. Meng and Rubin (1991) suggested a modified numerical differentiation and Oakes (1999) proposed to differentiate $Q(\cdot)$ directly. Our approach relates to Oakes' approach but can be seen as simplified version which makes use of the particular structure available in random effect models.

The paper is organized as follows. We embed the EM algorithm in the framework of estimating equations for GH approximation in Section 2 and for NPML estimation in Section 3. In Section 4 we apply the variance approximation to the data example and investigate its small sample behavior by a simulation study.

2 Gaussian Random Effects

2.1 EM Estimating Equations

In the t-th step of the EM algorithm function $Q(\cdot|\cdot)$ in (2) equals

$$Q(\beta|\beta^{(t)}) = \sum_{i=1}^{n} \frac{\int \log\{f(y_i|z_i;\beta)f(z_i)\}\, f(y_i|z_i;\beta^{(t)})f(z_i)dz_i}{\int f(y_i|z_i;\beta^{(t)})f(z_i)dz_i}. \tag{3}$$

Its calculation requires integration with respect to density $f(z)$ in both, the nominator and the denominator of (3). Assuming z_i as normally distributed allows to approximate the integrals by a quadrature formula like Gauß-Hermite. Taking $z_i \stackrel{iid}{\sim} N(0,1)$ we can model $\mu_{ij} = h(x_{ij}^t\beta + z_i\sigma_z)$ so that the unknown standard deviation σ_z enters the linear predictor as an additional regression parameter. Hinde (1982) used this technique in random effect models for overdispersed Poisson data, Anderson and Aitkin (1985) applied it to variance component models for binary response. For the denominator in (3) the GH quadrature yields the approximation

$$f(y_i;\theta) = \int f(y_i|z_i;\theta)\varphi(z_i)dz_i \approx \sum_{k=1}^{K} f(y_i|\zeta_k;\theta)\pi_k =: f_K(y_i;\theta), \tag{4}$$

which is well known in finite mixture models. Here φ denotes the standard normal density function, $\theta = (\beta^t, \sigma_z)^t$ and K is the number of approximation points. For given K the masspoints ζ_k and their associated masses π_k are fixed and known. Applying this quadrature to the nominator in (3) gives an approximation to $Q(\cdot)$ of

$$Q_K(\theta|\theta^{(t)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(t)} \log\{f(y_i|\zeta_k;\theta)\pi_k\} \tag{5}$$

with weights $w_{ik}^{(t)} = f(y_i|\zeta_k;\theta^{(t)})\pi_k / f_K(y_i;\theta^{(t)})$, the posterior masses for ζ_k, given the data. Formula (5) resumes the E step of the underlying EM algorithm. The M step is given by the $p+1$ dimensional estimating equation

$$\frac{\partial Q_K(\theta|\theta^{(t)})}{\partial\theta} = \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(t)} s_{ik} = 0 \tag{6}$$

with $s_{ik} = s_{ik}(\theta) = \partial \log f(y_i|\zeta_k;\theta)/\partial\theta$ denoting the i-th conditional score, given $z_i = \zeta_k$. For a canonical link we get for instance $s_{ik} = \sum_j^{n_i} \tilde{x}_{ijk}(y_{ij} - \mu_{ijk})$, where $\mu_{ijk} = h(x_{ij}^t\beta + \zeta_k\sigma_z)$ and $\tilde{x}_{ijk} = (x_{ij}^t, \zeta_k)^t$.

We embed the EM algorithm into the concept of estimating equations by defining $g_\theta(\theta) = \partial Q_K(\tilde\theta|\theta)/\partial\tilde\theta\Big|_{\tilde\theta=\theta}$ as estimating function for θ (see also Oakes, 1999). It is easily seen that the EM estimate solves $g_\theta(\hat\theta) = 0$. Assuming that the quadrature density $f_K(y;\theta)$ approximates the true density one finds for the true parameter θ

$$\begin{aligned} E\{g_\theta(\theta)\} &= \sum_{i=1}^n \sum_{k=1}^K \int w_{ik} s_{ik} f(y_i;\theta) dy_i \\ &\approx \sum_{i=1}^n \sum_{k=1}^K \int w_{ik} s_{ik} f_K(y_i;\theta) dy_i = \sum_{i=1}^n \sum_{k=1}^K \pi_k \int \frac{\partial f(y_i|\zeta_k;\theta)}{\partial\theta} dy_i = 0. \end{aligned}$$

Hence, we can define the true parameter as solution of $E_K\{g_\theta(\theta)\} = 0$, where subscript K indicates, that the expectation is calculated using $f_K(\cdot)$. As in the usual likelihood theory, we now expand $g_\theta(\hat\theta)$ about θ and find the first order approximation $\hat\theta - \theta = -\left(\partial g_\theta(\theta)/\partial\theta^t\right)^{-1} g_\theta(\theta)$. Differentiation of $g_\theta(\theta)$ has to take into account that the weights w_{ik} in (5) depend on θ. Assuming $f(y|z;\theta)$ to be of an exponential family form one gets $\partial w_{ik}/\partial\theta = w_{ik}(s_{ik} - \sum_l w_{il}s_{il})$ so that

$$\frac{\partial g_\theta(\theta)}{\partial\theta^t} = \sum_{i=1}^n \sum_{k=1}^K w_{ik}\left(s_{ik}s_{ik}^t + \frac{\partial s_{ik}}{\partial\theta^t}\right) - \sum_{i=1}^n \sum_{k=1}^K \sum_{l=1}^K w_{ik}w_{il}s_{ik}s_{il}^t. \tag{7}$$

Because of $E_K(w_{ik}s_{ik}s_{ik}^t) = -E_K(w_{ik}\partial s_{ik}/\partial\theta^t)$, the first component in (7) has zero expectation. Hence, we can approximate (7) in the usual likelihood fashion by

$$F_K(\theta) := E_K\left(-\frac{\partial g_\theta(\theta)}{\partial\theta^t}\right) = \sum_{i=1}^n \sum_{k=1}^K \sum_{l=1}^K E_K\Big(w_{ik}w_{il}s_{ik}s_{il}^t\Big) = E_K\Big\{g_\theta(\theta)g_\theta^t(\theta)\Big\}. \tag{8}$$

In the first order approximation one now finds $var(\hat\theta) \approx var_K(\hat\theta) = F_K^{-1}(\theta)$. The variance of the regression coefficient $\hat\beta$ is obtained by extracting the respective submatrix of $F_K^{-1}(\theta)$. This automatically takes into account that the random effect variance σ_z^2 is also estimated from the data (see Kackar and Harville, 1984, for an earlier discussion of this point).

Computation of $F_K(\cdot)$

Though $F_K(\cdot)$ above has a simple structure, analytic calculation is not straightforward. Weights w_{ik} and scores depend on y_i, which makes it complicated. A direct estimate would result by using its empirical version $\hat{F}_K(\hat{\theta}) = \sum_i \sum_k \sum_l \hat{w}_{ik}\hat{w}_{il}\hat{s}_{ik}\hat{s}^t_{il}$ where the hat notation refers to simple plug in estimates. However, there is some danger in using $\hat{F}_K^{-1}(\hat{\theta})$ as estimate for the variance. If the variance model is not appropriate this shows in large values of $\hat{s}_{ik}$ which in turn leads to a small variance estimate $\hat{F}_K^{-1}(\hat{\theta})$. Hence, we also calculate $F_K(\hat{\theta})$ by Monte Carlo integration. As emphasized in the Appendix, a rather small Monte Carlo sample size of 30-50 already provides stable results so that the additional numerical work is limited. Denoting the Monte Carlo estimate of $F_K(\hat{\theta})$ by $\tilde{F}_K(\hat{\theta})$ we finally suggest to estimate the variance by the sandwich $\widehat{var}(\hat{\theta}) := \tilde{F}_K^{-1}(\hat{\theta})\hat{F}_K(\hat{\theta})\tilde{F}_K^{-1}(\hat{\theta})n/df$, where n/df adjusts for the usual small sample bias with $df = n-(p+1)$ and $p+1$ as the dimension of θ. The sandwich estimate has the advantage of being robust against such misspecifications. Thus, $\widehat{var}(\hat{\theta})$ should be preferred to the direct use of $\tilde{F}_K^{-1}(\hat{\theta})$, which is also supported by the simulations given below.

2.2 Relation to Other Results

Relation to other Variance Estimates

Dietz and Böhning (1995) proposed a variance estimate based on the large sample equivalence of the likelihood ratio and the Wald statistics. They suggested to estimate the variance of a single component β_l, say, of β by fitting a reduced model with restriction $\beta_l \equiv 0$. The resulting likelihood ratio, the deviance, is then $dev := -2\{l_K(\hat{\beta}_{-l}) - l_K(\hat{\beta})\}$, where $l_K(\beta) = \sum_i \log f_K(y_i; \beta, \sigma)$ and $\hat{\beta}_{-l}$ is the EM estimate in the nested model. Large sample theory yields $var(\hat{\beta}_l) \approx \hat{\beta}_l^2/dev$. However, this approach requires refitting of a reduced model and provides only a single variance estimate at a time. Hence, if β has dimension p, their variance estimation requires p times refitting, which can be a numerical burden. Moreover, no covariances can be estimated by this approach.

Relation to Louis' Formula

Formula (7) can be related to the results given in Louis (1982). We rewrite the observed information as the difference of the complete and the missing information, again with respect to the approximate density $f_K(\cdot)$. Direct calculation provides $-\sum_i \sum_k w_{ik} \partial s_{ik}/\partial\theta^t$ as complete information, whereas the remaining components in (7) give the negative missing information. Important for this assignment is the property that the random effect distribution does not depend on θ. In this case, the conditional score s_{ik} equals the complete score evaluated at $z_i = \zeta_k$. Hence, $s_i(\theta) := E_K(s_{ik}|y_i) = \sum_k w_{ik} s_{ik}$ gives an approximation of the i-th observed score component, which is the conditional expectation of the respective complete score, given the data. Rewriting (8) as the sum of observed score variances, i.e. $F_K(\theta) = \sum_i E_K\{s_i(\theta)s_i(\theta)^t\} = \sum_i var_K\{s_i(\theta)\}$, leads to the idea suggested in Meilijson (1989). He used the empirical Fisher information matrix to estimate the Fisher information for identically distributed scores. Such relations do not exist at all, if the random effect distribution is totally unspecified. However, the estimating equation approach provides variance estimates also in this setting, as demonstrated in the next section.

3 Unknown Random Effect Distribution

3.1 EM Estimating Equations

Now let $f(z)$ be unknown so that β and $f(z)$ have to be estimated simultaneously. We use the NPML approach as suggested by Aitkin and Francis (1995) or Aitkin

(1996, 1999), see also Laird (1978). This approach directly generalizes (5), however now $\zeta = (\zeta_1, \ldots, \zeta_K)^t$ and $\pi = (\pi_1, \ldots, \pi_K)^t$ are both treated as unknown and are estimated from the data. Given $z_i = \zeta_k$ we can write the (conditional) model as $\mu_{ij} = h\{x_{ij}^t \beta + e_k^t \zeta\}$, with e_k as K dimensional vector of zeros except of a 1 at the k-th position. Hence, the unknown masspoints enter the model as additional regression coefficients. As in the previous section, $Q(\cdot|\cdot)$ as given in (5) is used to approximate (2). With $\theta = (\beta^t, \zeta^t)^t$ and π we denote the unknown parameters and get (6) as the M step for θ. Moreover, the masses are obtained from

$$\pi_k^{(t)} = \sum_{i=1}^{n} n_i w_{ik}^{(t)} / \sum_{i=1}^{n} n_i \tag{9}$$

and fulfill the restrictions $0 < \pi_k < 1$ and $\sum_k \pi_k = 1$. To avoid a restricted parameter space it is helpful to apply a technical trick and reparameterize π by the canonical multinomial representation $\pi_k = \exp\{\vartheta_k - \kappa(\vartheta)\}$ for $k = 1, \ldots, K-1$ and $\pi_K = \exp\{-\kappa(\vartheta)\}$, where $\partial\kappa(\vartheta)/\partial\vartheta_k = \pi_k$ and $\vartheta = (\vartheta_1, \ldots, \vartheta_{K-1})^t$. This guarantees the above restrictions for π for all $\vartheta \in \mathbb{R}^{K-1}$.

The vector θ has dimension $p + K$ and its estimating equation is

$$g_\theta(\theta, \vartheta) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} s_{ik}, \tag{10}$$

where $w_{ik} = w_{ik}(\theta, \vartheta)$ depend on both parameters, whereas $s_{ik} = s_{ik}(\theta)$ solely depend on θ. For canonical links we have $s_{ik} = \sum_j^{n_i} \tilde{x}_{ijk} w_{ij}(y_{ij} - \mu_{ijk})$ with $\tilde{x}_{ijk} = (x_{ij}^t, e_k^t)^t$. Additionally to (10) we have (9) as estimating equation which results in the $K-1$ dimensional estimating equation for ϑ

$$g_\vartheta(\theta, \vartheta) := \sum_{i=1}^{n} \sum_{k=1}^{K-1} n_i \tilde{e}_k (w_{ik} - \pi_k), \tag{11}$$

where $\tilde{e}_k$ consists of the first $K-1$ elements of e_k. The EM estimates are defined through $g(\hat{\theta}, \hat{\vartheta}) = 0$ with $g(\cdot) = \{g_\theta(\cdot)^t, g_\vartheta(\cdot)^t\}^t$. Again, we can argue that the true parameter (approximately) fulfills $E_K\{g(\theta,\vartheta)\} = 0$. A linear expansion about (θ,ϑ) gives the first order approximation

$$\begin{pmatrix} \hat\theta - \theta \\ \hat\vartheta - \vartheta \end{pmatrix} = - \begin{pmatrix} \frac{\partial g_\theta(\theta,\vartheta)}{\partial \theta^t} & \frac{\partial g_\theta(\theta,\vartheta)}{\partial \vartheta^t} \\ \frac{\partial g_\vartheta(\theta,\vartheta)}{\partial \theta^t} & \frac{\partial g_\vartheta(\theta,\vartheta)}{\partial \vartheta^t} \end{pmatrix}^{-1} \begin{pmatrix} g_\theta(\theta,\vartheta) \\ g_\vartheta(\theta,\vartheta) \end{pmatrix}.$$

The derivative $\partial g_\theta(\theta,\vartheta)/\partial\theta^t$ is found as in (7). Differentiation of the weights gives $\partial w_{ik}/\partial\vartheta_l = w_{ik}\delta_{(k=l)} - w_{ik}w_{il}$ with $\delta_{(k=l)} = 1$ for $k = l$ and 0 otherwise. This in turn provides the remaining derivatives of (10) and (11) by

$$\begin{aligned} \frac{\partial g_\theta(\theta,\vartheta)}{\partial\vartheta^t} &= \sum_{i=1}^{n}\sum_{k=1}^{K-1} w_{ik}s_{ik}\tilde e_k^t - \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K-1} w_{ik}w_{il}s_{ik}\tilde e_l^t = \frac{\partial g_\vartheta^t(\theta,\vartheta)}{\partial\theta} \\ \frac{\partial g_\vartheta(\theta,\vartheta)}{\partial\vartheta^t} &= \sum_{i=1}^{n}\sum_{k=1}^{K-1} (w_{ik}-\pi_k)n_i\tilde e_k\tilde e_k^t - \sum_{i=1}^{n}\sum_{k=1}^{K-1}\sum_{l=1}^{K-1} (w_{ik}w_{il}-\pi_k\pi_l)n_i\tilde e_k\tilde e_l^t. \end{aligned}$$

Let $F_K(\theta,\vartheta) = -E_K\{\partial g(\theta,\vartheta)/\partial(\theta^t,\vartheta^t)\}$ denote the entire Fisher matrix which consists of the submatrices $F_{K,\theta\theta}(\cdot)$, $F_{K,\theta\vartheta}(\cdot)$ and $F_{K,\vartheta\vartheta}(\cdot)$. As before, one gets with (8) the entries for $F_{K,\theta\theta}(\cdot)$. Using $E_K(w_{ik}) = \pi_k$, we find for the remaining components

$$\begin{aligned} F_{K,\theta\vartheta}(\theta,\vartheta) &= E_K\left(-\frac{\partial g_\theta(\theta,\vartheta)}{\partial\vartheta^t}\right) = E_K\Big(g_\theta(\theta,\vartheta)g_\vartheta^t(\theta,\vartheta)\Big) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K-1} E_K\Big(w_{ik}w_{il}s_{ik}\Big)\tilde e_l^t, \\ F_{K,\vartheta\vartheta}(\theta,\vartheta) &= E_K\left(-\frac{\partial g_\vartheta(\theta,\vartheta)}{\partial\vartheta^t}\right) = E_K\Big(g_\vartheta(\theta,\vartheta)g_\vartheta^t(\theta,\vartheta)\Big) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K-1}\sum_{l=1}^{K-1} n_i\tilde e_k\tilde e_l^t E_K\Big((w_{ik}-\pi_k)(w_{il}-\pi_l)\Big). \end{aligned} \tag{12}$$

The inverse of $F_K(\cdot)$ provides the first order approximation for the variance of $(\hat\theta,\hat\vartheta)$. The variance of $\hat\beta$, the parameter of interest, is obtained by extracting the corresponding submatrix of $F_K^{-1}(\cdot)$. This automatically takes the variability into account which stems from the estimation of ζ and π.

Computation of $F_K(\cdot)$

We suggest again the sandwich estimate $\widehat{var}_K\{(\hat{\theta},\hat{\vartheta})\} = \tilde{F}_K^{-1}(\hat{\theta},\hat{\vartheta})\hat{F}_K(\hat{\theta},\hat{\vartheta})\tilde{F}_K^{-1}(\hat{\theta},\hat{\vartheta})$, where $\tilde{F}_K(\cdot)$ is the Monte Carlo estimate of $F_K(\cdot)$ and $\hat{F}_K(\cdot)$ the plug in estimate. Details are given in the Appendix. In practice it frequently occurs that some estimated masspoints do not differ, i.e. their difference is smaller than a numerical lower bound, or some π_k are particularly small, i.e. smaller than a numerical lower bound. In this case the estimated model is not identifiable and the number of masspoints K has to be reduced. A new proposal for K is obtained by the number of disjunct estimated masspoints with sufficiently large masses. These masspoints together with their corresponding adjusted masses can be used as starting values for a fit of a reduced model. This means, if $\hat{\zeta}_k \approx \hat{\zeta}_l$, then a reduced model with masspoint $\hat{\zeta}_k$ and mass $\hat{\pi}_k + \hat{\pi}_l$ should be fitted. Generally, it seems to be an appropriate strategy to start with a larger value of K and reduce it until all masspoints are different and all masses are larger than a threshold. Reducing the dimensionality of the model in this way yields parameter estimates $\hat{\beta}$ which are distinctly similar to the parameter estimates obtained for a larger value of K, i.e. the EM algorithm converges extremely fast with estimates close to the starting values.

4 Example and Simulation

4.1 Example

Now we return to the example introduced in Section 1. The entire dataset is available from the data server at `http://www.stat.uni-muenchen.de`. We consider a conditional log-link Poisson model for the observed *cfu* counts ($n = 50$, cluster size $n_i = 3$, observations of one week have been neglected due to outliers). Stage 6 and the location without a composting system serve as reference category. The estimates obtained from a GH quadarature (K=10) and the NPML estimates (K=2) are given in Table 1. Pursuing the NPML procedure for $K > 2$ did not yield more than $K = 2$ distinguishable masspoints, which are $\hat{\zeta} = (0.50, 1.56)$ with $\hat{\pi} = (0.88, 0.12)$. Hence, in the NPML approach we have to estimate 7 effect parameters, 2 masspoints and one mass, whereas in the GH quadrature 8 parameters are estimated. The table also includes the estimated standard errors from the approximative Fisher matrix sandwich and the result of the deviance reduction technique.

(Table 1 about here)

The temperature t generally shows a quadratic influence. This means that the highest bacteria content is observed for middle range temperatures about 15 centigrade and the concentration decreases for higher or lower temperatures. The number of colonies formed on stage 4 and 5 is larger near the composting system than for the reference location. Interestingly this effect is reversed for bacteria collected on stage 6. The interaction effect between location and stage seems significant for stages 5 and 6, but not for stage 4. Hence, there is some evidence that the location effect differs among bacteria of different size. A similar result is also found if all 6 stages are analyzed simultaneously.

4.2 Simulation

Finally we run a simulation study to check the small sample behavior of the suggested variance approximation. Similar to the example before we consider the model $y_{ij}|z_i \sim \text{Poisson}\{\mu_{ij} = \exp(1 + z_i + x_{ij}\beta)\}$, $i = 1, \ldots, 50$ and $j = 1, 2$. We set $\beta = 0$ to mirror a testing context as found in the example above. The covariate x_{ij} is taken as binary factor, i.e. $x_{ij} \in \{0, 1\}$, with balanced design in the sense that $x_{i1} = x_{i2}$

and $x_{ij} = 1$ for half of the data. The z_i's are drawn from three different settings:

$$\begin{array}{rrl} \text{a)} & \text{normal:} & z_i \sim N(0, 0.4^2), \\ \text{b)} & \text{mixed:} & z_i \sim \begin{cases} N(0, 0.25^2) & \text{with probability } 0.5 \\ N(1, 0.25^2) & \text{with probability } 0.5, \end{cases} \\ \text{c)} & \text{contaminated:} & z_i \sim \begin{cases} N(0, 0.3^2) & \text{with probability } 0.8 \\ N(1.5, 0.1^2) & \text{with probability } 0.2. \end{cases} \end{array}$$

The model is fitted by NPML estimation starting with $K = 8$ masspoints and reducing K until all masspoints are different. Table 2 shows the number of masspoints definitely used in the simulations. We also apply a GH quadrature with $K = 12$ masspoints. Table 3 resumes the mean and standard deviation of all 500 simulated estimates. Both quadrature formulae provide unbiased estimates for all three settings. For a) and b) both approaches also show similar standard errors, in setting c) however the NPML estimate is clearly less variable than the GH estimate. In general, NPML estimation shows to be not less efficient than the GH counterpart, even if random effects are normally distributed where the GH procedure provides the right quadrature. Moreover, the NPML approach can cope for non-normality of the random effect distribution.

In each simulation we estimate the standard error based on the sandwich with the Fisher matrix $F_K(\cdot)$. For comparison we also calculate the standard error by the deviance reduction criteria (*deviance*). The mean and the standard error of these estimates is reported in Table 3, too. The mean estimated standard error is close to the standard error of the simulated NPML estimates. This does not generally hold for GH estimation, e.g. in setting c) the variance is clearly underestimated.

Finally, we also report the coverage probability of confidence intervals based on the derived variance estimates and appropriate t quantiles. The performance of both variance estimates is comparable. In general the estimate based on the Fisher matrix $F_K(\cdot)$ shows a slightly liberal character for the NPML estimation. Dependent on the random effect distribution, the GH approximation leads to liberal or conservative confidence intervals if the random effect distribution is in fact not normal. Hence, the variance approximation fails if the GH quadrature is not appropriate.

(Table 2+3 about here)

5 Results and Conclusions

We derive an estimate for the variance-covariance matrix of all fixed parameters in generalized linear models with random effects by embedding the EM algorithm into the concept of estimating equations. This estimate is based on the expected second derivative of the EM objective function. We do not propose to apply its empirical version directly but rather to use it in an information sandwich, where Monte Carlo integration gives the outer matrix. The fixed parameters are estimated by approximating the integrals in the EM function through quadrature methods. If the random effects can be assumed to follow a Gaussian distribution, a Gauß-Hermite quadrature yields the appropriate approximation. The NPML method is applied for models where the random effect distribution is totally unknown and can not be specified in advance. However, in this connection the number of masspoints K, which usually is a function of the data also, has to be fixed in advance. Therefore, we like to note that all of the following properties can be seen as conditional results only, given K. Moreover, the arguments used in the Gaussian case are extendable to other specific random effect distributions, where quadrature formulae can be obtained (see e.g. Thisted, 1988).

Acknowledgement

We wish to thank M. Aitkin and J. Hinde as well as both reviewers and the associate editor for discussion and helpful comments. We are grateful to the Information Center of Environmental Protection, Graz/Austria, for permission to use the data.

Appendix: Monte Carlo Integration of $F_K(\theta)$

In principle Monte Carlo (MC) integration of $F_K(\cdot)$ can be carried out by drawing n random effects z_i^* from the discrete distribution with masspoints ζ and masses π, where ζ and π are fixed for Gaussian random effects and estimated from the data in the NPML case. Given z_i^* one can then draw y_{ij}^* from $f(y|z_i^*, x_{ij}; \hat{\theta})$ and calculate the simulated Fisher matrix by replacing the expectations in (8) and (12) by the simulated empirical moments. Averaging over B simulations provides the MC estimate $\tilde{F}_K(\cdot)$, say. The MC variability occurring due to the simulation can be reduced by increasing the MC sample size B. This is however numerically time consuming and we therefore propose a modification where simulation of z_i^* is avoided. Reflecting (4) one finds that $f_K(\cdot)$ consists of a weighted sum of densities $f(\cdot|\zeta_k)$. This suggests to simulate directly from $f(\cdot|\zeta_k)$ and to use a weighted mean to obtain the MC estimate. To be more specific, we define with $B_k = [1 + B\pi_k]$ the MC sample size for masspoint ζ_k, where $[\cdot]$ is the integer fraction. We draw y_{ij}^* from $f(y|\zeta_k, x_{ij}; \hat{\theta})$ and calculate $F_k^*(\cdot) = \sum_i g_i(\hat{\theta}) g_i^t(\hat{\theta})$, where the star notation denotes that y_i^* is used for the calculation and $g_i(\cdot)$ is the i-th contribution of the estimating function. For instance for GH approximation we have $F_k^*(\cdot) = \sum_l \sum_m w_{il}^* w_{im}^* s_{il}^* s_{im}^{*^t}$ while for NPML we have with (12) additional but obvious components in $F_k^*(\cdot)$. Repeating this simulation B_k times yields the MC estimate $\tilde{F}_k(\cdot) = \sum_b^{B_k} F_{k,b}^*(\cdot)/B_k$ for the k-th masspoint, where subscript b gives the replicate number here. Applying this to all masspoints yields the final MC estimate $\tilde{F}(\cdot) = \sum_k \pi_k \tilde{F}_k(\cdot)$. The above sampling scheme allows to keep B rather limited. This means we obtain already for small values of B stable results which allows to neglect the MC variability. In our example and simulations we found that $30 \leq B \leq 50$ is already sufficient.

References

Abramowitz, M. and Stegun, I. A. (eds). (1964). *Handbook of Mathematical Functions*. Washington, D.C.: National Bureau of Standards.

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**, 218-234.

Aitkin, M. and Francis, B. J. (1995). Fitting overdispersed generalized linear models by nonparametric maximum likelihood. *The GLIM Newsletter*, **25**, 37-45.

Anderson, D. A. and Aitkin, M. (1985). Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society, Series B*, **47**, 203-210.

Böhning, D. (1999). *Computer-Assisted Analysis of Mixtures and Applications.* Boca Raton: Chapman & Hall/CRC.

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

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

Dietz, E. and Böhning, D. (1995). Statistical inference based on a general model of unobserved heterogeneity. In: *Statistical Modelling* (B. J. Francis, R. Hatzinger, G. U. H. Seeber and G. Steckel-Berger, eds), 75-82. New York: Springer-Verlag.

Diggle, P. J., Liang, K.-Y. and Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

Hinde, J. (1982). Compound Poisson regression models. In: *GLIM 82* (R. Gilchrist, ed.), 109-121. New York: Springer-Verlag.

Jewell, N. P. (1982). Mixtures of exponential distributions. *Annals of Statistics*, **10**, 479–484.

Kackar, R. N. and Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, **79**, 853-862.

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

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B*, **58**, 619-656.

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.

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York: John Wiley & Sons, Inc.

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society, Series B*, **51**, 127-138.

Meng, X. L. and van Dyk, D. (1997). The EM algorithm – An old folk-song sung to a fast new tune (with discussion). *Journal of the Royal Statistical Society, Series B*, **59**, 511-567.

Meng, X. L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, **86**, 899-909.

Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **61**, 479-482.

Simar, L. (1976). Maximum likelihood estimation of a compound Poisson process. *Annals of Statistics*, **4**, 1200–1209.

Thisted, R. A. (1988). *Elements of Statistical Computing*, London: Chapman & Hall.

Table 1: Estimated fixed parameters and standard errors with respective studentized values for bacteria data

		Variance estimate based on			
		$F_K(\cdot)$		$Deviance$	
Effect	$\hat{\beta}$	$\widehat{s.e.}(\hat{\beta})$	$\hat{\beta}/\widehat{s.e.}(\hat{\beta})$	$\widehat{s.e.}(\hat{\beta})$	$\hat{\beta}/\widehat{s.e.}(\hat{\beta})$
NPML ($K = 2$)					
Reference site					
stage 4	-0.13	0.27	-0.49	0.20	-0.67
stage 5	-0.19	0.28	-0.69	0.22	-0.87
Composting system site					
stage 4	0.32	0.38	0.86	0.24	1.36
stage 5	0.79	0.41	1.92	0.26	3.02
stage 6	-0.43	0.29	-1.48	0.34	-1.26
t	0.03	0.021	1.51	0.015	2.11
t^2	-0.001	0.0007	-1.67	0.0005	-2.17
GH ($K = 10$)					
Reference site					
stage 4	-0.14	0.25	-0.57	0.24	-0.59
stage 5	-0.19	0.20	-0.96	0.23	-0.82
Composting system site					
stage 4	0.37	0.38	0.96	0.26	1.38
stage 5	0.83	0.38	2.19	0.28	2.99
stage 6	-0.44	0.29	-1.49	0.23	-1.92
t	0.05	0.024	1.99	0.03	1.49
t^2	-0.002	0.0008	-2.15	0.0011	-1.53

Table 2: Final number of masspoints used in the simulation (in %)

	K					
Setting	2	3	4	5	6	7
a)	23.8	54.6	20.0	1.6	0	0
b)	23.8	54.2	21.4	0.6	0	0
c)	3.6	32.8	41.6	17.6	4.2	0.2

Table 3: MC mean and standard error of EM estimates (mean($\hat{\beta}$), $s.e.(\hat{\beta})$); MC mean and standard error of estimated standard errors (mean($\widehat{s.e.}(\hat{\beta})$), $s.e.(\widehat{s.e.}(\hat{\beta}))$) and corresponding coverage probability of confidence intervals

			Variance			Coverage	
Setting	mean($\hat{\beta}$)	$s.e.(\hat{\beta})$	Estimate	mean($\widehat{s.e.}(\hat{\beta})$)	$s.e.(\widehat{s.e.}(\hat{\beta}))$	90%	95%
NPML ($K \leq 8$)							
a)	-0.008	.167	$F_K(\cdot)$	.154	.015	86.0	94.4
			Deviance	.165	.015	89.8	95.4
b)	-0.030	.135	$F_K(\cdot)$	.132	.012	88.4	93.6
			Deviance	.152	.058	91.0	96.6
c)	0.007	.163	$F_K(\cdot)$	.138	.020	88.4	93.2
			Deviance	.134	.057	88.6	93.6
GH ($K = 12$)							
a)	0.011	.164	$F_K(\cdot)$	.163	.019	88.2	95.0
			Deviance	.162	.048	88.2	94.4
b)	-0.018	.122	$F_K(\cdot)$	.146	.018	95.2	97.4
			Deviance	.143	.032	94.8	97.6
c)	-0.013	.264	$F_K(\cdot)$	.162	.033	72.9	79.0
			Deviance	.186	.021	75.4	83.0