Biometrics 55, 117–128
March 1999

A General Maximum Likelihood Analysis of Variance Components in Generalized Linear Models

Murray Aitkin

Department of Statistics, University of Newcastle, U.K.
email: Murray.Aitkin@newcastle.ac.uk

Summary. This paper describes an EM algorithm for nonparametric maximum likelihood (ML) estimation in generalized linear models with variance component structure. The algorithm provides an alternative analysis to approximate MQL and PQL analyses (McGilchrist and Aisbett, 1991, *Biometrical Journal* **33,** 131–141; Breslow and Clayton, 1993; *Journal of the American Statistical Association* **88,** 9–25; McGilchrist, 1994, *Journal of the Royal Statistical Society, Series B* **56,** 61–69; Goldstein, 1995, *Multilevel Statistical Models*) and to GEE analyses (Liang and Zeger, 1986, *Biometrika* **73,** 13–22). The algorithm, first given by Hinde and Wood (1987, in *Longitudinal Data Analysis*, 110–126), is a generalization of that for random effect models for overdispersion in generalized linear models, described in Aitkin (1996, *Statistics and Computing* **6,** 251–262). The algorithm is initially derived as a form of Gaussian quadrature assuming a normal mixing distribution, but with only slight variation it can be used for a completely unknown mixing distribution, giving a straightforward method for the fully nonparametric ML estimation of this distribution. This is of value because the ML estimates of the GLM parameters can be sensitive to the specification of a parametric form for the mixing distribution. The nonparametric analysis can be extended straightforwardly to general random parameter models, with full NPML estimation of the joint distribution of the random parameters. This can produce substantial computational saving compared with full numerical integration over a specified parametric distribution for the random parameters. A simple method is described for obtaining correct standard errors for parameter estimates when using the EM algorithm. Several examples are discussed involving simple variance component and longitudinal models, and small-area estimation.

Key words: EM algorithm; Longitudinal data; Mixture model; Nonparametric maximum likelihood; Overdispersion; Random effects GLMs; Variance components.

1. Introduction

The literature on random effects in generalized linear models (GLMs) is now extensive (a small subset of relevant references is given later). Our concern in this paper is nonparametric maximum likelihood (NPML) in these models with shared random effects arising through variance component or repeated-measures structure, e.g., from longitudinal data. Several NPML algorithms have been proposed and used for these models, as noted here. The purpose of this paper is to set out the finite mixture algorithm, which can be readily implemented in standard GLM software (our implementation was in GLIM4). NPML in models with a unique random effect for each observation, leading to overdispersion, was discussed by Aitkin (1996), and this paper follows a similar development.

For simplicity of exposition, we begin with the simple two-level variance component model for a structure with upper- or second-level sampling units indexed by $j = 1, \ldots, r$ and lower- or first-level sampling units indexed by i sampled within each upper-level unit, where $i = 1, \ldots, n_j$. On each first-level unit, we measure or record a response y_{ij}, and we have explanatory variables x, which can be measured at both upper (x_j) and lower (x_{ij}) levels. We want to represent the distribution of the response y by an exponential family member, with a link function and linear predictor involving the explanatory variables at both levels and perhaps their cross-level interactions.

The nested structure of the responses y_{ij} induces an intraclass correlation between the lower-level responses on the same upper-level unit. A natural way (familiar from normal variance component modeling) of representing this common variation is by adding a common unobserved random effect to the linear predictor for each lower-level unit in the same upper-level unit. Thus, the common variation is modeled as an extra unobserved variable on the same scale as the linear predictor.

If the distribution of this random effect is conjugate to the exponential family distribution, then maximum likelihood (ML) is straightforward in principle from the marginal distribution of the observed data, as, e.g., in the negative binomial and beta-binomial distributions (Lee and Nelder, 1996). However, the conjugate approach lacks generality because a different conjugate distribution must be assumed for each exponential family distribution. A more appealing approach would be to assume a common distribution for the random effects across the exponential family; an obvious choice is the normal $N(0, \sigma^2)$ distribution (Breslow and Clayton, 1993; McGilchrist, 1994). This is especially natural for link functions giving an unbounded parameter space for the linear predictor.

However, exponential family models other than the normal with a normal random effect have been difficult and slow to fit by ML because the resulting likelihood does not have a closed form. A number of different approaches have been followed to deal with this problem.

First, the likelihood can be integrated numerically using some form of Gaussian quadrature (Anderson and Aitkin, 1985) to give full ML estimation. This approach is widely regarded as computationally intensive. Current quadrature methods use the EM algorithm (Hinde, 1982; Anderson and Hinde, 1988) for fitting the finite mixture distribution resulting from the discretization of the normal into K probability masses π_k at known mass points z_k. The EM algorithm can be slow to converge in mixture models and does not provide the correct (asymptotic) information-based standard errors for the ML estimates without additional computation.

Second, the log-likelihood function can be approximated by a quadratic, and standard computational methods for the normal variance component model can then be used, giving approximate ML or REML estimation (Laird and Ware, 1982). This approach has been implemented generally in slightly different forms by Breslow and Clayton (1993), McGilchrist (1994), Goldstein (1995), and Longford (1993), giving penalized quasi-likelihood (PQL) analyses. The success of the approximation depends on the closeness to normality of the observed data likelihood and might fail badly, e.g., for binary response data (Rodriguez and Goldman, 1995).

Third, the integrals required in the E-step of the EM algorithm can be avoided by Laplace approximations (Steele, 1996) or by Monte Carlo integration (Walker, 1996; McCulloch, 1997).

Fourth, the problem can be circumvented by the generalized estimating equation approach (Liang and Zeger, 1986; Diggle, Liang, and Zeger, 1994). Here the marginal distribution of y is assumed to be exponential family, and the repeated-measures structure is represented by a covariance matrix model whose parameters are estimated by a form of quasi-likelihood (marginal quasi-likelihood, or MQL), which does not require a full parametric specification for the random effect distribution. This approach is widely used.

Fifth, a fully Bayes approach can be followed, with the additional structure of a prior distribution on all the model parameters, and Markov chain Monte Carlo methods can be used to obtain marginal posterior distributions of the parameters. Gelman et al. (1995) give a detailed exposition of this approach, which is becoming increasingly popular with the widespread dissemination of Bayesian software.

A disadvantage of any approach using a specified parametric form for the mixing distribution of the unobserved random effects is the possible sensitivity of the conclusions to this specification. The influential paper by Heckman and Singer (1984) showed substantial changes in parameter estimates with quite small changes in mixing distribution specification; Davies (1987) showed similar effects. This difficulty can be avoided by NPML estimation of the mixing distribution concurrently with the structural model parameters; the NPML estimate is well known to be a discrete distribution on a finite number of mass points (Kiefer and Wolfowitz, 1956; Laird, 1978; Lindsay, 1983). Clayton and Kaldor (1987) gave an example of this approach, in the simpler framework of a single-level overdispersion model. The present paper gives a detailed discussion of this approach.

Finding the NPML estimate is widely regarded as computationally intensive, the particular difficulty being the location of the mass points. Current approaches use Gateaux or directional derivatives (Ezzet and Davies, 1988; Follmann and Lambert, 1989; Böhning, Schlattman, and Lindsay, 1992; Lesperance and Kalbfleisch, 1992) or optimal design algorithms (Mallet, 1986). Zackin, de Gruttola, and Laird (1996) gave an ECM (Gauss–Seidel) algorithm, alternately estimating the fixed effects and the mixture parameters. Schumitzky (1991) gave a general EM algorithm, but it is restricted to purely random models. Barry, Francis, and Davies (1989) remarked that "[NPML] is not a simplification of the parametric approach as the identification of the number, location and masses of these points of support present formidable computational problems."

The important paper by Hinde and Wood (1987) addressed the computational issues of NPML estimation in the framework of two-level variance component models. They showed that, quite generally, both the mass-point locations z_k and the masses π_k could be estimated very straightforwardly by ML within the framework of a finite mixture of GLMs, allowing the straightforward full NPML estimation of the mixing distribution. Convergence of the EM algorithm can then become very slow, as information in the data about the mixing distribution might be very limited, but the algorithm is easily programmed, e.g., in GLIM4 or S-plus.

In this paper, we give a simple exposition of this approach, extend it to general random coefficient regression models, and discuss a range of applications that make clear the value of the nonparametric approach. The computational approach we follow using GLIM4 is closely related to that of Dietz (1992) for finite mixtures of GLMs.

In Section 2 we describe the variance component GLM with a normal error term and in Section 3 the EM algorithm. A simple method due to Dietz and Böhning (1995) is given for obtaining correct standard errors for parameter estimates when using the EM algorithm. Section 4 relaxes the specification of the z_k and π_k and presents the EM algorithm for the more general case; it is especially simple. Section 5 extends the approach to general random coefficient regression models. Section 6 discusses three examples. Section 7 concludes. General GLIM4 macros for the simpler exponential family dispersion model were presented by Aitkin and Francis (1995), and corresponding macros for the variance component model were developed by Aitkin and Francis (1998). These are adaptations of the Hinde and Wood (1987) algorithm.

2. The Two-Level Variance Component GLM

We adopt the standard notation for GLMs (Aitkin et al., 1989; McCullagh and Nelder, 1989), and generalize the development in Anderson and Hinde (1988). We have a two-stage random sample y_{ij} with $i = 1, \ldots, n_j, j = 1, \ldots, r$ and $\Sigma n_j = n$, from an exponential family distribution $f(y \mid \theta)$ with canonical parameter θ and mean μ and explanatory variables $X = (x_{ij})$ related to μ through a link function $\eta_{ij} = g(\mu_{ij})$ with linear predictor $\eta_{ij} = \beta' x_{ij}$. Here the X matrix is understood to include both upper- and lower-level explanatory variables (and their interactions if any), with $x_{ij} = x_j$ for all i for an upper-level variable. Thus, the upper-level variable x_j is replicated n_j times for the n_j lower-level units in the jth upper-level unit.

In the extension to random effect models, we have an unobserved common random effect z_j for each lower-level unit in the jth upper-level unit, the z_j being initially assumed independently normally distributed $z_j \sim N(0, 1)$, and conditionally on z_j, the y_{ij} have independent GLMs with linear predictor $\eta_{ij} = \beta' x_{ij} + \sigma z_j$. The random effect is modeled as acting on the same scale as the linear predictor.

The likelihood is then

$$L(\beta, \sigma) = \prod_j \int \prod_i f(y_{ij} \mid \beta, \sigma, z_j) \pi(z_j) dz_j,$$

where $\pi(z)$ is the standard normal density function.

Because the integral does not have a closed form except for y normal, we approximate it by Gaussian quadrature: we replace the integral over the normal z_j by a finite sum over K Gaussian quadrature mass points z_k with masses π_k; the z_k and π_k are given in standard references (e.g., Abramowitz and Stegun, 1964). The likelihood is then

$$L(\beta, \sigma) \doteq \prod_{j=1}^{n} \sum_{k=1}^{K} \pi_k \prod_{i=1}^{n_j} f(y_{ij} \mid \beta, \sigma, z_k).$$

The likelihood is thus (approximately) the likelihood of a finite mixture of exponential family densities with known mixture proportions π_k at known mass points z_k, with the linear predictor for the ijth observation in the kth mixture component being

$$\eta_{ijk} = \beta' x_{ij} + \sigma z_k.$$

We can also regard this as the exact likelihood for this discrete mixing distribution for z. This is inherently of interest because the NPML estimate of the mixing distribution is well known (e.g., Laird, 1978; Lindsay, 1983) to be a discrete distribution on a finite number of mass points. In Section 4, we consider the joint estimation of β, the π_k and the mass points z_k, but for the moment consider the latter quantities as fixed.

3. ML Estimation for the Finite Mixture Model

We proceed as in Bock and Aitkin (1981), Hinde (1982), Anderson and Aitkin (1985), Anderson (1988), and Anderson and Hinde (1988). The log likelihood is

$$\ell(\beta, \sigma) = \sum_j \log \sum_k \pi_k f_{jk},$$

where for compactness we write

$$\begin{aligned} f_{jk} &= \prod_i f_{ijk}, \\ f_{ijk} &= f(y_{ij} \mid \beta, \sigma, z_k) \\ &= \exp\{\theta_{ijk} y_{ij} - b(\theta_{ijk}) + c(y_{ij})\} \end{aligned}$$

with

$$\begin{aligned} \mu_{ijk} &= b'(\theta_{ijk}), \\ V_{ijk} &= b''(\theta_{ijk}), \\ \eta_{ijk} &= g(\mu_{ijk}) = \beta' x_{ij} + \sigma z_k, \\ g'_{ijk} &= g'(\mu_{ijk}). \end{aligned}$$

Then

$$\frac{\partial \ell}{\partial \beta} = \sum_j \frac{\sum_k \pi_k f_{jk} \frac{\partial \log f_{jk}}{\partial \beta}}{\sum_k \pi_k f_{jk}} = \sum_j \sum_i \sum_k w_{jk} s_{ijk}(\beta),$$

where w_{jk} is the posterior probability that observation y_{ij} comes from component k,

$$w_{jk} = \frac{\pi_k f_{jk}}{\sum_\ell \pi_\ell f_{j\ell}},$$

and $s_{ijk}(\beta)$ is the β component of the score (the log-likelihood derivative with respect to β) for observation (ij) in component k:

$$s_{ijk}(\beta) = (y_{ij} - \mu_{ijk}) x_{ij} / V_{ijk} g'_{ijk}$$

(Aitkin et al., 1989, p. 323; McCullagh and Nelder, 1989). Similarly,

$$s_{ijk}(\sigma) = (y_{ij} - \mu_{ijk}) z_k / V_{ijk} g'_{ijk}.$$

Thus, z_k becomes another observable variable in the regression, with regression coefficient σ.

Equating the score to zero gives likelihood equations that are simple weighted sums of those for an ordinary GLM with weights w_{jk}; alternately solving these equations for given weights w_{jk} and updating these weights from the current parameter estimates is an EM algorithm (Dempster, Laird, and Rubin 1977; Aitkin and Tunnicliffe Wilson, 1980). The triple summation over (ij) and k is conveniently (if inefficiently) handled by expanding the data vectors to length Kn by replicating y and X K times and the Gaussian quadrature variable z n times (Hinde, 1982; Anderson and Aitkin, 1985). Model fitting is then identical to that of a single sample of Kn observations with prior weight vector w. Initial estimates for the first E-step for β are conveniently obtained from the ordinary GLM fit and for σ by arbitrary specification other than zero (e.g., $\sigma = 1$).

To obtain correct standard errors for the parameter estimates in the final model, we use the property (Dietz and Böhning, 1995) that in large samples from regular models for which the log likelihood is quadratic in the parameters, the likelihood ratio and Wald tests for the significance of an individual parameter are equivalent, so that the deviance change on omitting the variable is equal to the square of the t-statistic (parameter estimate/SE). Thus, the standard error can be calculated as the absolute value of the parameter estimate divided by the square root of the deviance change. This requires fitting a set of reduced models in which each variable in turn is omitted from the final model (these reduced models would often be fitted in any case to assess the significance of each variable by its deviance change). This property of regular models might not hold in small samples with skewed log likelihoods, but in the latter case the proposed standard error estimate is a more appropriate reflection of the significance of the variable than that based on the inverse information matrix, because it gives a squared t-statistic equal to the LR test statistic for each variable rather than the misleading Wald test statistic.

Standard errors based on the observed information could be calculated using the approach of Louis (1982), as described for two-component normal mixtures by Aitkin and Aitkin (1996), at the cost of substantially more computation or by numerical differentiation within the EM algorithm, using the SEM algorithm of Meng and Rubin (1991).

4. NPML Estimation of the Masses and Mass Points

A particular disadvantage of the modeling approach described above is the lack of information in the data about the mixing distribution (as this can come only from the marginal distribution of the data) and the possible sensitivity of conclusions to choice of a particular distributional form. A second disadvantage is the need to expand the data vector to length Kn; if K is large for accurate Gaussian quadrature, the time required for model fitting increases substantially. (In GLIM4 the data storage requirement can be obviated because the variables can be indexed instead of being expanded explicitly.) A third disadvantage is the possible inaccuracy of Gaussian quadrature, where even 20-point integration might not give high accuracy for the logistic/normal model when the variance component is large (Crouch and Spiegelman, 1990).

Because the model assumption for unobservable random variables cannot be directly assessed, we consider as a preferable modeling strategy the NPML estimation of the mixing distribution, together with the GLM parameters. Our aim is not to estimate this distribution—indeed the NPML estimate of it might be very poor, though consistent—but to avoid possibly misleading inferences from an inappropriate and unverifiable model assumption. Thus, the mixing distribution is a nuisance parameter or function and not the parameter of interest.

We now treat the masses and mass points as unknown parameters; the number K of mass points is also unknown but is treated as fixed and sequentially increased until the likelihood is maximized. Because the variance of the mixing distribution is a function of the unknown parameters, we drop the scale parameter σ and define the mass-point parameters as α_k, with linear predictor

$$\eta_{ijk} = \beta' x_{ij} + \alpha_k.$$

Thus, α_k functions as an intercept parameter for the kth component: It can immediately be estimated simply by including a component factor in the model with K levels instead of the variable z_k (Hinde and Wood, 1987). One of the α_k parameters will be aliased with the intercept term β_0; alternatively, the intercept can be removed from the model.

Differentiating the log likelihood with respect to π_k and using $\pi_K = 1 - \sum_1^{K-1} \pi_k$, we have directly

$$\frac{\partial \ell}{\partial \pi_k} = \sum_j \frac{f_{jk} - f_{jK}}{\sum_\ell \pi_\ell f_{j\ell}} = \sum_j \left\{ \frac{w_{jk}}{\pi_k} - \frac{w_{jK}}{\pi_K} \right\}.$$

Equating this to zero gives simply

$$\hat{\pi}_k = \sum_j w_{jk}/n,$$

a standard mixture ML result. The same EM algorithm applies with the additional calculation in each M-step of the estimate of π_k from the weights. A distinctive feature of the weights is that they are calculated for each upper-level unit in the E step but applied to all lower-level units in this upper-level unit in the M-step. Initial estimates of the α_k can be taken as the standard normal values z_k.

A GLIM4 implementation of the EM algorithm for the simpler overdispersion model is given in Aitkin and Francis (1995) and the implementation for the very similar variance component model in Aitkin and Francis (1998).

Hypothesis testing or model comparisons can be carried out through the likelihood ratio test using differences of deviances ($-2 \log L_{\max}$) in the usual way. Theoretical justifications of this approach seem to be lacking, though simulation studies by Davies (1987) support the usual asymptotic null χ^2 distribution for nested model comparisons. Presumably, AIC and other penalized likelihood ratio criteria could also be used for model comparisons in the usual way.

5. Random Coefficient Models

The analysis can be extended to general random coefficient models. The usual case of interest is when lower-level variables have slopes that vary across upper-level units. Consider a simple example with a lower-level variable, x_{1ij}, whose coefficient β_1 varies across upper-level units. We index it by $\beta_{1j} = \beta_1 + u_j$, where u_j represents variation about a mean β_1. The remaining regression coefficients β_2 are fixed. Then, conditional on u_j *and* z_j, the regression model is

$$\eta_{ij} = \beta_1' x_{1ij} + \beta_2' x_{2ij} + z_j + u_j' x_{1ij},$$

while marginally z_j and u_j have an unknown joint distribution $\pi(z, u)$. The likelihood is then

$$L(\beta) = \prod_j \int \prod_i f(y_{ij} \mid z_j, u_j) \pi(z_j, u_j) dz_j du_j.$$

If the distribution $\pi(z, u)$ were assumed Gaussian with unknown covariance matrix, we would need to numerically integrate over both parameters, at least doubling the computational load and rendering this approach unusable for many random parameters. However, by estimating the joint distribution of z_j and u_j nonparametrically, we again obtain the NPML as a discrete distribution on a finite number of points in the (z, u) plane, with an estimated mass π_k and estimated mass points z_k and u_k in the kth component. The likelihood is again that of a finite mixture, this time of regressions on x_1 with a different slope and intercept in each component of the mixture, and on x_2 with the same regression coefficient in each component. Because the components are indexed in the model by the component factor, the random coefficients can be handled in the computational implementation by including in the regression model, in addition to the main effect of the factor z_k, the interaction of this factor with the explanatory variable x_{1ij}. Again, the number of mass points must be determined by sequential increase from 1, and in general more mass points might be required than for the case of fixed β_1. This process is quite general, and in the GLIM macro implementation the user provides both a FIXED macro of the fixed-effect terms in the model and a RANDOM macro of terms with random regression coefficients (the intercept term is always included).

Upper-level variable slopes can also be allowed to vary over upper-level units by including them in interactions with the random factor in the same way. Thus, the single-level overdispersion models considered in Aitkin (1996) can be generalized to random coefficient models in exactly the same way, using the same FIXED and RANDOM model macros. However, incorporating both overdispersion and variance component structure requires two sets of random effects that are both modeled nonparametrically. This is beyond the scope of the present paper.

6. Examples

6.1 *Example 1*

The first example is a small-area estimation study by Tsutakawa (1985) of cancer mortality rates in the 84 largest cities in Missouri for males aged 45–54 over the period 1972–1981. The data are shown in Tables 1–3.

Most of the cities are small and several have no lung cancer deaths at all over the 10-year period. There are three large cities. Tsutakawa gave both full Bayes and empirical Bayes analyses based on a logistic model with an added normal city random effect. The sample design is two level, with men nested in cities. As there are no explanatory variables at the lower (man) level, the two-level model collapses to a single-level binomial logit model with the city random effect acting as an overdispersion variable. Tsutakawa gave posterior distributions for the individual city mortality rates using Gaussian quadrature to evaluate the likelihood. The Bayes and empirical Bayes posteriors are very similar, apart from slightly more concentrated empirical Bayes posteriors for the small cities. Posterior means by these two approaches are also shown in the tables.

Fitting the null logit model to all 84 cities gives a deviance of 176.18 on 83 d.f. Successively increasing the number of mass points gives deviances of 93.10 for $K = 2$ and 92.38 for $K = 3$, which is the NPML estimate. Mass points (and masses) on the logit scale are −4.215 (.140), −4.757 (.436), and −4.886 (.424), giving a mean of −4.736 and standard deviation 0.214, close to the ML estimates for the normal distribution of −4.733 and 0.238 quoted by Tsutakawa. The two-component solution is very little inferior to the NPML estimate; this has mass points and (masses) of −4.217 (.155) and −4.836 (.845). We base our interpretation for simplicity on this model. The first mass point corresponds to a probability of .0145 and is identified by the third-largest city (84) with a population of 22,514 and an observed death rate of .0153. The second mass point corresponds to a probability of .0079 and is identified by four cities (4, 8, 13, and 44) with populations of 54,155, 5756, 7137, and 28,937 and corresponding observed rates of .0074, .0073, .0077, and .0087. The posterior mean rates for the other cities are weighted means of these two rates, weighted by the posterior probabilities of the city belonging to these components.

The posterior probability of component 1 and the NPML estimates based on the two-point model are also given in the tables. Also shown are the posterior means based on four-point Gaussian quadrature; these differ somewhat from the corresponding values in Tsutakawa, presumably because of different numbers of mass points in the quadrature (they differ

Table 1

Lung cancer mortality in 84 Missouri cities, males 45–54, 1972–1981

City	Size	Deaths	Raw rate	Bayes estimate	NPML estimate	Gauss estimate	Posterior probability
1	1019	2	20	67	79	73	.001
2	1512	8	53	74	79	78	.001
3	1424	8	56	76	79	79	.002
4	54,155	402	74	75	79	77	0.0
5	447	1	22	77	80	82	.017
6	1907	12	63	77	79	79	.001
7	1755	11	63	77	79	80	.001
8	5756	42	73	77	79	78	0.0
9	509	2	39	79	80	83	.020
10	350	1	29	80	81	84	.031
11	473	2	42	80	80	84	.026
12	329	1	30	81	81	85	.036
13	7137	55	77	80	79	79	0.0
14	430	2	47	82	81	85	.034
15	304	1	33	82	82	86	.042
16	163	0	0	83	83	87	.058
17	163	0	0	83	83	87	.058
18	159	0	0	83	83	87	.059
19	281	1	36	83	82	86	.049
20	154	0	0	83	83	87	.061
21	889	6	68	82	80	85	.019
22	260	1	38	83	82	87	.056
23	371	2	54	84	82	87	.050
24	232	1	43	85	83	88	.067
25	228	1	44	85	83	88	.069
26	343	2	58	85	83	88	.059
27	454	3	66	85	82	88	.053
28	323	2	62	85	83	88	.067

Table 2

Lung cancer mortality in 84 Missouri cities, males 45–54, 1972–1981

City	Size	Deaths	Raw rate	Bayes estimate	NPML estimate	Gauss estimate	Posterior probability
29	311	2	64	86	84	89	.072
30	784	6	77	85	81	88	.037
31	426	3	71	86	83	89	.063
32	184	1	55	87	85	89	.090
33	181	1	55	87	85	89	.092
34	177	1	56	87	85	90	.094
35	177	1	56	87	85	90	.094
36	291	2	69	87	84	89	.082
37	170	1	59	87	85	90	.098
38	158	1	63	88	86	90	.105
39	274	2	73	87	85	90	.091
40	150	1	67	88	86	91	.111
41	265	2	76	88	85	90	.096
42	257	2	78	88	85	91	.101
43	254	2	79	88	86	91	.103
44	28,937	251	87	87	79	78	0.00
45	445	4	90	89	85	91	.099
46	447	4	90	89	85	91	.098
47	329	3	91	90	86	92	.114
48	206	2	97	90	88	92	.137
49	313	3	96	90	87	92	.125
50	314	3	96	90	87	92	.125
51	314	3	96	90	87	92	.125
52	202	2	99	91	88	92	.140
53	198	2	101	91	88	93	.143
54	183	2	109	91	89	93	.156
55	292	3	103	91	88	93	.142
56	178	2	112	92	89	93	.161

also by a scale factor, as Tables 1–3 give the annual rates per 10^5 population at risk, whereas Tsutakawa gave rates per 10^6 population). Qualitatively, all the posterior means imply heavy smoothing of the highly variable sample rates nearly all of which are based on small city sizes. The smoothing is greatest for the two-point NPML estimate. Figure 1 shows the observed and posterior mean NPML rates in increasing order of the latter (so the city numbers do not correspond to those in Tables 1–3). The heavy smoothing is strikingly evident.

6.2 *Example 2*

The second example is the 22-center clinical trial of beta-blockers for reducing mortality after myocardial infarction, described by Yusuf et al. (1985) and analyzed in detail in Gelman et al. (1995) by MCMC. The data are given in Table 4, adapted from Gelman et al. (p. 149), and are represented by a two-level model, with centers at the upper level and patients at the lower level. There is only one explanatory variable, the treatment assignment at the lower level. This is coded as 0 for control and as 1 for the beta-blocker treatment.

Fitting the logit regression model with treatment as a fixed effect, ignoring the center classification, gives a deviance of 305.76 with 42 d.f., indicating large variations in intercept (response under the standard treatment) among the centers. The treatment effect estimate is −0.257 with standard error 0.049. Adding a 22-level center fixed factor to the model reduces the deviance to 23.62 with 21 d.f. The treatment effect estimate changes slightly, to −0.261, with standard error 0.050. Fitting the fixed treatment model with a random intercept term and successively more mass points, and estimating the mixing distribution nonparametrically gives deviances of 145.23 ($K = 2$), 101.29 ($K = 3$), and 101.29 ($K = 4$). (These deviances are not comparable with that for the center fixed-factor model.) The NPML is a nearly symmetric three-point distribution located at the points −1.610, −2.250, and −2.834, with respective masses 0.249, 0.512, and 0.239. The fixed treatment effect estimate is −0.258 with standard error 0.050. The standard deviation of the mixing distribution is 0.43, representing substantial variation on the logit scale. This distribution is very close to normal—the mass points are located similarly to those in the three-point Gaussian quadrature analysis (which are at 0 and $\pm\sqrt{3}\sigma$ relative to the intercept), though the probability masses are somewhat different—2/3, 1/6, and 1/6 for the Gaussian model. The deviance for the three-point Gaussian quadrature analysis is 103.55, very close to that for the NPML analysis; the treatment effect and standard error are identical to those from the NPML analysis. The ML estimate of the standard deviation of the random effect using three-point Gaussian quadrature is 0.36.

Including the treatment effect in the random part of the model—i.e., fitting a full random slope and intercept model—gives a deviance of 99.12, which is a reduction of only 2.10

Table 3
Lung cancer mortality in 84 Missouri cities, males 45–54, 1972–1981

City	Size	Deaths	Raw rate	Bayes estimate	NPML estimate	Gauss estimate	Posterior probability
57	287	3	105	92	88	93	.146
58	282	3	106	92	89	93	.150
59	164	2	122	92	90	94	.174
60	164	2	122	92	90	94	.174
61	1923	18	94	91	81	93	.030
62	3672	34	93	91	79	93	.005
63	261	3	115	93	90	94	.169
64	581	6	103	93	87	94	.131
65	550	6	109	94	89	95	.157
66	431	5	116	94	91	95	.183
67	399	5	125	96	93	96	.217
68	286	4	140	96	95	97	.242
69	592	7	118	96	93	97	.207
70	246	4	163	99	98	98	.295
71	547	7	128	98	96	98	.261
72	438	6	137	99	98	99	.284
73	202	4	198	101	103	100	.360
74	790	10	127	100	99	100	.306
75	648	9	139	102	104	102	.382
76	354	6	169	103	106	102	.411
77	730	10	137	103	105	102	.398
78	144	4	277	105	109	102	.453
79	1093	14	128	104	106	103	.406
80	384	7	182	107	113	104	.514
81	278	6	216	107	114	105	.537
82	596	10	168	110	120	107	.619
83	1889	28	148	120	142	117	.949
84	22,514	334	153	148	145	146	1.00

relative to the fixed treatment model. The NPML estimate is a three-point distribution again, as in the fixed treatment model, located at the (intercept, slope) mass points (−1.580, −0.325), (−2.248, −0.263), and (−2.916, −0.081) with respective masses 0.249, 0.511, and 0.240. Although there is no satisfactory formal test for the degeneracy of the two-dimensional distribution of slope and intercept to a one-dimensional distribution of the intercept, it is clear that the deviance change of 2.10 for the two additional slope parameters cannot constitute evidence of real treatment effect variations over the centers, and we conclude that a fixed treatment effect over the centers is well supported. The intercept variation is consistent with a normal distribution with standard deviation about 0.4. This result is consistent with the fixed-effect model conclusions, though the distribution of the residual deviance might not be accurately represented by χ^2_{21}.

Similar results are given by Gelman et al., who give a posterior median for the treatment effect of −0.25 with an approximate posterior standard deviation of about 0.07, though the posterior median of the standard deviation for their Gaussian random effect model is substantially smaller, 0.13, than the standard deviation 0.36 of the NPML-estimated random effect distribution. This might be a consequence of the additional information in the prior or of the use of a normal approximation for the likelihood contributions of the individual centers.

6.3 *Example 3*

The third example is the longitudinal study of childhood obesity of Woolson and Clarke (1984), reanalyzed by Fitzmaurice, Laird, and Lipsitz (1994). The data come from

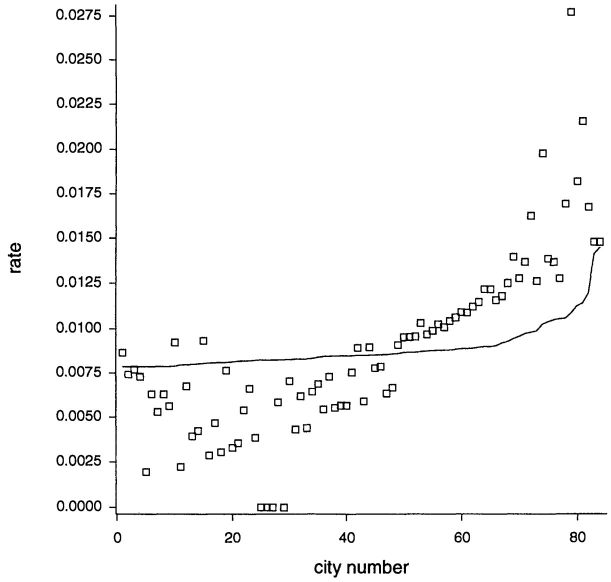

Figure 1. Observed and posterior mean rate.

Table 4

Results of 22 clinical trials of beta-blockers

Center j	Control deaths	Total	Treated deaths	Total
1	3	39	3	38
2	14	116	7	114
3	11	93	5	69
4	127	1520	102	1533
5	27	365	28	355
6	6	52	4	59
7	152	939	98	945
8	48	471	60	632
9	37	282	25	278
10	188	1921	138	1916
11	52	583	64	873
12	47	266	45	263
13	16	293	9	291
14	45	883	57	858
15	31	147	25	154
16	38	213	33	207
17	12	122	28	251
18	6	154	8	151
19	3	134	6	174
20	40	218	32	209
21	43	364	27	391
22	39	674	22	680

Table 5

Muscatine study: all data, males

Child's obesity status at	Age 8	Age 10	Age 12	Count
No missing	1	1	1	20
	1	1	0	7
	1	0	1	9
	1	0	0	8
	0	1	1	8
	0	1	0	8
	0	0	1	15
	0	0	0	150
Missing time 1	*	1	1	13
	*	1	0	3
	*	0	1	2
	*	0	0	42
Missing time 2	1	*	1	3
	1	*	0	1
	0	*	1	6
	0	*	0	16
Missing time 3	1	1	*	11
	1	0	*	1
	0	1	*	3
	0	0	*	38
Missing times 1, 2	*	*	1	14
	*	*	0	55
Missing times 1, 3	*	1	*	4
	*	0	*	33
Missing times 2, 3	1	*	*	7
	0	*	*	45

the Muscatine, Iowa, study of 1,014 children who were 7–9 years old in 1977 and were followed up in 1979 and 1981. Children were classified as obese if their weight was more than 110% of the median weight for their gender and height; about 20% of children were classified as obese. The repeated binary response of interest is whether the child is obese (1) or not (0) at each occasion. Data on many children are incomplete, and only 460 children had complete data from all three occasions. Tables 5 and 6, adapted from Fitzmaurice et al., give the child's obesity status at all three occasions for the 1014 children with complete or incomplete data.

Fitzmaurice et al. modeled the marginal probability of response by a logistic model with linear and quadratic age terms and their interactions with gender, and saturated the covariance matrix between occasions. This is similar to the GEE approach of Liang and Zeger, but the parameters are estimated by full maximum likelihood. They analyzed both the subset of children with complete data and the full sample with both complete and incomplete data and found substantial changes in the conclusions, demonstrating the need for inclusion of all the data in the analysis.

We repeat their analysis with the random effect model, with the two-level structure of children and occasions within child. The intraclass correlation structure on the logit scale is simple, though on the probability scale this corresponds to a complex correlation structure because of the nonlinear transformation. The linear and quadratic age effects are defined by AGE(L) = (AGE−10)/2, AGE(Q) = 3*AGE(L)^2 − 2, and the GENDER variable is defined as 0 for males, 1 for females. Interactions are defined as direct products of the GENDER and AGE(L), AGE(Q) effects.

Table 6

Muscatine study: all data, females

Child's obesity status at	Age 8	Age 10	Age 12	Count
No missing	1	1	1	21
	1	1	0	6
	1	0	1	6
	1	0	0	2
	0	1	1	19
	0	1	0	13
	0	0	1	14
	0	0	0	154
Missing time 1	*	1	1	8
	*	1	0	1
	*	0	1	4
	*	0	0	47
Missing time 2	1	*	1	4
	1	*	0	0
	0	*	1	16
	0	*	0	3
Missing time 3	1	1	*	11
	1	0	*	1
	0	1	*	3
	0	0	*	25
Missing times 1, 2	*	*	1	13
	*	*	0	39
Missing times 1, 3	*	1	*	5
	*	0	*	23
Missing times 2, 3	1	*	*	7
	0	*	*	47

Table 7
Parameter estimates from full model for complete data only

Parameter	Marginal ML estimate	SE	NPML estimate	SE
Intercept	−1.353	0.113	0.954	
GENDER	0.051	0.190	0.035	0.334
AGE(L)	0.106	0.077	0.228	0.168
AGE(Q)	0.045	0.047	0.095	0.098
GENDER.AGE(L)	0.230	0.119	0.442	0.236
GENDER.AGE(Q)	−0.149	0.065	−0.301	0.136

Table 7 gives the parameter estimates and standard errors for the full interaction model AGE(L) + AGE(Q) + GENDER + GENDER.AGE(L) + GENDER.AGE(Q), from both the Fitzmaurice et al. and the random effect model approaches, for the subset of complete data. For the random effect model, the NPML estimate of the mixing distribution is a symmetric three-point distribution, with masses of 0.251, 0.510, and 0.239 at +3.582, −0.162, and −3.414, respectively. This distribution has a standard deviation of 2.45 but has shorter tails than the three-point Gaussian quadrature distribution. Standard errors for the parameter estimates for this model are obtained by omitting each effect in turn from the full model, as described in Section 3.

Both approaches show a significant gender × quadratic age interaction, so the model cannot be further reduced. Table 8 gives the corresponding estimates for the full sample. Neither interaction term is now important, and the model can be reduced to the main effect model, also shown. The gender effect is clearly irrelevant, and the model can be reduced to a main effect age model, also shown, and further reduced to a linear age model. (Fitzmaurice et al. did not give estimates for this model.)

It is of interest that both ML analyses lead to the same models with nearly proportional parameter estimates and standard errors, though the models being fitted are different: a marginal logit model in the first case and a conditional model in the second. The inflation of estimates and standard errors for the NPML analysis is a consequence of its much greater variance on the logit scale, $\pi^2/3 + \sigma^2 = 9.29$ compared with $\pi^2/3 = 3.29$ for the marginal logit model. The corresponding intraclass correlation is 0.65. Neuhaus and Jewell (1990) report similar inflation in a paired study, and Neuhaus (1992) gives a general discussion of the relation between parameter estimates by these and other approaches.

7. Discussion

7.1 *Theoretical Issues*

The approach described here is very general. The emphasis is on model fitting while allowing for the random effect rather than on testing for a nonzero variance component. Because the mixing distribution is treated as a nuisance function and estimated nonparametrically, the distribution of the change in deviance on fitting the variance component model compared to an independence model is not of direct interest: the regression model parameters are the parameters of interest. (The distribution of this change is, however, of considerable theoretical interest, especially in random parameter models.)

Table 8
Parameter estimates from full model for full data

Parameter	Marginal ML estimate	SE	NPML estimate	SE
Intercept	−1.356	0.098	1.194	
GENDER	0.043	0.138	0.025	0.288
AGE(L)	0.142	0.063	0.339	0.150
AGE(Q)	0.014	0.035	0.032	0.088
GENDER.AGE(L)	0.162	0.096	0.345	0.209
GENDER.AGE(Q)	−0.089	0.049	−0.192	0.119
Deviance			1892.06	
Intercept	−1.370	0.097	1.151	
GENDER	0.073	0.137	0.002	*
AGE(L)	0.223	0.048	0.507	0.105
AGE(Q)	−0.032	0.025	−0.065	0.056
Deviance			1897.27	
Intercept	−1.321	0.094	1.151	
AGE(L)	0.220	0.059	0.507	0.105
AGE(Q)	−0.033	0.032	−0.065	0.059
Deviance			1897.27	
Intercept			1.165	
AGE(L)			0.507	0.105
Deviance			1898.50	

* No SE as the deviance change was 0.00.

It appears that deviance changes on omitting explanatory variables from the model can be treated as asymptotic χ^2 in the usual LRT framework even without conditioning on the number of mass points as in Follmann and Lambert (1989).

A limitation of the NPML analysis in the first example, and generally in single-level random effect models, is that it does not allow spatial dependence between neighboring units, as used, e.g., in Clayton and Kaldor (1987), Breslow and Clayton (1993), and other authors. A popular extension (e.g., Besag, York, and Mollie, 1991) of the simple random effect model for disease mapping is to include an additional spatial random effect for each area whose (conditional) mean is set equal to the mean of the random effects for neighboring areas (appropriately defined). Care is needed in such models as the joint distribution of all the random effects is likely to be singular unless it incorporates a regression parameter to reduce the very high intra-area correlation implied by the construction of the conditional means. Initial spatial examination of the posterior means from the model without spatial dependence should be carried out to establish whether such dependence actually exists; as Clayton and Kaldor (1987) note, "There is no a priori reason why geographic proximity should be reflected in correlated cancer rates." It should be noted that the simple variance component model provides consistent estimates of the regression model parameters, but these are inefficient if real spatial dependence exists, and extensions of the NPML approach to spatial modeling would be very useful.

In the second example, we establish that treatment effect variation over centers is negligible, an important practical issue, whereas variation in the standard treatment response is substantial. The ML estimate of this variation is somewhat different from the posterior median in the Bayes analysis, presumably a consequence of the additional information in the prior for this parameter, though it might also be a consequence of the normal distribution assumption in the Bayes analysis.

In the third example, we obtain essentially the same results as Fitzmaurice et al., though with a conditional rather than a marginal model.

7.2 *Computational Issues*

The EM algorithm for these variance component models is very stable and converged rapidly in every case. For the Tsutakawa data, the EM algorithm required 20 iterations for the two-point model and 81 for the three-point, using a convergence criterion of successive deviance change less than 0.001. The example from Gelman et al. required 8 iterations for the two-point, 11 for the three-point, and 24 for the four-point (fixed treatment) model and 11, 9, and 13 iterations, respectively, for the two-, three-, and four-mass-point random treatment models, with the same convergence criterion. The Muscatine obesity data required 40 iterations for the two-point model and 25 for the three-point, with the full regression model for the subset of complete data, and 49 and 26 iterations, respectively, for the full data, again with the same convergence criterion. The convergence rate is much faster than for the overdispersion models considered in Aitkin (1996) because there is much more information about the random effects when these are shared between multiple lower-level units. In overdispersion models, each observation has its own unique random effect. The NPML estimate was impressively stable and required at most three mass points in the small examples considered. This is a consequence of the small (upper-level) sample sizes in these examples; in large samples, substantially larger numbers of mass points are frequently required.

The discrete nature of the NPML estimate might be found unattractive if one believes *a priori* in the existence of a continuous mixing distribution. An alternative approach that assumed a smooth mixing density was described by Davidian and Gallant (1993). This approach requires numerical quadrature with library optimization algorithms. Magder and Zeger (1996) took as a mixing distribution a finite mixture of normals, guaranteeing a continuous mixing distribution estimate. They found that the likelihood for the model with equal variances approaches a maximum as the common variance tends to zero (reflecting the optimality of the NPML estimate over all mixing distributions) and is very flat in the variance parameter away from zero, so the information about distributional shape of the mixing distribution is inherently limited.

Local maxima of the likelihood are a possibility but were not found in the previous examples, though they have been in others. Local maxima may require variations in starting values for the EM algorithm to locate all the local maxima. In the GLIM4 macro used, this is most easily achieved by scaling the mass-point locations in the initial mass-point macros. In other examples not reported here, a considerable difference in local maxima sometimes occurred, depending on whether the number of mass points was odd or even. Reliable estimation of the true maximum in these cases was found by overfitting the number of mass points and identifying the location of the reduced number of points actually required for the NPML estimate.

A further computational issue in GLIM is that only six-figure accuracy can be obtained in the deviance, and this might require relaxing or changing the convergence criterion (of successive deviance changes) in large data sets to prevent roundoff error fluctuations in the value of the deviance.

Particular advantages of this approach compared to other approaches to NPML estimation are that no special computational effort is required to locate new mass points when K is increased and that the mass points need not be restricted to a grid.

Past experience (Hinde, 1982; Anderson and Aitkin, 1985) with mixture modeling for overdispersion and variance component analysis might have left the discouraging impression that the problem is computationally intensive. Although the Gauss–Newton algorithm might give more efficient model fitting (Aitkin and Aitkin, 1996, report a modest improvement with a hybrid EM/Gauss–Newton algorithm for normal mixtures), with present and projected future CPU speeds on personal computers this no longer seems such a serious issue, and the simplicity and generality of the random effect model and of the EM algorithm for full NPML estimation in exponential family variance component models make them powerful modeling tools. The lack of standard errors for the EM estimates can be rectified by a modest amount of further modeling or by explicit calculation of the observed information.

ACKNOWLEDGEMENTS

I am very much indebted to John Hinde, Andy Wood, and especially Brian Francis for helpful comments on computational issues and to the referees and an associate editor for references.

RÉSUMÉ

Cet article décrit un algorithme EM pour une estimation non paramétrique du maximum de vraisemblance dans le cadre des modèles linéaires généralisés avec structure sur les composantes de la variance. L'algorithme fournit une alternative aux analyses MQL et PQL (Goldstein 1995, McGilchrist et Aisbett 1991, Breslow et Clayton 1993, McGilchrist 1994), et aux GEE (Liang et Zeger 1986). L'algorithme, donné par Hinde et Wood (1987), est généralisé aux modèles de surdispersion à effets aléatoires dans les modèles linéaires généralisés, décrit par Aikin (1996). L'algorithme est initialement dérivé comme une forme de quadrature de Gauss supposant une distribution mélangeante de lois normales, avec une faible variation il peut être utilisé pour une distribution mélangeante de lois totalement inconnues, donnant une méthode directe de l'estimation non paramétrique complète du maximum de vraisemblance de cette distribution. Ceci est utile car les estimations du maximum de vraisemblance des paramètres des GLM peuvent être sensibles à la spécification de la forme paramétrique de la distribution mélangeante. L'analyse non paramétrique peut être étendue directement aux modèles à paramètres aléatoires, avec une estimation complète NPML de la distribution jointe des paramètres aléatoires. Ceci peut produire un gain de temps machine considérable comparer à l'intégration numérique complète pour une distribution paramétrique spécifiée des paramètres aléatoires. Une méthode simple est décrite pour obtenir des écarts type exactes des estimations des paramètres quand on utilise l'algorithme EM. Plusieurs exemples sont discutés, comportant des modèles simples pour les composants de variance, des modèles longitudinaux, et des estimations dans des petites zones.

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. and Aitkin, I. (1996). A hybrid EM/Gauss-Newton algorithm for maximum likelihood in mixture distributions. *Statistics and Computing* **6,** 127–130.

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

Aitkin, M. and Francis, B. J. (1998). Fitting generalized linear variance component models by nonparametric maximum likelihood. *The GLIM Newsletter*, in press.

Aitkin, M. and Tunnicliffe Wilson, G. T. (1980). Mixture models, outliers and the EM algorithm. *Technometrics* **22,** 325–331.

Aitkin, M., Anderson, D. A., Francis, B.J., and Hinde, J. P. (1989). *Statistical Modelling in GLIM.* Oxford: Clarendon Press.

Anderson, D. A. (1988). Some models for overdispersed binomial data. *Australian Journal of Statistics* **30,** 125–148.

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.

Anderson, D. A. and Hinde, J. P. (1988). Random effects in generalized linear models and the EM algorithm. *Communications in Statistics—Theory and Methods* **17,** 3847–3856.

Barry, J. T., Francis, B. J., and Davies, R. B. (1989). SABRE: Software for the analysis of binary recurrent events. In *Statistical Modelling*, A. Decarli, B. J. Francis, R. Gilchrist, and G. U. H. Seeber (eds). New York: Springer-Verlag.

Besag, J., York, J., and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics* **43,** 1–59.

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika* **46,** 443–459.

Böhning, D., Schlattman, P., and Lindsay, B. (1992). Computer-assisted analysis of mixtures (C.A.MAN): Statistical algorithms. *Biometrics* **48,** 285–303.

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

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* **43,** 671–681.

Crouch, E. A. C. and Spiegelman, D. (1990). The evaluation of integrals of the form $\int_{-\infty}^{+\infty} f(t)exp(-t^2)dt$: Application to logistic-normal models. *Journal of the American Statistical Association* **85,** 464–469.

Davidian, M. and Gallant, A. R. (1993). The nonlinear mixed effects model with a smooth random effects density. *Biometrika* **80,** 475–488.

Davies, R. B. (1987). Mass point methods for dealing with nuisance parameters in longitudinal studies. In *Longitudinal Data Analysis*, R. Crouchley (ed). Aldershot: Avebury.

Dempster, A. P., Laird, N. M., and Rubin, D. A. (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. (1992). Estimation of heterogeneity—A GLM approach. In *Advances in GLIM and Statistical Modelling*, L. Fahrmeir, B. Francis, R. Ailchrist, and G. Tutz (eds). New York: Springer-Verlag.

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). New York: Springer-Verlag.

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

Ezzet, F. and Davies, R. B. (1988). *A Manual for MIXTURE.* Lancaster, U.K.: Centre for Applied Statistics.

Fitzmaurice, G. M., Laird, N. M., and Lipsitz, S. R. (1994). Analysing incomplete longitudinal binary responses: a likelihood-based approach. *Biometrics* **50,** 601–612.

Follmann, D. A. and Lambert, D. (1989). Generalizing logistic regression by nonparametric mixing. *Journal of the American Statistical Association* **84,** 295–300.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis.* London: Chapman and Hall.

Goldstein, H. (1995). *Multilevel Statistical Models*, 2nd edition. London: Edward Arnold.

Heckman, J. J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models of duration. *Econometrica* **52,** 271–320.

Hinde, J. P. (1982). Compound Poisson regression models. In *GLIM82*, R. Gilchrist (ed). New York: Springer-Verlag.

Hinde, J. P. and Wood, A. T. A. (1987). Binomial variance component models with a non-parametric assumption concerning random effects. In *Longitudinal Data Analysis,* R. Crouchley (ed). Avebury, Aldershot, Hants.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *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.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38,** 963–974.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B* **58,** 619–678.

Lesperance, M. L. and Kalbfleisch, J. D. (1992). An algorithm for computing the nonparametric MLE of a mixing distribution. *Journal of the American Statistical Association* **87,** 120–126.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22.

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

Longford, N. T. (1993). *Random Coefficient Models.* Oxford: Clarendon Press.

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.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* London: Chapman and Hall.

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

McGilchrist, C. A. (1994). Estimation in generalized mixed models. *Journal of the Royal Statistical Society, Series B* **56,** 61–69.

McGilchrist, C. A. and Aisbett, C. W. (1991). Restricted BLUP for mixed linear models. *Biometrical Journal* **33,** 131–141.

Magder, L. S. and Zeger, S. L. (1996). A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *Journal of the American Statistical Association* **91,** 1141–1151.

Mallet, A. (1986). A maximum likelihood estimation method for random coefficient regression models. *Biometrika* **73,** 645–656.

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.

Neuhaus, J. M. (1992). Statistical methods for longitudinal and clustered designs with binary responses. *Statistical Methods in Medical Research* **1,** 249–273.

Neuhaus, J. M. and Jewell, N. P. (1990). Some comments on Rosner's multiple logistic model for clustered data. *Biometrics* **46,** 523–534.

Rodriguez, G. and Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A* **158,** 73–89.

Schumitzky A. (1991). Nonparametric EM algorithms for estimating prior distributions. *Applied Mathematics and Computation* **45,** 143–157.

Steele, B. M. (1996). A modified EM algorithm for estimation in generalized mixed models. *Biometrics* **52,** 1295–1310.

Tsutakawa, R. K. (1985). Estimation of cancer mortality rates: A Bayesian analysis of small frequencies. *Biometrics* **41,** 69–79.

Walker, S. (1996). An EM algorithm for nonlinear random effects models. *Biometrics* **52,** 934–944.

Woolson, R. F. and Clarke, W. R. (1984). Analysis of categorical incomplete longitudinal data. *Journal of the Royal Statistical Society, Series A* **147,** 87–99.

Yusuf, S., Peto, R., Lewis, J., Collins, R., and Sleight, P. (1985). Beta blockade during and after myocardial infarction: An overview of the randomized trials. *Progress in Cardiovascular Diseases* **27,** 335–371.

Zackin, R., de Gruttola, V., and Laird, N. (1996). Nonparametric mixed-effects models for repeated binary data arising in serial dilution assays: Application to estimating viral burden in AIDS. *Journal of the American Statistical Association* **91,** 52–61.

Received March 1996. *Revised October* 1997.
Accepted January 1998.