Statistical Methods in Medical Research 1997; **6**: 76–98

On the EM algorithm for overdispersed count data

GJ McLachlan Department of Mathematics, The University of Queensland, Queensland, Australia

In this paper, we consider the use of the EM algorithm for the fitting of distributions by maximum likelihood to overdispersed count data. In the course of this, we also provide a review of various approaches that have been proposed for the analysis of such data. As the Poisson and binomial regression models, which are often adopted in the first instance for these analyses, are particular examples of a generalized linear model (GLM), the focus of the account is on the modifications and extensions to GLMs for the handling of overdispersed count data.

1 Introduction

In medical research, data are often collected in the form of counts, corresponding to the number of times that a particular event of interest occurs. Because of their simplicity, one-parameter distributions for which the variance is determined by the mean are often used at least in the first instance to model such data. Familiar examples are the Poisson and binomial distributions, which are members of the one-parameter exponential family. However, there are many situations, where these models are inappropriate, in the sense that the mean–variance relationship implied by the one-parameter distribution being fitted is not valid. In most of these situations, the data are observed to be overdispersed; that is, the observed sample variance is larger than that predicted by inserting the sample mean into the mean–variance relationship. This phenomenon is called overdispersion. There are occasions in data analysis where the data are underdispersed; that is, the sample variance is smaller than that implied by the mean–variance relationship, called underdispersion. These phenomena are also observed with the fitting of regression models, where the mean (say of the Poisson or the binomial distribution), is modelled as a function of some covariates. If this dispersion is not taken into account, then using these models may lead to biased estimates of the parameters and consequently incorrect inferences about the parameters (Wang[1] and Wang *et al.*[2]). We focus here on the more common case of overdispersion.

Aitkin *et al.*,[3] Breslow,[4–6] Breslow and Clayton,[7] Brillinger,[8] Clayton,[9] Cox,[10] Efron,[11,12] Gelfand and Dalal,[13] Hinde,[14] Lawless[15,16] and McCullagh and Nelder[17] (Section 6.2), among many others, have discussed the analysis of count data when overdispersion is present. Score statistics for detecting extra-Poisson variation have been developed by Fisher[18], Collings and Margolin[19], Cameron and Trivedi,[20,21] Dean and Lawless[22] and Dean.[23] In addition to relevant work in these papers, tests for extra-

Address for correspondence: GJ McLachlan, Department of Mathematics, The University of Queensland, Queensland 4072, Australia. Email: gjm@maths.uq.edu.au

 0962-2802(97)SM126RA

binomial variation have been presented by Tarone,[24] Williams[25] and Prentice.[26] Lambert and Roeder[27] have considered overdispersion diagnostics for generalized linear models.

Various approaches to handling overdispersion have been suggested over the years, including the use of quasi-likelihood, continuous mixture models, and random effects models whose use has been broadened through the recent development of hierarchical generalized linear models by Lee and Nelder.[28] In recent times, much attention has been given to the use of finite mixture models for which the EM algorithm plays a central role in computing iteratively the maximum likelihood estimates (MLEs) of the parameters. The EM algorithm can also play useful role in the implementation of the random effects model. The Poisson and binomial regression models are particular examples of generalized linear models (GLMs) that can be fitted by maximum likelihood via an iteratively reweighted least-squares algorithm as in the GLIM program (Francis *et al.*[29]). We shall briefly review this approach to the fitting of a standard GLM model before proceeding to describe modifications and extensions to GLMs for the handling of overdispersed data.

2 Generalized linear models

2.1 Maximum likelihood approach

With the generalized linear model (GLM) approach originally proposed by Nelder and Wedderburn,[30] the log density of the response variable Y for a given subject has the form

$$\log f(y;\, \theta, \phi) = m\phi^{-1}\{\theta y - b(\theta)\} + c(y;\, \phi) \tag{2.1}$$

where θ is the natural or canonical parameter, ϕ is the dispersion parameter and m is the prior weight. In the case of discrete Y, we still view $f(y;\, \theta, \phi)$ as a density by the adoption of counting measure. We use f throughout the paper as a generic symbol for a density. The mean and variance of Y are given by

$$E(Y) = \mu = b'(\theta)$$

and

$$\text{var}(Y) = \phi b''(\theta)$$

respectively, where the prime denotes differentiation with respect to θ. In a GLM, it is assumed that

$$\begin{aligned} \eta &= g(\mu) \\ &= \mathbf{x}^T \boldsymbol{\beta} \end{aligned}$$

where $\mathbf{x}$ is a vector of covariates or explanatory variables and $\boldsymbol{\beta}$ is a vector of unknown parameters, and $g(\cdot)$ is a monotonic function known as the link function. Here the superscript T refers to vector transpose. If the dispersion parameter ϕ is known, then the distribution (2.1) is a member of the (regular) exponential family with natural or canonical parameter θ. The distribution may or may not be a member of the two-

parameter exponential family if ϕ is unknown. The variance of Y is the product of two terms, the dispersion parameter ϕ and the variance function $b''(\theta)$, which is usually written in the form as

$$V(\mu) = \partial\mu/\partial\theta$$

So-called natural or canonical links occur when $\eta = \theta$, which are respectively the log and logit functions for the Poisson and binomial distributions; see Table 2.1 in McCullagh and Nelder[17](Chapter 2). In the standard form of a GLM, μ is modelled as a function of the unknown parameter vector $\boldsymbol{\beta}$, assuming ϕ fixed and with $V(\mu)$ containing no unknown parameters. Of distributions of the form (2.1), the Poisson and binomial have $\phi = 1$ (that is, fixed a priori at 1). The negative binomial distribution, whose variance function can be written in the form

$$V(\mu) = \mu + \mu^2 k$$

is an example of a variance function containing an unknown parameter that is not a dispersion parameter. Suppose $y_1, \ldots y_n$ denote n independent observations on the response variable, where Y_j has prior weight m_j, canonical parameter θ_j, mean μ_j, and covariate vector $\mathbf{x}_j$ $(j = 1, \ldots, n)$. Then the log likelihood for $\boldsymbol{\beta}$ is given by

$$\log L(\boldsymbol{\beta}) = \sum_{j=1}^{n} [m_j \phi^{-1}\{\theta_j y_j - b(\theta_j)\} + c(y_j;\, \phi)] \tag{2.2}$$

On differentiation in (2.2) with respect to $\boldsymbol{\beta}$ using the chain rule (McCullagh and Nelder,[17] Section 2.5), the likelihood equation for $\boldsymbol{\beta}$ can be expressed as

$$\sum_{j=1}^{n} m_j w(\mu_j)(y_j - \mu_j)\eta'(\mu_j)\{\partial\eta(\mu_j)/\partial\boldsymbol{\beta}\} = \mathbf{0} \tag{2.3}$$

where $\eta'(\mu) = \mathrm{d}\eta/\mathrm{d}\mu$ and $w(\mu)$ is the weight function defined by

$$w(\mu) = 1/[\{\eta'(\mu)\}^2 V(\mu)]$$

In (2.3), $\partial\eta(\mu_j)/\partial\boldsymbol{\beta} = \mathbf{x}_j$, but we have left it as such in (2.3) for comparative purposes in Section 5 with a mixture of GLMs. It can be seen that for fixed ϕ, the likelihood equation for $\boldsymbol{\beta}$ is independent of ϕ. The likelihood equation (2.3) can be solved iteratively using Fisher's method of scoring, which for a GLM is equivalent to using iteratively reweighted least squares.[30] On the $(k+1)$th iteration, we form the adjusted response variable $\tilde{y}_j$ as

$$\tilde{y}_j^{(k)} = \eta(\mu_j^{(k)}) + (y_j - \mu_j^{(k)})\eta'(\mu_j^{(k)}) \tag{2.4}$$

These n adjusted responses are then regressed on the covariates $\mathbf{x}_1, \ldots, \mathbf{x}_n$ using weights $m_1 w(\mu_1^{(k)}), \ldots, m_n w(\mu_n^{(k)})$. This produces an updated estimate $\boldsymbol{\beta}^{(k+1)}$ for $\boldsymbol{\beta}$, and hence updated estimates $\mu_j^{(k+1)}$ for the μ_j, for use in the right-hand side of (2.4) to update the adjusted responses, and so on. This process is repeated until changes in the estimates are sufficiently small.

2.2 Quasi-likelihood approach

For all GLMs, we have the relation

$$\partial \log f(y;\, \theta, \phi)/\partial \mu = m(y - \mu)/\{\phi V(\mu)\}$$

so that this first derivative depends only on the first two moments of Y. This led Wedderburn[31] to define a quasi-likelihood approach by the relation

$$\partial q/\partial \mu = m(y - \mu)/\{\phi V(\mu)\}$$

The use of q as a criterion for fitting allows the class of GLMs to be extended to models defined only by the properties of the first two moments. The function q will correspond to a true log density if there is a distribution of the GLM type for which

$$\text{var}(Y) = \phi V(\mu) \tag{2.5}$$

For a fixed value of the dispersion parameter ϕ, the quasi-likelihood approach estimates $\boldsymbol{\beta}$ by the value of $\boldsymbol{\beta}$ that minimizes the sum of weighted squares

$$\sum_{j=1}^{n} m_j (y_j - \mu_j)^2 / \{\phi V(\mu_j)\} \tag{2.6}$$

A simple moment estimate of ϕ is obtained by the value of ϕ that makes the mean deviance equal to one or the expected value of the Pearson statistic equal to its degrees of freedom. With the latter, ϕ is obtained as a root of the equation

$$\sum_{j=1}^{n} m_j (y_j - \mu_j)^2 / \{\phi V(\mu_j)\} = (n - d) \tag{2.7}$$

where d is the number of parameters in the model. In the context of allowing for extra-Poisson variation, Breslow[4] suggested first fitting the ordinary Poisson model with $\phi = 1$ to obtain an initial estimate of μ_j for use in the left-hand side of (2.7). The value of ϕ obtained from (2.7) is then substituted into (2.8) to produce a new estimate of $\boldsymbol{\beta}$ and hence the μ_j, which are substituted into the left-hand side of (2.7) to produce a new estimate of ϕ, and so on. This process can be continued until convergence. A detailed review of the quasi-likelihood approach may be found in the book of McCullagh and Nelder.[17] It is well known that this approach leads to a consistent and efficient estimate of $\boldsymbol{\beta}$; see also Lawless,[16] Stirling[32] and Kim.[33]

3 Poisson regression model

3.1 Some standard modifications for overdispersed data

We consider now the Poisson regression model. We shall briefly review some modifications that can be made to it within a single-GLM framework for the modelling of overdispersed count data before proceeding to consider some methodology that can be implemented using a finite mixture of GLMs. The Poisson regression model is an example of a GLM in which the distribution of the response Y with covariate vector $\mathbf{x}$ is Poisson with density

$$f(y;\mu) = \{e^{-\mu}\mu^y/y!\}I_A(y), \tag{3.1}$$

which has mean $E(Y) = \mu$, and the natural link is the log function

$$\begin{aligned} g(\mu) &= \log \mu \\ &= \boldsymbol{\beta}^T \mathbf{x} \end{aligned}$$

In (3.1), $A = \{0, 1, 2, \ldots\}$ is the set of nonnegative integers and $I_A(y)$ is the indicator function, which is one if y belongs to the set A and is zero otherwise. In many situations in practice, the population size or, say, the time of exposure varies for each subject so that the mean of Y is given by $a\mu$, where a denotes the known population size or time of exposure, and μ now denotes the mean rate per unit size or time. This can be dealt with in the theory and software for GLMs by either declaring a as an 'offset' in the specification of the linear predictor or by redefining the response to be the observed rate y/a, with a^{-1} specified as the prior weight. Hence in the sequel we shall assume without loss of generality that $a = 1$ for all subjects. A consequence of using the Poisson regression model is that the variance equals the mean; that is

$$\begin{aligned} \text{var}(Y) &= E(Y) \\ &= \mu \end{aligned}$$

In practice, however, we often have overdispersed data; that is

$$\text{var}(Y) > \mu$$

When the Poisson regression model fits the count data poorly, overdispersion is often a cause of the problem. There are several ways to modify the Poisson regression model. Using the GLM formulation, we can modify it by either choosing an alternative link function or an alternative frequency distribution, or both. Since the log link has properties such as multiplicative effects of covariates on the Poisson mean, researchers have suggested the use of alternative link functions. On the other hand, there are a lot of studies of alternative frequency distributions for the Poisson distribution; see, for example, Breslow,[4] Efron[11,12] and Lawless.[16]

3.2 Gamma-Poisson mixture model

A classical approach is to use a continuous Poisson mixture model to adjust for extra-Poisson variation. In this framework in the nonregression case, the Poisson mean μ is taken to be a latent variable from a distribution, $H(\mu)$, so that the density of Y is modelled as

$$f(y) = \int_0^\infty \{e^{-\mu}\mu^y/y!\}I_A(y)\, \mathrm{d}H(\mu) \tag{3.2}$$

A common choice for $H(\mu)$ in (3.2) is the gamma (α, β) distribution, which has probability density function (pdf)

$$\{\beta^{\alpha}u^{\alpha-1}/\Gamma(\alpha)\}\exp(-\beta u)I_{[0,\infty)}(u), \quad (\alpha, \beta > 0) \tag{3.3}$$

This leads to the density of Y being modelled as

$$f(y; \alpha, \beta) = \binom{y+\alpha-1}{y}\left(\frac{\beta}{\beta+1}\right)^{\alpha}\left(\frac{1}{\beta+1}\right)^{y} I_A(y) \tag{3.4}$$

which is the negative binomial distribution $NB(\alpha, \beta/(\beta+1))$. This distribution is the model for the number of tails y, in independent flips of a coin with probability of heads equal to $\beta(\beta+1)$, until one observes α heads. On letting μ now denote the mean of this distribution, we have that

$$\mu = \alpha/\beta$$

while its variance is

$$\begin{aligned}\text{var}(Y) &= (\alpha/\beta)\{(\beta+1)/\beta\} \\ &= \mu + k\mu^2 \end{aligned} \tag{3.5}$$

where $k = 1/\alpha$. Hence this two-parameter model allows the variance to be greater than the mean, with the variance equal to the mean inflated multiplicatively by the factor $(1+\mu k)$. As k tends to zero in (3.4), we obtain the Poisson model. We can rewrite (3.4) as

$$f(y; \mu, k) = \binom{y+k^{-1}-1}{y}\left(\frac{k^{-1}}{\mu+k^{-1}}\right)^{k^{-1}}\left(\frac{\mu}{\mu+k^{-1}}\right)^{y} I_A(y) \tag{3.6}$$

This is the standard negative binomial model for extra-Poisson variation, and it can be seen that it arises by assuming that α is fixed as μ varies. If, however, we assume that μ varies with α and β remains constant, we obtain a negative binomial distribution with

$$\text{var}(Y) = \mu(1+k) \tag{3.7}$$

where $k = 1/\beta$. This distribution does not have the form of a standard GLM; see Nelder and Lee.[34] For the negative binomial distribution (3.6), the maximum likelihood estimate of $\boldsymbol{\Psi} = (\boldsymbol{\beta}^T, k)^T$ can be obtained as described in Lawless[16] for the log linear model

$$\log\mu = \boldsymbol{\beta}^T\mathbf{x}$$

It was noted there that the results for other regression specifications are qualitatively similar. Within the GLM framework, $\boldsymbol{\Psi}$ can be estimated as described in McCullagh and Nelder[17](Section 11.2). For fixed k, the negative binomial distribution (3.6) has the form of a GLM with canonical link

$$\eta(\mu) = \log\{\mu/(\mu+k^{-1})\}$$

and variance function

$$V(\mu) = \mu + k\mu^2$$

Ordinarily, k is unknown. Using the method of moments, it can be computed as a solution of

$$\sum_{j=1}^{n} \frac{(y_j - \hat{\mu}_j)^2}{\hat{\mu}_j(1 + k\hat{\mu}_j)} = n - d \tag{3.8}$$

where $\hat{\mu}_j$ is the current estimate of μ_j. Hence $\boldsymbol{\beta}$ can be estimated by a combined quasi-likelihood and method of moments approach. Alternatively within the GLM framework, $\boldsymbol{\beta}$ can be estimated by using the Poisson error function with the log link function, and defining the prior weights m_j as

$$m_j = (1 + k\hat{\mu}_j)^{-1}$$

The value of k is obtained iteratively from (3.8). The initial fit is made with unit prior weights $m_j = 1$; see Breslow.[4,5]

3.3 Multiplicative random effects model

Another way of viewing the gamma-Poisson mixture model (3.2) is to write the Poisson parameter as

$$\mu = u\mu_0$$

where μ_0 is an unknown parameter and u is a value of the random effect U taken to have some distribution $H(u)$, which without loss of generality, can be assumed to have mean one. This is the multiplicative random-effects model.[8,35]

If we take the random effect U to have the gamma (α, β) distribution with $\alpha = 1/\beta = k^{-1}$, we obtain the negative binomial distribution as given by (3.4). The mean–variance relationship (3.5) will hold for any mixing distribution H for U that has mean 1 and variance k. Other choices of the distribution of U include the inverse Gaussian[36] as adopted by Dean *et al.* [37] and the log normal.[14]

3.4 Additive random effects model

A random effect U can be introduced additively into a GLM on the same scale as the linear predictor. If for the log link function, the distribution of $\exp(U)$ is taken to be gamma, then it corresponds to the multiplicative random effects model given in the previous section for overdispersed Poisson data; if it has a log normal distribution, then it corresponds to the log normal model considered by Hinde.[14]

More generally, Aitkin[38] considered this approach for an arbitrary GLM. For an unobservable random effect u_j for the jth response on the same scale as the linear predictor, we have that

$$\eta_j = \boldsymbol{\beta}^T \mathbf{x}_j + \sigma u_j$$

where u_j is realization of a random variable U_j distributed $N(0, 1)$ independently of

the jth response Y_j $(j = 1, \ldots, n)$. The (marginal) log likelihood is thus

$$\log L(\boldsymbol{\Psi}) = \sum_{j=1}^{n} \log \int_{-\infty}^{\infty} f(y_j; \boldsymbol{\beta}, \sigma, u)\phi(u)\mathrm{d}u \tag{3.9}$$

where $\phi(u)$ denotes the pdf of a standard normal random variable.

The integral (3.9) does not exist in closed form except for a normally distributed response y_j. Following the development in Anderson and Hinde,[39] Aitkin[38] suggested that it be approximated by Gaussian quadrature, whereby the integral over the normal distribution of U is replaced by a finite sum of g Gaussian quadrature mass-points u_i with masses π_i; the u_i and π_i are given in standard references, for example, Abramowitz and Stegun.[40] The log likelihood so approximated thus has the form for that of a g-component mixture model

$$\sum_{j=1}^{n} \log \sum_{i=1}^{g} \pi_i f(y_j; \beta, \sigma, u_i)$$

where the masses $\pi_1, \ldots, \pi_g$ correspond to the (known) mixing proportions, and the corresponding mass points $u_1, \ldots, u_n$ to the (known) parameter values. The linear predictor for the jth response in the ith component of the mixture is

$$\eta_j = \boldsymbol{\beta}^T \mathbf{x}_j + \sigma u_i$$

Hence in this formulation, u_i becomes an observed covariate with regression coefficient σ.

The influential paper by Heckman and Singer[41] showed substantial changes in parameter estimates with quite small changes in the mixing distribution. As noted by Aitkin,[38] a particular disadvantage of the modeling approach described above is the possible sensitivity of conclusions the choice of a particular distributional form for the random effect U; there is a lack of information in the data about this distribution. A second disadvantage is the need to expand the data vector to length $g \times n$ as g may have to be large for accurate Gaussian quadrature. A third disadvantage is the possible inaccuracy of Gaussian quadrature, where even 20-point integration may not give high accuracy for the logistic/normal model (Crouch and Spiegelman[42]). As a consequence, Aitkin[38] suggested treating the masses $\pi_1, \ldots, \pi_g$ as g unknown mixing proportions and the mass points $u_1, \ldots, u_g$ as g unknown values of a parameter. This g-component mixture model is then fitted using the EM algorithm, as to be described in the next section. The value of g is increased sequentially until the increase in the likelihood is assessed to be nonsignificant. If $\boldsymbol{\beta}$ were known, then this approach would correspond to finding the nonparametric MLE of the distribution of U (the mixing distribution). The advantage of this approach is that it avoids having to specify the mixing distribution.

In this framework since now u_i is also unknown, we can drop the scale parameter σ and define the linear predictor for the jth response in the ith component of the mixture as

$$\eta_{ij} = \boldsymbol{\beta}^T \mathbf{x}_j + u_i$$

Thus u_i acts as an intercept parameter for the ith component. One of the u_i parameters will be aliased with the intercept term β_0; alternatively, the intercept can be removed from the model. Recently, Lee and Nelder[28] proposed hierarchical generalized linear models for which $\boldsymbol{\beta}$ is estimated by consideration of the likelihood formed on the basis of the joint distribution of the observed responses and the unobservable random effects. Their approach thus avoids the integration in (3.9) that is necessary with the use of the marginal likelihood; that is, the likelihood based on just the observed responses.

4 Logistic regression model

We consider now the binomial model for which the density of the jth response Y_j is given by

$$f(y_j;\, \theta_j) = \binom{N_j}{y_j} \theta_j^{y_j} (1-\theta_j)^{N_j - y_j} I_{A_j}(y_j) \tag{4.1}$$

where $A_j = \{0, 1, \ldots, N_j\}$. That is, the response y_j denotes the number of successes in a series of N_j independent Bernoulli trials on which the probability of success on each Bernoulli trial is θ_j.

In the case of logistic regression, θ_j is postulated to depend on the vector x_j of covariates through the logit function

$$\log\{\theta_j/(1-\theta_j)\} = \boldsymbol{\beta}^T \mathbf{x}_j, \quad j = 1, \ldots, n \tag{4.2}$$

or equivalently

$$\theta_j = \exp(\boldsymbol{\beta}^T \mathbf{x}_j)/\{1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_j)\}$$

It is given within the GLM framework by taking the response variable to be y_j/N_j, specifying the error function to be the binomial, and using the canonical logit link.

Logistic regression is a common method for analyzing the effect of a vector of covariates on the number of successes in a series of N_j independent Bernoulli trials. Overdispersion relative to the binomial distribution may occur if the N_j trials in a set are positively correlated, an important covariate is omitted, or $\mathbf{x}_j$ is measured with error. Other link functions include the probit, which gives similar results as the logit, and the complementary log–log function, which is limited to situations where it is appropriate to deal with the probability parameter θ in an asymmetric manner; see McCulloch and Nelder[17] for a comparison of these link functions. A classical approach in the case of no covariates is to use a continuous binomial mixture model

$$\int_0^1 \binom{N_j}{y_j} \theta^{y_j} (1-\theta)^{N_j - y_j} I_{A_j}(y_j)\, \mathrm{d}H(\theta) \tag{4.3}$$

where $H(\theta)$ is taken to be the beta (α, β) distribution, which has density

$$\{u^{\alpha-1}(1-u)^{\beta-1}/B(\alpha,\beta)\} I_{(0,1)}(u)$$

and

$$B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta)$$

This leads to the beta-binomial distribution

$$f(y_j; \alpha, \beta) = \binom{N_j}{y_j} \{B(\alpha + y_j, \beta + N_j - y_j)/B(\alpha, \beta)\} I_{A_j}(y_j) \tag{4.4}$$

see Williams.[43] If we now let $\theta = \alpha/(\alpha + \beta)$, we have that

$$E(Y_j) = N_j\theta$$

and

$$\text{var}(Y_j) = N_j\theta(1 - \theta)\{1 + (N_j - 1)\xi\}$$

where $\xi = (\alpha + \beta + 1)^{-1}$.

A beta-binomial regression model can be defined by postulating parametric forms for θ and ξ. Applications of this type of regression model appear to have been limited mainly to the special cases of one- and two-way ANOVA designs as in Crowder;[44] see Ochi and Prentice[45] and Anderson.[46]

As in the case of the Poisson distribution, a quasi-likelihood approach can be used to deal with overdispersion with the use of the binomial regression model (Williams[25]). With this approach, only the first two moments of the distribution of Y_j have to be specified. One such specification has

$$E(Y_j) = N_j\theta_j \tag{4.5}$$

and

$$\text{var}(Y_j) = N_j\theta_j(1 - \theta_j)\{1 + (N_j - 1)\phi\} \tag{4.6}$$

where

$$\log\{\theta_j/(1 - \theta_j\} = \boldsymbol{\beta}_j^T \mathbf{x}_j, \quad j = 1, \ldots, n$$

As Anderson[46] noted, it is interesting that the assumptions (4.5) and (4.6) for the first two moments of Y_j are satisfied by the beta-binomial distribution if $\alpha = (1 - \phi)\phi^{-1}\theta_j$ and $\beta = (1 - \phi)\phi^{-1}(1 - \theta_j)$.

5 Finite mixtures of GLMs

5.1 Specification of finite mixture model

We have seen in the last section that using an additive random effects model leads to the fitting of a finite mixture of GLMs. Also, if we work with an arbitrary distribution in (3.2) and consider the nonparametric MLE of it, we are led to the fitting of a finite mixture mixture of Poisson regression models. It provides additional motivation to adopt in the first instance a finite mixture of GLMs to handle overdispersion when present with the use of a single GLM. For a mixture of g component distributions of GLM form in proportions $\pi_1, \ldots, \pi_g$, we have that the log density of the response

variable Y is given by

$$f(y; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f_i(y; \theta_i, \phi_i) \tag{5.1}$$

where for a fixed dispersion parameter ϕ_i

$$\log f_i(y; \theta_i, \phi_i) = \phi_i^{-1}\{\theta_i y - b_i(\theta_i)\} + c_i(y; \phi_i) \tag{5.2}$$

for $i = 1, \ldots, g$. For the ith component GLM, we let μ_i be the mean of Y, $g_i(\mu_i)$ the link function, and $\eta_i = g_i(\mu_i) = \boldsymbol{\beta}_i^T \mathbf{x}$ the linear predictor ($i = 1, \ldots, g$).

Typically in practice, the components of the mixture will be from the same GLM, so that the log density for the ith component can be written as

$$\log f(y; \theta_i, \phi_i) = \phi_i^{-1}\{\theta_i y - b(\theta_i)\} + c(y; \phi_i) \tag{5.3}$$

for $i = 1, \ldots, g$.

In some applications, the mixing proportions may be modelled as functions of some vector $\mathbf{x}_m$ of covariates associated with the response. This vector of covariates $\mathbf{x}_m$ may or may not have some elements in common with the vector of covariates $\mathbf{x}$ on which the component means of the mixture depend. A common model for expressing the ith mixing proportion π_i as a function of $\mathbf{x}_m$ is the logistic for which

$$\begin{aligned} \pi_i &= \pi_i(\boldsymbol{\alpha}; \mathbf{x}_m) \\ &= \frac{\exp(\boldsymbol{\alpha}_i^T \mathbf{x}_m)}{1 + \sum_{h=1}^{g-1} \exp(\{\boldsymbol{\alpha}_h^T \mathbf{x}_m)}, \quad i = 1, \ldots, g-1 \end{aligned}$$

where

$$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_{g-1}^T)^T$$

contains the logistic regression coefficients. The first element of $\mathbf{x}_m$ is usually taken to be one, so that the first element of each $\boldsymbol{\alpha}_i$ is an intercept. We let $\boldsymbol{\Psi}$ be the vector of unknown parameters, given by

$$\boldsymbol{\Psi} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$$

where $\boldsymbol{\beta}$ contains the elements of $\boldsymbol{\beta}_1$, ..., $\boldsymbol{\beta}_g$ known a priori to be distinct.

5.2 Maximum likelihood estimation via the EM algorithm

As the mixing proportions are modelled to depend on the covariate vector $\mathbf{x}_m$, which may have some elements in common with those of $\mathbf{x}$, it means that there may be identifiability problems with some of the parameters in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, in particular with the intercept terms of the $\boldsymbol{\alpha}_i$ and the elements of $\boldsymbol{\beta}$; see Wang.[1] The question of identifiability is to be examined more closely later in the specific cases of Poisson and binomial components. We let $y_1, \ldots, y_n$ denote n independent observations of the

response variable with covariates $(\mathbf{x}_{m1}^T, \mathbf{x}_1^T)^T, \ldots, (\mathbf{x}_{mn}^T, \mathbf{x}_n^T)^T$, respectively, and with all prior weights unity. The log likelihood for $\boldsymbol{\Psi}$ that can be formed from these data under the mixture model (5.1) is given by

$$\log L(\boldsymbol{\Psi}) = \sum_{j=1}^{n} \log \sum_{i=1}^{g} \pi_{ij} f_i(y_j;\, \theta_{ij}, \phi_i) \tag{5.4}$$

where θ_{ij} is the canonical parameter for Y_j in its ith component density and

$$\pi_{ij} = \pi_i(\boldsymbol{\alpha};\, \mathbf{x}_{mj}) \tag{5.5}$$

for $(i = 1, \ldots, g; j = 1, \ldots, n)$.

The EM algorithm of Dempster *et al.*[47] can be applied to obtain the MLE of $\boldsymbol{\Psi}$ as in the case of a finite mixture of arbitrary distributions; see also McLachlan and Basford[48] and McLachlan and Krishnan.[49] More precisely, the EM algorithm can be used to find solutions of the likelihood function corresponding to local maxima. In the complete-data framework for the application of the EM algorithm to this problem, each response y_j is viewed as having arisen from one of the g components of the postulated mixture model of GLMs (5.1). Accordingly, for each y_j, the vector $\mathbf{z}_j$ is introduced as missing data, where $z_{ij} = 1$ or zero according as y_j does or does not belong to the ith component of the mixture model $(i = 1, \ldots, g; j = 1, \ldots, n)$. The unobservable indicator vector $\mathbf{z}_j$ is taken to be the realization of a random sample of size one from a multinomial distribution, consisting of a single draw on g categories with probabilities $\pi_{1j}, \ldots, \pi_{gj}$ $(j = 1, \ldots, n)$; $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are independently distributed. The complete-data log likelihood is given by

$$\log L_c(\boldsymbol{\Psi}) = \sum_{i=1}^{g} \sum_{j=1}^{n} z_{ij} \{\log \pi_{ij} + \log f_i(y_j;\, \theta_{ij}, \phi_i)\} \tag{5.6}$$

5.3 E-step

On the $(k+1)$th iteration of the EM algorithm, the E-step is easily affected to give the Q-function

$$Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)}) = \sum_{i=1}^{g} \sum_{j=1}^{n} \tau_{ij}(y_j;\, \boldsymbol{\Psi}^{(k)}) \{\log \pi_{ij} + \log f_i(y_j;\, \theta_{ij}, \phi_i)\} \tag{5.7}$$

where

$$\tau_{ij}(y_j;\, \boldsymbol{\Psi}^{(k)}) = \frac{\pi_{ij}^{(k)} f_i(y_j;\, \theta_{ij}^{(k)}, \phi_i)}{\sum_{h=1}^{g} \pi_{hj}^{(k)} f_h(y_j;\, \theta_{hj}^{(k)}, \phi_h)} \tag{5.8}$$

is the current estimate of the posterior probability that the jth response belongs to the ith component given y_j with covariate vectors $\mathbf{x}_{mj}$ and $\mathbf{x}_j$ $(i = 1, \ldots, g; j = 1, \ldots, n)$.

5.4 M-step

The M-step on the $(k+1)$th iteration involves solving the two system of equations

$$\sum_{i=1}^{g}\sum_{j=1}^{n}\tau_{ij}(y_i;\,\boldsymbol{\Psi}^{(k)})\partial\log\pi_{ij}/\partial\boldsymbol{\alpha}=\mathbf{0} \tag{5.9}$$

and

$$\sum_{i=1}^{g}\sum_{j=1}^{n}\tau_{ij}(y_j;\,\boldsymbol{\Psi}^{(k)})\partial\log f_i(y_j;\,\theta_{ij},\phi_i)/\partial\boldsymbol{\beta}=\mathbf{0} \tag{5.10}$$

assuming that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ have no elements known a priori to be in common. This will often be the case in practice. An application where this is not the case concerns the zero-inflated Poisson regression model of Lambert,[50] which is to be discussed later.

Equation (5.8) can be solved using a standard algorithm for logistic regression to produce the updated estimate $\boldsymbol{\alpha}^{(k+1)}$ for the logistic regression coefficients. For $g=2$, $\boldsymbol{\alpha}^{(k+1)}$ can be computed using the GLIM macro for a binomial error structure with the canonical logit transformation as the link.

In the situation where the mixing proportions $\pi_1,\ldots,\pi_g$ do not depend on any covariates, the updated estimate of π_i is given by

$$\pi_i^{(k+1)}=\sum_{j=1}^{n}\tau_{ij}(y_j;\,\boldsymbol{\Psi}^{(k+1)})/n \tag{5.11}$$

where then

$$\tau_{ij}(y_j;\,\boldsymbol{\Psi}^{(k)})=\frac{\pi_i^{(k)}f_i(y_j;\,\theta_{ij}^{(k)},\phi_i)}{\sum_{h=1}^{g}\pi_h^{(k)}f_h(y_j;\,\theta_{hj}^{(k)},\phi_h)} \tag{5.12}$$

Concerning the computation of $\boldsymbol{\beta}^{(k+1)}$, it follows from earlier work on the ML fitting of a single GLM that (5.10) can be written as

$$\sum_{i=1}^{g}\sum_{j=1}^{n}\tau_{ij}(y_j;\,\boldsymbol{\Psi}^{(k)})w(\mu_{ij})(y_j-\mu_{ij})\eta_i'(\mu_{ij})\{\partial\eta_i(\mu_{ij})/\partial\boldsymbol{\beta}\}=\mathbf{0} \tag{5.13}$$

where, for the ith component, μ_{ij} is the mean of Y_j. If the $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_g$ have no elements in common a priori, then

$$\begin{aligned}\partial\eta_i(\mu_{ij})/\partial\boldsymbol{\beta}_h&=\mathbf{x}_j,\quad \text{if } h{=}i\\&=0,\quad \text{otherwise}\end{aligned}$$

In this case, (5.13) reduces to solving

$$\sum_{j=1}^{n}\tau_{ij}(y_j;\,\boldsymbol{\Psi}^{(k)})w(\mu_{ij})(y_j-\mu_{ij})\eta_i'(\mu_{ij})\mathbf{x}_j=\mathbf{0} \tag{5.14}$$

separately for each $\boldsymbol{\beta}_i$ to produce $\boldsymbol{\beta}_i^{(k+1)}$ $(i=1,\ldots,g)$.

On contrasting (5.4) with (2.3), it can be seen that it has the same form as for a single GLM fitted to the responses $y_1, \ldots, y_n$ with prior weights $m_1 = \tau_{i1}(y_1; \boldsymbol{\Psi}^{(k)}), \ldots, m_n = \tau_{in}(y_n; \boldsymbol{\Psi}^{(k)})$ and fixed dispersion parameter ϕ_i.

In the general case where $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_g$ may have some elements in common, we can still solve (5.3) using the iteratively reweighted least-squares approach for a single GLM. The double summation over i and j in (5.11) can be handled by expanding the response vector to have dimension $g \times n$ by replicating each original observation $(y_j, \mathbf{x}_{mj}^T, \mathbf{x}_j^T)^T$ g times, with prior weights $\tau_{1j}(y_j; \boldsymbol{\Psi}^{(k)}), \ldots, \tau_{gj}(y_j; \boldsymbol{\Psi}^{(k)})$, fixed dispersion parameters $\phi_1, \ldots, \phi_g$, and linear predictors $\mathbf{x}_j^T\boldsymbol{\beta}_1, \ldots, \mathbf{x}_j^T\boldsymbol{\beta}_g$.

Although there are more efficient methods of solving (5.13), this approach has the advantage that it is easily done in a GLM fitting program, such as GLIM or GENSTAT. Dietz[51] has provided a GLIM-macro for the computation of $\boldsymbol{\beta}$. Previously, Hinde[14] provided the GLIM code for a Poisson model and the modifications needed for the binomial model; see also Anderson and Hinde,[39] Anderson[46] and Aitkin.[38] Wang *et al.*[2] have available FORTRAN codes for algorithms that fit finite mixtures of Poisson regression models.

The response for each subject has been taken to be unvariate in the above. The results generalize in a straightforward manner to the case of multivariate responses $\mathbf{Y}_j = (Y_{1j}, \ldots, Y_{pj})^T$, if it is assumed that $Y_{1j}, \ldots, Y_{pj}$ are independently distributed when conditioned on their component membership of the mixture model; see Wedel and DeSarbo[52] for the details. The case of component multivariate GLMs where $Y_1, \ldots, Y_p$ are not necessarily independent has been considered by Dietz.[51] The reader is referred to McLachlan and Krishnan[49] for a detailed discussion of the computation of standard errors of MLEs obtained via the EM algorithm and of the selection of suitable starting values for this algorithm. As the likelihood function tends to have multiple local maxima for mixture models, the choice of starting values for the EM algorithm is an important consideration with its use.

5.5 Multicycle ECM algorithm

We have seen above in the computation of the updated estimate of

$$\boldsymbol{\Psi} = (\boldsymbol{\Psi}_1^T, \boldsymbol{\Psi}_2^T)^T$$

where $\boldsymbol{\Psi}_1 = \boldsymbol{\alpha}$ and $\boldsymbol{\Psi}_2 = \boldsymbol{\beta}$ that $\boldsymbol{\alpha}^{(k+1)}$ and $\boldsymbol{\beta}^{(k+1)}$ are computed independently of each other on the M-step of the EM algorithm. Therefore, the latter is the same as the expectation-conditional maximization (ECM) algorithm with two CM-steps, where on the first CM-step, $\boldsymbol{\Psi}^{(k+1)}$ is calculated with $\boldsymbol{\Psi}_2$ fixed at $\boldsymbol{\Psi}_2^{(k)}$, and where on the second CM-step, $\boldsymbol{\Psi}_2^{(k+1)}$ is calculated with $\boldsymbol{\Psi}_1$ fixed at $\boldsymbol{\Psi}_1^{(k+1)}$. In order to improve convergence, a multicycle version of the ECM algorithm can be used, where an E-step is performed after the computation of $\boldsymbol{\alpha}^{(k+1)}$ and before the computation of $\boldsymbol{\beta}^{(k+1)}$; see Meng and Rubin[53] and McLachlan and Krishnan[49] for further details of the ECM algorithm. The multicycle E-step is effected here by updating $\boldsymbol{\alpha}^{(k)}$ with $\boldsymbol{\alpha}^{(k+1)}$ in $\boldsymbol{\Psi}^{(k)}$ in the right-hand side of the expression (5.8) for $\tau_{ij}(y_j; \boldsymbol{\Psi}^{(k)})$.

5.6 Choice of the number of components

Up to now, we have considered the fitting of a finite mixture of GLMs for a given value of the number of components g in the mixture model. Typically, in practice

where the mixture model is being used to handle overdispersion, the value of g has to be inferred from the data. A guide to the final choice of g can be obtained from monitoring the increase in the log likelihood as g is increased from a single component. Unfortunately, it is difficult to carry out formal tests at any stage of this sequential process for the need of an additional component, since, as is well known, regularity conditions fail to hold for the likelihood ratio statistic λ to have its usual asymptotic null distribution. There is the resampling approach of McLachlan,[54] which was used by Schlattmann and Böhning[55] to decide on g in their application of Poisson mixtures to disease mapping. Also, Pauler *et al.*[56] used this method to decide on the number of Poisson components in the finite mixture modelling of anticipatory saccade counts from schizophrenic patients and controls. In the context of the fitting of mixtures of Poisson regression components to overdispersed count data, Wang *et al.*[2] have reported encouraging results for the selection of g based on Akaike's[57] information criterion (AIC) and the Bayesian information criterion (BIC) of Schwarz.[58] With these criteria, the choice of g is that value of g that maximizes $2\log L(\hat{\boldsymbol{\Psi}}) - ad$, where d denotes the number of parameters in the model, and $a = 2$, and $\log n$, for AIC and BIC, respectively. For mixture problems in general, AIC often leads to too many components being fitted. Concerning the significance of the covariates in the mixture of GLMs, Wang *et al.*[2] considered the deletion of covariates from the model only after the choice of g had been essentially finalized.

6 Finite mixture of Poisson regression models

6.1 Mean and variance

We consider now the finite mixture model (5.1) of arbitrary component GLMs to the case where the component GLMs are Poisson regression models with means specified by a log linear model. That is

$$f_i(y_j;\, \theta_{ij}) = \{e^{-\mu_{ij}} \mu_{ij}^{y_j}/y_j!\} I_A(y_j)$$

where

$$\log \mu_{ij} = \boldsymbol{\beta}_i^T \mathbf{x}_j, \quad i = 1, \ldots, g$$

For this g-component mixture of Poisson regression models, the mean and variance of Y_j is equal to

$$E(Y_j) = \sum_{i=1}^{g} \pi_i \mu_{ij}$$

and

$$\begin{aligned} \text{var}(Y_j) &= E\{\text{var}(Y_j \mid \mathbf{Z}_j\} + \text{var}\{E(Y_j \mid \mathbf{Z}_j\} \\ &= E(Y_j) + v_{ij} \end{aligned}$$

respectively, where

$$v_{ij} = \sum_{i=1}^{g} \pi_i \mu_{ij}^2 - \left(\sum_{i=1}^{g} \pi_i \mu_{ij} \right)^2$$

Here

$$\mu_{ij} = \exp(\beta_i^T \mathbf{x}_j)$$

is the mean of the jth response conditional on its membership of the ith component of the mixture, and $\mathbf{Z}_j$ is the component-indicator vector of zeros and ones, where $Z_{ij} = (\mathbf{Z}_j)_i$ is one or zero, according as y_j is viewed as having come from the ith component or not ($i = 1, \ldots, g; j = 1, \ldots, n$). Obviously, $v_{ij} = 0$ if and only if

$$\mu_{1j} = \mu_{2j} = \cdots = \mu_{gj}$$

Hence the mixture model is able to cope better than the one-component (homogeneous) model with excess variation among $Y_1, \ldots, Y_n$.

6.2 Identifiability

In order to be able to estimate Ψ, we require the mixture to be identifiable; that is, two sets of parameters which do not agree after permutation cannot yield the same mixture distribution. Without covariates, Teicher[59] proved that the class of finite mixtures of Poisson distributions is identifiable. As noted by Wang *et al.*,[2] a sufficient condition for the class of Poisson regression mixtures to be identifiable is that the matrices $(\mathbf{x}_{m1}, \ldots, \mathbf{x}_{mn})$ and $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ each be of full rank. Wang *et al.*[2] also considered the residual analysis and goodness-of-fit statistics for this class of mixture regression models. For examples of applications of finite mixtures of Poisson regression models to biological data sets, the reader is referred to Wang *et al.*[2] who used this methodology to analyse epileptic seizure frequency and Ames salmonella assay data.

6.3 Count data with excess zeros

Two-component mixture models are frequently used to model data that appear to have an excess of zeros. In a medical context, a possible explanation for the excess of zeros might be due to the fact that the patient is cured after the treatment and so no realization of the symptom being monitored will occur. This phenomenon can be handled by a two-component mixture where one of the components is taken to be a degenerate distribution, having mass 1 at $y = 0$. The other component is a Poisson (or binomial) regression model, depending on the situation. This model formed the basis of the zero-inflated Poisson (ZIP) regression technique proposed by Lambert[50] for the handling of zero-inflated count data with covariates; see also Yip[60,61] and Fong and Yip.[62,63] The fascinating history of this model (in the absence of covariates) has been given elsewhere in this issue by Meng.[64]

6.4 Components and mixing proportions without covariates

Böhning *et al.*[65] have provided an excellent account of the use of Poisson mixture models where the mixing proportions and the components do not depend on any covariates. They also gave several examples of applications of the Poisson mixture

model in medical problems; see also Lindsay.[66] Albert,[67] Leroux and Puterman[68] and Lee *et al*.[69] have presented examples in the context of modelling epileptic seizure counts and fetal movements by Poisson mixtures where the assumption of independence of the data has been relaxed. In the absence of covariates, we can write the Poisson mixture model as

$$f(y; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f(y; \mu_i) = \int f(y; \theta)\, \mathrm{d}H(\theta)$$

where

$$f(y;\theta) = \{e^{-\theta}\theta^y/y!\}I_A(y)$$

and $H(\theta)$ is the measure that puts mass π_i at the point $\theta = \mu_i\,(i = 1, \ldots, g)$.

In the work up to now, we have effectively considered the estimation of the mixing distribution H in the fixed support case. However, we can treat g itself as unknown, which is the flexible support size case. In this latter case, we can approach the problem by considering the so-called nonparametric maximum likelihood estimator (NPMLE) of H, $\hat{H}$, which is the probability measure that maximizes

$$l(H) = \sum_{j=1}^{n} \log \int f(y_j; \theta)\, \mathrm{d}H(\theta)$$

where $H(\theta)$ is now allowed to be any mixing distribution. Lindsay[70] showed that $\hat{H}$ is a discrete measure with at most a finite number of support points; see also Lindsay[66] and Lindsay and Roeder[71] who have considered residual diagnostics for mixture models.

6.5 Algorithms for NPMLE of a mixing distribution

Böhning *et al*.[65] have provided the computer package C.A.MAN (computer-assisted mixture analysis) for computing $\hat{H}$. It includes an algorithmic menu with choices of the EM algorithm, the vertex exchange algorithm, a combination of both, as well as the vertex direction method. The package C.A.MAN has the option to work with fixed support size; that is, when the number of components is known a priori. In the latter case, the EM algorithm is used. More recently, Böhning[72] has reviewed reliable algorithms for the ML fitting of mixture models.

6.6 Disease mapping

Poisson mixtures have played a very useful role in disease mapping. The analysis of the geographic variation of disease and its representation on a map is an important topic in epidemiological research. Identification of high-risk groups provides valuable hints for possible experience and targets for subsequent analytical studies; see Schlattmann and Böhning[55] and Schlattmann *et al*.[73] A measure often used is the standardized mortality rate (SMR). For a given area, SMR_j is defined as

$$\begin{aligned} \text{SMR}_j &= y_j/e_j \\ &= y_j/\sum_h A_{hj}\omega_h \end{aligned} \tag{6.1}$$

where for the jth regional area, y_j is the number of observed cases, e_j is the expected number based on an external reference, A_{hj} is the person years in the hth age stratum, and w_h is the age-specific mortality rate, which is assumed to be known. A common approach to map construction in the literature is based on the assumption that y_j is the realization of the random variable Y_j which has a Poisson distribution with parameter $\mu_j = \lambda e_j$. Here λ denotes the relative risk of disease due to living within the study area.

The assumption (6.1) implies that all geographical area have the same relative risk λ. This homogeneous model of a single Poisson distribution is often too simple with overdispersion frequently occurring. One approach for more flexibility has been to be adopt a random effects model, where λ is gamma or log normal; see Clayton and Kaldor,[74] Mollie and Richardson[75] and Breslow and Clayton.[7]

Schlattman and Böhning[55] modelled the distribution of Y_j by the Poisson mixture distribution

$$f(y_j) = \sum_{i=1}^{g} \pi_i f(y_j;\ \lambda_i e_j)$$

where the relative risk λ_i is specific to the ith component of the mixture $(i = 1, \ldots, g)$. More recently, Schlattmann *et al.*[73] proposed the Poisson mixture regression model

$$f(y_j) = \sum_{i=1}^{g} \pi_i f(y_j;\ \mu_{ij})$$

for the distribution of Y_j, where

$$\mu_{ij} = \lambda_i e_j \exp(\boldsymbol{\beta}_i^T \mathbf{x}_j)$$

and $\mathbf{x}_j$ is a vector of covariates associated with the jth region, and the parameters λ_i and $\boldsymbol{\beta}_i$ are specific to the ith component of the mixture $(i = 1, \ldots, g)$.

7 Finite mixtures of logistic regressions

7.1 Mean and variance

We consider now the mixture of GLMs model (5.1) in the case where the component GLMs belong to the binomial family. That is, the density $f_i(y_j;\ \theta_{ij})$ for the jth response Y_j is given by

$$f_i(y_j;\ \theta_{ij}) = \binom{N_j}{y_j} \theta_{ij}^{y_j} (1 - \theta_{ij})^{N_j - y_j} I_{A_j}(y_j) \tag{7.1}$$

where $A_j = \{0, 1, \ldots, N_j\}$. Under the logistic regression model, θ_{ij} is postulated to depend on the covariates so that

$$\log\{\theta_{ij}/(1-\theta_{ij})\} = \boldsymbol{\beta}_i^T \mathbf{x}_j, \quad i = 1, \ldots, g; j = 1, \ldots, n \tag{7.2}$$

We consider now the mean and variance of the logistic regression mixture model

$$\sum_{i=1}^{g} \pi_{ij} f_i(y_j;\, \theta_{ij}) \tag{7.3}$$

The mean of (7.3) is

$$\sum_{i=1}^{g} \pi_{ij}\theta_{ij}$$

and its variance is

$$\begin{aligned}\operatorname{var}(Y_j) &= E\{\operatorname{var}(Y_j \mid \mathbf{Z}_j\} + \operatorname{var}\{E(Y_j \mid \mathbf{Z}_j)\} \\ &= N_j(\sum_{i=1}^{g} \pi_{ij}\theta_{ij})(1 - \sum_{i=1}^{g} \pi_{ij}\theta_{ij}) \\ &\quad + \{(N_j - 1)/N_j\}\operatorname{var}\{E(Y_j \mid \mathbf{Z}_j)\}\end{aligned}$$

where

$$\operatorname{var}\{E(Y_j \mid \mathbf{Z}_j)\} = N_j^2\{\sum_{i=1}^{g} \pi_{ij}\theta_{ij}^2 - (\sum_{i=1}^{g} \pi_{ij}\theta_{ij})^2\}.$$

For $N_j > 1$, $\operatorname{var}\{E(Y_j \mid \mathbf{Z}_j)\} = 0$ holds if and only if $E(Y_j \mid \mathbf{Z}_j)$ is constant. Hence for each j $(j = 1, \ldots, n)$,

$$\operatorname{var}(Y_j) = N_j\left(\sum_{i=1}^{g} \theta_{ij}\right)\left(1 - \sum_{i=1}^{g} \theta_{ij}\right)$$

if and only if

$$\theta_{1j} = \theta_{2j} = \cdots = \theta_{gj}$$

for $1 \le j \le n$. This implies the proposed mixture model is able to cope better than the one-component model with extra-binomial variation among $Y_1, \ldots, Y_n$ due to heterogeneity in the population.

7.2 Mixing at the binary level

For binary data ($N_j = 1$), we can rewrite (7.1) as

$$\left(\sum_{i=1}^{g} \pi_{ij}\theta_{ij}\right)^{y_j}\left(1 - \sum_{i=1}^{g} \pi_{ij}\theta_{ij}\right)^{1-y_j}$$

so that the g-component mixture of Bernoulli distributions is itself a Bernoulli distribution with probability parameter

$$\sum_{i=1}^{g} \pi_{ij}\theta_{ij} \tag{7.4}$$

This model is given within the GLM framework by still specifying the binomial as the error function, but now specifying the link function according to (7.2) and (7.4). In the parlance of generalized linear models, mixing at the binary level changes the link but not the frequency function or dispersion, whereas mixing at the binomial level ($N_j > 1$) changes both the link and the frequency function and introduces overdispersion. Caution has to be exercised in using (7.3) with $N_j = 1$, as the model may not be identifiable without imposing some unrealistic restrictions on the covariates (Follman and Lambert[76,77] and Wang[1]).

7.3 Identifiability

Teicher,[59,78] Blischke[79] and Margolin *et al.*[80] have given necessary and sufficient conditions for the identifiability of the finite binomial mixture

$$f(y; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i \binom{N}{y} \theta_i^y (1-\theta_i)^{N-y} I_{A_N}(y) \tag{7.5}$$

where $A_N = \{0, 1, \ldots, N\}$. Their results may be summarized as follows. The g-component binomial mixture model (7.5) with $0 < \theta_i < 1$ $(i = 1, \ldots, g)$ is identifiable if and only if

$$g \leq \frac{1}{2}(N+1)$$

Wang[1] has considered the identifiability of the collection of logistic regression models

$$f(y_j; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_{ij} \binom{N_j}{y_j} \theta_{ij}^{y_j} (1-\theta_{ij})^{N_j - y_j} I_{A_j}(y_j), \quad j = 1, \ldots, n \tag{7.6}$$

where the π_{ij} and the θ_{ij} are specified as functions of the covariates by (5.5) and (7.2), respectively. In the case where the number of Bernoulli trials N_j are all equal ($N_j = N$ for all j), sufficient conditions for the identifiability of (7.6) are that $g \leq \frac{1}{2}(N+1)$ and that the matrices $(\mathbf{x}_{m1}, \ldots, \mathbf{x}_{mn})$ and $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ are each of full rank. In the case of unequal N_j, the restriction on g in these sufficient conditions is specified in terms of the minimum number of trials for all proper subsets of the observations. Previously, Follman and Lambert[76,77] had considered sufficient conditions for the identifiability of (7.6) in the special case where the mixing proportions are not functions of any covariates and $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_g$ have common elements apart from the the first; that is, the logistic components differ only in their intercepts. They showed that for a binary response the number of components g in the mixture must be bounded by a function of the number of covariate vectors that agree except for one element; and for a binomial response, g must satisfy the same bound or be bounded by a function of the largest

number of trials per response. Examples on the fitting of mixtures of logistic regressions to biological data may be found in Follman and Lambert[77] and Wang,[1] while Farewell and Sprott[81] gave an example on the fitting of a mixture of binomial distributions. Overdispersion in the case of the multinomial distribution has been considered by Mosiman,[82] Paul *et al*.,[83] Kim and Margolin[84] and Morel and Nagaraj,[85] among others.

References

1 Wang P. Mixed regression models for discrete data. PhD dissertation, University of British Columbia, Vancouver, 1996.
2 Wang P, Puterman ML, Cockburn I, Le N. Mixed Poisson regression models with covariate dependent rates. *Biometrics* 1996; **52**: 381–400.
3 Aitkin M, Anderson D, Francis BJ, Hinde J. *Statistical modelling in GLIM*. Oxford: Oxford University Press, 1989.
4 Breslow NE. Extra-Poisson variation in log-linear models. *Applied Statistics* 1984; **33**: 38–44.
5 Breslow NE. Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *Journal of the American Statistical Association* 1987; **85**: 565–71.
6 Breslow NE. Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *Journal of the American Statistical Association* 1990; **85**: 565–71.
7 Breslow NE, Clayton, DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 1993; **88**: 9–25.
8 Brillinger DR. The natural variability of vital rates and associated statistics (with discussion). *Biometrics* 1986; **42**: 693–734.
9 Clayton DG. Some approaches to the analysis of recurrent event data. *Statistical Methods in Medical Research* 1994; **3**: 244–62.
10 Cox DR. Some remarks on overdispersion. *Biometrika* 1983; **70**: 269–74.
11 Efron B. Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association* 1986; **81**: 709–21.
12 Efron B. Poisson overdispersion estimates based on the method of asymmetric maximum likelihood. *Journal of the American Statistical Association* 1992; **87**: 98–107.
13 Gelfand AE, Dalal SR. A note on overdispersed exponential families. *Biometrika* 1990; **77**: 55–64.
14 Hinde JP. Compound Poisson regression models. In: Gilchrist R ed. *GLIM 82*. New York: Springer, 1982: 109–21.
15 Lawless, JF. Regression methods for Poisson process data. *Journal of the American Statistical Association* 1987; **82**: 808–15.
16 Lawless, JF. Negative binomial and mixed Poisson regression *Canadian Journal of Statistics* 1987; **15**: 209–25.
17 McCullagh PA, Nelder J. *Generalised linear models*, 2nd edition. London: Chapman & Hall, 1989.
18 Fisher RA. The significance of deviation from the expectation in a Poisson series. *Biometrics* 1950; **6**: 17–24.
19 Collings BJ, Margolin BH. Testing of goodness-of-fit for the Poisson assumption when observations are not indentically distributed. *Journal of the American Statistical Association* 1985; **80**: 411–18.
20 Cameron AC, Trivedi PK. Econometric models based on count data: comparisons and applications of some estimators and tests. *Journal of Applied Economics* 1986; **1**: 29–53.
21 Cameron AC, Trivedi PK. Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics* 1990; **46**: 347–64.
22 Dean C, Lawless JF. Tests for detecting overdispersion in Poisson regression model. *Journal of the American Statistical Association* 1989; **84**: 467–72.
23 Dean C. Testing for overdispersion in Poisson and binomial regression models. *Journal of the American Statistical Association* 1992; **87**: 451–57.
24 Tarone RE. Testing the goodness of fit of the binomial distribution. *Biometrika* 1979; **66**: 585–90.
25 Williams DA. Extra-binomial variation in logistic linear models. *Applied Statistics* 1982; **31**: 144–48.
26 Prentice RL. Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association* 1986; **81**: 321–27.
27 Lambert D, Roeder K. Overdispersion diagnostics for generalized linear models. *Journal of the American Statistical Association* 1995; **90**: 1225–36.

28 Lee Y, Nelder JA. Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society Series B* 1996; **58**: 619–78.

29 Francis B, Green M , Payne C. *The GLIM system. Release 4 manual.* Oxford: Oxford University Press, 1993.

30 Nelder JA, Wedderburn, RWM. Generalized linear models. *Journal of the Royal Statistical Society Series A* 1972; **135**: 370–84.

31 Wedderburn RWM. Quasilikelihood functions, generalized linear models and the Gauss–Newton method. *Biometrika* 1974; **61**: 439–517.

32 Stirling WD. Iteratively reweighted least squares for models with a linear part. *Applied Statistics* 1984; **33**: 7–17.

33 Kim C. Dispersion statistics in overdispersed mixture models. *Communications in Statistics – Theory and Methods*1994; **23**: 27–45.

34 Nelder JA, Lee Y. Likelihood, quasi-likelihood a pseudolikelihood: some comparisons. *Journal of the Royal Statistical Society Series B* 1992; **54**: 273–84.

35 Manton KG, Woodbury MA, Stallard E. A variance components approach to categorical data models with heterogeneous cell populations: analysis of spatial gradients in lung cancer mortality rates in North Carolina counties. *Biometrics* 1981; **37**: 259–69.

36 Folks JL, Chhikara RS. The inverse Gaussian distribution and its statistical application – a review. *Journal of the Royal Statistical Society Series B* 1978; **40**: 263–89.

37 Dean C, Lawless JF, Willmot GE. A mixed Poisson-inverse-Gaussian regression model. *Canadian Journal of Statistics* 1989; **17**, 171–81.

38 Aitkin M. A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing* 1996; **6**: 251–62.

39 Anderson DA, Hinde JP. Random effects in generalized linear models and the EM algorithm. *Communications in Statistics – Theory and Methods* 1988; **17**: 3847–56.

40 Abramowitz M, Stegun IA eds. *Handbook of mathematical functions*. Washington, DC: National Bureau of Standards, 1984.

41 Heckman JJ, Singer B. A method for minimizing the impact of distributional assumptions in econometric models of duration. *Econometrika* 1984; **52**: 271–320.

42 Crouch EAC, Spiegelman D. The evaluation of integrals of the form $\int_{-\infty}^{\infty} f(t) \exp(-t^2)\, dt$: an application to logistic-normal models. *Journal of the American Statistical Association* 1990; **85**: 464–69.

43 Williams DA. The analysis of binary responses from toxological experiments involving reproduction and teratogenicity. *Biometrika* 1975; **61**: 439–47.

44 Crowder MJ. Beta-binomial ANOVA for proportions. *Applied Statistics* 1978; **27**: 34–47.

45 Ochi Y, Prentice RL. Likelihood inference in a correlated probit regression model. *Biometrika* 1984; **71**: 531–54.

46 Anderson DA. Some models for overdispersed binomial data. *Australian Journal of Statistics* 1988; **30**: 125–48.

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

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

49 McLachlan GJ, Krishnan T. *The EM algorithm and extensions*. New York: John Wiley, 1997.

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

51 Dietz E. Estimation of heterogeneity – a GLM-approach. In: Fahrmeir F, Francis F, Gilchrist R, Tutz G eds. *Advances in GLIM and statistical modelling*. Berlin: Springer, 1992: 66–72.

52 Wedel M, DeSarbo WS. A mixture likelihood approach for generalized linear models. *Journal of Classification* 1996; **12**: 21–55.

53 Meng XL, Rubin D. Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* 1993; **80**: 267–78.

54 McLachlan GJ. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics* 1987; **36**: 318–24

55 Schlattmann P, Böhning D. Mixture models and disease mapping. *Statisitics in Medicine* 1993; **12**: 1943–50.

56 Pauler DK, Escobar MD, Sweeney JA, Greenhouse J. Mixture models for eye-tracking data: a case study. *Statistics in Medicine* 1996; **15**: 1365–76.

57 Akaike H. Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F eds. *Second international symposium on information theory*. Budapest: Akademiai Kiado, 1973: 267–81.

58 Schwarz G. Estimating the dimension of a model. *Applied Statistics* 1978; **6**: 461–64.

59 Teicher H. Identifiability of mixtures. *Annals of Mathematical Statistics* 1961; **32**; 244–48.

60 Yip P. Inference about the mean of a Poisson

distribution in the presence of a nuisance parameter. *Australian Journal of Statistics* 1988; **30**: 299–306.
61 Yip P. Conditional inference for a mixture model for the analysis of count data. *Communications in Statistics – Theory and Methods* 1991; **20**: 2045–57.
62 Fong DYT, Yip P. An EM algorithm for a mixture model of count data. *Statistics & Probability Letters* 1993; **17**: 53–60.
63 Fong DYT, Yip PSF. A note on information loss in analyzing a mixture model for count data. *Communications in Statistics – Theory and Methods* 1995; **24**: 3197–209.
64 Meng XL. The EM algorithm and medical studies: a historical link. *Statistical Methods in Medical Research* 1997; **6**: 3–23.
65 Böhning D, Schlattmann P, Lindsay B. Computer-assisted analysis of mixtures (C.A.MAN): statistical algorithms. *Biometrics* 1992; **48**: 283–303.
66 Lindsay BG. *Mixture models: theory, geometry and applications, NSF-CBMS regional conference series in probability and statistics*, Vol. 5. Alexandria, VA: Institute of Mathematical Statistics and the American Statistical Association, 1995.
67 Albert PS. A two-state Markov mixture model for a time series of epileptic seizure counts. *Biometrics* 1991; **47**: 1371–81.
68 Leroux BG, Puterman ML. Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics* 1992; **48**: 545–58.
69 Lee, ND, Leroux, BG, Puterman ML. Exact likelihood evaluation in a Markov mixture model for time series of seizure counts. *Biometrics* 1992; **48**: 317–23.
70 Lindsay BG. The geometry of likelihoods, Part I: a general theory. *Applied Statistics* 1983; **11**: 86–94.
71 Lindsay BG, Roeder K. Residual diagnostics for mixture models. *Journal of the American Statistical Association* 1992; **87**: 785–94.
72 Böhning D. A review of reliable maximum likelihood algorithms for semiparametric mixture models. *Journal of Statistical Planning and Inference* 1995; **47**: 5–28.
73 Schlattmann P, Dietz E, Böhning D. Covariate adjusted mixture models and disease mapping with the program DismapWin. *Statistics in Medicine* 1996; **15**: 919–29.
74 Clayton DG, Kaldor J. Empirical Bayes estimates for age-standardized relative risks. *Biometrics* 1987; **43**: 671–81.
75 Mollie A, Richardson L. Empirical Bayes estimates of cancer mortality rates using spatial models. *Statistics in Medicine* 1991; **10**; 95–112.
76 Follmann DA, Lambert D. Identifiability of finite mixture of mixture of logistic regression models. *Journal of Statistical Planning and Inference* 1991; **27**: 375–81.
77 Follmann, DA, Lambert D. Generalizing logistic regression by nonparametric mixing. *Journal of the American Statistical Association* 1989; **84**: 295–300.
78 Teicher H. Identifiability of finite mixtures. *Annals of Mathematical Statistics* 1963; **34**: 1265–69.
79 Blischke R. Mixtures of distributions. In: Kruskal WH, Tanur JM eds. *International encyclopedia of statistics*, Vol. 1. New York: Free Press, 1978: 174–80.
80 Margolin BH, Kim BS, Risko KJ. The Ames salmonella/microsome mutagenicity assay: issues of inference and validation. *Journal of the American Statistical Association* 1989; **84**: 651–61.
81 Farewell VT, Sprott D. The use of a mixture model in the analysis of count data. *Biometrics* 1988; **44**: 1191–94.
82 Mosimann JE. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika* 1962; **49**: 65–82.
83 Paul SR, Liang KY, Self S. On testing departure from the binomial and multinomial assumptions. *Biometrics* 1989; **45**: 231–36.
84 Kim BS, Margolin BH. Testing goodness of fit of a multinomial against overdispersed alternatives. *Biometrics* 1992; **48**; 711–719.
85 Morel JG, Nagaraj NK. A finite mixture distribution for modelling multinomial extra variation. *Biometrika* 1993; **80**: 363–371.