Stat Comput (2012) 22:79–92
DOI 10.1007/s11222-010-9207-y

Multivariate generalized linear mixed models with semi-nonparametric and smooth nonparametric random effects densities

Georgios Papageorgiou · John Hinde

Received: 8 March 2010 / Accepted: 7 October 2010 / Published online: 3 November 2010

Abstract We extend the family of multivariate generalized linear mixed models to include random effects that are generated by smooth densities. We consider two such families of densities, the so-called semi-nonparametric (SNP) and smooth nonparametric (SMNP) densities. Maximum likelihood estimation, under either the SNP or the SMNP densities, is carried out using a Monte Carlo EM algorithm. This algorithm uses rejection sampling and automatically increases the MC sample size as it approaches convergence. In a simulation study we investigate the performance of these two densities in capturing the true underlying shape of the random effects distribution. We also examine the implications of misspecification of the random effects distribution on the estimation of the fixed effects and their standard errors. The impact of the assumed random effects density on the estimation of the random effects themselves is investigated in a simulation study and also in an application to a real data set.

Keywords Longitudinal data · Mixed models · Multinomial responses · Random effects · Semi-nonparametric densities · Smooth nonparametric densities

1 Introduction

Generalized linear mixed models (GLMMs) provide a very general framework for the analysis of clustered data (Breslow and Clayton 1993). This kind of data occur in many areas of research including clinical trials, where subjects are measured over time, and the educational research area, where measurements are obtained on children of the same classroom. In each case, observations on the same cluster tend to be correlated and the GLMMs take this correlation into account by including random effects in the linear predictor of the model. Usually, the random effects are assumed to follow a specific parametric family of distributions, such as the multivariate normal, although, unlike in linear mixed models, this assumption does not provide much of a mathematical convenience in the class of GLMMs as the random effects enter the model in a nonlinear way.

Our focus here is on regression models for multinomial responses. Multinomial random effects models have previously been treated as special cases of multivariate generalized linear mixed models (MGLMMs) (Tutz and Hennevogl 1996; Hartzel et al. 2001). We also adopt this general approach as it provides unified fitting and inferential procedures for a broad class of models. Model specification and model fitting procedures, described in Sects. 2 and 3, are for a general MGLMM and we only make specific assumptions about the model form for simulation and application, described in Sects. 4 and 5 respectively.

In mixed effects models the main interest is inference about the vector of fixed effects. Variance components are also of interest as they provide information about the intraclass correlation and, also, the degree of heterogeneity of a population. The overly restrictive normality assumption, however, does not allow the capturing of features such as multimodality and skewness which, if present, may provide additional information about the form of heterogeneity of a population and suggest failure to measure important explanatory factors.

Furthermore, misspecification of the random effects distribution can have harmful effects on the estimation of

G. Papageorgiou (✉) · J. Hinde
School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway, Ireland
e-mail: georgios.papageorgiou@nuigalway.ie

J. Hinde
e-mail: john.hinde@nuigalway.ie

the model parameters. Indeed, Neuhaus et al. (1992) in a logistic-normal regression analysis, remarked that under misspecification of the random effects distribution, the estimates of the model parameters, including the ones in the mean structure, typically are asymptotically biased. They also pointed out, however, that the magnitude of this bias is usually small. In a more extensive study, also on mixed effects logistic regression models, Heagerty and Kurland (2001) considered several misspecification scenarios and found high bias in the fixed effects estimates when fitting simple random intercept models when, in reality, the distribution of the random effects depends on explanatory variables and when there are autoregressive random effects. These authors also reported small relative bias when fitting Gaussian random intercept models when, in reality, the random intercepts have a skewed distribution. Under this scenario, high relative bias was observed only when the true distribution of the random effects was highly skewed with large between cluster heterogeneity. Related is the work of Chen et al. (2002), who reported low relative bias in the fixed effects estimates, under misspecification of the random effects distribution. Moreover, Agresti et al. (2004) reported a serious drop in efficiency in the estimation of the fixed parameters from assuming a normal random intercepts model when the true distribution was a two-point mixture with large variance. This situation can arise in practice when subjects vary considerably according to an unmeasured binary factor. More recently, Litière et al. (2008) also showed that in the presence of misspecification, the mean parameters could be subject to considerable bias, especially when there is large between cluster heterogeneity.

In addition, the assumption about the distribution of the random effects is a very important factor in the estimation of the random effects themselves. As was shown by Verbeke and Lesaffre (1996), under a normality assumption, the resulting empirical Bayes estimates of the random effects can have a symmetric, unimodal distribution even if they arise from a bimodal population. Thus, there has been considerable interest in methods that avoid the normality assumption.

Nonparametric (NP) maximum likelihood estimation and the resulting discrete estimate of the random effects distribution (Laird 1978; Lindsay 1983) is not always satisfactory, especially when the random effects distribution is of primary interest, as it is more likely to be continuous than discrete. Another drawback of this approach is the extend of data required in order to estimate the NP mixing distribution precisely (Carroll and Hall 1988). Nonetheless, the NP approach yields fairly efficient estimates of the effects of the explanatory variables and it has been used extensively by statisticians. For instance, Aitkin (1999) used this approach in the context of generalized linear models and Hartzel et al. (2001) for modeling correlated multinomial responses. General fitting algorithms have been provided by Laird (1978), Lindsay (1983), Follmann and Lambert (1989) and Lesperance and Kalbfleisch (1992).

In addition to the NP approach, here we utilize flexible random effects densities that avoid the restrictive normality assumption but allow some degree of smoothness in the estimate of the random effects distribution, unlike the NP approach. Proposals for flexible random effects distributions include that of Gallant and Nychka (1987) who provided a representation of densities that belong to a class of smooth densities. In their representation, the normal density is multiplied by the square of an infinite series expansion. Several authors have approximated these smooth densities using truncated series expansions, treating the number of terms in the resulting polynomial as a tuning parameter. Specifically, Davidian and Gallant (1993) used these semi-nonparametric (SNP) densities in the context of nonlinear models, Zhang and Davidian (2001) in the context of linear models, and again, more recently, Chen et al. (2002) in the family of generalized linear models. We describe these flexible SNP densities in Sect. 2.1.

Magder and Zeger (1996) proposed smooth nonparametric (SMNP) random effects densities which they represented by arbitrary mixtures of Gaussians constraining the minimum variance (for the univariate case) of the mixture components to be greater than or equal to some value h, which they treated as a tuning parameter. The extension to the multivariate case constrains the determinants of the covariance matrices of the components to be greater than or equal to h. A related proposal is that of Verbeke and Lesaffre (1996) who proposed to represent the density of random effects using a mixture of K Gaussians with common covariance matrix. These proposals are described in Sect. 2.2. A related approach, as it is also based on mixtures of Gaussians, is that of Ghidey et al. (2004). This approach uses ideas from the P-spline smoothing literature (Eilers and Marx 1996) in order to obtain a smooth estimate of the random effects density. However, we do not further pursue this approach here.

In this paper, we extend the class of MGLMMs to include flexible random effects densities, such as the SNP and SMNP densities. A unified and automated Monte Carlo EM algorithm for fitting MGLMMs with either the SNP or SMNP mixing densities is presented. We also compare the two flexible densities in terms of their performance in capturing the true underlying shape of the random effects distribution, in terms of prediction accuracy of the random effects, and in estimating the vector of regression parameters and the corresponding standard errors. An application to a real data set, in which using flexible random effects densities allows the capturing of complex features of the underlying distribution, stresses the importance of these densities.

The remainder of this paper is arranged as follows. In Sect. 2 we introduce the family of MGLMMs and the SNP, SMNP and NP densities for the random effects. In Sect. 3

an algorithm for fitting MGLMMs with random effects that have either the SNP or the SMNP density is presented. Section 4 presents a simulation study while Sect. 5 presents an application to a real data set. The paper concludes with a brief discussion.

2 Model specification

Let $\mathbf{y}_{ij}$ denote the jth k-dimensional response vector for cluster i, $j = 1, \ldots, n_i$, $i = 1, \ldots, m$. Conditionally on the cluster specific random effect $\mathbf{u}_i (q \times 1)$, the responses $\mathbf{y}_{ij}$, $j = 1, \ldots, n_i$, are assumed to be independent with densities that belong to the multivariate exponential family (Fahrmeir and Tutz 2001, Chap. 3)

$$f(\mathbf{y}_{ij}|\mathbf{u}_i;\boldsymbol{\beta},\phi)=\exp\left\{\frac{w_{ij}}{\phi}[\mathbf{y}_{ij}^T\boldsymbol{\theta}_{ij} - b(\boldsymbol{\theta}_{ij})]+c(\mathbf{y}_{ij};\phi,w_{ij})\right\},$$

where the conditional mean $\boldsymbol{\mu}_{ij}^c = E(\mathbf{y}_{ij}|\mathbf{u}_i) = \partial b(\boldsymbol{\theta}_{ij})/\partial\boldsymbol{\theta}_{ij}$ is related to the linear predictor $\boldsymbol{\eta}_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{u}_i$ through the link function $\mathbf{g}(\boldsymbol{\mu}_{ij}^c) = \boldsymbol{\eta}_{ij}$. Further, $\boldsymbol{\beta}$ is a p-vector of fixed effects, $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ are $k \times p$ and $k \times q$ matrices of covariates associated with the (i, j)th response, where the columns of $\mathbf{Z}_{ij}$ are usually a subset of those of $\mathbf{X}_{ij}$. Moreover, w_{ij} is a known weight, $c()$ is a known function and ϕ is a dispersion parameter. For the multinomial probability mass function, which is our main focus here, $\phi = 1$ and it will thus be dropped in the remainder of this paper.

The random effects $\mathbf{u}_i, i = 1, \ldots, m$, are assumed to be independently and identically distributed. We now describe the SNP and SMNP densities for the random effects. These densities will be denoted by $g_{K,q}(\mathbf{u}_i|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a vector of parameters, K is a tuning parameter that controls the flexibility of the random effects density and q is the dimension of $\mathbf{u}_i$.

2.1 Semi-nonparametric (SNP) densities

Let $\mathbf{Z}$ be a random vector with a density that belongs to a class of smooth densities. Gallant and Nychka (1987) provided a characterization of such a class and a representation of the corresponding densities using an infinite series expansion. Using truncated series expansions, in order to approximate densities in this class, the standard SNP density of a q-dimensional vector $\mathbf{Z}$ has the representation

$$h_{K,q}(\mathbf{z}) \propto \{P_{K,q}(\mathbf{z})\}^2\phi_q(\mathbf{z}) = \left\{\sum_{|\boldsymbol{\lambda}|=0}^{K} a_{\boldsymbol{\lambda}}\mathbf{z}^{\boldsymbol{\lambda}}\right\}^2 \phi_q(\mathbf{z}), \qquad (1)$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q)$ is a vector of nonnegative integers, $|\boldsymbol{\lambda}| = \sum_{r=1}^q \lambda_r$, $\mathbf{z}^{\boldsymbol{\lambda}} = \prod_{r=1}^q z_r^{\lambda_r}$ and $\phi_q(.)$ is the probability density function (pdf) of a $N_q(\mathbf{0}, \mathbf{I})$ vector. The truncation parameter K controls the flexibility of the shape of $h_{K,q}(.)$. For instance, by setting $K = 1$, the resulting density can have up to two modes, while $K = 2$ allows up to three modes. Skewed and also distributions with fatter or thinner tails than the normal distribution can also be incorporated. It is worth noting that the normal distribution is a special case of $h_{K,q}(.)$ with $K = 0$.

An example that nicely illustrates $P_{K,q}(.)$ is the case where $q = 2$ and $K = 2$. Under this specification, $P_{2,2}(\mathbf{z}) = a_{00} + a_{10}z_1 + a_{01}z_2 + a_{11}z_1z_2 + a_{20}z_1^2 + a_{02}z_2^2$.

The constraint $\int h_{K,q}(\mathbf{z})d\mathbf{z} = 1$, is imposed by suitable choice of the coefficients $a_{\boldsymbol{\lambda}}$. Zhang and Davidian (2001) and Chen et al. (2002) treat the problem of normalizing (1) by imposing $E[\{P_{K,q}(\mathbf{V})\}^2] = 1$, where $\mathbf{V} \sim N_q(\mathbf{0}, \mathbf{I})$, and they show how this can be achieved for general K and q.

The resulting normalized density of vector $\mathbf{Z}$ does not have a mean of zero, a usual assumption in random effects models. The transformation $\mathbf{u} = \mathbf{R}\{\mathbf{Z} - \boldsymbol{\gamma}\}$, where $\boldsymbol{\gamma} = E(\mathbf{Z})$ and $\mathbf{R}$ is a lower triangular matrix, sets the mean to zero and, also, allows for more flexibility in the covariance matrix of the random effects. Thus, with $\boldsymbol{\theta}_1 = (\boldsymbol{\psi}^T, \mathbf{R}_d^T)^T$, where $\boldsymbol{\psi}$ includes the parameters in $\boldsymbol{\gamma}$ and in the coefficients $a_{\boldsymbol{\lambda}}$, and $\mathbf{R}_d$ includes the parameters in matrix $\mathbf{R}$, the density of the random effects can be expressed as

$$g_{K,q}(\mathbf{u};\boldsymbol{\theta}_1) = \{P_{K,q}(\mathbf{R}^{-1}\mathbf{u} + \boldsymbol{\gamma})\}^2\phi_q(\mathbf{u}; -\mathbf{R}\boldsymbol{\gamma}, \mathbf{R}\mathbf{R}^T), \quad (2)$$

where, for $\mathbf{V} \sim N_q(\mathbf{0}, \mathbf{I})$ we require $E\{P_{K,q}(\mathbf{V})\}^2 = 1$, and $\phi_q(.; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the pdf of a $N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ vector.

We consider the example where $q = 1$ and $K = 2$. In this case, $P_{2,1}(z) = a_0 + a_1z + a_2z^2$, and a reparametrization that ensures that the corresponding $h_{2,1}$ integrates to one, is $a_0 + a_2 = \cos(\psi_1)$, $a_1 = \sin(\psi_1)\cos(\psi_2)$ and $\sqrt{2}a_2 = \sin(\psi_1)\sin(\psi_2)$, where, for all r, $\psi_r \in (-\pi/2, \pi/2]$. The corresponding normalized density of a mean-zero variable u is

$$g_{2,1}(u;\psi_1,\psi_2,\sigma) = \{a_0 + a_1(\gamma + u/\sigma) + a_2(\gamma + u/\sigma)^2\}^2 \times \phi_1(u; -\gamma\sigma, \sigma^2), \qquad (3)$$

where $\gamma = 2\sin(\psi_1)\cos(\psi_2)\{\cos(\psi_1) + \sqrt{2}\sin(\psi_1)\sin(\psi_2)\}$.

2.2 Smooth nonparametric (SMNP) densities

The smooth nonparametric (SMNP) densities are defined in terms of mixtures of Gaussian distributions. Two classes of SMNP distributions for the random effects $\mathbf{u}_i$ have been proposed by Magder and Zeger (1996) and Verbeke and Lesaffre (1996).

First we describe the class of densities proposed by Magder and Zeger (1996). Using these authors' own notation, let $\Gamma_{h,q}$ be the class of q-variate distributions that can be expressed as mixtures of Gaussians with covariance matrices that have determinants greater than or equal to h. An elegant proof, which partly relies on the results of Laird (1978) and Lindsay (1983) on the discreetness and support

size of nonparametric mixing distributions, shows that the likelihood is maximized over $\Gamma_{h,q}$ by a mixture of Gaussians that has at most m components (m is the number of clusters) where each component has a covariance matrix with determinant exactly equal to h.

Verbeke and Lesaffre (1996) proposed a similar model. They assumed that the random effects arise from a mixture of K Gaussians with a common covariance matrix. We will denote this class of distributions by $\Gamma_{K,q}$. In practice K is unknown but in light of the discussion of the previous paragraph, it is easily seen that the likelihood is maximized by a mixture of at most m Gaussians with common covariance matrix.

For random intercept models, where the random effects are univariate, the family $\Gamma_{h,1}$ includes those distributions that can be expressed as mixtures of Gaussians that have common pre-specified variance h, while the family $\Gamma_{K,1}$ includes those distributions that can be expressed as mixtures of Gaussians that have components with common, but otherwise unspecified, variance. This is the main reason why in this paper, for the purposes of simulation and data analysis, we consider distributions in $\Gamma_{K,q}$ only. It is worth mentioning, however, that for models with multivariate random effects, $\Gamma_{K,q}$ requires a common covariance matrix in the components of the mixture, while $\Gamma_{h,q}$ requires the components to have covariance matrices with common determinant h, but not necessarily common covariance matrix.

With $\boldsymbol{\theta}_2$ representing the model parameters, the densities in $\Gamma_{K,q}$ can be expressed as

$$g_{K,q}(\mathbf{u}_i;\boldsymbol{\theta}_2)=\sum_{g=1}^{K} p_g N_q(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}), \tag{4}$$

where $\sum_{g=1}^{K} p_g = 1$. We set $\boldsymbol{\mu}_K = \sum_{g=1}^{K-1} p_g\boldsymbol{\mu}_g/(\sum_{g=1}^{K-1} p_g - 1)$ in order to satisfy the constraint $E(\mathbf{u}_i)=0$. Thus, there are $K-1$ independent means, $\boldsymbol{\mu}=(\boldsymbol{\mu}_1^T,\ldots,\boldsymbol{\mu}_{K-1}^T)^T$, and $K-1$ independent probabilities $\mathbf{p}=(p_1,\ldots,p_{K-1})^T$. Letting $\boldsymbol{\Sigma}_d$ be the vector of parameters in matrix $\boldsymbol{\Sigma}$, we have that $\boldsymbol{\theta}_2=(\boldsymbol{\mu}^T,\boldsymbol{\Sigma}_d^T,\mathbf{p}^T)^T$.

2.3 Nonparametric densities

The nonparametric approach uses arbitrary random effects, that is, it does not make any assumptions about the form of the distribution of the random effects. The resulting discrete estimate (Laird 1978; Lindsay 1983) of the random effect distribution is represented by K mass points, $m_1,\ldots,m_K$, and the corresponding probabilities, $\pi_1,\ldots,\pi_K$, where $\sum_{i=1}^{K}\pi_i=1$. In order to satisfy the constraint that $E(\mathbf{u}_i)=0$, we set $m_K=\sum_{i=1}^{K-1}\pi_i m_i/(\sum_{i=1}^{K-1}\pi_i-1)$. Fitting algorithms have been described by Laird (1978), Lindsay (1983), Follmann and Lambert (1989) and Lesperance and Kalbfleisch (1992). Also, for the special case of multinomial responses, Hartzel et al. (2001) provided a model fitting procedure. Thus, no more details will be provided here.

3 Model fitting

3.1 Log-likelihood, EM algorithm and an MC approximation of the E-step

A Monte Carlo EM (MCEM) algorithm that can be used for fitting MGLMMs with either semi or smooth nonparametric random effects is now described. First let $\boldsymbol{\delta}^T=(\boldsymbol{\beta}^T,\boldsymbol{\theta}^T)$, where $\boldsymbol{\theta}$ is either $\boldsymbol{\theta}_1=(\boldsymbol{\psi}^T,\mathbf{R}_d^T)^T$ of the SNP density or $\boldsymbol{\theta}_2=(\boldsymbol{\mu}^T,\boldsymbol{\Sigma}_d^T,\mathbf{p}^T)^T$ of the SMNP density. The marginal log-likelihood, $l(\boldsymbol{\delta}|\mathbf{y})$, where $\mathbf{y}=\{\mathbf{y}_i : i=1,\ldots,m\}$, can be written as

$$\begin{aligned} l(\boldsymbol{\delta}|\mathbf{y}) &= \log f(\mathbf{y}|\boldsymbol{\delta}) \\ &= \log\left\{\prod_{i=1}^{m}\int\cdots\int f(\mathbf{y}_i|\mathbf{u}_i;\boldsymbol{\beta})g_{K,q}(\mathbf{u}_i;\boldsymbol{\theta})d\mathbf{u}_i\right\}. \end{aligned} \tag{5}$$

Since no further evaluation of (5) is possible, we maximize $l(\boldsymbol{\delta}|\mathbf{y})$ by employing the EM algorithm of Dempster et al. (1977). This algorithm uses $(\mathbf{y},\mathbf{u})=\{(\mathbf{y}_i,\mathbf{u}_i) : i=1,\ldots,m\}$ as the complete data and it maximizes $l(\boldsymbol{\delta}|\mathbf{y})$ indirectly, by iteratively maximizing

$$\begin{aligned} Q(\boldsymbol{\delta}|\boldsymbol{\delta}^{(r)}) &= E\{\log f(\mathbf{y},\mathbf{u})|\mathbf{y},\boldsymbol{\delta}^{(r)}\} \\ &= \sum_{i=1}^{m}\int\cdots\int\{\log f(\mathbf{y}_i|\mathbf{u}_i;\boldsymbol{\beta})+\log g_{K,q}(\mathbf{u}_i|\boldsymbol{\theta})\} \\ &\quad\times h(\mathbf{u}_i|\mathbf{y}_i;\boldsymbol{\delta}^{(r)})d\mathbf{u}_i, \end{aligned} \tag{6}$$

where $\boldsymbol{\delta}^{(r)}$ is the current value of the vector parameter and $h(\mathbf{u}_i|\mathbf{y}_i;\boldsymbol{\delta}^{(r)})$ is the conditional distribution of the missing data given the observed data and current estimates. Let $\boldsymbol{\delta}^{(r+1)}=\boldsymbol{\phi}(\boldsymbol{\delta}^{(r)})$ be the vector parameter that maximizes $Q(\boldsymbol{\delta}|\boldsymbol{\delta}^{(r)})$. Under regularity conditions, at convergence $\hat{\boldsymbol{\delta}}=\boldsymbol{\phi}(\hat{\boldsymbol{\delta}})$ maximizes both the complete and the observed data likelihoods.

However, the E-step of (6) cannot be evaluated analytically as the normalizing constant of density $h(\mathbf{u}|\mathbf{y};\boldsymbol{\delta}^{(r)})$ is the marginal likelihood function. The MCEM algorithm of Booth and Hobert (1999) overcomes this difficulty by replacing $Q(\boldsymbol{\delta}|\boldsymbol{\delta}^{(r)})$ with a Monte Carlo (MC) approximation. Given random samples $\mathbf{u}_{i1},\ldots,\mathbf{u}_{iM}$ from $h(\mathbf{u}_i|\mathbf{y}_i;\boldsymbol{\delta}^{(r)})$, $i=1,\ldots,m$, an approximation to $Q(\boldsymbol{\delta}|\boldsymbol{\delta}^{(r)})$ is given by

$$\begin{aligned} Q_M(\boldsymbol{\delta}|\boldsymbol{\delta}^{(r)}) &= Q_M(\boldsymbol{\beta})+Q_M(\boldsymbol{\theta}) \\ &= M^{-1}\sum_{l=1}^{M}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log f(\mathbf{y}_{ij}|\mathbf{u}_{il};\boldsymbol{\beta}) \\ &\quad + M^{-1}\sum_{l=1}^{M}\sum_{i=1}^{m}\log g_{K,q}(\mathbf{u}_{il}|\boldsymbol{\theta}), \end{aligned} \tag{7}$$

where $Q_M(\boldsymbol{\beta})$ and $Q_M(\boldsymbol{\theta})$ have the obvious definitions.

The method of Booth and Hobert (1999) uses independent samples in order to construct the approximation Q_M. As these authors discuss, a significant advantage of independent over dependent samples that arise from Markov chains (see McCulloch 1997 and for other fitting algorithms therein), is the simplicity in resolving each MCEM step into the true EM step and the MC error which, in turn, allows us to automatically increase the MC sample size, M, as the algorithm approaches convergence. Indeed, as was also noted by Wei and Tanner (1990), it would be very wasteful to start the algorithm with a large M as the estimates at the beginning of the algorithm might be far from the true maximum likelihood (ML) estimates. On the other hand, it is sensible to increase M when the algorithm approaches convergence as this will free the ML estimates from MC error. As it was shown by Booth and Hobert (1999), ad hoc methods for increasing the MC sample size are not as efficient as methods that separate the true EM step from the MC error. We thus adopt the automated MCEM algorithm of Booth and Hobert (1999), as was also done by Hartzel et al. (2001) and Chen et al. (2002).

3.2 Rejection sampling

The random samples $\mathbf{u}_{i1},\ldots,\mathbf{u}_{iM}$ from $h(\mathbf{u}_i|\mathbf{y}_i;\boldsymbol{\delta}^{(r)})$, $i=1,\ldots,m$, are generated using a rejection sampling algorithm (Geweke 1996) in which the marginal distribution $g_{K,q}(\mathbf{u}_i|\boldsymbol{\theta}^{(r)})$, such as (2) and (4), is used as the candidate. Specifically, suppose that it is easy to generate random samples from $g_{K,q}(\mathbf{u}_i|\boldsymbol{\theta}^{(r)})$. Then, the following algorithm can be used to generate a sample from $h(\mathbf{u}_i|\mathbf{y}_i;\boldsymbol{\delta}^{(r)})$:

1. Generate $\mathbf{s}$ from $g_{K,q}(\mathbf{u}_i|\boldsymbol{\theta}^{(r)})$ and w_1 from uniform $(0,1)$.
2. If $w_1 \le f(\mathbf{y}_i|\mathbf{s};\boldsymbol{\beta}^{(r)})/\zeta_i$, where $\zeta_i=\sup_{\mathbf{u}_i} f(\mathbf{y}_i|\mathbf{u}_i;\boldsymbol{\beta}^{(r)})$, accept $\mathbf{s}$ as a sample from density $h(\mathbf{u}_i|\mathbf{y}_i;\boldsymbol{\delta}^{(r)})$. Otherwise return to 1.

When $g_{K,q}(\mathbf{u}_i|\boldsymbol{\theta}^{(r)})$ is the SMNP that appears in (4), Step 1 of the above algorithm can be achieved by simply: (a) generating w_2 from uniform $(0,1)$ and (b) generating $\mathbf{s}$ from $N_q(\boldsymbol{\mu}_\xi,\boldsymbol{\Sigma})$, where ξ is the smallest positive integer such that $w_2 \le \sum_{i=1}^{\xi} p_i^{(r)}$.

When $g_{K,q}(\mathbf{u}_i|\boldsymbol{\theta}^{(r)})$ is the SNP that appears in (2), the algorithm proposed by Chen et al. (2002) can be used. Their development, which for the sake of completeness we describe briefly here, takes advantage of the developments of Gallant and Tauchen (1992). Specifically, if we can generate a random sample $\mathbf{s}_h$ from normalized $h_{K,q}(\mathbf{x}|\boldsymbol{\theta}^{(r)})$ in (1), then a sample from density (2) can be obtained as $\mathbf{s}_g=\mathbf{R}^{(r)}(\mathbf{s}_h-\boldsymbol{\gamma}^{(r)})$. Generating from $h_{K,q}(\mathbf{x}|\boldsymbol{\theta}^{(r)})$ is done using an accept-reject algorithm which requires finding an integrable function $b_{K,q}(\mathbf{x}|\boldsymbol{\theta}^{(r)})$ that satisfies $0 \le h_{K,q}(\mathbf{x}|\boldsymbol{\theta}^{(r)}) \le b_{K,q}(\mathbf{x}|\boldsymbol{\theta}^{(r)})$ for all $\mathbf{x}$. Gallant and Tauchen (1992) proposed using the following function

$$b_{K,q}(\mathbf{x}|\boldsymbol{\theta}^{(r)}) = \{P^*_{K,q}(|\mathbf{x}|)\}^2\phi_q(\mathbf{x}) = \left\{\sum_{|\boldsymbol{\lambda}|=0}^{K} |a_{\boldsymbol{\lambda}}^{(r)}||\mathbf{x}|^{\boldsymbol{\lambda}}\right\}^2 \phi_q(\mathbf{x}),$$

where $|a_{\boldsymbol{\lambda}}^{(r)}|$ denotes the absolute value of coefficient $a_{\boldsymbol{\lambda}}^{(r)}$ and $|\mathbf{x}|^{\boldsymbol{\lambda}}=\prod_{i=1}^{q}|x_i|^{\lambda_i}$. By expanding $\{P^*_{K,q}\}^2$ and restricting the support of $b_{K,q}(\mathbf{x}|\boldsymbol{\theta}^{(r)})$ (which is symmetric around zero) to $\{\mathbf{x}: x_r>0, r=1,\ldots,q\}$, we observe that $b_{K,q}$ can be written as a weighted sum of products of q density functions of χ random variables, where a χ density function is obtained as the density of the square root of a χ^2 random variable. After normalizing the weights, this sum can be interpreted as a mixture distribution of independent components, each of them consisting of q independent χ random variables, with weights $w_t^{(r)}$, where t indexes the components of the mixture. The algorithm of Monahan (1987) can be used to generate a random sample from a χ distribution.

Thus, generating a sample from $h_{K,q}(\mathbf{x}|\boldsymbol{\theta}^{(r)})$ can be achieved by the following algorithm: (a) generate w_3 from uniform$(0,1)$ and determine ν: the smallest positive integer such that $w_3 \le \sum_{t=1}^{\nu} w_t^{(r)}$, for some arbitrary put predetermined ordering of the components of the dominating function, (b) generate a sample v_i from each of the χ distributions that appear in component ν, $i=1,\ldots,q$, (c) Change the sign of v_i with probability 50% and assign the result to the ith element of the q-vector $\mathbf{s}$, $i=1,\ldots,q$ and (d) generate w_4 from uniform$(0,1)$ and accept $\mathbf{s}$ as a sample from $h_{K,q}(\mathbf{x}|\boldsymbol{\theta}^{(r)})$ if $w_4 \le h_{K,q}(\mathbf{x}|\boldsymbol{\theta}^{(r)})/b_{K,q}(\mathbf{x}|\boldsymbol{\theta}^{(r)})$; otherwise return to (a).

3.3 Maximization step

Returning now to (7), maximizing $Q_M(\boldsymbol{\beta})$ is equivalent to maximizing the log-likelihood of an MGLM with independent observations and an offset term, $\mathbf{Z}_{ij}\mathbf{u}_i$, in the linear predictors. Thus, a Fisher scoring algorithm can be used for maximizing with respect to the regression parameters. The score function and Fisher information matrix take the following form (Fahrmeir and Kaufmann 1985)

$$\sum_{ijl}\mathbf{X}_{ij}^T\mathbf{D}_{ijl}^T\mathbf{S}_{ijl}^{-1}\{\mathbf{y}_{ij}-\boldsymbol{\mu}_{ijl}\} \quad \text{and}$$

$$\sum_{ijl}\mathbf{X}_{ij}^T\mathbf{D}_{ijl}^T\mathbf{S}_{ijl}^{-1}\mathbf{D}_{ijl}\mathbf{X}_{ij},$$

where $\boldsymbol{\mu}_{ijl}\equiv\boldsymbol{\mu}_{ijl}(\boldsymbol{\beta}^{(r)})=\mathbf{h}(\boldsymbol{\eta}_{ijl}^{(r)})=\mathbf{h}(\mathbf{X}_{ij}\boldsymbol{\beta}^{(r)}+\mathbf{Z}_{ij}\mathbf{u}_{il})$ and suppressing the dependence on $\boldsymbol{\beta}^{(r)}$, $\mathbf{D}_{ijl}=\partial\mathbf{h}(\mathbf{s})/\partial\mathbf{s}$ evaluated at $\mathbf{s}=\boldsymbol{\eta}_{ijl}^{(r)}$ and $\mathbf{S}_{ijl}$ is the covariance matrix of $\mathbf{y}_{ij}$.

Maximization of $Q_M(\boldsymbol{\theta})$ depends on the choice of the random effects density. For the SNP density, we maximize $Q_M(\boldsymbol{\theta}_1)$, $\boldsymbol{\theta}_1 = (\boldsymbol{\psi}^T, \mathbf{R}_d^T)^T$, using an implementation of the Nelder-Mead algorithm in C++ (library ASA047). For the SMNP density, maximization $Q_M(\boldsymbol{\theta}_2)$, $\boldsymbol{\theta}_2 = (\boldsymbol{\mu}^T, \boldsymbol{\Sigma}_d^T, \mathbf{p}^T)^T$, is achieved by an EM algorithm for fitting mixtures of Gaussians to independent data, as follows:

First notice that $Q_M(\boldsymbol{\theta}_2) = M^{-1}\sum_{i,l} \log g_{K,q}(\mathbf{u}_{il}|\boldsymbol{\theta}_2)$ is the log-likelihood of independent $\mathbf{u}_{il}$, $i = 1, \ldots, m, l = 1, \ldots, M$. We regard $\{\mathbf{u}_{il} : i = 1, \ldots, m, l = 1, \ldots, M\}$ as the observed and $\{G_{il} : i = 1, \ldots, m, l = 1, \ldots, M\}$ as the missing data, where $G_{il} = g$, $g = 1 \ldots, K$, denotes the event that $\mathbf{u}_{il}$ comes from the gth component of (4). It is then clear that maximization of $Q_M(\boldsymbol{\theta}_2)$ with respect to $\boldsymbol{\theta}_2$ is equivalent to maximizing

$$\begin{aligned} &Q_{M,2}(\boldsymbol{\theta}_2|\boldsymbol{\theta}_2^{(r)}) \\ &\quad = Q_{M,2}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{p}) + Q_{M,2}(\mathbf{p}) \\ &\quad = \frac{1}{M}\sum_{gli} \pi(G_{il} = g|\mathbf{u}_{il}; \boldsymbol{\theta}_2^{(r)}) \\ &\qquad \times \{\log f(\mathbf{u}_{il}|G_{il} = g) + \log P(G_{il} = g)\}, \end{aligned} \tag{8}$$

where $f(\mathbf{u}_{il}|G_{il} = g)$ denotes the $N_q(\boldsymbol{\mu}_g, \boldsymbol{\Sigma})$ density and $\pi(G_{il} = g|\mathbf{u}_{il}; \boldsymbol{\theta}_2^{(r)})$ is the posterior probability that $\mathbf{u}_{il}$ arises from the gth component of the mixture. Notice that since $\boldsymbol{\mu}_K = \sum_{i=1}^{K-1} p_i \boldsymbol{\mu}_i / (\sum_{i=1}^{K-1} p_i - 1)$, as discussed after (4), the parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{p}$ do not separate in $Q_{M,2}$. Even so, given the representation in (8), finding the maximum likelihood estimate of $\boldsymbol{\theta}_2$ is a straightforward iterative procedure.

3.4 The MCEM algorithm

The MC sample size M is automatically increased as the algorithm approaches convergence. Using Taylor series methods along with the central limit theorem, Booth and Hobert (1999) showed that after the $(r + 1)$th iteration, $\boldsymbol{\delta}^{(r+1)}$, the maximizer of (7), is approximately normally distributed around the maximizer of (6), $\boldsymbol{\delta}_Q^{(r+1)}$ say. These authors also provided an estimator of the corresponding covariance matrix. Thus, if convergence cannot be declared at the end of the $(r + 1)$th iteration, an approximate $100(1 - \alpha)\%$ confidence ellipsoid for $\boldsymbol{\delta}_Q^{(r+1)}$ is constructed and it is used to infer whether or not the MCEM algorithm is close to convergence. If the previous observed maximizer, $\boldsymbol{\delta}^{(r)}$, lies in this confidence ellipsoid, then we may conclude that convergence cannot be declared due mostly to the MC error and thus M should be increased; otherwise M is not increased. In the simulation studies and the data analysis we performed, we set the initial value for $M = 100$, $\alpha = 25\%$ and each time $\boldsymbol{\delta}^{(r)}$ was in the confidence region of $\boldsymbol{\delta}_Q^{(r+1)}$, we increased M by 33%.

When fitting the SNP and SMNP models, we start by setting $K = 0$, that is, by fitting the normal random effects model. Starting values for the regression coefficients of this model are taken to be the estimates from the NP model (Laird 1978; Lesperance and Kalbfleisch 1992) with the largest number of mass points, while the covariance matrix of the random effects is taken to be a diagonal matrix with elements the variances implied by the NP random effects model. For the SNP model, the extra parameters that result by increasing K are given a starting value of zero. For the SMNP model, when $K = d > 1$, the estimated masses and mass points of the NP model with d mass points are taken to be the starting values for the probabilities and means of the components.

In summary, the MCEM algorithm for fitting the SNP or SMNP models, for a fixed value of K, starts by setting starting values $\boldsymbol{\delta}^{(0)}$ and selecting a starting MC sample size M. Iteration $(r + 1)$ consists of generating samples of random effects from $h(\mathbf{u}_i|\mathbf{y}_i; \boldsymbol{\delta}^{(r)})$, $i = 1, \ldots, m$, finding the maximizer $\boldsymbol{\delta}^{(r+1)}$ of $Q_M(\boldsymbol{\delta}|\boldsymbol{\delta}^{(r)})$, given in (7), and either stopping if convergence is achieved or considering increasing the MC sample size. Our rule for stopping the fitting algorithm is that the maximum absolute change in the parameters from successive iterations is less than ϵ, where we set $\epsilon = 0.002$ for our data analysis application and $\epsilon = 0.005$ for our simulation studies.

To find the optimal value of K for each model, we used the Akaike's information criterion. For fixed K, $\mathrm{AIC}(K) = -l(\hat{\boldsymbol{\delta}}; \mathbf{y}) + \dim(\hat{\boldsymbol{\delta}}|K)$, where $\dim(\hat{\boldsymbol{\delta}}|K)$ is the dimension of vector $\hat{\boldsymbol{\delta}}$. As the log-likelihood $l(\hat{\boldsymbol{\delta}}; \mathbf{y})$ cannot be evaluated analytically, we use the MC method of Chen et al. (2002) in order to find an approximate value for it. Specifically, given random samples $\mathbf{u}_{il}, i = 1, \ldots, m, l = 1, \ldots, M$, from $g_{K,q}(\mathbf{u}_{il}; \hat{\boldsymbol{\theta}})$ the log-likelihood is approximated by $l(\hat{\boldsymbol{\delta}}; \mathbf{y}) \approx \sum_{i=1}^{m} \log\{M^{-1}\sum_{l=1}^{M} f(\mathbf{y}_{ij}|\mathbf{u}_{il}; \hat{\boldsymbol{\beta}})\}$.

At convergence, in addition to the log-likelihood, we calculate the standard errors of the estimates by inverting the observed information matrix of all estimated parameters. We calculate the information matrices as $I_1 = -\sum_{i=1}^{m} \partial^2 l(\hat{\boldsymbol{\delta}}; \mathbf{y}_i)/\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}^T$ but, due to MC error, it was not always positive definite. This problem has also been reported by Chen et al. (2002) and, in our experience, is more pronounced for the SNP than the SMNP densities. We thus also use the method proposed by Chen et al. (2002), $I_2 = \sum_{i=1}^{m} \{\partial l(\hat{\boldsymbol{\delta}}; \mathbf{y}_i)/\partial\boldsymbol{\delta}\}\{\partial l(\hat{\boldsymbol{\delta}}; \mathbf{y}_i)/\partial\boldsymbol{\delta}^T\}$.

4 A simulation study

A simulation study is carried out in order to evaluate the performance of the SNP and SMNP densities in a case of an ordinal response with three categories. We consider two values for the number of clusters m, namely $m = 250$ and

$m = 75$, corresponding to large and moderate sample size. The number of observations per cluster is $n_i = 5$ for all i. Data is generated in a two stage process, the first step of which consists of generating univariate random effects $u_i, i = 1, \ldots, m$, from a distribution G. For all choices of G that we consider, the implied variance of the random effects is fixed to $\text{var}(u_i) = 4$. Specific choices of G include (a) discrete distribution with probabilities equally split between two mass points, (b) centered at zero and scaled χ^2 distribution with $df = 5$: $(\chi^2_5 - 5)/\sqrt{(10/4)}$, and (c) normal distribution.

The ordinal data is generated by a random-effects proportional odds model. This model describes the ordinal responses in the three categories in terms of two cumulative probabilities that are related to linear predictor through a logit link. In proportional odds models, the effects of covariates are assumed to be equal across the logits of cumulative probabilities and only the intercepts are specific to these logits. Specifically, given random effects $u_i, i = 1, \ldots, m$, each of the two components of the linear predictor, $r = 1, 2$, for each time point $j = 1, \ldots, 5$, is generated according to

$$\eta_{ijr} = \theta_r + \beta_1 D_i + \beta_2 T_j + u_i,$$

where $D_i = 1$ if $i < m/2$ and $D_i = 0$ otherwise, that is, D_i is a treatment indicator and $\beta_1 = 0.75$ is the treatment effect. Further, T_j is the jth element of $(-0.2, -0.1, 0.0, 0.1, 0.2)^T$, that is, T_j is a time varying covariate and its effect is $\beta_2 = 2.0$. The two intercepts are set to $\theta_1 = -1.0$ and $\theta_2 = 1.0$. With these specifications, we can obtain values for $\boldsymbol{\eta}_{ij} = (\eta_{ij1}, \eta_{ij2})^T$ for all i and j. Based on $\boldsymbol{\eta}_{ij}$, the response vector $\mathbf{y}_{ij}$ is then obtained as a realization of a multinomial variable with cell probabilities obtained as the solution to

$$\log\left(\frac{\pi_{ij1}}{1 - \pi_{ij1}}\right) = \eta_{ij1} \quad \text{and}$$

$$\log\left(\frac{\pi_{ij1} + \pi_{ij2}}{1 - \pi_{ij1} - \pi_{ij2}}\right) = \eta_{ij2}.$$

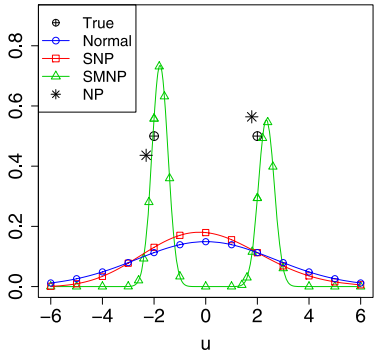
We simulate 200 data sets for each combination of random effects distribution and sample size m and fit the random-effects proportional odds models, with correct mean structure, and the SNP and SMNP densities with $K = 2$, in which case both densities have three parameters. The SNP density for $K = 2$ is given in (3), while the SMNP density is $g_{2,1}(u|p, \mu, \sigma) = p\phi(u; \mu, \sigma^2) + (1 - p)\phi(u; p\mu/(p - 1), \sigma^2)$. We also fit the models with normal and NP densities for the random effects, but our main focus is on the flexible densities.

The performance of each of these models in capturing the characteristic features of the true underlying random effects distribution is evaluated by calculating the integrated squared error (ISE) between the estimated and true cumulative distribution functions, given by $\int\{\hat{G}(u) - G(u)\}^2 du$, where $\hat{G}$ is obtained by replacing the parameters of G with parameter estimates. Bias in estimates of the regression parameters is summarized by reporting the relative bias: for any parameter ϑ, with a corresponding estimate $\hat{\vartheta}$, relative bias is calculated as $RB(\hat{\vartheta}) = 100 * (\hat{\vartheta} - \vartheta)/\vartheta$. We investigate the impact of the assumed random effects distribution on the standard errors (SEs) by reporting the ratio of the SEs under all considered models to the SE under the normal model: $R_1\{\text{SE}(\hat{\vartheta})\}$. We also report the ratio of the estimated random effects standard deviation (SD) to the true SD: $R_2\{\hat{\text{SD}}(u_i)\} = \hat{\text{SD}}(u_i)/2$. Finally, the prediction accuracy of the random effects is examined by calculating the average prediction error: $\text{PE}(G) = m^{-1}\sum_{i=1}^{m}(u_i - \hat{u}_i^G)^2$, where $\hat{u}_i^G$ is the empirical Bayes estimate of u_i, $i = 1, \ldots, m$, for the random effects distribution G.

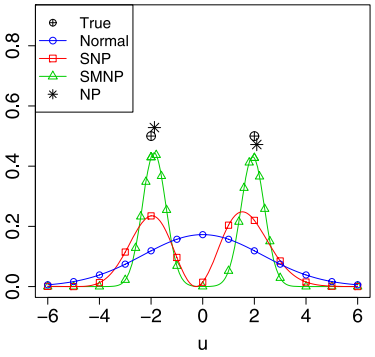
Table 1 displays the results for the case where the true random effects distribution G is a two point discrete distribution which splits the masses equally between the two mass points. With the mean constrained to be zero and the variance fixed to four, the only choice of mass points is $m_1 = -2$ and $m_2 = 2$. Some interesting issues emerge out of this table. We first notice that average ISE under an SNP density is about 70% of that under a normal density, for both values of m. Also, average ISE under an SMNP density is about 55% of that under an SNP density, again for both values of m. It is also obvious that ISE decreases for the SNP, SMNP and NP densities as the number of clusters increases. This, however, does not hold true for the normal model because

Table 1 Simulation Results: Discrete uniform case. ISE: integrated squared error; β_1, β_2: effects of cluster-level and within-cluster covariates; RB: relative bias; R_1: ratio of the SE of the estimate to the SE under the normal model, R_2: ratio of the estimated SD of the random effects distribution to the true one, PE: prediction error of the random effects

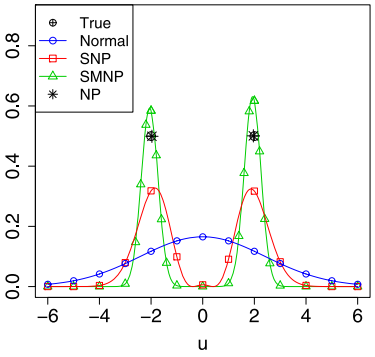
Sample size	$m = 250$				$m = 75$			
Model	Nor.	SNP	SMNP	NP	Nor.	SNP	SMNP	NP
Mean ISE	0.191	0.123	0.059	0.063	0.196	0.134	0.086	0.115
$\text{RB}(\hat{\beta}_1)(\%)$	2.07	4.73	2.13	−1.43	17.44	10.20	5.52	1.44
$R_1\{\text{SE}(\hat{\beta}_1)\}$	1.00	0.68	0.46	0.45	1.00	0.72	0.46	0.44
$\text{RB}(\hat{\beta}_2)(\%)$	2.42	4.32	3.02	0.77	8.13	10.06	8.39	5.70
$R_1\{\text{SE}(\hat{\beta}_2)\}$	1.00	1.02	1.00	0.98	1.00	1.01	0.98	0.95
$R_2\{\hat{\text{SD}}(u_i)\}$	1.19	1.13	1.04	0.99	1.20	1.16	1.05	1.01
$\text{PE}(G)$	0.914	0.364	0.196	0.170	1.017	0.480	0.270	0.244



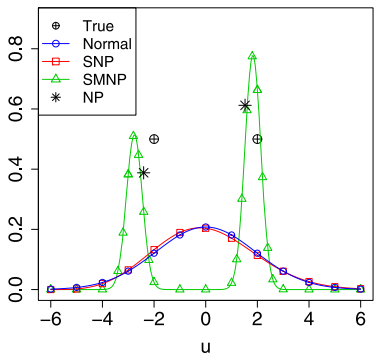
(a) ISE = 0.195, 0.192, 0.084, 0.128



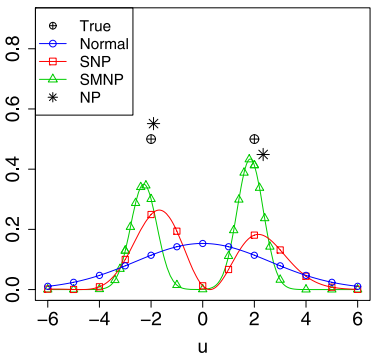
(b) ISE = 0.191, 0.105, 0.057, 0.056



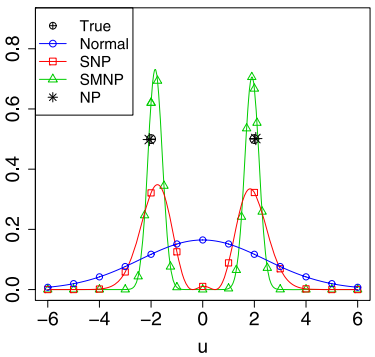
(c) ISE = 0.190, 0.073, 0.040, 0.018



(d) ISE = 0.212, 0.209, 0.147, 0.224



(e) ISE = 0.193, 0.118, 0.077, 0.103



(f) ISE = 0.190, 0.074, 0.045, 0.031

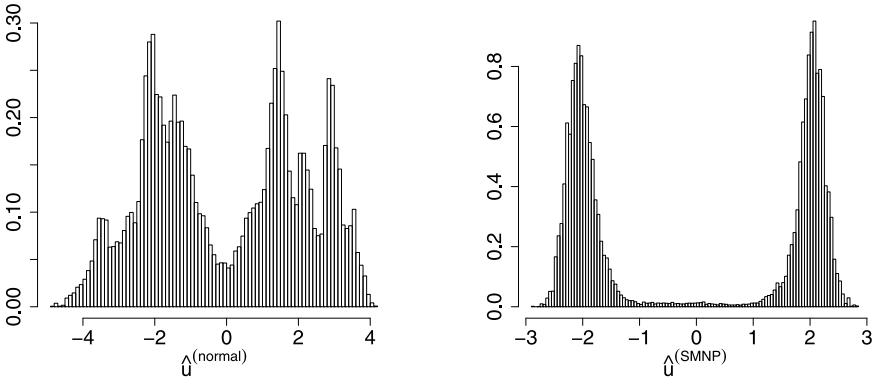
Fig. 1 True discrete distribution and plots of the estimated Normal, SNP, SMNP and NP densities that correspond to (**a**) & (**d**) 95th percentile of ISE, (**b**) & (**e**) 50th percentile of ISE, and (**c**) & (**f**) 5th percentile of ISE. (**a**)–(**c**) correspond to $m = 250$ and (**d**)–(**f**) to $m = 75$. ISE provided for the Normal, SNP, SMNP and NP densities respectively

the normal estimated density is always unimodal and symmetric. In order to get an understanding of what the values of ISE reported in Table 1 mean, we plot in Fig. 1 the estimated densities that correspond to the 95th, 50th and 5th percentile of ISE, for both $m = 250$ and $m = 75$. The corresponding values of ISE are provided in the legends of these graphs. We can see that, at the 95th percentile of ISE, the corresponding SNP density does not provide a bimodal estimate, as the SMNP density does. In fact, for both values of m, the estimated SMNP density is bimodal for all simulated datasets, while the estimated SNP density is bimodal for 73% of the simulated datasets.

For $m = 250$, the estimated SMNP density always results in smaller ISE than the corresponding ISE of the SNP density. For this value of m, AIC, described in Sect. 3.4, or equivalently any other selection criterion, always selects the model that assumes an SMNP density for the random effects over the model that assumes an SNP density. For $m = 75$, ISE of the estimated SNP density is smaller than that of the SMNP density for only 7.5% of the simulated datasets. However, when this happens, AIC selects the SMNP model and in 2% of the simulated datasets it selects the SNP model even though in those instances ISE of the SMNP model is smaller than that of the SNP.

It is also interesting to observe that the fitted nonparametric densities result in an average ISE that is larger than the average ISE of the SMNP densities. In fact, this happens for 51.5% and 78.5% of the simulated data sets with $m = 250$ and $m = 75$, respectively. Even though the nonparametric model is the true model in this case, a great deal of data is needed in order to obtain a precise estimate of the nonparametric density.

Returning to Table 1, we further see that for the large sample size case, bias in the estimation of the effect of cluster varying covariate is negligible, under all considered models. However, for the moderate sample size case, the relative bias can sometimes be quite high. For instance, under the normal density, the relative bias is as high as 17.4%. It decreases, however, considerably under the SNP and SMNP densities. The inefficient estimation under a normal model is also evident in the third line of Table 1. First, notice that there is little difference between the large and moderate sam-



(a) Normal (b) SMNP

Fig. 2 Histograms of the empirical Bayes estimates of the random effects from all 200 simulations with $m = 250$ when the true random effects are discrete with equal masses at $m_1 = -2$ and $m_2 = 2$ based on (**a**) normal and (**b**) SMNP random effects distributions

Table 2 Simulation Results: Centered and scaled χ^2_5 distribution case. ISE: integrated squared error; β_1, β_2: effects of cluster-level and within-cluster covariates; RB: relative bias; R_1: ratio of the SE of the estimate to the SE under the normal model, R_2: ratio of the estimated SD of the random effects distribution to the true one, PE: prediction error of the random effects. * Not calculated as the NP density allows infinite mass points

Sample size	$m = 250$				$m = 75$			
Model	Nor.	SNP	SMNP	NP	Nor.	SNP	SMNP	NP
Mean ISE	0.022	0.019	0.018	*	0.024	0.024	0.029	*
RB($\hat{\beta}_1$)(%)	−0.31	−2.89	−5.43	−0.04	−10.96	−7.72	−10.01	−5.85
$R_1\{\text{SE}(\hat{\beta}_1)\}$	1.00	1.06	1.02	0.92	1.00	1.01	0.98	1.09
RB($\hat{\beta}_2$)(%)	2.23	1.60	1.37	−1.11	−3.28	−3.54	−4.37	−2.40
$R_1\{\text{SE}(\hat{\beta}_2)\}$	1.00	0.99	0.99	1.06	1.00	1.00	0.99	0.99
$R_2\{\widehat{\text{SD}}(u_i)\}$	1.00	1.03	0.94	*	0.99	1.01	0.93	*
PE(G)	1.384	1.318	1.418	*	1.382	1.425	1.467	*

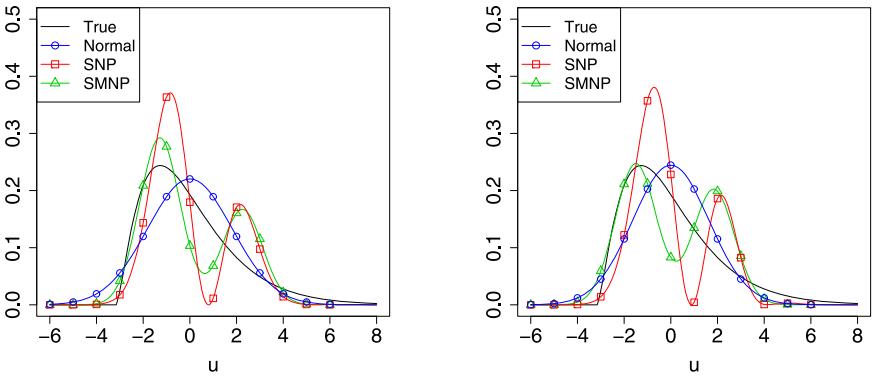
ple size cases. We will thus describe only the large sample size case. The SE of $\hat{\beta}_1$ under the normal model is, on average, $1.47 = 1/0.68$ and $2.17 = 1/0.46$ times larger than that under the SNP and SMNP models respectively. In addition, the SE under the SNP model is 1.48 times bigger than that under the SMNP model. Generally speaking, for the cluster varying covariate, the relative bias and SE are decreasing as ISE decreases. For the time varying covariate, however, estimation, both in terms of relative bias and SE, does not depend on the assumption about the random effects distribution: entries in the relative bias line, within each sample size, are approximately equal for all random effects distributions, and entries in the SE ratio line is approximately equal to one. Finally, estimation of the random effects SD is poor under the normal and SNP models.

Concerning the prediction error (PE) of the random effects, we notice from the last row of Table 1 that PE under an SNP density, for the case where $m = 250$, is about 40% of the PE assuming a normal density. For $m = 75$, the ratio of the PE under an SNP to that under a normal density is 47%. Further, PE under an SMNP density is about 55% of that under an SNP density, for both values of m. Finally, comparing the PEs for $m = 250$ to those obtained for $m = 75$, it is clear that PE decreases as the number of independent clusters increases under all assumed models for the random effects densities. However, this decrease is slow even for the true NP model.

Figure 2 displays the EB estimates of the random effects obtained in all 200 simulations with $m = 250$. Although both graphs are bimodal, it is clear that the estimates of the random effects are driven by the estimated random effects density. Under the SMNP density the separation between the two modes is very clear (as the two modes are captured by the SMNP density) but this separation is blurred for the restrictive normality assumption.

When the true random effects distribution is continuous, as it is in the second scenario we consider, differences in ISE, estimation of the unknown parameters and the accompanying SEs and, also, the PEs of the random effects, are minimal. Table 2 displays the results for the case where the random effects have a χ^2_5 distribution centered at zero and scaled to have SD equal to two. On average, within each sample size considered, ISE is approximately equal among the three densities considered. It is also clear that it decreases at a very slow rate as the sample size increases. The behavior of the PEs is very similar: they are approximately equal over the three densities considered and they decreases at a very slow rate as the number of clusters increases.

Figure 3 displays the estimated densities that correspond to the median value of ISE for $m = 250$ and $m = 75$. It is

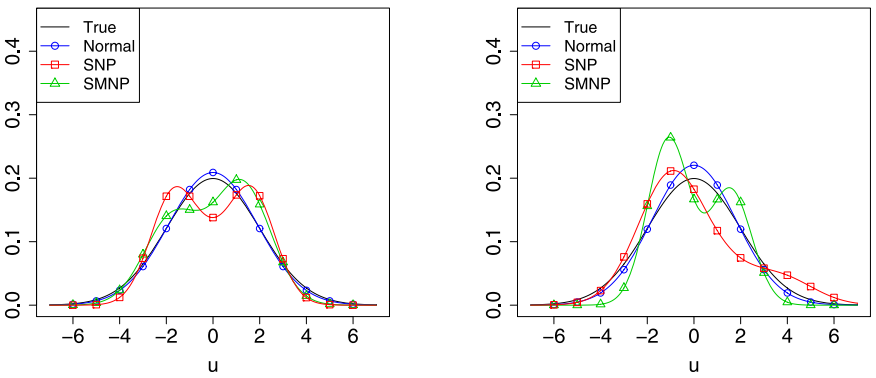


(a) ISE = 0.020, 0.017, 0.017 (b) ISE = 0.021, 0.020, 0.025

Fig. 3 True skewed density and plots of the estimated densities under the Normal, SNP and SMNP models that correspond to the 50th percentile of ISE for (**a**) $m = 250$ and (**b**) $m = 75$. ISE provided for the Normal, SNP and SMNP densities respectively

Table 3 Simulation Results: Normal distribution case. ISE: integrated squared error; β_1, β_2: effects of cluster-level and within-cluster covariates; RB: relative bias; R_1: ratio of the SE of the estimate to the SE under the normal model, R_2: ratio of the estimated SD of the random effects distribution to the true one, PE: prediction error of the random effects. * Not calculated as the NP density allows infinite mass points

Sample size	$m = 250$				$m = 75$			
Model	Nor.	SNP	SMNP	NP	Nor.	SNP	SMNP	NP
Mean ISE	0.002	0.017	0.003	*	0.005	0.021	0.014	*
RB($\hat{\beta}_1$)(%)	4.05	1.89	−3.92	1.41	−1.16	−2.63	−5.41	−3.35
$R_1\{\text{SE}(\hat{\beta}_1)\}$	1.00	1.10	0.98	1.10	1.00	1.07	0.93	1.09
RB($\hat{\beta}_2$)(%)	−2.96	−3.82	2.04	2.04	−0.20	−0.62	5.33	−1.56
$R_1\{\text{SE}(\hat{\beta}_2)\}$	1.00	1.00	1.18	1.14	1.00	0.99	1.01	1.01
$R_2\{\widehat{\text{SD}}(u_i)\}$	0.99	1.05	0.98	*	1.01	1.06	0.98	*
PE(G)	0.887	0.989	0.919	*	0.963	1.142	1.051	*



(a) ISE = 0.001, 0.005, 0.003 (b) ISE = 0.003, 0.016, 0.013

Fig. 4 True normal density and plots of the estimated densities under the Normal, SNP and SMNP models that correspond to the 50th percentile of ISE for (**a**) $m = 250$ and (**b**) $m = 75$. ISE provided for the Normal, SNP and SMNP densities respectively

clear that the SNP and SMNP densities with three parameters cannot really capture a skewed distribution. This, however, is not necessarily bad news as one can approximate distributions of any shape by increasing the number of parameters in the SNP and SMNP densities.

Table 2 makes obvious that the relative bias in the estimation of both β_1 and β_2 is negligible for $m = 250$. Relative bias is larger for $m = 75$, especially for the cluster varying covariate, but results do not differ for the three densities considered. Furthermore, SEs and estimation of the SD of the random effects is equally good for all densities considered.

In the third and final scenario that we consider, the true random effects distribution is normal with variance equal to four. Results are displayed in Table 3. Not surprisingly, average ISE, for both values of m, is smallest when fitting a model that assumes a normal random effects distribution. On average it is largest, again for both values of m, when fitting an SNP density to the random effects. Figure 4 shows the estimated densities that correspond to the median value of ISE for $m = 250$ and $m = 75$. For some simulated datasets, the two flexible densities 'identify' more complex than just unimodal, symmetric densities. However, AIC selects the

normal over the SNP (SMNP) model for 99% (97%) of the simulated datasets, for the large sample size case. For the moderate sample size case, these percentages are 88% and 92% for the SNP and SMNP densities respectively.

For both values of m, and for all models considered, relative bias in the estimation of the regression parameters is negligible. In addition, SEs of parameter estimates are approximately equal for all models considered and, also, estimation of the SD of the random effects does not differ by much for each of the models fitted. Thus, our conclusion is very similar to the one of Chen et al. (2002): the estimation efficiency is not compromised when fitting more flexible than needed random effects densities.

For the PE, again, not surprisingly, we see that it is smallest when fitting a model that assumes a normal random effects distribution, for both values of m. The differences between the PEs under SNP and SMNP densities are quite small, with the SMNP density doing slightly better than the SNP.

5 Application: NIMH schizophrenia data

We apply the methods described here to data from the National Institute of Mental Health Schizophrenia Collaborative Study. Specifically, the response variable of interest measures 'Severity of Illness'. Hedeker and Gibbons (2006) provide a detailed description and also several analyses of these data. Some of the important features of the study, as described by Hedeker and Gibbons (2006), are as follows:

In this parallel group study, $m_1 = 329$ patients were randomly assigned to the anti-psychotic drug group and $m_2 = 108$ patients to the placebo group. 'Severity of Illness' was measured, at weeks $j = 0, 1, \ldots, 6$, on a four category ordered scale: 1. normal or borderline mentally ill, 2. mildly or moderately ill, 3. markedly ill, and 4. severely or among the most extremely ill. There are very few observations at weeks 2, 4 and 5 as the plan was for subjects to be measured during weeks 0, 1, 3 and 6. Due to missing data, the total number of observations on these subjects, over all weeks, was $N = 1603$. In the placebo and drug group, 65% and 81% of the subjects completed the study.

As in Hedeker and Gibbons (2006), we model the logits of the cumulative probabilities, $\log[P(Y_{ij} \leq r)/\{1 - P(Y_{ij} \leq r)\}]$. In order to linearize the relationship between time and the observed cumulative logits, time is expressed as the square root of week. We thus have the following linear predictor

$$\eta_{ijr} = \theta_r + \beta_1 D_i + \beta_2 \sqrt{j} + \beta_3 D_i \sqrt{j} + u_i,$$

where θ_r is the rth intercept, $r = 1, 2, 3$, $D_i = 1$ if the ith subject was assigned to the drug group and $D_i = 0$ otherwise, β_1 is the difference, on logit scale, between the two groups at baseline, $j = 0, \ldots, 6$ is the week, β_2 is the effect of time on the logit of the placebo patients, and β_3 is the differential effect of time for the drug group relative to the control group. Lastly, u_i is the subject specific random effect that allows a location shift to the latent distribution. Hedeker and Gibbons (2006) assumed that $u_i \sim N(0, \sigma_u^2)$. Here, in addition to the normal density for the random effects, we fit models assuming a NP density for these unobservable quantities and, also, SNP and SMNP densities. For each of these families of densities we choose K according to the AIC criterion described in Sect. 3.4.

Results are presented in Table 4. We first observe that the estimates of the intercepts, $\theta_r, r = 1, 2, 3$, under different model assumptions, are quite different. This is not surprising since, according to Neuhaus et al. (1992), the shape of the mixing density is important in the estimation of the intercepts. This sensitivity is also observable in the studies by Heagerty and Kurland (2001) and Litière et al. (2008). It is possible that it is due to the fact that these parameters serve as locations of the distributions of the random effects, but under each model, the random effects have a different distribution.

Furthermore, the estimate of the effect of the cluster varying covariate, β_1, is more sensitive to the assumption about the random effects distribution than the estimates of the effect of the time varying covariate, β_2, and the effect of the interaction term, β_3, which is usually the parameter of main interest. As far the estimation of β_1 goes, the conjecture of Chen et al. (2002), who reported similar results for a binary response variable, is that it is more affected than the estimation of β_2 because it is a parameter that describes between cluster variation and so do the random effects. On the other hand, parameters that describe within cluster variation, such as β_2, are less affected. Estimation of the interaction term also seems to be unaffected by the assumption about the random effects distribution. These observations, about the sensitivity in the estimation of β_1 and the relative robustness in the estimation of β_2 and β_3, concur with those of Heagerty and Kurland (2001) on results reported when fitting a logit model with normal random effects when, in reality, the random effects have a skewed distribution.

Table 4 makes clear that the subject heterogeneity in the intercepts is not adequately captured by the normal random effects models. The estimated random effects variance is smallest under the normal random effects assumption. This is also evident in Fig. 5. The normal random effects model does not capture the long left tail in the distribution of the random intercepts.

Based on AIC, the SMNP 2 model is preferred, providing evidence that the random effects distribution is non-normal. The value of AIC does not change much among the normal and SNP 2 models, although the estimated densities under

Table 4 Results for the NIMH schizophrenia study—severity of illness. Columns refer to the NP density with 4 mass points, the normal density, the SNP density with $K = 2$ and the SNMP density with $K = 2$

	NP 4 estimate (se)	Normal estimate (se)	SNP 2 estimate (se)	SMNP 2 estimate (se)
θ_1	−5.940 (0.336)	−5.854 (0.343)	−5.629 (0.325)	−5.806 (0.339)
θ_2	−2.938 (0.285)	−2.822 (0.295)	−2.729 (0.272)	−2.818 (0.287)
θ_3	−0.785 (0.258)	−0.706 (0.270)	−0.629 (0.246)	−0.681 (0.262)
β_1	−0.084 (0.288)	0.057 (0.311)	−0.184 (0.264)	−.069 (0.293)
β_2	0.779 (0.120)	0.765 (0.120)	0.768 (0.122)	0.775 (0.120)
β_3	1.210 (0.133)	1.206 (0.133)	1.175 (0.134)	1.207 (0.133)
Var(u_i)	4.078 (0.498)	3.764 (0.495)	4.069 (0.570)	4.117 (0.572)
AIC	1703.45	1708.67	1707.84	1703.13

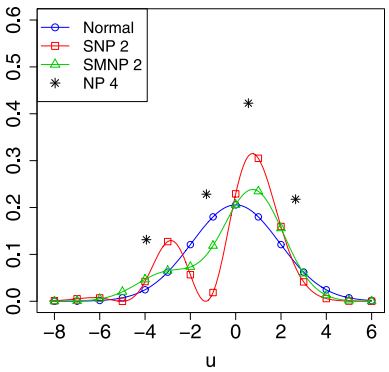


Fig. 5 Results for the NIMH schizophrenia study: estimated random effects densities

these two models are quite different. A possible explanation for this is that the SNP 2 model forces the corresponding probability density function to become zero at a range that seems to have nonzero probability according to all other models. There is, however, a considerable drop in the AIC value when considering the SNMP 2 model. We have also fitted the SNP and SMNP models with $K = 3$ but AIC did not suggest that these models provide a better fit than the corresponding models with $K = 2$ and thus the results of these models are not provided here.

The empirical Bayes (EB) estimates of the random effects are clearly driven by the estimated random effects density. This is evident in Figs. 6(a) and (b) which display the EB estimates under the normal and SMNP 2 models. Figure 6(c) displays the differences between these two sets of estimated random effects. Under the normality assumption, the clusters with large negative EB estimates of the random effects are missed. These clusters have relatively small probability of responding to a smaller category at all measurement times. For instance, under the normal model, the minimum EB estimate of the random effects is $\hat{u}_i^{(0)} = -2.49$. Thus, under this model, at week 6, the end of the clinical trial and usually the time point of main interest, the minimum probabilities of responding at either of the first two categories, for subjects that are on drug ($D_i = 1$) and placebo ($D_i = 0$) are 39.2% and 3.1% respectively. Under the SMNP 2 model, however, 10% of the EB estimates of the random effects are less than $\hat{u}_i^{(0.1)} = -3$ and 2.5% less than $\hat{u}_i^{(0.025)} = -4.2$. With $\hat{u}_i = -3$, the probability of responding at either of the first two categories, at week 6, for subjects that are on drug and placebo are 26.3% and 1.9% respectively. Setting $\hat{u}_i = -4.2$ these probabilities become 9.7% and 0.6%, much smaller than the probabilities obtained under the normal model.

Although in our example big reductions in the estimated probabilities of response in the first two categories happens only for a small fraction of the subjects, it is, from a public health point of view, important to be able to identify groups of clusters that are at high risk or that are on treatment which most likely will not be of any benefit to them.

6 Discussion

We have extended the family of multivariate generalized linear mixed models to include flexible random effects densities. Specifically, we have considered two such families of densities, the semi-nonparametric (SNP) and smooth nonparametric (SMNP) densities. We have proposed an algorithm for fitting these models and we have examined their performance through simulation studies and application to real a dataset.

In a simulation study, reported by Agresti et al. (2004), considerable loss of efficiency was observed from assuming that the random intercepts have a normal distribution when the true distribution is a two point mixture with large variance. We have observed similar results in the simulation studies we have conducted but, also, that the problem of misspecification, in this scenario, can be greatly alleviated by the use of the SNP and SMNP densities.

Our simulation study and studies of other authors (Heagerty and Kurland 2001; Chen et al. 2002), suggest that when the distribution of a random intercept is misspecified, there is more loss of efficiency in the estimation of the between than the within cluster covariate effects. In studies by

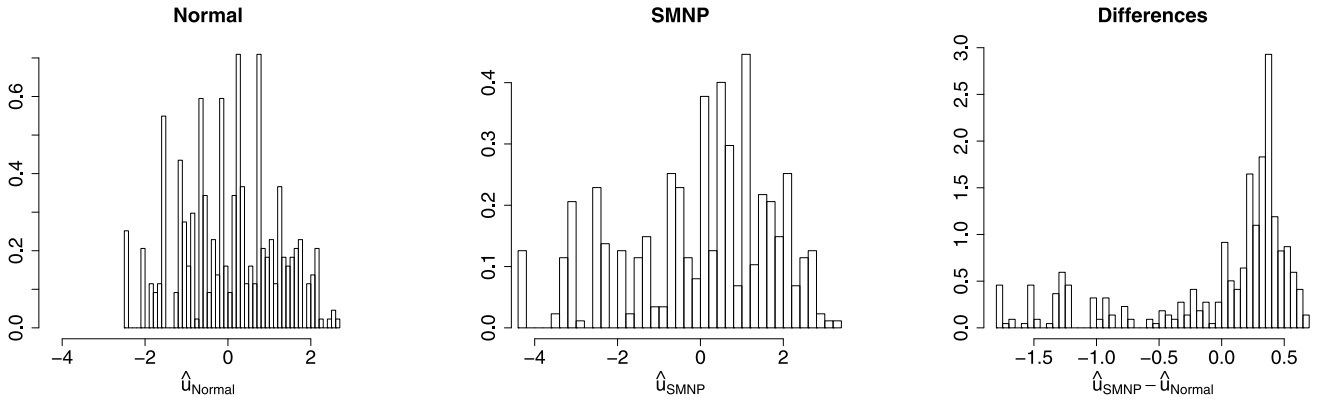


Fig. 6 Results for the NIMH schizophrenia study: histogram of (**a**) the estimated random effects under the assumption of normal random effects, (**b**) the estimated random effects under the SMNP 2 model, and (**c**) the differences between these estimates

Heagerty and Kurland (2001) and Litière et al. (2008) it was further shown that when the distribution of a random within cluster effects is misspecified, estimation of the corresponding effect also suffers loss of efficiency. It is thus important to evaluate the SNP and SMNP densities, in the family of MGLMMs, when there are multivariate random effects. This is one of our plans for future research.

Of course, in addition to the increased efficiency that is gained by avoiding misspecification and the little price to pay in the unlikely event that the parametric assumption about normality of the random effects is satisfied, flexible random effect densities may provide important information about the nature of heterogeneity of a population. Such an example has been provided in Sect. 5.

All our programs were written in C. We have found that it is computationally expensive to fit these models. In particular, many times we could not declare convergence of our automated algorithm due to the MC error, in which case we would expect the algorithm to automatically increase the sample size, but this often times did not happen. We feel, however, that by simply reducing the significance level α (described in Sect. 3.4) the algorithm can speed up.

Acknowledgements Papageorgiou's research was supported by Science Foundation Ireland Research Frontiers grant 07/RFP/MATF448.

References

Agresti, A., Caffo, B., Ohman-Strickland, P.: Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. Comput. Stat. Data Anal. **47**(3), 639–653 (2004)

Aitkin, M.: A general maximum likelihood analysis of variance components in generalized linear models. Biometrics **55**(1), 117–128 (1999)

Booth, J.G., Hobert, J.P.: Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. J. R. Stat. Soc., Ser. B Stat. Methodol. **61**, 265–285 (1999)

Breslow, N.E., Clayton, D.G.: Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc. **88**, 9–25 (1993)

Carroll, R.J., Hall, P.: Optimal rates of convergence for deconvolving a density. J. Am. Stat. Assoc. **83**, 1184–1186 (1988)

Chen, J., Zhang, D., Davidian, M.: A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution. Biostatistics **3**(3), 347–360 (2002)

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

Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc., Ser. B, Methodol. **39**, 1–22 (1977)

Eilers, P.H.C., Marx, B.D.: Flexible smoothing with B-splines and penalties. Stat. Sci. **11**(2), 89–121 (1996)

Fahrmeir, L., Kaufmann, H.: Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. Ann. Stat. **13**, 342–368 (1985)

Fahrmeir, L., Tutz, G.: Multivariate Statistical Modelling Based on Generalized Linear Models. Springer, Berlin (2001)

Follmann, D.A., Lambert, D.: Generalizing logistic regression by nonparametric mixing. J. Am. Stat. Assoc. **84**, 295–300 (1989)

Gallant, AR, Nychka, D.W.: Semi-nonparametric maximum likelihood estimation. Econometrica **55**, 363–390 (1987)

Gallant, R., Tauchen, G.: A nonparametric approach to nonlinear time series analysis: estimation and simulation. In: Brillinger, D., Caines, P., Geweke, J., Parzen, E., Rosenblatt, M., Taqqu, M. (eds.) New Directions in Time Series Analysis, Part II, pp. 71–92. Springer, Berlin (1992)

Geweke, J.: Monte Carlo simulation and numerical integration. In: Amman, H.M., Kendrick, D.A., Rust, J. (eds.) Handbook of Computational Economics, pp. 731–800. Elsevier, Amsterdam (1996)

Ghidey, W., Lesaffre, E., Eilers, P.: Smooth random effects distribution in a linear mixed model. Biometrics **60**(4), 945–953 (2004)

Hartzel, J., Agresti, A., Caffo, B.: Multinomial logit random effects models. Stat. Model. **1**(2), 81–102 (2001)

Heagerty, P.J., Kurland, B.F.: Misspecified maximum likelihood estimates and generalised linear mixed models. Biometrika **88**(4), 973–985 (2001)

Hedeker, D., Gibbons, R.: Longitudinal Data Analysis. Wiley, Palo Alto (2006)

Laird, N.: Nonparametric maximum likelihood estimation of a mixing distribution. J. Am. Stat. Assoc. **73**, 805–811 (1978)

Lesperance, M.L., Kalbfleisch, J.D.: An algorithm for computing the nonparametric MLE of a mixing distribution. J. Am. Stat. Assoc. **87**, 120–126 (1992)

Lindsay, B.G.: The geometry of mixture likelihoods, part II: the exponential family. Ann. Stat. **11**, 783–792 (1983)

Litière, S., Alonso, A., Molenberghs, G.: The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. Stat. Med. **27**(16), 3125–3144 (2008)

Magder, L.S., Zeger, S.L.: A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. J. Am. Stat. Assoc. **91**, 1141–1151 (1996)

McCulloch, C.E.: Maximum likelihood algorithms for generalized linear mixed models. J. Am. Stat. Assoc. **92**, 162–170 (1997)

Monahan, J.F.: An algorithm for generating chi random variables. ACM Trans. Math. Softw. **13**, 168–172 (1987)

Neuhaus, J.M., Hauck, W.W., Kalbfleisch, J.D.: The effects of mixture distribution misspecification when fitting mixed-effects logistic models. Biometrika **79**, 755–762 (1992)

Tutz, G., Hennevogl, W.: Random effects in ordinal regression models. Comput. Stat. Data Anal. **22**, 537–557 (1996)

Verbeke, G., Lesaffre, E.: A linear mixed-effects model with heterogeneity in the random-effects population. J. Am. Stat. Assoc. **91**, 217–221 (1996)

Wei, G.C.G., Tanner, M.A.: A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. J. Am. Stat. Assoc. **85**, 699–704 (1990)

Zhang, D., Davidian, M.: Linear mixed models with flexible distribution of random effects for longitudinal data. Biometrics **57**(3), 795–802 (2001)