STATISTICS IN MEDICINE
Statist. Med. **17**, 2579–2596 (1998)

A SEMI-PARAMETRIC BAYESIAN APPROACH TO GENERALIZED LINEAR MIXED MODELS

KEN P. KLEINMAN[1] AND JOSEPH G. IBRAHIM[2]*

[1]*New England Research Institutes, 9 Galen St., Watertown, MA 02172, U.S.A.*
[2]*Department of Biostatistics, Harvard School of Public Health, 677 Huntington Ave., Boston, Massachusetts 02115, U.S.A.*

SUMMARY

The linear mixed effects model with normal errors is a popular model for the analysis of repeated measures and longitudinal data. The generalized linear model is useful for data that have non-normal errors but where the errors are uncorrelated. A descendant of these two models generates a model for correlated data with non-normal errors, called the generalized linear mixed model (GLMM). Frequentist attempts to fit these models generally rely on approximate results and inference relies on asymptotic assumptions. Recent advances in computing technology have made Bayesian approaches to this class of models computationally feasible. Markov chain Monte Carlo methods can be used to obtain 'exact' inference for these models, as demonstrated by Zeger and Karim.[6] In the linear or generalized linear mixed model, the random effects are typically taken to have a fully parametric distribution, such as the normal distribution. In this paper, we extend the GLMM by allowing the random effects to have a non-parametric prior distribution. We do this using a Dirichlet process prior for the general distribution of the random effects. The approach easily extends to more general population models. We perform computations for the models using the Gibbs sampler. © 1998 John Wiley & Sons, Ltd.

1. INTRODUCTION

Generalized linear models (McCullagh and Nelder,[1] Nelder and Wedderburn[2]) are a unified approach to regression methods. They apply to a wide array of discrete, continuous, and censored outcomes, and are most commonly used when the outcomes are independent. However, in many applications this independence is not a reasonable assumption. This is particularly obvious in longitudinal studies, where multiple measurements made on the same individual are likely correlated.

One recent technique for the analysis of such general correlated data is the generalized estimating equation approach introduced by Liang and Zeger[3] and Zeger and Liang.[4] This approach has the desirable quality that it allows for independence between subjects while introducing a correlation structure within subjects. A drawback to this approach, however, is that it assumes all subjects have the same covariance structure.

* Correspondence to: Joseph G. Ibrahim, Department of Biostatistics, Harvard School of Public Health, 677 Huntington Ave, Boston, Massachusetts 02115, U.S.A. E-mail: ibrahim@jimmy.harvard.edu

Contract/grant sponsor: NIH
Contract/grant number: MH 17119 and CA 70101-01

CCC 0277–6715/98/222579–18$17.50

Received July 1997
Accepted February 1998

For continuous outcomes with normal errors, Laird and Ware[5] present the random effects model. In this model, a subject-specific covariance structure is generated by assuming that each individual has a unique set of regression coefficients, the random effects, distributed around the mean regression coefficients for the population, also known as the fixed effects. There may also be regression coefficients that are equal for all individuals. Conditional on the random effects, repeated observations on a subject are considered independent, while marginalizing over the random effects, a unique covariance structure for the observations within each subject is obtained.

Zeger and Karim[6] present a generalization of the normal random effects model to the class of generalized linear models, generating a generalized linear mixed model (GLMM). They frame the model from the Bayesian perspective and fit it using a Gibbs sampler. They point out that attempts to fit this model using classical (frequentist) techniques are limited by the need for multi-dimensional numerical integrations, except in special cases. As a result of these analytically intractable integrations, classical analysis of GLMMs has relied on approximations to maximum likelihood techniques (Breslow and Clayton[7]).

Bush and MacEachern[8] describe a semi-parametric Bayesian version of the normal random effects model, where the normal assumption on the random effects is relaxed. Kleinman and Ibrahim[9] show a more general covariance structure. In this article, we extend our approach to the class of GLMMs. We present a semi-parametric Bayesian model for generalized linear models with correlated data and random effects, where the random effects have a non-parametric prior distribution. We note that attempts to fit models of this class using classical techniques are hindered by the need for additional multi-dimensional numerical integrations, except in special cases (Aitkin[10]).

The desirability of this approach from a Bayesian perspective is immediately clear, since restricting the model to normally distributed random effects may be contrary to our prior beliefs. In any event, the non-parametric prior is robust to misspecification at this stage of the model. For the normal random effects model, Kleinman and Ibrahim[9] present an example where inference about the regression coefficients is sensitive to the assumption of normality about the random effects. From the classical perspective, it has been shown that large changes in parameter estimates can be caused by small changes in the distribution of the random effects (Heckman and Singer,[11] Davies[12]). Verbeke and Lesaffre[13] show that the normal random effects model can also perform poorly when the random effects have a mixture distribution. Finally, the examples presented herein demonstrate the advantages of the proposed technique.

The non-parametric Bayesian approach for the random effects is to specify a prior distribution on the space of all possible distribution functions. We apply this prior to the general prior distribution for the random effects. We do this with a Dirichlet process prior distribution. Thus, for random effects models, this means that we replace the usual Normal prior on the random effects with a non-parametric prior, followed by a Dirichlet process prior on that general distribution. This approach applies to any parametric model, and is not limited to GLMMs. The foundation of this technology is discussed by Ferguson[14] which also discusses the Dirichlet process and its usefulness as a prior distribution. The practical application of such models, using the Gibbs sampler, has been pioneered by Doss,[15] MacEachern,[16] Escobar,[17] Bush and MacEachern,[8] Liu[18] and Müller *et al.*[19] Other important work in this area has been done by West *et al.*,[20] Escobar and West,[21] MacEachern and Müller,[22] and Neton *et al.*[23] An application of the Dirichlet process prior to the generalized linear model can be found in Mukhopadhyay and Gelfand.[24]

Of these, the most similar to the present article is Mukhopadhyay and Gelfand.[24] Among the more important distinctions between their work and ours are the following. First, they do not discuss prior distributions for the fixed effects. Second, they do not compare their results to the fully parametric case. Their set-up is slightly more general, in the sense that they allow modelling from an overdispersed exponential family. They do not extend their model to correlated data. Finally, they do not focus interest on the random effects themselves.

The rest of this article is organized as follows. In Section 2 we provide a more detailed description of the generalized linear model with random effects. In Section 3 we describe the mixture of Dirichlet process (MDP) structure that we propose for our model. In Section 4 we show how to apply the MDP structure to the generalized linear mixed model. In Section 5 we illustrate our methodology with real data, and, in addition, compare our approach to the fully parametric Bayesian model. In Section 6 we discuss our results and propose directions for future research.

2. GENERALIZED LINEAR MIXED MODELS

First, we define the normal linear random effects model, then introduce random effects into generalized linear models. For individual i, with n_i repeated measurements, the Normal linear random effects model for outcome vector y_i is given by

$$y_i = X_i\beta + Z_i b_i + e_i, \quad i, \ldots, N$$

where y_i is $n_i \times 1$, X_i is an $n_i \times p$ matrix of fixed covariates, β is a $p \times 1$ parameter vector of regression coefficients, commonly referred to as fixed effects in these models, Z_i is an $n_i \times v$ matrix of covariates for the $v \times 1$ vector of random effects b_i, and e_i is an $n_i \times 1$ vector of errors. It is standard in implementations of this model to assume e_i and b_i are independent and that both are distributed Normal, with $e_i \sim \mathrm{N}_{n_i}(0, \sigma^2 I_{n_i})$ and $b_i \sim \mathrm{N}_v(0, D)$, where I_s is the $s \times s$ identity matrix and $\mathrm{N}_s(\mu, \Sigma)$ denotes the s-dimensional multivariate normal distribution with mean μ and covariance matrix Σ. Under these assumptions,

$$[y_i|\beta, b_i] \sim \mathrm{N}_{n_i}(X_i\beta + Z_i b_i, \sigma^2 I_{n_i}). \tag{1}$$

Throughout, we denote the conditional distribution of A given B by $[A|B]$. Notice that marginally

$$[y_i|\beta, \sigma^2, D] \sim \mathrm{N}_{n_i}(X_i\beta, Z_i D Z_i^{\mathrm{T}} + \sigma^2 I_{n_i}) \tag{2}$$

which shows the unique covariance structure for subject i. For the sake of convenience, we call model (1) the normal random effects model and refer to the regression coefficients β as the population-mean effects.

Suppose the sampling distribution of y_{it}, $t = 1, \ldots, n_i$ is from the exponential family, so that

$$p(y_{it}|\theta_{it}, \tau) = \exp\{\tau[y_{it}\theta_{it} - a(\theta_{it})] + c(y_{it}, \tau)\}$$

where

$$\mu_{it} = E(y_{it}|\theta_{it}, \tau) = \frac{\mathrm{d}a(\theta_{it})}{\mathrm{d}\theta_{it}}$$

and

$$v_{it} = \text{var}(y_{it}|\theta_{it}, \tau) = \tau^{-1} \frac{\mathrm{d}^2 a(\theta_{it})}{\mathrm{d}\theta_{it}^2}$$

where τ is a scalar dispersion parameter.

In the generalized linear mixed model, the canonical parameter θ_{it} is related to the covariates by

$$h(\theta_{it}) = \eta_{it} = x_{it}^{\mathrm{T}}\beta + z_{it}^{\mathrm{T}}b_i$$

where x_{it} and z_{it} are rows of the X_i and Z_i matrices, $h(\cdot)$ is a monotonic differentiable function, often referred to as the θ-link, and η_{it} is called the linear predictor. Throughout, we write

$$p(y_{it}|\theta_{it}, \tau) \equiv p(y_{it}|\beta, b_i, \tau)$$

where

$$p(y_{it}|\beta, b_i, \tau) = \exp\{\tau[y_{it}h^{-1}(\eta_{it}) - a(h^{-1}(\eta_{it}))] + c(y_{it}, \tau)\}. \tag{3}$$

When $h(\theta_{it}) = \theta_{it} = \eta_{it}$, then we say the link is the canonical link.

For example, in GLMM logistic regression, we have

$$p(y_{it}|\beta, b_i, \tau) = \exp\{y_{it}(x_{it}^{\mathrm{T}}\beta + z_{it}^{\mathrm{T}}b_i) - \log(1 + \mathrm{e}^{x_{it}^{\mathrm{T}}\beta + z_{it}^{\mathrm{T}}b_i})\}$$

where $\theta_{it} = \eta_{it} = x_{it}^{\mathrm{T}}\beta + z_{it}^{\mathrm{T}}b_i$ and $\tau \equiv 1$.

Note that the GLMM imitates the normal random effects model in that we assume that, conditional on the random effect b_i, the repeated observations on subject i are independent. Thus the likelihood for N subjects in the GLMM is

$$p(y|\beta, b, \tau) \propto \prod_{i=1}^{N} \prod_{t=1}^{n_i} p(y_{it}|\beta, b_i, \tau) \tag{4}$$

where $b = (b_1, \ldots, b_N)^{\mathrm{T}}$ and $y = (y_{11}, \ldots, y_{Nn_N})^{\mathrm{T}}$.

There are several attractive properties of (4). First, it takes within-subject correlation into account while allowing each individual to have a unique correlation structure and maintaining independence between subjects. Second, the model accommodates unbalanced data, in that response vectors need not be of the same length. Similarly, we can fit irregularly timed measurements with this model without any adjustment. Finally, posterior distributions or estimates of the random effects have interpretive value when the trend of the mean function of individuals is of interest.

Zeger and Karim[6] assume

$$\beta \sim \mathrm{N}_p(\mu_0, \Sigma_0)$$

where (μ_0, Σ_0) are considered known and fixed and there are p elements in β. Typically, one assumes that

$$b_i \sim N_v(0, D)$$

where there are v random effects. In the Bayesian framework, D^{-1} is commonly assumed to have a Wishart prior. In this article, we relax the normal assumption for the b_i, and allow

$$b_i \sim G$$

where G is a general distribution. However, before we accomplish that goal, we introduce the machinery that makes it possible, the Dirichlet process prior.

3. MIXTURE OF DIRICHLET PROCESS MODELS

The mixture of Dirichlet process model arises in cases of the following general situation.

Suppose an $n_i \times 1$ random vector x_i has a parametric distribution indexed by the $w \times 1$ vector θ_i, $i = 1, \ldots, N$. Then suppose the θ_i themselves have a prior distribution with known hyper-parameters Ψ_0. Thus

$$\begin{aligned} \text{Stage 1: } & [x_i|\theta_i] \sim D_{n_i}(h_1(\theta_i)) \\ \text{Stage 2: } & [\theta_i|\Psi_0] \sim D_w(h_2(\Psi_0)) \end{aligned} \tag{5}$$

where $D_s(\cdot)$ is a generic label for an s-dimensional parametric multivariate distribution and $h_1(\cdot)$ and $h_2(\cdot)$ are functions. The MDP model (Escobar,[17] MacEachern[16]) removes the assumption of a parametric prior at the second stage, and replaces it with a general distribution G. The distribution G then in turn has a Dirichlet process prior (Ferguson[14]), leading to

$$\begin{aligned} \text{Stage 1: } & [x_i|\theta_i] \sim D_{n_i}(h_1(\theta_i)) \\ \text{Stage 2: } & \theta_i|G \overset{\text{i.i.d.}}{\sim} G \\ \text{Stage 3: } & [G|M, \Psi_0] \sim \text{DP}(MG_0(h_2(\Psi_0))) \end{aligned} \tag{6}$$

where G_0 is a w-dimensional parametric distribution and M is a positive scalar. The parameters of a Dirichlet process are $G_0(\cdot)$, a probability measure, and M, a positive scalar. The parameter $MG_0(\cdot)$, often called the base measure, contains a distribution, $G_0(\cdot)$, which approximates the true non-parametric shape of G, and the scalar M, which reflects our prior belief about how similar the non-parametric distribution G is to the base measure $G_0(\cdot)$.

There are two special cases in which the MDP model leads to the fully parametric case. As $M \to \infty$, $G \to G_0(\cdot)$, so that the base measure is the prior distribution for θ_i. Also, if $\theta_i \equiv \theta$ for all i, the same is true. For a more hierarchical modelling approach, it is possible to place prior distributions on (M, Ψ_0). In Section 4 we place a prior on Ψ_0, but we do not do so for M. The specification in (6) results in a semi-parametric specification in that a fully parametric distribution is given in Stage 1 and a non-parametric distribution is given in Stages 2 and 3.

The Polya urn representation of the Dirichlet process was developed by Blackwell and MacQueen[25] and is useful for sampling purposes. We describe it as follows. The draw of θ_1 is always from the base measure. The draw of θ_2 is equal to θ_1 with probability p_1 and is from the base measure with probability $p_0 = 1 - p_1$. The draw of θ_3 is equal to θ_1 with probability p_1, equal to θ_2 with probability p_2, and is a draw from the base measure with probability $p_0 = 1 - (p_1 + p_2)$. The values of the ps change with each new draw. This process continues until θ_N is equal to each of the preceding θs with probability p_i, $i \in 1, \ldots, N-1$ and is a draw from the base measure with probability $p_0 = 1 - \sum_{i=1}^{N-1} p_i$. We determine the values of p_i, $i = 0, \ldots, N-1$ from the Dirichlet process parameters. In other words, the θs are actually drawn from a mixture distribution where the mixing probabilities are determined by the Dirichlet process of Stage 3, thus giving rise to the MDP label. From this representation, it is clear that if all of the $\theta_i \equiv \theta$ for all i, then we draw θ from the base measure with probability 1 and thus the base measure is the prior.

The MDP model is simplified in practice by the Polya urn representation, using the fact that marginally, the θ_i are distributed as the base measure along with the added property that $P(\theta_i = \theta_j, i \neq j) > 0$. The Dirichlet process prior results in what MacEachern[16] calls a 'cluster structure' among the θ_is. This cluster structure partitions the N θ_is into k sets or clusters,

$0 < k \leqslant N$. All of the observations in a cluster share an identical value of θ and subjects in different clusters have differing values of θ.

As described by Escobar,[17] conditional on the other θs, θ_i has the following mixture distribution:

$$p(\theta_i | x, \theta_{-i}) \propto \sum_{j \neq i} q_j \delta_{\theta_j} + M q_0 g_0(\theta_i) p(x_i | \theta_i) \tag{7}$$

where $x = (x_1, \ldots, x_N)$, $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_N)$ and $p(x_i|\theta_i)$ is the sampling distribution of the x_is. We normalize the values q_j and Mq_0 to obtain the selection probabilities p_i, $i = 0, \ldots, N-1$ in the Polya urn scheme described above. In addition δ_s is a degenerate distribution with point mass at s, and $g_0(\cdot)$ is the density corresponding to the probability measure $G_0(\cdot)$. Finally, $q_j = p(x_i|\theta_j)$, $j = 1, \ldots, i-1, i+1, \ldots, N$ and $q_0 = \int p(x_i|\theta) g_0(\theta)\, \mathrm{d}\theta$.

To demonstrate the MDP model, consider the seminal example of Escobar[17] and Escobar and West.[21] Suppose that x_i has the univariate Normal distribution with unknown mean θ_i and known variance σ_x^2. In this case we have $n_i = 1$, $i = 1, \ldots, N$. Also assume that each θ_i has the univariate Normal distribution. Then (5) becomes

$$\text{Stage 1: } [x_i | \theta_i, \sigma_x] \sim \mathrm{N}(\theta_i, \sigma_x^2)$$

$$\text{Stage 2: } [\theta_i | \mu, \sigma_\theta] \sim \mathrm{N}(\mu, \sigma_\theta^2).$$

The MDP model removes the assumption of normality at the second stage, resulting in

$$\text{Stage 1: } [x_i | \theta_i, \sigma_x] \sim \mathrm{N}(\theta_i, \sigma_x^2)$$

$$\text{Stage 2: } \theta_i | G \overset{\text{i.i.d.}}{\sim} G$$

$$\text{Stage 3: } [G | M, \Psi_0] \sim \mathrm{DP}(M G_0(h_2(\Psi_0))). \tag{8}$$

3.1. Conjugate MDP models

Suppose $G_0 = \mathrm{N}(\mu, \sigma_\theta^2)$ in (8) so that $\Psi_0 = (\mu, \sigma_\theta^2)$. In this case, the unnormalized selection probability q_j is equal to $p(x_i|\theta_j) = \phi(x_i|\theta_j, \sigma_x^2)$, where $\phi(\cdot|\mu, \sigma^2)$ denotes the Normal density with mean μ and variance σ^2. With probability proportional to q_j, $\theta_i \sim \delta_{\theta_j}$, which means that $\theta_i = \theta_j$ with probability 1. The unnormalized selection probability q_0 is given by

$$q_0 = \int p(x_i|\theta, \sigma_x^2) g_0(\theta|\Psi_0)\, \mathrm{d}\theta = \int \phi(x_i|\theta, \sigma_x^2) \phi(\theta|\mu, \sigma_\theta^2)\, \mathrm{d}\theta.$$

With probability proportional to $M \times q_0$

$$[\theta_i | x_i] \sim g_0(\theta) p(x_i|\theta) = \mathrm{N}(\theta|\mu, \sigma_\theta^2) \mathrm{N}(x_i|\theta, \sigma_x^2)$$

where $\mathrm{N}(s|\mu, \sigma^2)$ indicates that s has a normal distribution with mean μ and variance σ^2. Then

$$[\theta_i | x_i] \sim \mathrm{N}\left([(\sigma_\theta^2 + \sigma_x^2)^{-1} \sigma_\theta^2 \sigma_x^2] \left(\frac{\mu}{\sigma_\theta^2} + \frac{x_i}{\sigma_x^2} \right),\ (\sigma_\theta^2 + \sigma_x^2)^{-1} \sigma_\theta^2 \sigma_x^2 \right).$$

In the example above, selecting G_0 to be Normal when the sampling distribution of the data is Normal emulates the conjugate relationship between sampling distribution and prior in the usual Bayesian hierarchy. In the MDP case, the sampling distribution is conjugate to the base measure. MacEachern[16] calls MDP models with base measures and sampling distributions that are

conjugate in this fashion 'conjugate MDP models'. The computational advantages of the conjugate MDP model are clear from the example. First, q_0 has a closed form. Secondly, the distribution of θ_i corresponding to q_0 is from the same exponential family as the base measure. As a result, Gibbs sampling in the conjugate model described above can proceed in a relatively straightforward fashion, as described in detail in Kleinman and Ibrahim.[9]

3.2. Non-conjugate MDP models

When we do not assume conjugacy, the integral needed for q_0 typically has no closed-form solution. Since we must evaluate this integral N times within each Gibbs cycle, the cost in time of numerical integration is compounded, as is the cost in accuracy of approximations. Several attempts to avoid this integration have been made. For example, West *et al.*[20] approximate q_0 with $p(y_{it}|b')$ where $b' \sim G_0(\cdot)$. This is certainly simple, but unfortunately the stationary distribution underlying the Gibbs sampler is no longer the posterior distribution we desire. In fact, the stationary distribution may be quite different from the posterior. We employ a technique described by MacEachern and Müller[22] whereby one can fit non-conjugate MDP models without numerical integration or approximation. Another technique has been suggested by Walker and Damien.[26]

Some additional notation is necessary for the exposition of this method. Recall that when the θ_is are known, the observations are grouped into clusters which have equal θ_is. There will be some number k, $0 < k \leqslant N$ of unique values among the θ_is. Denote these unique values by γ_l, $l = 1, \ldots, k$ and recall from the Polya urn scheme that the γ_l are independent observations from $G(\cdot)$. Let n_l be the number of observations that share the value γ_l. Additionally, let l represent the set of subjects with common random effect γ_l. Note that knowing the θ_is is equivalent to knowing k, γ_l, n_l and the cluster memberships l, $l = 1, \ldots, k$.

The routine of MacEachern and Müller[22] is closely intertwined with the Gibbs sampler it generates. Thus the following discussions is in terms of the Gibbs sampler, rather than general model terms. The method relies on the augmentation of the k independent γ_ls with an additional $N - k$ independent samples from $G_0(\cdot)$ at the start of each loop of the Gibbs sampler. Label these additional draws $\gamma_{k+1}, \ldots, \gamma_N$.

Then the routine proceeds in the following fashion. If $n_l > 1$, $i \in l$, meaning that at least one other subject has the same value of θ as subject i, then θ_i has the distribution

$$p(\theta_i | x, \theta_{-i}) \propto \sum_{l=1}^{k} \eta_l^- q_l \delta_{\gamma l} + \frac{M}{k^* + 1} q_{k+1} \delta_{\gamma_{k+1}} \tag{9}$$

where $k^* = k$ and η_l^- is the number of observations sharing γ_l when we exclude observation i. Note that this means $n_l^- = n_l$, except when $i \in l$, in which case $n_l^- = n_l - 1$. Also, $q_l = p(x_i|\gamma_l)$, $l = 1, \ldots, k+1$. In other words, with probability proportional to $n_l^- p(x_i|\gamma_l)$, θ_i is equal to γ_l with probability 1, $l = 1, \ldots, k$. With probability proportional to $[M/(k^* + 1)]\, p(x_i|\gamma_{k+1})$, θ_i is distributed $\delta_{\gamma_{k+1}}$, meaning that $\theta_i = \gamma_{k+1}$ with probability 1. If $n_l = 1$, $i \in l$, then only subject i has the value θ_i. In this case, we do the following. With probability $k^*/(k^* + 1)$, leave θ_i unchanged. Otherwise, with probability $1/(k^* + 1)$, θ_i is distributed according to equation (9), with the modification that $k^* = k - 1$.

If it should occur that this routine causes a cluster to disappear, meaning that $n_{l'} = 0$ for some $l' \leqslant k$, switch the cluster labels of l' and k. Notice that k decreases as a result of this process. Another important point in the above is that the value $\gamma_{l'}$ is not removed, but becomes γ_{k+1} in the

distribution of θ_{i+1}. Once we have completed iteration of the Gibbs sampler, we discard the augmentary values $\gamma_{k+1}, \ldots, \gamma_N$.

4. DP PRIORS IN THE GENERALIZED LINEAR MIXED MODEL

In this section we describe how one can apply the MDP model to the generalized linear mixed model. Assume that the base measure for the b_is is Normal. Then any exponential family sampling distribution completes an MDP GLMM. This is a non-conjugate MDP model except when the data have a normal sampling distribution. Denote the distribution of the outcome y_{it} for subject i at time t as $p(y_{it}|\beta, b_i, \tau)$ as given in (3). The prior specifications for the parameters of the MDP GLMM are

$$\begin{aligned} \tau &\sim \text{Gamma}(\alpha_0, \lambda_0) \\ \beta &\sim \text{N}_p(\mu_0, \Sigma_0) \\ b_i &\overset{\text{i.i.d.}}{\sim} G \\ G &\sim \text{DP}(M\text{N}_v(0, D)). \end{aligned} \tag{10}$$

The model (10) implies that there are p population-mean effects and v random effects.

When G is a fully parametric prior, we can write down the joint posterior. Suppose G is a v-dimensional normal distribution with mean 0 and covariance matrix D, as in the standard (fully parametric) GLMM. Given D, the joint posterior for the parameters is

$$p(\beta, \tau, b|y) \propto p(y|\beta, b, \tau)\pi(\beta, b, \tau) \propto$$

$$\exp\left\{\sum_{i=1}^{N}\sum_{t=1}^{n_i} \log p(y_{it}|\beta, b_i, \tau) - \frac{1}{2}(\beta-\mu_0)^{\text{T}}\Sigma_0^{-1}(\beta-\mu_0) - \tau\lambda_0 - \frac{1}{2}\sum_{i=1}^{N} b_i^{\text{T}} D^{-1} b_i\right\}\tau^{\alpha_0-1} \tag{11}$$

where $\pi(\cdot)$ denotes the joint prior density and $b = (b_1, \ldots, b_n)$. However, if G has the form of (10), it is impossible to write down the joint posterior density of the parameters, because there is not a common dominating measure. We include here the special case where $G = \text{N}_v(0, D)$ because we can find the conditional distributions of β and τ through it in the usual way.

From equation (11) and the discussion in Section 2, we can obtain the full conditional distributions needed for Gibbs sampling.

Following usual algebraic routes, we get

$$p(\beta|b, \tau, y) \propto \exp\left(\sum_{i=1}^{N}\sum_{t=1}^{n_i} \log p(y_{it}|\beta, b_i, \tau) - \frac{1}{2}(\beta-\mu_0)^{\text{T}}\Sigma_0^{-1}(\beta-\mu_0)\right).$$

Unless y_{it} has the normal distribution, sampling from this full conditional is not straightforward, but it can still be accomplished, using for example a Metropolis step (see Metropolis *et al.*[27] and Hastings[28]). The full conditional distribution of τ is

$$p(\tau|\beta, b, y) \propto \tau^{\alpha_0-1}\exp\left\{\sum_{i=1}^{N}\sum_{t=1}^{n_i} c(y_{it}, \tau)\right\}\exp\left\{\sum_{i=1}^{N}\sum_{t=1}^{n_i}[y_{it}\theta_{it} - a(\theta_{it})] - \tau\lambda_0\right\}.$$

Sampling from the full conditional distribution of τ can also be accomplished through a Metropolis step, unless $\log(c(y_{it}, \tau))$ takes a form proportional to $\tau^{f(y_{it})}$. In such a case

$$[\tau | \beta, b, y] \sim \text{Gamma}\left(\alpha_0 + \sum_{i=1}^{N}\sum_{t=1}^{n_i} f(y_{it}), \lambda_0 - \sum_{i=1}^{N}\sum_{t=1}^{n_i}[y_{it}\theta_{it} - a(\theta_{ij})]\right).$$

From the discussion of MDP models in Section 2, we find

$$p(b_i | \beta, \tau, y, b_{-i}) \propto \sum_{j \neq i}^{N} \exp\left\{\sum_{t=1}^{n_i} \log p(y_{it} | \beta, b_j, \tau)\right\} \delta_{b_j} + \left[M \int \exp\left\{\sum_{t=1}^{n_i} \log p(y_{it} | \beta, b_i, \tau)\right\} \times \phi(b_i | 0, D)\, \mathrm{d}b_i\right] \phi(b_i | 0, D) \prod_{t=1}^{n_i} p(y_{it} | \beta, b_i, \tau) \tag{12}$$

where b_{-i} denotes the random effects for the subjects excluding subject i. Also, as in Section 2, δ_s is a degenerate distribution with point mass at s. An important subtlety in equation (12) is that the terms $p(y_{it} | \beta, b_j, \tau)$ in the first summation use the data for the subject i and the random effects for each of the other subjects. That is, we evaluate the likelihood for subject i using the other subjects' random effects. The better the fit of subject j's random effect, that is, the greater the likelihood, then the more likely it is that δ_j is the distribution from which b_i is drawn.

When the sampling distribution is not normal, q_0 generally will not have a closed form. To avoid numerical integrations or approximation, we use the algorithm of MacEachern and Müller,[22] described in Section 3.2. Recall that there are $k \leqslant N$ unique random effects among the N subjects. These random effects, which we label γ_l, $l = 1, \ldots, k$ are independent draws from $G_0(\cdot)$ which in this case is $\mathrm{N}_v(0, D)$. The algorithm requires that we sample an additional $N - k$ values from $\mathrm{N}_v(0, D)$; we label these values $\gamma_{k+1}, \ldots, \gamma_N$. Then, if $n_i > 1$, $i \in l$ we sample from the following full conditional:

$$p(b_i | y, \gamma, \beta, \tau) = \sum_{l=1}^{k} n_l^- \sum_{t=1}^{n_i} p(y_{it} | \beta, \tau, \gamma_l)\delta_{\gamma_l} + \frac{M}{k^* + 1} \sum_{t=1}^{n_i} p(y_{it} | \beta, \tau, \gamma_{k+1})\delta_{\gamma_{k+1}} \tag{13}$$

where $k^* = k$ and n_l^- is the number of subjects who share the random effect γ_l excluding subject i. The value of $\prod_{t=1}^{n_i} p(y_{it} | \beta, \tau, \gamma_l)$ is the likelihood of subject i's data using the random effect that belongs to some group of subjects. Thus the effect of the distribution (13) has a sensible interpretation. The greater subject i's likelihood with random effect γ_l and the greater the number of other subjects who share that random effect, the more likely it is that we will select γ_l as subject i's random effect. On the other hand, the scalar parameter M regulates the probability that subject i gets a new random effect, meaning that they start a new cluster. If $n_l = 1$, $i \in l$, then with probability $k^*/(k^* + 1)$ we leave b_i unchanged. Otherwise, we let b_i be distributed according to equation (13) except with $k^* = k - 1$.

We may end up with one fewer cluster after drawing b_i from (13), though this can only happen when $n_l = 1$, $i \in l$. In other words, we may have the case that $n_l = 0$ after drawing from the conditional distribution of b_i. If this occurs we switch the values γ_l for the empty group and γ_k for the last group. We also switch the set memberships l and k. Before drawing from the full conditional in such a case, the number of clusters is k, and afterwards there are $k - 1$ clusters. Before the draw, there were n_k subjects in set k who shared the random effect γ_k. After the draw and the switching, the set in which these subjects are simply has a different label l, $l < k$, and we label the random effect formerly held by subject i as γ_k. In the full conditional of b_{i+1}, there are

only $k-1$ clusters, thus we use the value γ_k from the full conditional of b_i as γ_{k+1} in drawing b_{i+1} from (13).

Typically, the covariance matrix D in the base measure of the Dirichlet process in model (10) is unknown, and therefore we must specify a suitable for it. Note that once we have done this, the base measure is no longer marginally Normal. For convenience, suppose

$$D^{-1} \sim \text{Wishart}(d_0, c_0 R_0)$$

where $d_0 \geqslant v$, $c_0 > 0$ and R_0 is a $v \times v$ positive definite matrix. Then *a priori*

$$p(D^{-1} | d_0, c_0, R_0) \propto |D^{-1}|^{\frac{d_0 - v - 1}{2}} \exp\left\{ -\frac{1}{2} \text{tr}((c_0 R_0)^{-1} D^{-1}) \right\}$$

where $\text{tr}(\cdot)$ denotes trace.

The γ_l are k independent observations from $\text{N}_v(0, D)$. Thus

$$p(D^{-1} | b, \beta, y, \tau) \equiv p(D^{-1} | \gamma, l, \beta, y, \tau)$$

$$\propto |D^{-1}|^{\frac{d_0 + k - v - 1}{2}} \exp\left\{ -\frac{1}{2} \text{tr}\left((c_0 R_0)^{-1} D^{-1} - \frac{1}{2} \sum_{l=1}^{k} \gamma_l^{\text{T}} D^{-1} \gamma_l \right) \right\}$$

so that

$$[D^{-1} | b, \beta, y, \tau] \sim \text{Wishart}\left(d_0 + k, \left((c_0 R_0)^{-1} + \sum_{l=1}^{k} \gamma_l \gamma_l^{\text{T}} \right)^{-1} \right). \tag{14}$$

Bush and MacEachern[8] recommend one additional step for the model as an aid to convergence for the Gibbs sampler. To speed mixing over the entire parameter space, they suggest moving around the γs after determining how the b_is are grouped. The conditional density of γ_l is

$$p(\gamma_l | \beta, \tau, b, D, y) \propto \phi(\gamma_l | 0, D) \left\{ \prod_{i \in l} \prod_{t=1}^{n_i} p(y_{it} | \beta, \gamma_l, \tau) \right\}.$$

One must apply a Metropolis sampler or some other technique to draw a sample from this distribution as well.

5. APPLICATIONS

Here we present examples of analyses of the two most common types of GLMMs, the logistic and Poisson models. These particular models are simpler than the general framework described above, in that $\tau \equiv 1$. The purpose of the analyses is to make comparisons between the MDP and fully parametric analyses of the same data. We also demonstrate the computations and inference for these models.

5.1. Correlated Binary Data

This section presents an analysis of longitudinal repeated binary measurements. Zeger and Karim[6] analyze a subset of data from a study of respiratory infections in Indonesian children. An analysis of the full data set appears in Sommer *et al.*[29] The children, all pre-schoolers, were seen quarterly for up to six quarters. At each examination, the presence or absence of respiratory infection was noted and is the outcome in this analysis. The covariates modelled by Zeger and

Karim are an intercept, age in months, presence/absence of xeropthalmia, cosine and sine terms for the annual cycle, height for age as a percentage of the National Center for Health Statistics standard, and presence/absence of stunting, defined as being below the 85th percentile in height for age. Age in months was centred at 36 and height for age was centred at 90 per cent. Xeropthalmia is a symptom of chronic vitamin A deficiency and height for age is an indicator of long-term nutritional status. In addition, Zeger and Karim model a random intercept for each child. To facilitate comparison with the results of Zeger and Karim, we use the same model that they use. The original study followed 3000 children. Zeger and Karim use a subset of 250 of these children. To speed computations we take a random subset of 50 from the set of Zeger and Karim.

Since no prior information regarding respiratory infections in this population was available, we chose parameters for the Wishart prior on D^{-1} in the following fashion. First, we chose d_0 to be 10. Though the prior is proper if d_0 is smaller, Gibbs samplers for random effects models frequently fail to converge when d_0 is too small (see Cowles *et al.*[30] for an example). Based on experience with fully parametric logistic GLMM, we wanted $d_0^{-1}R_0$, the prior expected value of D^{-1} to be 0·5. Thus we chose $c_0 = 1$ and $R_0 = 5$. Note that since D is scalar, this is equivalent to a Gamma prior on D^{-1}.

We chose relatively flat priors for the other parameters. We let $\mu_0 = (0\ 0\ 0\ 0\ 0\ 0\ 0\ 0)^{\mathrm{T}}$ and $\Sigma_0 = 10{,}000 I_8$. This is equivalent to saying that the *a priori* probability of respiratory infection is 0·5 for all subjects across all seasons, but that great uncertainty exists as to the accuracy of this assumption. Without previous experience in this population, this vague prior seems appropriate.

We chose three values of the parameter M to reflect large, moderate, and small departures from normality for the distribution of the random effects. A value of $M = 0{\cdot}75$ reflects a large departure from normality with the average number of clusters $\bar{k} \approx 5$. A value of $M = 200$ reflects a moderate departure from normality with $\bar{k} \approx 20$. A value of $M = 10^8$ suggests that the distribution of the random effects is very nearly normal with $\bar{k} \approx 45$. Finally, we also modelled the fully parametric case by choosing M large enough that $\bar{k} = 50$.

In Table I, we present results from the MDP GLMM analysis along with the fully parametric GLMM results. We ran Gibbs samplers for 25,000 iterations with the first 3000 discarded as a burn-in. In addition, due to high autocorrelation, we used only every tenth iterate, with the remainder discarded. This makes for a total sample size of 2200. We assessed convergence by the methods of Geweke[31] and Raftery and Lewis[32] using the CODA (Best *et al.*[33]) suite of diagnostics in S-plus. Most of the parameters in each of the four samplers had Geweke statistics within $\pm 1{\cdot}96$, indicating that convergence is plausible. The more appropriate convergence diagnostic in this case may be that of Raftery and Lewis, which evaluates the accuracy of estimates of percentiles of posterior distributions. Under the conditions of burn-in and thinning described above, the 2·5 percentile of the posterior is within 0·01 of the observed 2·5 percentile with probability 0·8.

We interpret the results in Table I as follows. In general, many of the medians and 95 per cent highest posterior density (HPD) regions for the population-mean effects are strikingly similar. The exceptions to this rule are the effects for gender, xeropthalmia, stunting, and the intercept. The posterior median of the intercept is similar in all four models, but the 2·5 percentile is progressively more negative as the distribution of the random effects becomes less normal. The same may be said for the effects of xeropthalmia. There is no discernible pattern for the effects of gender or stunting. In no case does the 95 per cent HPD region change from excluding 0 to

Table I. Posterior 2·5, 50 and 97·5 percentiles for various parameters from MDP logistic GLMM and the fully parametric Poisson GLMM. β_0 is the intercept, β_1 is the effect of age, β_2 is the effect of xeropthalmia, β_3 is the gender effect, β_4 is the effect of the seasonal cosine, β_5 is the effect of the seasonal sine, β_6 is the effect of height for age, and β_7 is the effect of stunting. b_i is the intercept for subject i. For the MDP models, $D(0, 0)$ is the variance of the random effects in the base measure. For the fully parametric model, D is the variance of the random effects. $\bar{k}$ is the average number of clusters observed in the course of sampling

Parameter	MDP GLMM model			FP GLMM
	$M = 0{\cdot}75$	$M = 200$	$M = 10^8$	$M \approx \infty$
β_0	(−7·03, −3·76, −1·97)	(−5·06, −3·47, −2·26)	(−4·56, −3·27, −2·25)	(−4·56, −3·29, −2·24)
β_1	(−0·088, −0·047, −0·014)	(−0·094, −0·048, −0·011)	(−0·092, −0·048, −0·013)	(−0·091, −0·047, −0·01)
β_2	(−1·08, 1·15, 3·12)	(−0·84, 1·44, 3·43)	(−0·65, 1·50, 3·48)	(−0·54, 1·32, 3·30)
β_3	(−1·58, −0·26, 0·86)	(−1·94, −0·52, 0·77)	(−1·69, −0·31, 0·93)	(−1·65, −0·34, 0·91)
β_4	(−2·18, −1·08, −0·11)	(−2·23, −1·14, −0·191)	(−2·14, −1·08, −0·14)	(−2·17, −1·09, −0·13)
β_5	(−2·01, −0·97, −0·05)	(−2·16, −0·99, −0·001)	(−2·07, −0·94, 0·01)	(−2·08, −0·97, −0·08)
β_6	(−0·36, −0·15, 0·03)	(−0·35, −0·14, 0·04)	(−0·36, −0·16, 0·01)	(−0·36, −0·16, 0·02)
β_7	(−2·43, −0·19, 2·1)	(−2·20, 0·05, 2·20)	(−2·49, −0·12, 1·92)	(−2·24, −0·04, 1·86)
D	(0·36, 0·79, 2·22)	(0·50, 1·26, 3·41)	(0·45, 0·98, 2·34)	(0·48, 0·99, 2·32)
b_1	(−1·02, 1·22, 4·58)	(−0·87, 1·09, 3·45)	(−2·16, 0·31, 3·34)	(−2·06, 0·26, 2·46)
b_2	(−2·32, 0·46, 4·24)	(−3·16, −0·01, 2·66)	(−3·06, 0·02, 2·31)	(−2·36, −0·05, 2·28)
b_3	(−0·78, 1·38, 4·56)	(−0·73, 1·26, 3·53)	(−1·78, 0·52, 2·91)	(−1·73, 0·57, 3·05)
$\bar{k}$	4·47	19·0	43·2	50

including it across the models, with the exception of the seasonal cosine, where the 95 per cent HPD region hovers around 0 in all the models. The posterior distributions are most affected by the changing model if the covariates are binary. These parameters are the ones with the highest correlation with the intercept, the continuous covariates being roughly centred at 0. Since it is the distribution of the random intercepts that is directly affected by the changing models, it is not surprising that the posterior distributions of parameters highly correlated with the population-mean intercept are the most affected.

The largest effect of introducing the non-parametric piece of the model is on the intercept, where it indicates that an even more extreme value may be the population mean. In this model, this means that the probability of respiratory infection when all the other covariates are 0 may plausibly be smaller under the MDP model than under the fully parametric GLMM. This is also demonstrated in the kernel estimates of the posterior distributions presented in Figure 1. In comparing the distributions in Figure 1, we see that the marginal posterior distribution of β_0 based on the semi-parametric model (Figure 1(b)) has much heavier tails than the one based on the fully parametric model (Figure 1(a)). Also, we see that the distribution in Figure 1(a) appears symmetric, while the distribution in Figure 1(b) is quite asymmetric. However, the medians and modes of the two distributions are similar.

The introduction of the MDP stage in the model introduces larger changes to the posterior distributions of the random effects. For the three sample random effects tabulated, the medians differ markedly under each of the models. If one were predicting the probability that a particular child would have a respiratory infection, the MDP models would give results very unlike those from the fully parametric model. Finally, the posterior distribution of D seems different under the MDP models, but the role of this parameter is not the same under the MDP and fully parametric models and therefore no straightforward comparison can be made.

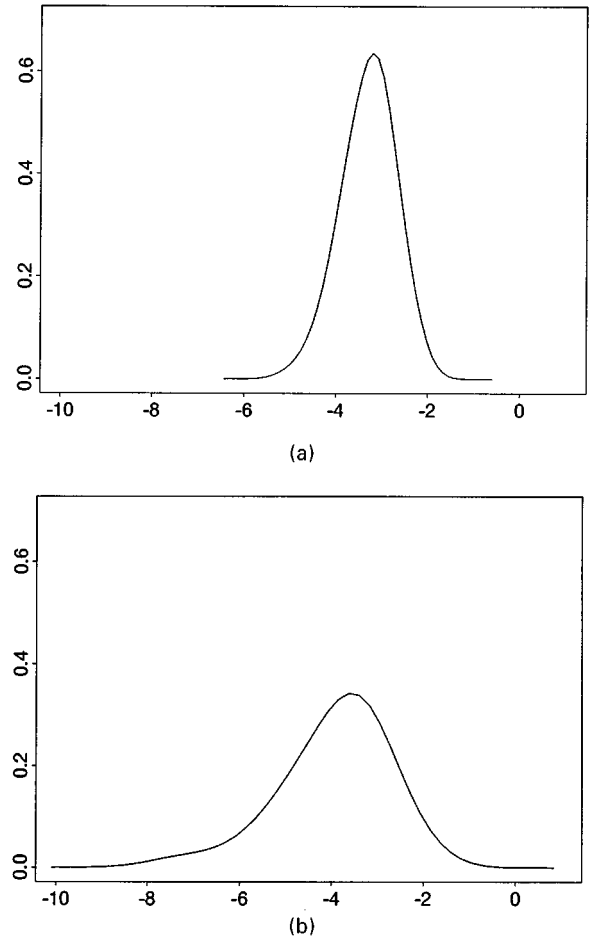

Figure 1. Posterior distribution of the intercept: (*a*) fully parametric model; (*b*) MDP model with $M = 0{\cdot}75$

5.2. Correlated Count Data

This section presents an analysis of longitudinal repeated count measurements. Thall and Vail[34] (their Table 2) present data from a study of seizures in epileptic patients. They use a classical analysis to fit a random intercept to each individual's data and a random effect for each visit across patients. These data were also analysed by Breslow and Clayton[7] using an approximate classical method called penalized quasi-likelihood (PQL). We will fit a model described by Breslow and Clayton.

The study included 59 epileptic patients randomized to either a treatment or a placebo as an adjuvant treatment to standard chemotherapy. Subjects were evaluated every two weeks to determine the number of seizures that occurred during the previous two-week period. This process was repeated for 4 measurements, or 8 weeks after randomization. Baseline data available

Table II. Posterior 2·5, 50 and 97·5 percentiles for various parameters from MDP Poisson GLMM, fully parametric Poisson GLM, and the 2·5 per cent confidence limit, the estimated value, and the 97·5 per cent confidence limit for Breslow and Clayton's PQL analysis. β_0 is the intercept, β_1 is the Base effect, β_2 is the Trt effect, β_3 is the Base × Trt effect, β_4 is the Age effect, and β_5 is the slope over Visit. For the MDP models, $D(0,0)$ is the variance of the intercepts in the base measure, $D(1,1)$ is the variance of the slopes, and $D(0,1)$ is their covariance. For the fully parametric model and the PQL model, $D(\cdot,\cdot)$ are the elements of the covariance matrix of the random effects. $\bar{k}$ is the average number of clusters observed in the course of sampling

Parameter	MDP GLMM model		FP GLMM	PQL
	$M = 1{\cdot}5$	$M = 100$		
β_0	(−7·0, −3·13, 0·51)	(−6·84, −2·87, 0·48)	(−4·57, −0·72, 2·86)	(−3·6, −1·27, 1·08)
β_1	(0·54, 0·97, 1·4)	(0·62, 0·98, 1·36)	(0·59, 0·90, 1·14)	(0·60, 0·87, 1·14)
β_2	(−2·9, −1·32, 0·39)	(−2·85, −1·35, −0·01)	(−1·78, −0·81, −0·09)	(−1·71, −0·91, −0·11)
β_3	(−0·15, 0·55, 1·18)	(0·00, 0·55, 1·15)	(−0·09, 0·26, 0·74)	(−0·08, 0·33, 0·74)
β_4	(−0·16, 0·88, 1·74)	(−0·09, 0·92, 2·04)	(−0·76, 0·32, 1·34)	(−0·25, 0·46, 1·17)
β_5	(−2·54, −0·44, 1·74)	(−1·49, −0·37, 0·78)	(−0·58, −0·27, 0·05)	(−0·57, −0·26, 0·05)
$D(0,0)$	(0·28, 0·48, 0·94)	(0·18, 0·35, 0·82)	(0·23, 0·35, 0·52)	(0·40, 0·52, 0·64)
$D(0,12)$	(−0·37, 0·02, 0·51)	(−0·34, 0·02, 0·41)	(−0·17, −0·00, 0·17)	(−0·07, −0·01, 0·05)
$D(1,1)$	(0·41, 0·84, 2·79)	(0·42, 0·98, 2·38)	(0·31, 0·55, 1·02)	(0·43, 0·74, 1·05)
$\bar{k}$	10·6	17·7	–	–

as covariates include the number of seizures in the 8 weeks prior to enrolment in the study and age in years. For consistency with the previous analysis, we use the log of one-fourth of the number of seizures (Base) and the log of age in years (Age) as population-mean covariates. The full set of population-mean effects includes an intercept (Int), Base, a treatment effect (Trt), an interaction effect between Base and Trt (Base × Trt), and Age. Finally, we centre the visit time in weeks and divide by 10 (Visit) to give the final population-mean covariate. The random effects include Int and a Visit effect. Our model thus corresponds to model IV of Breslow and Clayton.

In the absence of similar analyses of this treatment, we obtained parameters for the Wishart prior on D^{-1} in the following manner. First, we chose d_0 to be 10. Based on previous experience with fully parametric Poisson GLMM, we wanted $d_0^{-1}R_0$, the prior expected value of D^{-1}, to be diagonal with values of 0·5 on the diagonal. Thus we chose $c_0 = 1$ and

$$R_0 = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}.$$

We chose relatively flat priors for the population-mean parameters. In particular, $\mu_0 = (0\ 0\ 0\ 0\ 0\ 0)^{\mathrm{T}}$ and $\Sigma_0 = 10{,}000 I_6$. We chose two values of the parameter M to reflect moderate and large departures from normality for the distribution of the random effects. A value of $M = 100$ reflects a moderate departure from normality, with the average number of clusters $\bar{k} \approx 20$. A value of $M = 1{\cdot}5$ reflects a large departure from normality, with $\bar{k} \approx 10$.

In Table II, we present results from the MDP GLMM analysis of the model along with the fully parametric Bayesian GLMM analysis and the PQL results of Breslow and Clayton. Recall that the fully parametric model is equivalent to the MDP model when $M \to \infty$. We generated the fully parametric Bayesian results using the BUGS program (Gilks *et al.*[35]). We ran Gibbs

samplers for 25,000 iterations, with the first 3000 discarded as a burn-in. In addition, due to high autocorrelation among the samples, we used every tenth iterate to make posterior inference, with the rest discarded. This makes for a total sample size of 2200. We assessed convergence of the Gibbs sampler by the methods of Geweke[31] and Raftery and Lewis[32] using the CODA (Best *et al.*[33]) suite of diagnostics in S-plus. Most of the parameters in each model had Geweke statistics within $\pm$ 1·96, indicating that convergence is plausible. Due to the lower autocorrelation in the iterates from this analysis than from the binary outcome example, the Raftery and Lewis diagnostic shows that the burn-in and the use of every tenth iterate suggest that the 2·5 percentile of the posterior is within 0·01 of the observed 2·5 percentile with probability 0·9.

The Bayesian results differ from the PQL results, though in all cases the 95 per cent highest posterior density (HPD) regions overlap with the 95 per cent confidence intervals. In general, the posterior medians of the fully parametric GLMM results are very close to the estimates from the PQL analysis. The main differences are that the 95 per cent HPD regions for Int and Age are much wider than the 95 per cent confidence intervals (CI). In addition, the PQL estimates of the elements of the covariance matrix of the random effects differ from the posterior means; both the variance of the intercept and the variance of the slopes are smaller under the fully parametric Bayesian model. These results are evidence that PQL can be a fairly accurate technique.

Comparing the MDP GLMM models to the fully parametric model, we notice that among the population-mean effects, the 95 per cent HPD region for the Base effect excludes 0, just as the 95 per cent CI does. However, an apparent Trt effect, seen in the PQL and fully parametric Bayesian GLMM is much more tenuous under the MDP model. When $M = 1{\cdot}5$, that is, when the distribution of the random effects is least similar to a normal distribution, 0 is well within the 95 per cent HPD region for the Trt effect. Conversely, the effect of Age seems less likely to be 0 in the MDP models, while 0 is well within the CI for the PQL model and the HPD region for the fully parametric Bayesian model. Other parameters seem more or less equivalent across models, although the 95 per cent HPD regions are generally wider under the MDP models than under the fully parametric or PQL fits. Finally, the elements of the D matrix seem different under the MDP models, but we must note that the role of this matrix is not the same under the MDP models and the others and therefore no straightforward comparison can be made. Based on the MDP GLMM analysis, we conclude that there is a slightly larger effect of Base, that the treatment may not be effective, and that Age may be a useful predictor.

One focus of earlier analyses was outlier detection. This is an area where our semi-parametric method can differ substantially from fully parametric techniques. In Figure 2, we present a graphical display of the posterior means of the random effects. This display shows markedly different results than those of Breslow and Clayton[7] (their Figure 2), which are reproduced as Figure 3. They identified five patients with unusual random effects, who are labelled here with their ID numbers from Thall and Vail.[35] In contrast to the PQL results, we notice that four of the five (subjects 112, 225, 227 and 234) no longer seem so unusual. This calls to greater attention the unusual improvement over time of subject 135 in both analyses, given the high initial seizure rate. However, even this subject is drawn toward the central group, and is no longer the subject with the greatest improvement.

6. DISCUSSION

In this article we applied a general technique for Bayesian non-parametrics to an important class of models, the generalized linear mixed model (GLMM). We showed that the GLMM can be

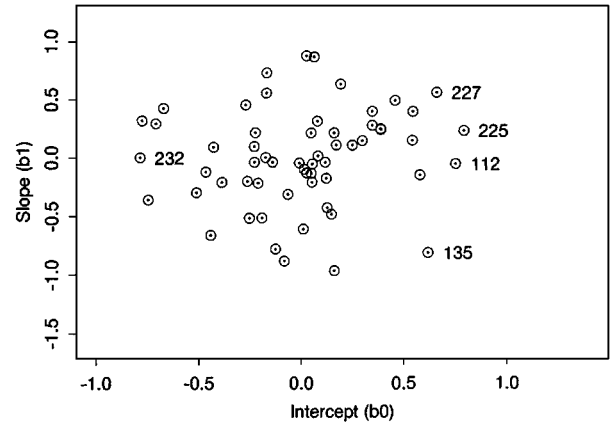

Figure 2. Posterior means of random slopes and intercepts from the epilepsy example: MDP GLMM with $M = 1{\cdot}5$

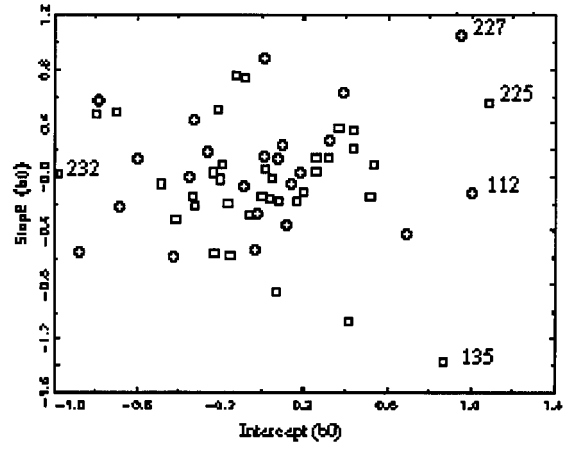

Figure 3. PQL estimates of random slopes and intercepts from the epilepsy example. Reproduced from Breslow and Clayton[7]

freed from the parametric assumption for the random effects. Our technique involved specifying a non-parametric prior for the distribution of the random effects, and a Dirichlet process prior on the space of prior distributions for that non-parametric prior. We then fit the resulting model with the Gibbs sampler. The approach extends quite easily to population models more generally. The proposed procedure represents a new application of the MDP model to correlated data.

We also demonstrated how one can effectively and usefully apply the model to longitudinal data that come from the Bernoulli and Poisson families, which correspond to popular study designs in the medical literature. In each case we show that results based on MDP models can be substantially and meaningfully different from the fully parametric and/or classical approaches to the same data. This effect is more pronounced when interest centres on the random effects, but is also observed in the fixed effects. We conclude that application of the technique can

result in different conclusions or at least valuable additional insight in cases where GLMMs are appropriate.

The important contributions of this article revolve around the semi-parametric model for the random effects. The implementation of the MDP model for the GLMM has not been laid out in the detail presented here. In addition, the interpretation of the model may be of some use to applied biostatisticians, who often use random effects models. There has been little direct focus on this model, and little data analysis of semi-parametric random effects models. The computational implementation of this model is new. Also, the application to and discussion of the data sets is helpful for understanding the importance and utility of this model. Finally, we demonstrate how to make Bayesian inference for all of the parameters in our model.

For our examples, the Gibbs sampler takes about 1 hour to run for 8000 iterations on a Sparcstation 4. The time needed increases dramatically with the number of subjects and the number of observations. A potential drawback of the Gibbs sampler for this problem is that one needs more iterations for small values of M. However, reparameterization techniques, such as those suggested in Gilks and Roberts[36] and Gelfand *et al.*[37] may speed up convergence.

Future work suggested by this article includes possible use of other base measures; in particular a Uniform base measure. In our experience with both the MDP linear random effects model and the MDP GLMM, autocorrelation among the iterates has been a cause for concern. Another area of future interest is to determine to what degree this problem stems from the MDP model, and if there is any way to address it.

ACKNOWLEDGEMENTS

This work was supported in part by NIH grant MH 17119 and NIH grant CA 70101-01. The authors thank Scott Zeger, Joanne Katz and Alfred Sommer for help in obtaining and permission to use the data for the respiratory infection example, which comes from a study funded by the USAID office of Nutrition.

REFERENCES

1. McCullagh, P. and Nelder, J. A. *Generalized Linear Models*, 2nd edn, Chapman & Hall, London, 1989.
2. Nelder, J. A. and Wedderburn, R. W. M. 'Generalized linear models', *Journal of the Royal Statistical Society*, *Series A*, **135**, 370–384 (1972).
3. Liang, K-Y. and Zeger, S. L. 'Longitudinal data analysis using generalized linear models', *Biometrika*, **73**, 13–22 (1986).
4. Zeger, S. L. and Liang, K-Y. 'Longitudinal data analysis for discrete and continuous outcomes', *Biometrics*, **42**, 121–130 (1986).
5. Laird, N. M. and Ware, J. H. 'Random-effects models for longitudinal data', *Biometrics*, **38**, 963–974 (1982).
6. Zeger, S. L. and Karim, M. R. 'Generalized linear models with random effects: a Gibbs sampling approach', *Journal of the American Statistical Association*, **86**, 79–86 (1991).
7. Breslow, N. E. and Clayton, D. G. 'Approximate inference in generalized linear mixed models', *Journal of the American Statistical Association*, **88**, 9–25 (1993).
8. Bush, C. A. and MacEachern, S. N. 'A semi-parametric Bayesian model for randomized block designs', *Biometrika*, **33**, 275–285 (1996).
9. Kleinman, K. P. and Ibrahim, J. G. 'A Semi-parametric Bayesian approach to the random-effects model', *Biometrics*, **54**, 265–278 (1998).
10. Aitkin, M. 'A general maximum likelihood analysis of variance components in generalized linear models', Personal communication.
11. Heckman, J. J. and Singer, B. 'A method for minimizing the impact of distributional assumptions in econometric models of duration', *Econometrica*, **52**, 271–320 (1984).

12. Davies, R. B. 'Mass point methods for dealing with nuisance parameters in longitudinal studies', *in* Crouchley, R.(ed.), *Longitudinal Data Analysis*, Avebury, Aldershot, Hants, 1987.
13. Verbeke, G. and Lesaffre, E. 'A linear mixed-effects model with heterogeneity in the random-effects population', *Journal of the American Statistical Association*, **91**, 217–221 (1996).
14. Ferguson, T. S. 'A Bayesian analysis of some non-parametric problems', *Annals of Statistics*, **1**, 209–230 (1973).
15. Doss, H. 'Bayesian nonparametric estimation for incomplete data via successive substitution sampling', *Annals of Statistics*, **22**, 1763–1786 (1994).
16. MacEachern, S. N. 'Estimating normal means with a conjugate style Dirichlet process prior', *Communications in Statistics*, **23**, 727–741 (1994).
17. Escobar, M. D. 'Estimating Normal means with a Dirichlet process prior', *Journal of the American Statistical Association*, **89**, 268–277 (1994).
18. Liu, J. 'Nonparametric hierarchical Bayes via sequential imputation', *Annals of Statistics*, **24**, 911–930 (1996).
19. Müller, P., Erkanli, A. and West, M. 'Bayesian curve fitting using multivariate normal mixtures', *Biometrika*, **83**, 67–79 (1996).
20. West, M., Müller, P. and Escobar, M. D. 'Hierarchical priors and mixture models, with applications in regression and density estimation', In Smith, A. F. M. and Freeman, P. B. (eds), *Aspects of Uncertainty*: *A Tribute to D. V. Lindley*, Wiley, London, 1994.
21. Escobar, M. D. and West, M. 'Bayesian density estimation and inference using mixtures', *Journal of the American Statistical Association*, **90**, 578–588 (1995).
22. MacEachern, S. N. and Müller, P. 'Estimating mixture of Dirichlet process models', *Journal of Computational and Graphical Statistics*, in press (1998).
23. Newton, M. A., Czado, C. and Chappell, R. 'Bayesian inference for semiparametric binary regression', *Journal of the American Statistical Association*, **91**, 142–153 (1996).
24. Mukhopadhyay, S. and Gelfand, A. E. 'Dirichlet process mixed generalized linear models', *Journal of the American Statistical Association*, **92**, 633–647 (1995).
25. Blackwell, D. and MacQueen, J. B. ' Ferguson distributions via Polya urn schemes', *Annals of Statistics*, **1**, 353–355 (1973).
26. Walker, S. G. and Damien, P. 'Sampling a Dirichlet-process mixture model', submitted for Publication, University of Michigan Business School Working Paper.
27. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. 'Equations of state calculations by fast computing machine', *Journal of Chemical Physics*, **21**, 1087–1091 (1953).
28. Hastings, W. K. 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika*, **57**, 97–109 (1970).
29. Sommer, A., Katz, J. and Tarwotjo, I. 'Increased mortality in children with vitamin A deficiency', *American Journal of Clinical Nutrition*, **40**, 1090–1095 (1983).
30. Cowles, M. K., Carlin, B. P. and Connett, J. E. 'Bayesian tobit modeling of longitudinal ordinal clinical trial compliance data with non-ignorable missingness', *Journal of the American Statistical Association*, **91**, 86–98 (1996).
31. Geweke, J. 'Evaluating the accuracy of sampling-based approaches to calculating posterior moments', *in* Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (eds), *Bayesian Statistics* **4**, Clarendon Press, Oxford, 1992.
32. Raftery, A. L. and Lewis, S. 'How many iterations in the Gibbs sampler?' *in* Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (eds), *Bayesian Statistics 4*, Clarendon Press, Oxford, 1992.
33. Best, N. G., Cowles, M. K. and Vines, S. K. 'CODA: Convergence diagnostics and output analysis software for Gibbs sampling output', Version 0.3, MRC Biostatistics Unit, Cambridge, 1995.
34. Thall, P. F. and Vail, S. C. 'Some covariance models for longitudinal count data with overdispersion', *Biometrics*, **46**, 657–671 (1990).
35. Gilks, W. R., Thomas, A. and Spiegelhalter, D. J. 'A language and program for complex Bayesian modeling', *Statistician*, **43**, 169–178 (1994).
36. Gilks, W. R. and Roberts, G. O. 'Strategies for improving MCMC', *in* Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds), *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London, 1996.
37. Gelfand, A. E., Sahu, S. K. and Carlin, B. P. 'Efficient parametrisations for normal linear mixed models', *Biometrika*, **82**, 479–488 (1995).

