INFERENCE FOR THE RANDOM EFFECTS IN BAYESIAN GENERALIZED LINEAR MIXED MODELS

Hemant Ishwaran[*], **Cleveland Clinic Foundation**
Department of Biostatistics Wb4, Cleveland Clinic Foundation,
9500 Euclid Avenue, Cleveland, OH 44195

Key Words: Blocked Gibbs sampler, Dirichlet process, Hybrid Monte Carlo, Random probability measure, Truncation level.

ABSTRACT

The use of a truncation approximation to the Dirichlet process is utilized in a Gibbs sampling scheme for fitting a semiparametric generalized linear mixed model. A novel aspect of this approach is that it generates draws directly from the posterior random measure, thus giving a method for directly inferring the random effects distribution. The usual problems associated with non-conjugacy in Dirichlet process computing is resolved in this approach, and computationally, the Gibbs sampler is simpler to implement than traditional methods. Furthermore, the adequacy of the truncation level can be monitored from the output of the Gibbs sampler to ensure that the approximating process is almost indistinguishable from the Dirichlet process. The methodology is illustrated by examples.

1. Introduction

This paper presents a novel Gibbs sampling method for fitting a Bayesian generalized linear mixed model (GLMM) containing approximate Dirichlet process nonparametric random effects. The GLMM model considered is a useful mechanism for studying non-normal as well as normal data that is assumed to involve some form of clustering or correlation. This type of clustering may arise from a longitudinal study in which repeated measurements are collected on an individual over time, or it may arise from a cross-sectional study in which a primary unit is sub-sampled repeatedly. Our models also apply to the univariate case arising when only one observation is recorded per individual, when it is anticipated that individuals are clustered into groups. We take a nonparametric approach to each of these problems by modeling the random effects by an approximate Dirichlet process random measure. Model fitting is carried out by a Gibbs sampling scheme whose key feature is that it can directly draws values from the random effects distribution. Moreover, the proposed Gibbs sampler is relatively straightforward to implement computationally, and avoids the complications arising from non-conjugacy seen with traditional Dirichlet process Gibbs sampling methods.

An important application of the GLMM are longitudinal studies involving repeated measurements where analysis is based on mixed effects models using covariates for fixed and random effect terms. Although the GLMM will be applied in other contexts, it will be easier if we present our approach so that it most closely resembles this important application. We will assume, therefore, that the data X is made up of observations $X_i = (X_{i,1}, \ldots, X_{i,n_i})$ representing n_i repeated observations for an individual (or unit) i, where $i = 1, \ldots, n$, and n is the number of distinct units. In addition, for each individual i and repeat measurement j, we observe a p-dimensional fixed effects covariate $U_{i,j}$ and a q-dimensional random effects covariate $V_{i,j}$, for $j = 1, \ldots, n_i$. The GLMM arises under the assumption that the data has an underlying exponential family density. More precisely, we assume that conditional on the canonical variable $\theta_{i,j}$, and the dispersion parameter ϕ, the $X_{i,j}$ are conditionally independent with a density from an exponential family of the form

$$f(X_{i,j}|\theta_{i,j}, \phi) = \exp\left[\frac{X_{i,j}\theta_{i,j} - \mu(\theta_{i,j})}{a(\phi)} + c(X_{i,j}, \phi)\right].$$

Our models are based on a canonical link, and assume that

$$\theta_{i,j} = U'_{i,j}\beta + V'_{i,j}Y_i,$$

where β represents the p-dimensional fixed effects parameter and Y_i is the q-dimensional random effects parameter with unknown distribution.

*The author thanks Tom Greene for providing access to the MDRD data and for helpful discussion surrounding its analysis.

1.1. Intractability of the likelihood

The analysis of the GLMM is complicated due to the intractable form of the likelihood. If Y_i are iid values drawn from some random effects distribution P, then the GLMM likelihood, $\mathcal{L}(\beta, \phi, P)$, equals

$$\prod_{i=1}^{n} \int \prod_{j=1}^{n_i} f(X_{i,j}|U'_{i,j}\beta + V'_{i,j}Y_i, \phi)\, dP(Y_i).$$

This is of course intractable except in the special case when P is conjugate. Without conjugacy, evaluation of the likelihood can be quite formidable, especially in high dimensions.

Classical frequentist approaches for handling this problem have involved the application of various types of integral approximations (see Breslow and Clayton 1996 and Fahrmeir and Tutz 1994, Chapter 7). Such approximations are typically applied when P is assumed to be parametric. More recent methods allowing for nonparametric random effects take a different approach. See Liang and Zeger (1986), Follmann and Lambert (1989) and Aitkin (1999) for examples.

1.2. Likelihoods based on Y

A different method for fitting the GLMM works by including Y_i as parameters in the likelihood. Applied within a Bayesian framework this can be used in combination with the Gibbs sampler as a powerful method for fitting the GLMM and avoids the need to evaluate complex integrals. This trick apparently was first noticed by Zeger and Karim (1991) who studied the GLMM with parametric random effects. More recently, Gibbs sampling methods have been developed for fitting the GLMM with nonparametric random effects by using the Ferguson (1973) Dirichlet process. For example, Bush and MacEachern (1996) considered a semiparametric randomized block design, Kleinman and Ibrahim (1998) studied the semiparametric normal linear random effects model, while Ibrahim and Kleinman (1998) looked in general at the semiparametric GLMM.

In fitting the GLMM with the Dirichlet process, the papers mentioned above all work with the likelihood

$$\mathcal{L}(\beta, \phi, Y) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} f(X_{i,j}|U'_{i,j}\beta + V'_{i,j}Y_i, \phi).$$

The methods then work by using a generalized Pólya urn prior for $Y = (Y_1, \ldots, Y_n)$, which is the marginal law for Y subject to Y_i sampled from a Dirichlet process (Blackwell and MacQueen 1973). Doing so makes it possible to fit the GLMM by the use of the Escobar-West-MacEachern Pólya urn Gibbs sampler. See Escobar (1988, 1994), Escobar and West (1995) and MacEachern (1994).

1.3. Approximate Dirichlet processes and the blocked Gibbs sampler

We will take a different approach and work with the full likelihood

$$\mathcal{L}(\beta, \phi, Y, P) = \prod_{i=1}^{n} dP(Y_i) \prod_{j=1}^{n_i} f(X_{i,j}|U'_{i,j}\beta + V'_{i,j}Y_i, \phi).$$

However, rather than working with the Dirichlet process, we will use a Dirichlet process truncation approximation as our prior for P. Doing so will allow us to exploit a new type of Gibbs sampling approach, the *blocked Gibbs sampler* developed by Ishwaran and Zarepour (2000a) and Ishwaran and James (2000a).

There are several important features of the blocked Gibbs approach. One is that it allows for straightforward handling of non-conjugacy which is a serious concern in the GLMM. This is one of its key advantages over Pólya urn approaches which require the use of specialized techniques like those discussed in West, Müller, and Escobar (1994), MacEachern and Müller (1998) and Walker and Damien (1998). Unfortunately, these methods complicate the description of the Gibbs sampler and tend to make the overall approach less accessible for use by the non-expert. Furthermore, the resulting Markov chain may not always mix efficiently. See Neal (2000) for performance comparisons.

The blocked Gibbs also gives direct inference for the random effects distribution, as the procedure is designed to draw values directly from the posterior of the random measure. Thus, one can directly study the random effects distribution from posterior sampled values. Another novel aspect of the proposed Gibbs sampler is its use of hybrid Monte Carlo (HMC) in the update for fixed effects parameters. The “leapfrog” algorithm of Duane et al. (1987) that will be used has been observed to work well in the generalized linear model and promotes rapid mixing of the Markov chain even in complex problems such as logistic regression with quasicomplete separation (Ishwaran, 1999). We anticipate that these same attributes will be inherited in the GLMM. Also see Neal (1994, 1996) and Gustafson (1997) for further applications of HMC.

In summary, the paper begins in Section 2 with a discussion of the hierarchical model and the Dirichlet process approximation. Section 3 presents the Gibbs sampling details, including the HMC step, while Sections 4 and 5 present examples illustrating the methods. We end with a brief discussion in Section 6.

2. Hierarchical Priors

We begin describing our Bayesian model by first presenting the hierarchical priors for the p-dimensional fixed effects, β. Writing $b = (b_1, \ldots, b_p)$ and $\sigma = (\sigma_1, \ldots, \sigma_p)$, we choose the priors

$$\begin{aligned} (\beta|b,\sigma) &\sim \mathrm{N}_p(b, \mathrm{Diag}(\sigma)) \\ (b_l|\sigma_b) &\stackrel{\text{iid}}{\sim} \mathrm{N}(0,\sigma_b) \\ (\sigma_l^{-1}|s_1,s_2) &\stackrel{\text{iid}}{\sim} \mathrm{Ga}(s_1,s_2), \qquad l = 1,\ldots,p, \end{aligned} \tag{1}$$

where the notation $\mathrm{Diag}(\sigma)$ is used to indicate a diagonal matrix with diagonal elements $\sigma_1, \ldots, \sigma_p$.

In (1) we use a flat prior for b by choosing a large value for its variance ($\sigma_b = 1,000$). We also use a noninformative prior for σ by selecting small values for the shape and scale parameters in the inverse-gamma prior for σ_l ($s_1 = s_2 = 0.0001$). In the case where the data is normal, we include the use of a dispersion parameter ϕ, with the prior

$$\phi^{-1} \sim \mathrm{Ga}(\rho_1, \rho_2),$$

where $\rho_1 = \rho_2 = 0.0001$ is chosen for noninformativeness. For non-normal data, such as in Poisson and logistic regression examples, we usually fix the dispersion parameter by setting $a(\phi) = 1$. In these examples we focus on modeling heterogeneity through the use of random effects, although the method could be extended to include ϕ as well.

2.1. Truncated Dirichlet processes

The random effects distribution for Y_i will involve a truncated Dirichlet process (Ishwaran and Zarepour 2000a,b). Specifically, we assume that

$$\begin{aligned} (Y_i|P) &\stackrel{\text{iid}}{\sim} P \\ P &\sim \mathcal{P}_N, \end{aligned}$$

where $\mathcal{P}_N$ is the random probability measure defined by

$$\mathcal{P}_N(\cdot) = \sum_{k=1}^{N} p_k \, \delta_{Z_k}(\cdot) \qquad 1 < N < \infty. \tag{2}$$

In (2) we write $\delta_{Z_k}(\cdot)$ to denote a discrete measure concentrated at Z_k, while $p_1 = \xi_1$ and

$$p_k = (1-\xi_1)(1-\xi_2)\cdots(1-\xi_{k-1})\,\xi_k, \tag{3}$$

for $k = 2, \ldots, N$, where $\xi_1, \xi_2, \ldots \xi_{N-1}$ are iid $\mathrm{Beta}(1,\alpha)$ random variables and we set $\xi_N = 1$ to ensure that $\sum_{k=1}^N p_k = 1$. Furthermore, it is assumed that the ξ_k (and hence p_k) are independent of the Z_k which are assumed to be iid from a distribution H over the measurable space $(\mathcal{Y}, \mathcal{B})$. In the applications considered here, $\mathcal{Y} = \Re^q$ and $\mathcal{B}$ is the corresponding Borel algebra.

By the construction given in Sethuraman (1994), it easily follows that the random probability measure described in (2) is a truncated version of the Ferguson (1973) Dirichlet process. In fact, $\mathcal{P}_N$ converges almost surely to the random measure $\mathcal{P}_\infty$, which is the Dirichlet process with finite measure αH.

In applications to semiparametric hierarchical models, the parameters in the Dirichlet process typically involve hyperparameters. See for instance, West, Müller and Escobar (1994), Escobar and West (1998), and MacEachern (1998). In a similar fashion, we allow the random probability weights p_k and the points Z_k defined in $\mathcal{P}_N$ to have a hierarchical prior structure. This introduces greater model flexibility by allowing for an approximate Dirichlet process mixture. In particular, it is assumed that

$$\begin{aligned} (Z_k|M,\tau) &\stackrel{\text{iid}}{\sim} \mathrm{N}_q(M, \mathrm{Diag}(\tau)), \qquad k = 1,\ldots,N \\ (M_l|\sigma_M) &\stackrel{\text{iid}}{\sim} \mathrm{N}(0,\sigma_M) \\ (\tau_l^{-1}|t_1,t_2) &\stackrel{\text{iid}}{\sim} \mathrm{Ga}(t_1,t_2), \qquad l = 1,\ldots,q \\ (\alpha|a_1,a_2) &\sim \mathrm{Ga}(a_1,a_2), \end{aligned} \tag{4}$$

where $M = (M_1, \ldots, M_q)$ and $\tau = (\tau_1, \ldots, \tau_q)$.

In (4) we use a flat prior for M by selecting a large value for its variance ($\sigma_M = 1,000$). We will also use a noninformative prior for τ by working with small shape and scale parameters for its inverse-gamma prior ($t_1 = t_2 = 0.0001$). The values for a_1 and a_2 in the prior for α *should be selected with care* because they directly influence the value for α, which controls the clustering behavior of Y_i. For example, smaller values for α in the Dirichlet process will encourage less distinct values in Y, which directly affects not only the inference for the random effects, but the fixed effects as well. In fact, because α can have such a great impact on the final inference, it is wise to generally include it as a Bayesian parameter in the model fitting. With this word of caution, it was decided to use values $a_1 = a_2 = 2$ as they will encourage both small and large values for α.

2.2. Appropriate truncation values for N

As noted by Ishwaran and Zarepour (2000a), the key to working with the approximation $\mathcal{P}_N$, instead of its limit $\mathcal{P}_\infty$, is that it allows us to update our nonparametric parameters directly in the Gibbs sampler. In the GLMM this method is used to produce a direct draw from the posterior (see Section 3), and involves blocked multivariate updates for parameters that encourages the Markov chain to mix efficiently even with non-conjugacy.

This technique works for any finite value for N, and with a suitably chosen truncation level, the method will produce a posterior almost indistinguishable from one obtained using a Dirichlet process. With such a complex hierarchical structure, however, it is difficult to compute exactly how close the approximating posterior is to its limit. Ultimately, though, the level of this accuracy is tied to the level of precision in which $\mathcal{P}_N$ approximates $\mathcal{P}_\infty$, which can be quantified relatively easily. Let

$$D(\mathbb{P}_1, \mathbb{P}_2) = \sup_A |\mathbb{P}_1(A) - \mathbb{P}_2(A)|$$

be the total variation distance between two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ built over the same measurable space. Consider the following theorem that measures the distance between $\mathcal{P}_N$ and $\mathcal{P}_\infty$.

THEOREM 1. *Let p_k be the random weights defined by (3). Then,*

$$\begin{aligned} D(\pi_N, \pi_\infty) &\leq 2\left[1 - \mathbb{E}\left\{\left(\sum_{k=1}^{N-1} p_k\right)^n\right\}\right] \\ &\approx 2n\exp(-(N-1)/\alpha), \end{aligned}$$

where π_N and π_∞ are the marginal distributions for $Y = (Y_1, \ldots, Y_n)$ under $\mathcal{P}_N$ and $\mathcal{P}_\infty$ respectively.

PROOF. This follows using a similar argument as in Ishwaran and James (2000b). □

Theorem 1 shows that a precise approximation to $\mathcal{P}_\infty$ is available for reasonably sized truncation levels N. For example, if $\alpha = 1$, then the total variation distance between π_N and π_∞ is approximately 1.05×10^{-18} if $N = 50$ and $n = 1000$ (notice that the sample size will make almost no impact on the near zero distance). Although this calculation is based on a fixed value for α, we can still assess the accuracy of the truncation when α is fit as a parameter in the model. To do so, monitor the posterior values for α as the Gibbs sampler iterates and check that the N chosen is adequate over the range of observed α values.

3. Gibbs Sampling

In implementing the Gibbs sampler, we need to sample the posterior

$$(\beta, b, \sigma, \phi, p, Z, Y, M, \tau, \alpha | X) \tag{5}$$

where $p = (p_1, \ldots, p_N)$ and $Z = (Z_1, \ldots, Z_N)$. This is done by first drawing values for the fixed effects parameters $(\beta, b, \sigma, \phi | Y, X)$, which is then followed by an update for the random effects conditioned on the current fixed effects values.

Repeating this many times eventually produces a draw from the posterior (5), and each value Z^* and p^* in a cycle of the Gibbs sampler defines a random probability measure

$$P^*(\cdot) = \sum_{k=1}^N p_k^* \, \delta_{Z_k^*}(\cdot)$$

that is a draw from the posterior distribution of the random effects.

3.1. Gibbs sampling for fixed effects

In the case where the data is normal, we draw β and its hyperparameters exactly due to conjugacy. In particular, $(\beta | b, \sigma, \phi, Y, X) \sim \mathrm{N}_p(b^*, \Sigma^*)$ where

$$b^* = \Sigma^* \Big(\mathrm{Diag}(\sigma)^{-1} b + \frac{1}{\phi} \sum_{i,j} U_{i,j}(X_{i,j} - V_{i,j})' Y_i\Big)$$

and

$$\Sigma^* = (\mathrm{Diag}(\sigma)^{-1} + \sum_{i,j} U_{i,j} U'_{i,j} / \phi)^{-1}.$$

Furthermore, it also follows that

$$(b | \beta, \sigma) \sim \mathrm{N}_p(\Lambda^* \mathrm{Diag}(\sigma)^{-1} \beta, \Lambda^*),$$

where $\Lambda^* = (\mathrm{Diag}(\sigma)^{-1} + \sigma_b^{-1} I)^{-1}$. Also,

$$(\sigma_l^{-1} | \beta, b) \overset{\text{ind}}{\sim} \mathrm{Ga}\Big(s_1 + 1/2, s_2 + (\beta_l - b_l)^2/2\Big).$$

With normal data we use an inverse-gamma prior for ϕ. Thus, by conjugacy, given the value for $\theta = (\theta_{1,1}, \theta_{1,2}, \ldots)$

$$(\phi^{-1} | \theta) \sim \mathrm{Ga}\Big(\rho_1 + n./2, \rho_2 + \sum_{i,j} (X_{i,j} - \theta_{i,j})^2/2\Big),$$

where $n. = n_1 + \cdots + n_n$ is the total number of observations.

3.1.1. *Hybrid Monte Carlo (HMC)*

All updates for hyperparameters are the same as described above in the case where the data is non-normal (recall that we fix ϕ by assuming that $a(\phi) = 1$). This leaves only the update for β, which we complete using the leapfrog hybrid Monte Carlo method of Duane et al. (1987). Theoretical details concerning the algorithm can be found in Duane et al. (1987), Kennedy (1990) and Neal (1994; 1996, Chapter 3). Here we will only be concerned with the description of the algorithm and it's implementation within the GLMM. If β_* is the current value for β, then the leapfrog algorithm generates a new value for β by performing the following steps:

1. Simulate $\zeta_* \sim \mathrm{N}_p(0, I)$. Let $\widetilde{\beta}_0 = \beta_*$ and $\widetilde{\zeta}_0 = \zeta_* + \delta\Delta(\beta_*)/2$.
2. For $l = 1, \ldots, L$, let
$$\begin{aligned} \widetilde{\beta}_l &= \widetilde{\beta}_{l-1} + \delta\widetilde{\zeta}_{l-1} \\ \widetilde{\zeta}_l &= \widetilde{\zeta}_{l-1} + \delta_l \Delta(\widetilde{\beta}_l), \end{aligned}$$
where $\delta_l = \delta$ for $l < L$, otherwise $\delta_L = \delta/2$.
3. Move from the current value of β_* to the new candidate $\widetilde{\beta}_L$ with probability
$$\left[\frac{\pi(\widetilde{\beta}_L)}{\pi(\beta_*)} \exp\left(-\frac{1}{2}(\widetilde{\zeta}_L'\widetilde{\zeta}_L - \zeta_*'\zeta_*)\right)\right] \wedge 1.$$

As noted by Ishwaran (1999), a key to the computational success of the leapfrog algorithm in the generalized linear model is the simplicity of Δ, the gradient of the log of the posterior. In particular, with a normal prior for β as specified in (1), we end up with a p-dimensional log-gradient

$$\Delta(\beta) = \sum_{i,j}\Big(X_{i,j} - \mu'(\theta_{i,j})\Big)U_{i,j} - \mathrm{Diag}(\sigma)^{-1}(\beta - b),$$

which is easily computed, especially if we keep track of the value for θ when running the algorithm.

A good strategy for selecting the number of leapfrog steps L and the step size δ used in the algorithm is to choose values that encourage as large a trajectory as possible for β, while leading to as a high an acceptance rate as possible. In the examples we have considered we find a good choice is to select $L = 100$, while tuning δ in order to get approximately an 80% acceptance rate. Also, a good strategy for tuning δ is to start the Gibbs sampler off for a few iterations (such as 50) using random walk Metropolis-Hastings for β. This will bring us to a high probability region of the posterior where we can tune δ with minimal fuss.

3.2. Gibbs sampling for random effects

The key to efficient sampling for $\mathcal{P}_N$ is to recast the definition of Y_i in terms of a latent variable K_i which records its cluster membership. We redefine Y_i as Z_{K_i}, where K_i is a random integer, independent of Z_k, with the discrete distribution

$$(K_i|p) \overset{\text{iid}}{\sim} \sum_{k=1}^{N} p_k \delta_k(\cdot), \qquad i = 1, \ldots, n.$$

For example, this means that we can reexpress the canonical parameter as

$$\theta_{i,j} = U_{i,j}'\beta + V_{i,j}' Z_{K_i}.$$

Write $K = (K_1, \ldots, K_N)$. To complete the update for $\mathcal{P}_N$, we will need to sample

$$(p, Z, K, M, \tau, \alpha|\beta, \phi, X),$$

which is implemented in the following five steps within the blocked Gibbs sampler:

$$\begin{aligned} &(p, Z|K, M, \tau, \alpha, \beta, \phi, X) \\ &(K|p, Z, \beta, \phi, X) \\ &(M|Z, \tau) \\ &(\tau|Z, M) \\ &(\alpha|p). \end{aligned}$$

As mentioned earlier in Section 2, the draws from the nonparametric parameters p, Z and K are blocked, with each step completed simultaneously in a multivariate draw to encourage mixing of the Markov chain. The issue of conjugacy only arises in the draw for Z, which is done exactly in the conjugate model and by using random walk Metropolis-Hastings in the non-conjugate case. This will be straightforward. The exact details are as follows:

(a) As worked out by Ishwaran and Zarepour (2000a), the conditional distribution for $(p|K, \alpha)$ is a generalized Dirichlet distribution, and can be sampled efficiently by $p_1 = \xi_1^*$ and

$$p_k = (1 - \xi_1^*)(1 - \xi_2^*)\cdots(1 - \xi_{k-1}^*)\,\xi_k^*, \qquad (6)$$

where

$$\xi_k^* \overset{\text{ind}}{\sim} \mathrm{Beta}\Big(1 + m_k, \alpha + \sum_{j=k+1}^{N} m_j\Big)$$

for $k = 1, \ldots, N-1$ and m_k records the number of K_i values which equal k.

(b) As mentioned, sampling the value for Z is exact for the case of normal data (see below). Without conjugacy, however, its conditional density is

$$f(Z|K,M,\tau,\beta,\phi,X) \propto f(Z^K|M,\tau) \times \prod_{l=1}^{m} f(Z_{K_l^*}|M,\tau) \prod_{\{i:K_i=K_l^*\}} f(X_i|Y_i = Z_{K_l^*},\beta,\phi),$$

where $K_1^*,\ldots,K_m^*$ are the unique set of K_i values and Z^K corresponds to those values in Z which exclude $Z_K = (Z_{K_1^*},\ldots,Z_{K_m^*})$. The components of Z^K are iid $N_q(M,\mathrm{Diag}(\tau))$ and are straightforward to sample, but each of the $Z_{K_j^*}$ will have to be sampled using some other method, such as random walk Metropolis-Hastings for example. Note that the number of these terms, m, will usually be much smaller than the truncation level N, and so this step can be done efficiently.

In the normal case, conjugacy allows us to update Z_K exactly. Specifically, $(Z_{K_l^*}|M,\tau,\beta,\phi,X) \overset{\text{ind}}{\sim} \mathrm{N}_q(M_l^*,T_l^*)$, where

$$M_l^* = T_l^*\Big(\mathrm{Diag}(\tau)^{-1}M + \frac{1}{\phi}\sum_{\{i:K_i=K_l^*\}}\sum_{j=1}^{n_i} V_{i,j}X_{i,j}^*\Big)$$

$$T_l^* = \Big(\mathrm{Diag}(\tau)^{-1} + \frac{1}{\phi}\sum_{\{i:K_i=K_l^*\}}\sum_{j=1}^{n_i} V_{i,j}V_{i,j}'\Big)^{-1}$$

and $X_{i,j}^* = X_{i,j} - U_{i,j}'\beta$.

(c) It is obvious that the conditional distribution for K_i is

$$(K_i|p,Z,\beta,\phi,X) \overset{\text{ind}}{\sim} \sum_{k=1}^{N} p_{k,i}^*\,\delta_k(\cdot),$$

where $(p_{1,i}^*,\ldots,p_{N,i}^*)$ is proportional to

$$\big(p_1 f(X_i|Z_1,\beta,\phi),\ldots,p_N f(X_i|Z_N,\beta,\phi)\big).$$

(d) The remaining parameters can all be sampled exactly due to conjugacy. Indeed, it easily follows that,

$$(M|Z,\tau) \sim \mathrm{N}_q\Big(T^*\mathrm{Diag}(\tau)^{-1}\sum_{k=1}^{N} Z_k\,,\,T^*\Big)$$

where $T^* = (N\times \mathrm{Diag}(\tau)^{-1} + \sigma_M^{-1}I)^{-1}$. Also,

$$(\tau_l^{-1}|Z,M) \overset{\text{ind}}{\sim} \mathrm{Ga}\Big(t_1 + \frac{N}{2}, t_2 + \sum_{k=1}^{N}(Z_{k,l}-M_l)^2/2\Big).$$

(e) Finally, as discussed by Ishwaran and Zarepour (2000a),

$$(\alpha|p) \sim \mathrm{Ga}\left(N + a_1 - 1, a_2 - \sum_{k=1}^{N-1}\log(1-\xi_k^*)\right),$$

for the *same values* of ξ_k^* used in the simulation (6) for p.

4. Poisson Mixed Model

For our first example, we re-analyzed the ships data presented in McCullagh and Nelder (Chapter 6.3, 1989) consisting of different characteristics of a ship: its type (A, B, C, D, E), year of construction (y60, y65, y70, y75), and period of operation (p60, p75); and an offset variable, months of operation, as well as the number of damage incidents incurred to the ship. We discarded observations with zero months of operation, leaving a total of $n = 34$ cases. As in McCullagh and Nelder (1989), the number of damage incidents was considered to be a Poisson random variable with a mean depending upon the characteristics of the ship and the log-months of operation (time).

	MLE		HMC		Random Int.	
Par	Est	StEr	Mean	StDv	Mean	StDv
Int.	-4.58*	0.73	-4.69*	0.75	-4.97[a]*	2.66[a]
A	-0.27	0.24	-0.27	0.24	-0.32	0.39
B	-0.62	0.32	-0.63	0.33	-0.69	0.39
C	-1.04*	0.34	-1.06*	0.35	-1.16*	0.47
D	-0.41	0.31	-0.42	0.32	-0.59	0.49
E	-	-	-	-	-	-
y60	-0.37	0.25	-0.37	0.25	-0.51	0.54
y65	0.29	0.21	0.29	0.21	0.26	0.31
y70	0.39	0.20	0.39	0.20	0.37	0.29
y75	-	-	-	-	-	-
p60	-0.37*	0.12	-0.37*	0.12	-0.39*	0.13
p75	-	-	-	-	-	-
time	0.90*	0.10	0.91*	0.10	0.95*	0.12
α	-	-	-	-	1.06	0.75

Table 1: Parameter estimates and standard errors for ships data via maximum likelihood estimation (MLE) using SAS PROC GENMOD (SAS Institute Inc., 1996). Mean values and standard deviations are via hybrid Monte Carlo (HMC) for the fixed effects model and via the blocked Gibbs sampler for the random intercept model. [a]Mean value and standard deviation for the hypermean M given in place of the intercept. *Parameters significant at the 5% level.

We fit a fixed effects model for our first GLMM, and compared these results to maximum likelihood estimation (MLE). In a purely fixed effects model with non-normal data, our MCMC approach reverts solely to HMC. Note that with Poisson data and a log-link, this means that the gradient used in the leapfrog algorithm equals

$$\Delta(\beta) = \sum_{i,j}\Big(X_{i,j} - \exp(\theta_{i,j})\Big)U_{i,j} - \mathrm{Diag}(\sigma)^{-1}(\beta - b).$$

Inference was based on 5,000 sampled values obtained after an initial 500 iteration burn-in, with the first 50 iterations computed using Metropolis-Hastings (following the strategy discussed earlier). The parameter estimates from the Gibbs sampler are indicated in Table 1 and are almost identical to those derived from MLE.

A second model included a random intercept term as a method for analyzing heterogeneity between ships. We used the same strategy as before in fitting this mixed effects model, although a longer burn-in period of 2,500 iterations was used. A truncation level of $N = 100$ was used, and given the small sample size, the Z_K values updated by Metropolis-Hastings were sampled 100 times for each cycle in the Gibbs sampler. This can help to promote mixing for these terms. We also used a large value of $\tau = 10$ in the hypervariance for M defined in (4). A large variance is useful in reducing clustering, and is helpful in this example in isolating similar groups.

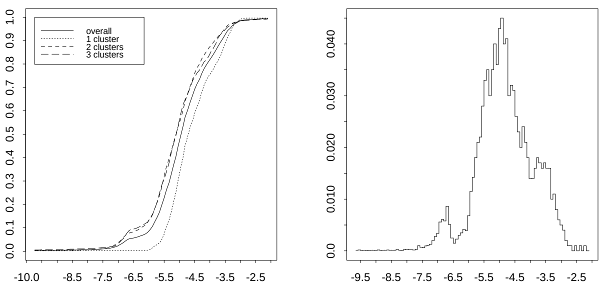

Figure 1: Averaged posterior cdf for the random intercept indexed by the number of clusters (left-hand side). Only the overall averaged model and one cluster, two cluster and three cluster models are presented (corresponding to over 90% of the posterior models). Right-hand side plot is overall averaged cdf converted into a histogram with vertical axis measuring probability.

Reconsidering Table 1 again, we see that the greatest difference in our two GLMM models is between the intercept parameter, and most of this difference can be attributed to one observation, which appears to be an outlier. In fact, the left-most mode seen in the right-hand side plot of Figure 1 for the random intercept is due to this particular observation. Probably the presence of 2 groups in the random intercept is what contributes to the larger standard deviations seen for the fixed effects in this model.

As expected, HMC mixes very well in the fixed effects model, but there appears to be a high, but apparently vanishing, autocorrelation in two of the fixed effects parameters for the random intercept model. See Figure 2.

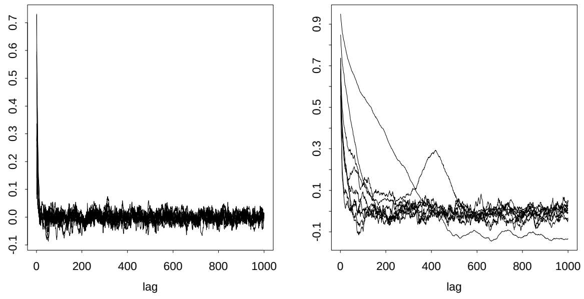

Figure 2: Autocorrelations for β parameters from fixed effects model (left-hand side) and random intercept model (right-hand side) from ships data.

5. Chronic renal disease

Here, we analyze data collected from the Modification of Diet in Renal Disease (MDRD) study, a longitudinal study investigating the effects of dietary protein restriction and strict blood pressure control on the rate of renal disease progression (Klahr et al. 1994). A key measurement in the study for evaluating renal disease progression are sequential measurements of glomerular filtration rate (GFR). We consider data on GFR measurements from the planned follow up at four month intervals over 1.6 to 4.4 years. The data we considered consisted of $n = 840$ patients.

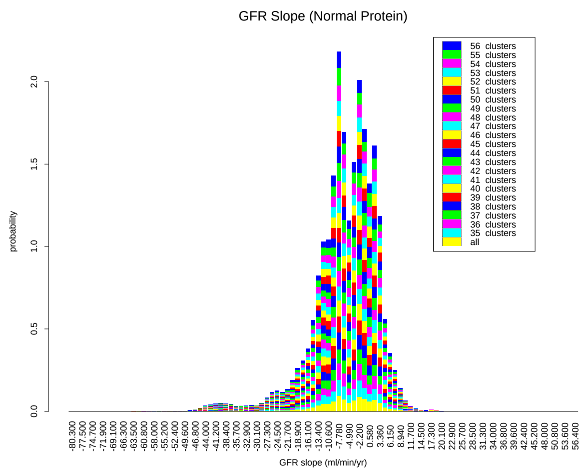

Figure 3: Posterior estimates of GFR slope densities for the normal diet treatment group from the MDRD study. Probabilities for densities are stacked and indexed by the number of clusters. This allows us to compare the shape of the densities as a function of smoothing.

Analyses of MDRD data conducted after the completion of the study indicated that the distribution of the GFR slopes deviated from normality due to

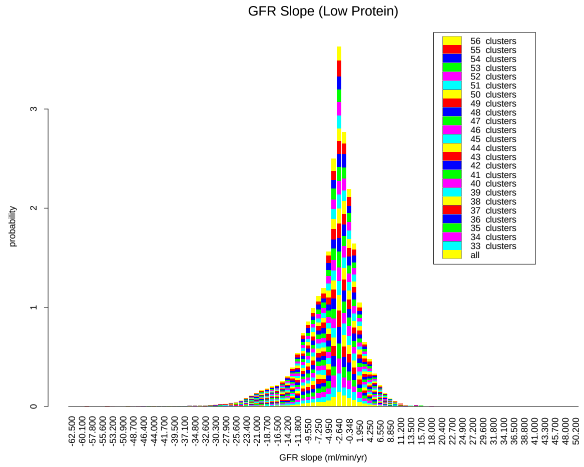

Figure 4: Posterior estimates of GFR slope densities for the protein restricted treatment group.

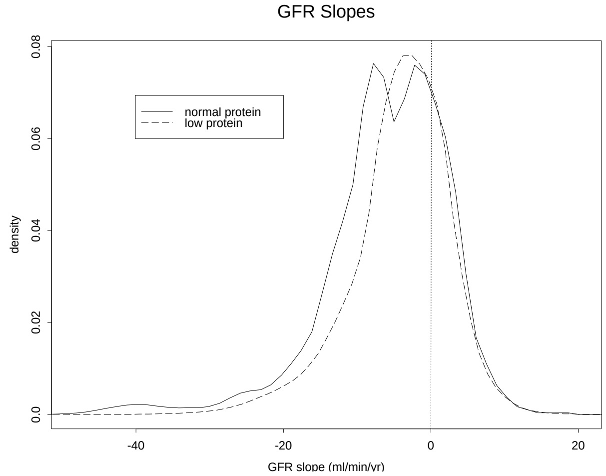

Figure 5: Averaged posterior GFR slope densities.

negative skewness and positive kurtosis. Thus a non-parametric analysis is quite justified here. We fit a normal slope/intercept model to the data, with separate sets of slopes and intercepts for the two treatment groups (protein restricted versus normal diet). The results of the analysis are presented in Figures 3, 4 and 5, and are based on 5,000 sampled values from the blocked Gibbs sampler following an initial 2,500 iteration burn-in.

Figures 3 and 4 present posterior estimates for the slope distribution of GFR for the two treatment groups. Note that a negative value indicates a more rapid progression of disease. It is clear from the plots that both distributions are non-normal. The protein restricted group has a slope distribution that appears skewed, while the normal dietary group appears to have a bimodal slope distribution.

Of greater interest is a direct comparison of the two slope distributions. See Figure 5. Some very interesting results are revealed from this plot. As anticipated from our previous plots we find a bimodal distribution for the normal diet group and a skewed distribution for the protein restricted group. However, what is quite interesting is the almost identical behavior of the two distributions for positive values. This indicates that the treatment appears to have little or no effect on a subset of patients who appear to be "non-progressors"; i.e. patients without disease progression. Moreover, comparing the left-tail behavior of the two distributions we find a thinner tail for the protein restricted group, confirming the effect of restricted diet in slowing disease progression for "progressors". A similar conclusion was found in Greene (2000) who analyzed the same data using a generalized estimating equations approach. A comprehensive discussion of the data and related issues can be found there.

6. Discussion

The blocked Gibbs sampler we have presented offers several important computational advantages in fitting the GLMM. It handles non-conjugacy straightforwardly, mixes well, and involves simple updates for most of its parameters, including the important concentration parameter α. We also discussed the use of hybrid Monte Carlo in sampling the fixed effects, which can help to improve efficiency.

The ability to draw values directly from the random effects posterior distribution is another key feature of our approach. In the examples of Sections 4 and 5 we have illustrated graphical methods for translating this information into analysis for the random effects distributions. In particular, in the study of renal disease in Section 5 we have discussed methods for interpreting slope distributions and how they relate to disease progression.

REFERENCES

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

Blackwell, D. and MacQueen, J.B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1** 353-355.

BUSH, C.A. AND MACEACHERN, S.N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika* **83** 275-285.

DUANE, S., KENNEDY, A.D., PENDLETON, B.J. AND ROWETH, D. (1987). Hybrid Monte Carlo. *Physics Letters B* **195** 216-222.

ESCOBAR, M.D. (1988). Estimating the means of several normal populations by nonparametric estimation of the distribution of the means. Unpublished dissertation, Yale University.

ESCOBAR, M.D. (1994). Estimating normal means with a Dirichlet process prior. *J. Amer. Stat. Assoc.* **89** 268-277.

ESCOBAR, M.D. AND WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Stat. Assoc.* **90** 577-588.

ESCOBAR, M.D. AND WEST, M. (1998). Computing nonparametric hierarchical models. In *Practical Nonparametric and Semiparametric Bayesian Statistics* (D. Dey, P. Mueller and D. Sinha, eds.) 1-22. Springer Lecture Notes.

FAHRMEIR, L., AND TUTZ, G. (1994). *Multivariate statistical modelling based on generalized linear models.* New York: Springer-Verlag.

FERGUSON, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209-230.

FOLLMANN, D.A. AND LAMBERT, D. (1989). Generalized logistic regression by nonparametric mixing. *J. Amer. Stat. Assoc.* **84** 295-300.

GREENE, T. (2000). A model for a proportional treatment effect on disease progression. To appear in Biometrics.

GUSTAFSON, P. (1997). Large hierarchical Bayesian analysis of multivariate survival data. *Biometrics* **53** 230-242.

IBRAHIM, J.G. AND KLEINMAN, K.P. (1998). Semiparametric Bayesian methods for random effects models. In *Practical Nonparametric and Semiparametric Bayesian Statistics* (D. Dey, P. Mueller and D. Sinha, eds.). Springer Lecture Notes.

ISHWARAN, H. (1999). Applications of hybrid Monte Carlo to Bayesian generalized linear models: quasicomplete separation and neural networks. *J. Comp. Graph. Statist.* **8** 779-799.

ISHWARAN, H. AND JAMES, L.F. (2000a). Gibbs sampling methods for stick-breaking priors. To appear in *J. Amer. Stat. Assoc.* (March 2001).

ISHWARAN, H. AND JAMES, L.F. (2000b). Approximate Dirichlet process computing for finite normal mixtures: smoothing and prior information. Submitted.

ISHWARAN, H. AND ZAREPOUR, M. (2000a). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika* **87** 371-390.

ISHWARAN, H. AND ZAREPOUR, M. (2000b). Exact and approximate sum-representations for the Dirichlet process. Conditionally accepted by *Statist. Sinica.*

KENNEDY, A.D. (1990). The theory of hybrid stochastic algorithms. In *Probabilistic methods in quantum field theory and quantum gravity*, Ed. Damgaard P. H. et al., pp. 209-223. New York: Plenum Press.

KLAHR, S., LEVEY, A., BECK, G., CAGGIULA, A., HUNSIKCER, L., KUSEK J. AND STRIKER, G. (1994). The effects of dietary protein restriction and blood-pressure control on the progression of chronic renal disease. *New England Journal of Medicine* **330** 877-884.

KLEINMAN, K.P. AND IBRAHIM, J.G. (1998). A semiparametric Bayesian approach to the random effects model. *Biometrics* **54** 921-938.

LIANG, K.-Y., AND ZEGER, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13-22.

MACEACHERN, S.N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Commun. Statist.—Simula.* **23** 727-741.

MACEACHERN, S.N. (1998). Computational methods for mixture of Dirichlet process models In *Practical Nonparametric and Semiparametric Bayesian Statistics* (D. Dey, P. Mueller and D. Sinha, eds.). Springer Lecture Notes.

MACEACHERN, S.N AND MÜLLER, P. (1998). Estimating mixture of Dirichlet process models. *J. Comp. Graph. Statist.* **7** 223-238.

MCCULLAGH, P. AND NELDER, J.A. (1989). *Generalized linear models (second edition).* London: Chapman and Hall.

NEAL, R.M. (1994). An improved acceptance procedure for the hybrid Monte Carlo algorithm. *J. Computational Physics* **111** 194-203.

NEAL, R.M. (1996). *Bayesian learning for neural networks*. New York: Springer Verlag.

NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comp. Graph. Stat.* **9** 249-265.

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639-650.

WALKER, S. AND DAMIEN, P. (1998). Sampling methods for Bayesian nonparametric inference involving stochastic processes. In *Practical Nonparametric and Semiparametric Bayesian Statistics* (D. Dey, P. Mueller and D. Sinha, eds.) 243-254. Springer Lecture Notes.

WEST, M., MÜLLER, P., AND ESCOBAR, M.D. (1994). Hierarchical priors and mixture models, with applications in regression and density estimation. In *A tribute to D. V. Lindley* (A.F.M Smith and P.R. Freeman). John Wiley and Sons.

ZEGER, S, L.AND KARIM, M.R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *J. Amer. Stat. Assoc.* **86** 79-86.