Posterior Simulation in the Generalized Linear Model with Semiparmetric Random Effects

Subharup Guha*

*Harvard University, sguha@hsph.harvard.edu

Posterior Simulation in the Generalized Linear Model with Semiparmetric Random Effects

Subharup Guha

Abstract

Generalized linear mixed models with semiparametric random effects are useful in a wide variety of Bayesian applications. When the random effects arise from a mixture of Dirichlet process (MDP) model, normal base measures and Gibbs sampling procedures based on the Pólya urn scheme are often used to simulate posterior draws. These algorithms are applicable in the conjugate case when (for a normal base measure) the likelihood is normal. In the non-conjugate case, the algorithms proposed by MacEachern and Müller (1998) and Neal (2000) are often applied to generate posterior samples. Some common problems associated with simulation algorithms for non-conjugate MDP models include convergence and mixing difficulties.

This paper proposes an algorithm based on the Pólya urn scheme that extends the Gibbs sampling algorithms to non-conjugate models with normal base measures and exponential family likelihoods. The algorithm proceeds by making Laplace approximations to the likelihood function, thereby reducing the procedure to that of conjugate normal MDP models. To ensure the validity of the stationary distribution in the non-conjugate case, the proposals are accepted or rejected by a Metropolis-Hastings step. In the special case where the data are normally distributed, the algorithm is identical to the Gibbs sampler.

Posterior simulation in the generalized linear mixed model with semiparametric random effects

Subharup Guha[*]
sguha@hsph.harvard.edu

Abstract

Generalized linear mixed models with semiparametric random effects are useful in a wide variety of Bayesian applications. When the random effects arise from a mixture of Dirichlet process (MDP) model, normal base measures and Gibbs sampling procedures based on the Pólya urn scheme are often used to simulate posterior draws. These algorithms are applicable in the conjugate case when (for a normal base measure) the likelihood is normal. In the non-conjugate case, the algorithms proposed by MacEachern and Müller (1998) and Neal (2000) are often applied to generate posterior samples. Some common problems associated with simulation algorithms for non-conjugate MDP models include convergence and mixing difficulties.

This paper proposes an algorithm based on the Pólya urn scheme that extends the Gibbs sampling algorithms to non-conjugate models with normal base measures and exponential family likelihoods. The algorithm proceeds by making Laplace approximations to the likelihood function, thereby reducing the procedure to that of conjugate normal MDP models. To ensure the validity of the stationary distribution in the non-conjugate case, the proposals are accepted or rejected by a Metropolis-Hastings step. In the special case where the data are normally distributed, the algorithm is identical to the Gibbs sampler.

[*] Subharup Guha is Postdoctoral Research Fellow, Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, SPH2, 4th Floor, Boston, MA 02115. The author thanks Professor Steven MacEachern for introducing him to Dirichlet process models and for many insightful comments that helped improve the focus of the paper.

The performance of the technique is investigated using a Poisson regression example with semiparametric random effects. The algorithm is found to perform efficiently and reliably, even in problems where large sample results do not guarantee the success of the Laplace approximation. This is further demonstrated by a simulation study where most of the count data consist of small numbers. The technique is associated with substantial benefits relative to existing methods, both in terms of convergence properties and computational cost.

Keywords: Gibbs sampling, MCMC, Dirichlet process models, non-conjugate models, Pólya urn scheme, semiparametric Bayesian methods.

1 INTRODUCTION

This paper proposes a novel simulation algorithm for generating posterior samples from a generalized linear mixed model (GLMM) with semiparametric random effects that follow a mixture of Dirichlet process (MDP) model with normal base measure. The model has found application in such diverse areas of Bayesian analysis as survival analysis, spatial statistics, economics and modeling of physical systems in engineering.

1.1 Generalized linear mixed models

To motivate a description of the model, we begin with the normal linear random effects model of Laird and Ware (1982). For case $i = 1, \dots, n$, the model assumes the likelihood function:

$$Y_i \quad \sim \quad N(\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{\theta}_i, \sigma^2) \tag{1}$$

where the random effects are distributed as

$$\boldsymbol{\theta}_i \quad \overset{i.i.d.}{\sim} \quad N_q(\mathbf{0}, D), \quad i = 1, \dots, n \tag{2}$$

When the data consist of counts, a Poisson likelihood could be used:

$$Y_i \mid \mu_i \overset{ind}{\sim} Po(\mu_i), \quad \text{where } \log \mu_i = o_i + \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{\theta}_i \tag{3}$$

where o_i is a known offset that is possibly equal to zero. Some investigators prefer to include an additional, independent area-specific random effect in the expression for $\log \mu_i$. However, (3) is also reasonable because the stochastic mechanism of the Poisson model can be regarded as replacing the independent errors.

More generally, the data Y_i could represent integer outcomes, binary outcomes (presence or absence of a particular condition) or continuous measurements for which the normal assumption is not valid even after transformation. Zeger and Karim (1991) recommend the use of a generalized linear mixed model (GLMM) that replaces the normal likelihood with an exponential family distribution: $Y_i \mid \omega_i, \varsigma \overset{ind}{\sim} h(Y_i, \varsigma) \cdot \exp\left\{ \left(Y_i \, \omega_i - b(\omega_i) \right) / a(\varsigma) \right\}$, where ς is a dispersion parameter and for which the conditional expectation is given by $E[Y_i \mid \omega_i, \varsigma] = \mu_i = b'(\omega_i)$. Refer to McCullagh and Nelder, 1999, for the details. The conditional variance is $Var[Y_i \mid \omega_i, \varsigma] = \Upsilon(\omega_i) \, a(\varsigma)$, with the *variance function* $\Upsilon(\omega_i)$ defined as $b''(\omega_i)$. For an appropriate link function $g(\cdot)$, the *linear predictor* η_i is related to the mean μ_i as $\eta_i = g(\mu_i)$, and is defined as

$$\eta_i = o_i + \boldsymbol{x}_i' \boldsymbol{\beta} + \boldsymbol{z}_i' \boldsymbol{\theta}_i \tag{4}$$

where o_i is a known (and possibly zero) offset. The likelihood function can be regarded as a function of η_i and dispersion parameter ς:

$$Y_i \mid \eta_i, \varsigma \overset{ind}{\sim} h(Y_i, \varsigma) \cdot \exp\left\{ \left(Y_i \, \omega(\eta_i) - b(\eta_i) \right/ a(\varsigma) \right\} \tag{5}$$

Linear regression (1) is a special case of this class of models with identity link and $\varsigma = \sigma^2$. Poisson regression (3) corresponds to the log link and dispersion parameter $\varsigma = 1$. Logistic regression corresponds to a Bernoulli likelihood, logit (or probit) link and dispersion parameter $\varsigma = 1$ (McCullagh and Nelder, 1999, p. 30).

A normal prior is typically assumed for the GLMM fixed effects: $\boldsymbol{\beta} \sim N_p(\mu_\beta, \Sigma_\beta)$. A prior for the precision matrix of the random effects is $D^{-1} \sim Wishart\,(d_0, R_0)$, where the positive definite matrix R_0 is of order q and $d_0 \geq q$.

1.2 GLMMs with mixture of Dirichlet process random effects

Theoretical properties of the Dirichlet process are developed, among other sources, in Ferguson (1973), Blackwell and MacQueen (1973), Antoniak (1974), Sethuraman (1994) and Ishwaran and Zarepour (2002). General features of the MDP model and Gibbs sampling methods are investigated in Escobar (1994), MacEachern (1994), Escobar and West (1995), West et al. (1994) and Bush and MacEachern (1996).

Kleinman and Ibrahim (1998a and 1998b) introduce a semiparametric version of the GLMM by replacing assumption (2) for the random effects by a mixture of Dirichlet processes (MDP) model with normal base measure:

$$\begin{aligned} \boldsymbol{\theta}_i \mid P &\overset{i.i.d.}{\sim} P \\ P &\sim DP\,(M \cdot N_q(\mathbf{0}, D)) \end{aligned}$$

where $DP\,(M \cdot N_q(\mathbf{0}, D))$ denotes a Dirichlet process with base measure $N_q(\mathbf{0}, D)$ and mass parameter M. An overview of Dirichlet process models and the MCMC techniques is provided below.

To summarize, the GLMM with semiparametric random effects is:

$$\begin{aligned} Y_i \mid \eta_i, \varsigma &\overset{ind}{\sim} h(Y_i, \varsigma) \cdot \exp\left\{(Y_i\, \omega(\eta_i) - b(\eta_i)\,/a(\varsigma)\right\} \\ \eta_i &= o_i + \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{\theta}_i \\ \boldsymbol{\beta} &\sim N_p(\mu_\beta, \Sigma_\beta) \\ \boldsymbol{\theta}_i \mid P &\overset{i.i.d.}{\sim} P \\ P &\sim DP\,(M \cdot N_q(\mathbf{0}, D)) \\ D^{-1} &\sim Wishart\,(d_0, R_0) \end{aligned} \qquad (6)$$

A prior on ς is not needed for Poisson or logistic regression, but may be necessary in other situations. Except for the special case of the normal likelihood (1), model (6) assumes a non-conjugate likelihood for the normal base measure.

1.3 The Dirichlet process

Let α be a finite measure on R^m such that $\alpha = M \cdot G_0$, where G_0 is a probability measure and M is a positive real number. Suppose that $\boldsymbol{\theta}_i \mid P \overset{i.i.d.}{\sim} P$ for $i = 1, \dots, n$. The Dirichlet process $DP(M \cdot G_0)$ is a prior on the space of all distributions P on $(R^m, \Re^m)$. Given any measurable partition $\{A_1, \dots, A_k\}$ of R^m, the random vector has the distribution

$$(P(A_1), \dots, P(A_k)) \sim \mathcal{D}\left(M \cdot G_0(A_1), \dots, M \cdot G_0(A_k)\right)$$

where $\mathcal{D}(\cdot)$ represents the Dirichlet distribution. The distribution G_0 is called the *base measure* and M is called the *mass parameter* of the Dirichlet process. An MDP model assumes a prior on the base measure G_0, for example, by assuming that G_0 belongs to a parametric family and assigning appropriate priors to the hyperparameters. Refer to Freedman (1963), Ferguson (1973), Blackwell and MacQueen (1973) and Antoniak (1974) for further details.

Let δ_x denote a point mass at x and $\{V_j\}_{j=1}^{\infty}$ be i.i.d. beta$(1, M)$ random variables. Sethuraman (1994) gives a constructive definition of the Dirichlet process: $P \overset{a.s.}{=} \sum_{j=1}^{\infty} p_j \delta_{\boldsymbol{\theta}_j^*}$ where the $\boldsymbol{\theta}_j^*$'s are i.i.d. draws from the base measure G_0, and the probability masses $\{p_j\}_{j=1}^{\infty}$ are defined as $p_1 = V_1$ and as $p_j = V_j(1 - \sum_{i=1}^{j-1} p_i)$ for $j \geq 2$. The a.s. representation implies that P is almost surely discrete so that multiple cases share the same value of $\boldsymbol{\theta}_j^*$. We refer to this set as a *cluster*. The *cluster structure* can be inferred from $(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n)$ because of the a.s. discreteness of P. Additional theoretical properties of the Dirichlet process are discussed in Ishwaran and Zarepour (2002). The set of all possible cluster structures increases exponentially with n, making simulation-based computational techniques necessary for posterior inference.

The Dirichlet process induces a prior on the set of cluster structures. This can be easily seen from the Pólya urn scheme representation of $DP(M \cdot G_0)$:

$$
\begin{aligned}
\boldsymbol{\theta}_1 \quad & \sim G_0 \\
\boldsymbol{\theta}_j \quad & \mid \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{j-1} \begin{cases} \sim G_0 & \text{with probability } M/(M+j-1) \\ = \boldsymbol{\theta}_t & \text{with probability } 1/(M+j-1), \quad \text{for } t = 1, \ldots, j-1 \end{cases}
\end{aligned}
$$

where $j = 2, \ldots, n$. The Pólya urn scheme marginalizes over P and so the $\boldsymbol{\theta}_j$'s are not independent under this representation though they are identically distributed as G_0 (Blackwell and MacQueen, 1973; Ferguson, 1973).

The basic Gibbs sampler for MDP models (Escobar, 1994; MacEachern, 1994; West et al., 1994; Escobar and West, 1995; Bush and MacEachern, 1996) relies on the Pólya urn scheme to update for cases $i = 1, \ldots, n$, the random effect $\boldsymbol{\theta}_i$ conditional on the vector $\boldsymbol{\theta}_{-i} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_n)$. The algorithm proceeds as follows. Let vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ represent the outcomes. Let the vector $\boldsymbol{\theta}_{-i}$ consist of k^- clusters that are respectively associated with values $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{k^-}^*$ and number of cases $n_1^-, \ldots, n_{k^-}^-$, where $\sum_{j=1}^{k^-} n_j^- = n - 1$. Under the Pólya urn scheme, the full conditional of $\boldsymbol{\theta}_i$ is

$$
\boldsymbol{\theta}_i \quad \mid \boldsymbol{\theta}_{-i}, \boldsymbol{Y} \begin{cases} \sim G_0 \mid Y_i & \text{with probability } \propto M \cdot E_i \\ = \boldsymbol{\theta}_j^* & \text{with probability } \propto n_j^- \, [Y_i \mid \boldsymbol{\theta}_j], \quad j = 1, \ldots, k^- \end{cases} \tag{7}
$$

where $E_i = \int [Y_i \mid \boldsymbol{\theta}] \, dG_0(\boldsymbol{\theta})$. The first line in (7) corresponds to case i starting its own cluster. The second line corresponds to case i joining one of the k^- clusters obtained after excluding case i from the data set.

Following an update of all n cases, Bush and MacEachern (1996) recommend adding an extra step that generates, conditional on the cluster structure, the distinct $\boldsymbol{\theta}_j^*$'s associated with the k clusters. This step considerably improves the mixing of the sampler. The reader is referred to Dey, Müller and Sinha (1998) for Gibbs sampling strategies developed for conjugate models.

1.4 Non-conjugate MDP models

The Gibbs sampler is easy to implement in conjugate MDP models because the integral E_i in (7) can be exactly computed. For non-conjugate models, the integral does not have a computationally closed form and generation of a new value from the posterior distribution $[G_0 \mid Y_i]$ is less straightforward.

The "no gaps" algorithm of MacEachern and Müller (1998) extends the basic Gibbs sampling algorithm to non-conjugate MDP models. A description of the algorithm involves quantities that will be used throughout the rest of this paper. Given the cluster memberships, the *allocation variable* c_i is defined as equal to j if case i belongs to cluster j, where $i = 1, \ldots, n$ and $j = 1, \ldots, k$. For updating the random effect $\boldsymbol{\theta}_i$, we exclude the i^{th} case from the data set and define the variables $\boldsymbol{\theta}^*_1, \ldots, \boldsymbol{\theta}^*_{k^-}$ and $n^-_1, \ldots, n^-_{k^-}$ as in (7). Additionally, we define n_j as the number of cases *including* the case i that belong to cluster j, so that $\sum_{j=1}^{k} n_j = n$. This implies that $k = k^- + 1$ if $n_{c_i} = 1$ and $k = k^-$ if $n_{c_i} > 1$. Given an integer k^* and values $\boldsymbol{\theta}^*_1, \ldots, \boldsymbol{\theta}^*_{k^*+1}$, let the random variable $\boldsymbol{\phi}_{k^*}$ be defined as

$$\boldsymbol{\phi}_{k^*} = \begin{cases} \boldsymbol{\theta}^*_{k^*+1} & \text{with probability} \propto \frac{M}{k^*+1} \, [Y_i \mid \boldsymbol{\theta}_{k^*+1}] \\ \boldsymbol{\theta}^*_j & \text{with probability} \propto n^-_j \, [Y_i \mid \boldsymbol{\theta}_l], \quad j = 1, \ldots, k^* \end{cases} \tag{8}$$

With this notation, the "no gaps" algorithm can be described as follows. At the start of the cycle of updates for the random effects $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$, augment the set of values $\boldsymbol{\theta}^*_1, \ldots, \boldsymbol{\theta}^*_k$ by generating $n - k$ additional values, $\boldsymbol{\theta}^*_{k+1}, \ldots, \boldsymbol{\theta}^*_n$, from the base measure $N_q(\mathbf{0}, D)$. For cases $i = 1, \ldots, n$:

(i) If $n_{c_i} > 1$, set $k^* = k$ in (8) and generate $\boldsymbol{\theta}_i \sim \boldsymbol{\phi}_{k^*}$.

(ii) If $n_{c_i} = 1$, leave $\boldsymbol{\theta}_i$ unchanged with probability $(k-1)/k$. Otherwise, swap the labels of the c_i^{th} and k^{th} clusters (i.e. $c_i \rightleftharpoons k$) and the associated $\boldsymbol{\theta}^*_j$ values (i.e. $\boldsymbol{\theta}^*_{c_i} \rightleftharpoons \boldsymbol{\theta}^*_k$), set $k^* = k - 1$ in (8), and generate $\boldsymbol{\theta}_i \sim \boldsymbol{\phi}_{k^*}$.

The "no gaps" algorithm avoids the integral E_i in (7), sometimes at the cost of slower convergence and mixing properties (MacEachern, 1998). The simulation study presented in Neal (2000) suggests that the “no gaps” algorithm has difficulty starting new clusters. Walker, Damien, Laud and Smith (1999), Neal (2000) and Ishwaran and James (2001) discuss alternative approaches that do not rely on the Pólya urn scheme. Recently, Papaspiliopoulos and Roberts (2006) have investigated retrospective MCMC methods for MDP models.

The auxiliary Gibbs algorithm of Neal (2000) proceeds as follows. Given an integer m, for cases $i = 1, \ldots, n$:

(i) Let $h = k^- + m$.

(ii) If $n_{c_i} > 1$, exclude case i from the data to get $k^- = k$ clusters. Label these clusters using $\{1, \ldots, k^-\}$ and their associated $\boldsymbol{\theta}_j^*$ values as $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{k^-}$. Independently sample $h - k^-$ additional draws from the base measure $N_q(\mathbf{0}, D)$ and label them as $\boldsymbol{\phi}_{k^-+1}, \ldots, \boldsymbol{\phi}_h$. Go to step *(iv)*.

(iii) If $n_{c_i} = 1$, label the c_i^{th} cluster as $(k^- + 1)$. Label the clusters that remain after excluding the i^{th} case using $\{1, \ldots, k^-\}$ and their associated $\boldsymbol{\theta}_j^*$ values as $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{k^-}$. If $h \geq (k^- + 2)$, independently sample $(h - k^- 1)$ additional draws from the base measure $N_q(\mathbf{0}, D)$ and label them as $\boldsymbol{\phi}_{k^-+2}, \ldots, \boldsymbol{\phi}_h$.

(iv) Sample a new value for c_i as follows

$$P(c_i = j) \propto \begin{cases} \frac{n_j^-}{n-1+M} \cdot [Y_i \mid \boldsymbol{\phi}_j], & j = 1, \ldots, k^- \\ \frac{M/m}{n-1+M} \cdot [Y_i \mid \boldsymbol{\phi}_j], & j = (k^- + 1), \ldots, h \end{cases} \tag{9}$$

Drop all $\boldsymbol{\phi}_j$'s not associated with a cluster.

In this paper, I propose an algorithm based on the Pólya urn scheme for generating posterior draws from the non-conjugate MDP model in (6). The key idea is to make Laplace approximations to likelihood function (5) and to use the resulting normal posterior for the random effects as the proposal distribution in a Metropolis-Hastings step. The details of this technique are provided in Section 2. Section 3.1 presents an example where Poisson regression with semiparametric random effects is applied to analyze heart disease incidence rates in New South Wales, Australia. Section 3.2 uses a transformed version of the data generated by Neal (2000) to investigate the benefits of the algorithm relative to some of the existing ones for non-conjugate MDP models. The simulation study demonstrates the effectiveness of the algorithm in problems where asymptotic results do not guarantee the success of the Laplace approximation. For the examples in Section 3 where the approximation's accuracy is somewhat greater (e.g. moderately large counts in Poisson regression), the efficiency of the algorithm approaches that of conjugate MDP models, and it substantially outperforms the "no gaps" and auxiliary Gibbs algorithms.

2 A NEW ALGORITHM

2.1 The Laplace approximation

For models belonging to the exponential family, the Laplace approximation applies a linearized version of the link function $g(\cdot)$ to the data, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. Conditional on the model parameters, the *working value* for each case is defined as

$$y_i = \eta_i + \frac{\partial \eta_i}{\partial \mu_i} \cdot (Y_i - \mu_i), \quad i = 1, \ldots, n \tag{10}$$

The *working weight* is defined as $w_i = \{\Upsilon(\mu_i)\}^{-1} \left(\partial \mu_i / \partial \eta_i\right)^2$ where the variance function $\Upsilon(\cdot)$ is defined in Section 1. We obtain (Harville, 1977):

$$y_i \overset{indep}{\sim} N\left(\eta_i, w_i^{-1}\right) \tag{11}$$

where $(y_1, \ldots, y_n)$ represents the vector of working values, and not the data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. Refer to McCullagh and Nelder (1999, p. 40) for an explanation. For the special case of normal likelihoods (1), the Laplace approximation is exact with $Y_i = y_i$, and the MCMC strategy proposed in this section is identical to the Gibbs sampler (7) for conjugate normal MDP models.

The Laplace approximation forms the basis of many well-known numerical and simulation-based methods. The approximation is not restricted to problems where approximate normality is achieved due to large sample sizes. For the exponential family, one of the reasons for the remarkable success of the technique is that the likelihood function (regarding as usual the data as fixed and the parameters as random) is guaranteed to be log-concave for canonical link functions. Since the MLE belongs to the interior of the parameter space under fairly mild conditions (e.g. if $Y > 0$ when Y is Poisson with log link, or if $0 < Y < m$ when Y is binomial with logit link and m, the number of trials, exceeds one), the log-likelihood is often approximated reasonably well by a quadratic function even when asymptotic results are not applicable. For example, Figure 1 plots the Poisson log-likelihood of $Y = 1$ versus the logarithm of the mean.

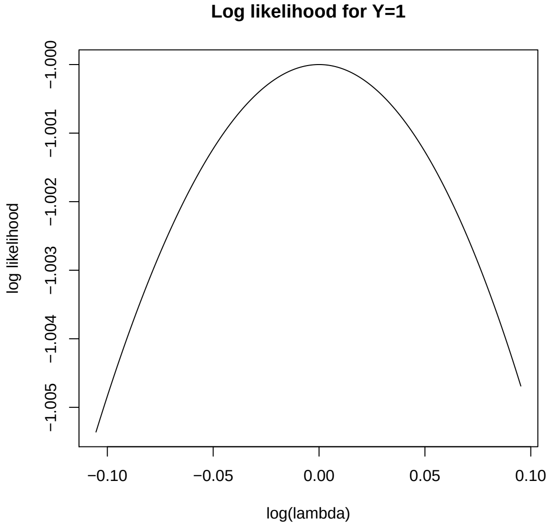

Figure 1: Log-likelihood function for $Y \sim Po(\lambda)$ versus $\theta = \log \lambda$ when $Y = 1$.

2.2 Description of the algorithm

I have assumed below that the mass parameter M is fixed. However, a prior on M can be easily accommodated by making minor changes to the algorithm. See MacEachern (1998) for an approach that marginalizes over M, and Escobar and West (1995) for an approach that assumes a gamma prior for the mass parameter.

Step (A): Generating the allocation variables and random effects

Let $\boldsymbol{\varphi}$ denote the set of parameters $(\boldsymbol{\beta}, D, \varsigma)$. Applying the Pólya urn scheme representation (7) to the model (6), we obtain an exact expression for the full conditional of $\boldsymbol{\theta}_i$:

$$\boldsymbol{\theta}_i \quad | \ \boldsymbol{\theta}_{-i}, \boldsymbol{\varphi}, \boldsymbol{Y} \begin{cases} \sim [\boldsymbol{\theta} \mid Y_i, \boldsymbol{\varphi}] \ \text{ with probability} \propto M \cdot E_i \\ \\ = \boldsymbol{\theta}_j^* \ \text{ with probability} \propto n_j^- \cdot [Y_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}_j^*, \varsigma] \quad \text{for } j = 1, \ldots, k^- \end{cases} \tag{12}$$

where $E_i = [Y_i \mid \boldsymbol{\varphi}] = \int [Y_i \mid \boldsymbol{\theta}, \boldsymbol{\varphi}] \cdot N(\boldsymbol{\theta} \mid \mathbf{0}, D) \cdot d\boldsymbol{\theta}$. The second line in (12) can be easily computed for the non-conjugate case. However, for a normal base measure, the integral E_i and distribution $[\boldsymbol{\theta} \mid Y_i, \boldsymbol{\varphi}]$ that appear on the first line (corresponding to case i starting its own cluster) are generally not analytically available for the likelihood (5). We apply theoretical properties of the exponential family to approximate the integral E_i via the normal approximation (11). This simplifies the procedure for proposals to that of the basic Gibbs sampler, for which E_i is known and for which $\boldsymbol{\theta}_{k^-+1}^*$ corresponding to a new cluster can be easily generated using the normal approximation to $[\boldsymbol{\theta} \mid Y_i, \boldsymbol{\varphi}]$. A Metropolis-Hastings acceptance-rejection step is then applied to compensate for the approximation.

Before describing the Step (A) procedure in detail, it will be helpful to define the quantities that are iteratively evaluated. As before, y_i denotes working value (10) and Y_i denotes the outcome. Define the residual $\xi_i = y_i - o_i - \boldsymbol{x}_i'\boldsymbol{\beta}$. Applying normal approximation (11) to the working value likelihood, it can be shown that E_i is approximated by

$$\hat{E}_i = \left(\frac{\partial \eta_i}{\partial \mu_i} \right) \cdot N \left(\xi_i \mid 0, w_i^{-1} + \boldsymbol{z}_i' D \boldsymbol{z}_i \right) \tag{13}$$

where $\partial \eta_i / \partial \mu_i$ is the Jacobian of the transformation that maps the data Y_i to the working value y_i. $N_r(\boldsymbol{\xi} \mid \boldsymbol{a}, B)$ denotes the density of the r-variate normal distribution with mean $\boldsymbol{a}$ and variance matrix B evaluated at the point $\boldsymbol{\xi}$.

Remove the i^{th} case from the data set and label the remaining clusters using $\{1, \ldots, k^-\}$. Imagine that case i now begins a new, $(k^- + 1)^{th}$ cluster. Then the posterior precision $\Omega_{k^-+1}^{-1}$ of the normal approximation to the full conditional of $\boldsymbol{\theta_i}$ is

$$\Omega_{k^-+1}^{-1} = D^{-1} + \boldsymbol{z}_i' w_i \boldsymbol{z}_i \tag{14}$$

and its mean $\boldsymbol{\mu}_{k^-+1}^*$ equals

$$\boldsymbol{\mu}_{k^-+1}^* = \Omega_{k^-+1} \cdot \boldsymbol{z}_i' w_i \left(y_i - o_i - \boldsymbol{x}_i' \boldsymbol{\beta} \right). \tag{15}$$

An approximation to the Pólya urn scheme update (12) is then:

$$\boldsymbol{\theta}_i \quad | \, \boldsymbol{\theta}_{-i}, \boldsymbol{\varphi}, \boldsymbol{Y} \begin{cases} \overset{approx}{\sim} N_q \left(\boldsymbol{\mu}_{k^-+1}^*, \, \Omega_{k^-+1} \right) \text{ with probability } \propto M \cdot \hat{E}_i \\ = \boldsymbol{\theta}_j^* \text{ with probability } \propto n_j^- \cdot [Y_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}_j^*, \varsigma] \quad \text{for } j = 1, \ldots, k^- \end{cases} \tag{16}$$

where $\hat{E}_i$ is defined in (13).

<u>**Procedure**</u> For case $i = 1, \ldots, n$:

(i) Remove the i^{th} case from the data set and label the clusters using $\{1, \ldots, k^-\}$. Let $c_{i,0}$ be the current allocation of the case i under this labeling scheme, so that $c_{i,0} \leq k^-$ if $n_{c_{i,0}} > 1$, and $c_{i,0} = k^- + 1$ if $n_{c_{i,0}} = 1$. If $c_{i,0} = k^- + 1$, evaluate the current conditional mean $\boldsymbol{\theta}_{(k^-+1),0}^*$ and conditional precision $\Omega_{(k^-+1),0}$ using (14) and (15).

(ii) A new value of $\boldsymbol{\theta}_i$, denoted by $\boldsymbol{\theta}_{(k^-+1),1}^*$, is proposed using approximation (16). Let the corresponding allocation variable be $c_{i,1}$ so that $c_{i,1} = k^- + 1$ if case i begins a new cluster. If $c_{i,1} = k^- + 1$, use $\boldsymbol{\theta}_{(k^-+1),1}^*$ to compute the conditional mean $\boldsymbol{\theta}_{(k^-+1),1}^*$ and conditional precision $\Omega_{(k^-+1),1}$ and evaluate

$$\rho_1 = \frac{[Y_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}_{(k^-+1),1}^*, \varsigma] \cdot N_q(\boldsymbol{\theta}_{(k^-+1),1}^* \mid \mathbf{0}, D)}{\hat{E}_{i,1} \cdot N_q \left(\boldsymbol{\theta}_{(k^-+1),1}^* \mid \boldsymbol{\mu}_{(k^-+1),0}^*, \, \Omega_{(k^-+1),0} \right)}$$

If the old allocation variable $c_{i,0} = k^- + 1$, also evaluate

$$\rho_0 = \frac{[Y_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}_{(k^-+1),0}^*, \varsigma] \cdot N_q(\boldsymbol{\theta}_{(k^-+1),0}^* \mid \mathbf{0}, D)}{\hat{E}_{i,0} \cdot N_q \left(\boldsymbol{\theta}_{(k^-+1),0}^* \mid \boldsymbol{\mu}_{(k^-+1),1}^*, \, \Omega_{(k^-+1),1} \right)}$$

(iii) Accept the proposal $\boldsymbol{\theta}_i = \boldsymbol{\theta}^*_{(k^-+1),1}$ with a probability of $\min\{1, \varpi_i\}$, where

$$\varpi_i = \begin{cases} 1 \quad \text{if } c_{i,0} \le k^- \text{ and } c_{i,1} \le k^- \\ \rho_1 \quad \text{if } c_{i,0} \le k^- \text{ and } c_{i,1} = k^- + 1 \\ 1/\rho_0 \quad \text{if } c_{i,0} = k^- + 1 \text{ and } c_{i,1} \le k^- \\ \rho_1/\rho_0 \quad \text{if } c_{i,0} = k^- + 1 \text{ and } c_{i,1} = k^- + 1 \end{cases} \tag{17}$$

See the Appendix for a proof that this strategy has the right stationary distribution.

Step (B): Generating $\{\boldsymbol{\theta}^*_j\}_{j=1}^k$ conditional on the allocation variables

This step generates the $\{\boldsymbol{\theta}^*_j\}_{j=1}^k$ associated with the clusters without changing the cluster memberships. The move was originally proposed for the conjugate MDP model by Bush and MacEachern (1996). Approximation (11) allows us to apply it in a straightforward manner to the non-conjugate model (6) to further improve the mixing of the chain.

For cluster $j = 1, \ldots, k$, let Z_j be the n_j by q matrix formed by subsetting the rows of matrix Z that correspond to cluster j (i.e. rows i for which $c_i = j$). Let W_j be the n_j by n_j diagonal matrix of the working weights of these subsetted cases. After applying the Laplace approximation, the precision matrix of the full conditional of $\boldsymbol{\theta}^*_j$ is

$$\Omega_j^{-1} = D^{-1} + Z_j' W_j Z_j$$

Let $\boldsymbol{\theta}^*_{j,0}$ be the current value of $\boldsymbol{\theta}^*_j$ and $\Omega_{j,0}$ be the current value of Ω_j. We apply an overdispersed random walk proposal (Gelman, Roberts and Gilks, 1995):

$$\boldsymbol{\theta}^*_{j,1} \sim N_q(\boldsymbol{\theta}^*_{j,0},\, a_q^2\, \Omega_{j,0}), \quad \text{where } a_q \approx 2.4/\sqrt{q}$$

The generated $\boldsymbol{\theta}^*_{j,1}$ is used to compute the updated conditional variance $\Omega_{j,1}$. The procedure is repeated for the clusters $j = 1, \ldots, k$ to obtain the set of proposals $\{\boldsymbol{\theta}^*_{j,1}\}_{j=1}^k$. The set of proposals is jointly accepted with the Metropolis-Hastings probability

$$
\begin{aligned}
&A\left(\{\boldsymbol{\theta}^*_{j,0}\}_{j=1}^{k}, \{\boldsymbol{\theta}^*_{j,1}\}_{j=1}^{k}\right) = \\
&\qquad \min\left\{1, \frac{\prod_{i=1}^{n}[Y_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}^*_{c_i,1}, \varsigma] \cdot \prod_{j=1}^{k} N_q(\boldsymbol{\theta}^*_{j,1} \mid \mathbf{0}, D) \cdot \prod_{j=1}^{k} N_q(\boldsymbol{\theta}^*_{j,0} \mid \boldsymbol{\theta}^*_{j,1},\, a_q^2\Omega_{j,1})}{\prod_{i=1}^{n}[Y_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}^*_{c_i,0}, \varsigma] \cdot \prod_{j=1}^{k} N_q(\boldsymbol{\theta}^*_{j,0} \mid \mathbf{0}, D) \cdot \prod_{j=1}^{k} N_q(\boldsymbol{\theta}^*_{j,1} \mid \boldsymbol{\theta}^*_{j,0},\, a_q^2\Omega_{j,0})}\right\} \qquad (18)
\end{aligned}
$$

Step (C): Generating $\boldsymbol{\beta}$

A new value of $\boldsymbol{\beta}$ is proposed using a normal random-walk proposal. Define the matrix Q with typical element q_{ij} equal to $\boldsymbol{z}_i' D \boldsymbol{z}_j$ if $c(i) = c(j)$, and equal to zero otherwise. Set the n by n matrix $T = W^{-1} + Q$, where W is the diagonal matrix of working weights. After applying approximation (11), the covariance matrix Ω_β of the full conditional of $\boldsymbol{\beta}$ satisfies

$$\Omega_\beta^{-1} = \Sigma_\beta^{-1} + X'T^{-1}X \qquad (19)$$

Let $\boldsymbol{\beta}_0$ be the current value of $\boldsymbol{\beta}$ and let Ω_{β_0} be the corresponding value of the covariance matrix. We propose a new value $\boldsymbol{\beta}_1$ using an overdispersed random walk proposal:

$$\boldsymbol{\beta}_1 \sim N_p(\boldsymbol{\beta}_0, a_p^2\,\Omega_{\beta_0}), \text{ where } a_p \approx 2.4/\sqrt{p}$$

The covariance matrix Ω_{β_1} is computed using $\boldsymbol{\beta}_1$ and the proposed move accepted with the probability:

$$A\left(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1\right) = \min\left\{1, \frac{[Y_i \mid \boldsymbol{\beta}_1, \boldsymbol{\theta}_i, \varsigma] \cdot N_p(\boldsymbol{\beta}_1 \mid \boldsymbol{\mu}_\beta, \Sigma_\beta) \cdot N_p(\boldsymbol{\beta}_0 \mid \boldsymbol{\beta}_1, a_p^2\,\Omega_{\beta_1})}{[Y_i \mid \boldsymbol{\beta}_0, \boldsymbol{\theta}_i, \varsigma] \cdot N_p(\boldsymbol{\beta}_0 \mid \boldsymbol{\mu}_\beta, \Sigma_\beta) \cdot N_p(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_0, a_p^2\,\Omega_{\beta_0})}\right\}$$

Remark: (i) Steps (B) and (C) can be merged to jointly update $\boldsymbol{\beta}$ and $\{\boldsymbol{\theta}^*_j\}_{j=1}^{k}$. *(ii)* The Appendix contains a note on the computation of Ω_β^{-1} that avoids the inversion of non-diagonal matrices of order n.

Step (D): Generating the matrix D

We simulate from the full conditional

$$D^{-1} \mid \boldsymbol{\theta}^*_1, \ldots, \boldsymbol{\theta}^*_j, \boldsymbol{Y} \sim Wishart\left(d_0 + k, \left\{R_0^{-1} + \sum_{j=1}^{k} \boldsymbol{\theta}^*_j \cdot (\boldsymbol{\theta}^*_j)'\right\}^{-1}\right)$$

Step (E): Generating the dispersion parameter ς

The details of this step depends on the particular form of the exponential family and the prior on ς. For Poisson and logistic regression models, this step is not needed as mentioned in Section 1 because $\varsigma = 1$.

3 APPLICATIONS

3.1 Ischemic heart disease in New South Wales, Australia

The Spatial Environmental Epidemiology in New South Wales (SEE NSW) project yielded outcome data on ischemic heart disease (IHD) abstracted from daily separation records from all public and private hospitals in New South Wales, Australia, during the period July 1, 1996 to June 30, 2001. Patient reported residential postcode was used to assign the geographical location of hospitalization for IHD. Population data were obtained from census information collected by the Australian Bureau of Statistics (ABS) and inter-censal estimates, called Estimated Residential Populations (ERPs), provided for July 1st of each non-census year.

The goal of this study is to explore the association of IHD with an index of socioeconomic disadvantage, SEIFA (Socio-Economic Indexes for Areas), provided by the Australian Bureau of Statistics for each postal area. This score reflects relatively low educational attainment and income, high unemployment, and jobs in relatively unskilled occupations. The higher an area's SEIFA value the less disadvantaged the area is compared with other areas (Breslow and Day, 1987). The SEIFA scores were re-centered around zero to justify the specification of independent priors on the fixed effects, and also scaled by a factor of 10^3. For an analysis accounting for the spatial association of all 591 postcodes using CAR models (Besag et al., 1991; also see Banerjee et al., 2004), refer Burden et al. (2005), Guha and Ryan (2006) and Guha et al. (2006).

To alleviate concerns that the proposed algorithm is effective only in problems where the marginal distributions of the random effects are approximately normal due to the central limit theorem, I chose a random subset of 50 postcodes and analyzed only the IHD hospitalizations for the first year of the study (July 1, 1996 to June 30, 1997). The five-number summary of the data is provided in Table 1 and Figure 2 displays the histogram of the IHD counts. Although the graph reveals some large values, most of the values are small to moderately large. Section 3.2 discusses an example where asymptotic theory plays an even smaller role, with all the outcomes being less than 18 and only four of them exceeding 5.

With subscript i denoting postcode, let Y_i denote the number of IHD hospitalizations in the i^{th} postcode ($i = 1, \ldots, 50$), among the N_i subjects at risk during the first year. Let x_i denote the SEIFA index of postcode i. Because IHD is relatively rare, I assumed the Poisson approximation to the binomial and fit the model:

$$Y_i \sim Po(\mu_i), \text{ where } \eta_i = \log(\mu_i) = \log(N_i) + \beta_1 + x_i\beta_2 + \theta_i \tag{20}$$

and where β_1 is the intercept, β_2 is the coefficient associated with the SEIFA index and θ_i is the random effect associated with postcode i. For the priors, I assumed:

$$\begin{aligned} \boldsymbol{\beta} = (\beta_1, \beta_2)' &\sim N_p(\mu_\beta, \Sigma_\beta) \\ \theta_i \mid P &\overset{i.i.d.}{\sim} P \\ P &\sim DP\left(M \cdot N_q(0, D)\right) \\ D^{-1} &\sim Wishart\left(d_0, R_0\right) \end{aligned} \tag{21}$$

where D is one-dimensional in this example. For the hyperparameters of the inverse-Wishart prior, I chose R_0 based on a parametric analysis using a different subset of postcodes, and set $d_0 = 10$. A relatively non-informative prior was assumed for the fixed effect $\boldsymbol{\beta}$.

The algorithm described in Section 2 was used to generate posterior samples. As indicated in Section 2, Steps (B) and (C) were combined into a single step that jointly generated the fixed effects $\boldsymbol{\beta}$ and the k distinct random effects $\{\boldsymbol{\theta}_j^*\}_{j=1}^k$ conditional on the cluster memberships. An initial set of

	Minimum	1^{st} quartile	Median	3^{rd} quartile	Maximum
IHD cases	0	8	20	78	452

Table 1: Five-number summary of the IHD hospitalizations in the 50 randomly selected postcodes.

5,000 samples was discarded as burn-in and the subsequent 100,000 samples used for posterior inferences. The acceptance rate of the cluster proposals in Step (A), averaged over the 50 postcodes and conditional on either $c_{i,0}$ or $c_{i,1}$ being equal to $k^- + 1$ (otherwise, the proposals are always accepted) was approximately 36%.

Based on the MCMC samples and the the subsetted data, an estimate of the posterior mean of the SEIFA coefficient, $E[\beta_2|\boldsymbol{Y}]$, is -1.625 with an estimated standard error of 0.0018. A 95% posterior credible interval for β_2 is $(-2.22, -1.011)$. The interval excludes zero, suggesting that the risk of heart disease increases with socioeconomic disadvantage. The conclusion confirms the relationship between socioeconomic status and risk of IHD observed by previous studies (e.g. Marmot et al., 1997).

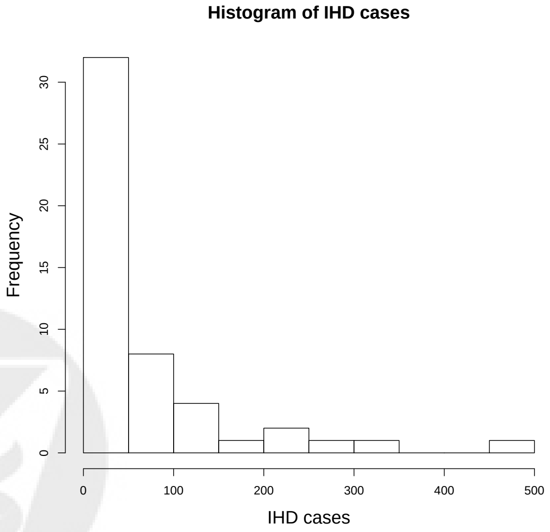

Figure 2: Number of IHD hospitalizations in the 50 randomly selected postcodes during the period July 1, 1996 to June 30, 1997.

3.2 Simulation study

The performance of the proposed algorithm was tested using a framework similar to that of Neal (2000), where the data consist of nine numbers generated from the standard normal distribution: −1.48, −1.40, −1.16, −1.08, −1.02, +0.14, +0.51, +0.53, +0.78. Neal's paper uses the conjugate normal MDP model to analyze these data. I added the constant $\beta_1 = 2$ to these numbers and exponentiated the values to get 1.68, 1.82, 2.32, 2.51, 2.66, 8.5, 12.3, 12.55, 16.12. I then generated Poisson variables with these numbers as the means to obtain the data $(Y_1, \ldots, Y_9) = (1, 1, 2, 5, 1, 12, 17, 13, 12)$. The model used to analyze the data was:

$$\begin{aligned} Y_i \mid \eta_i &\overset{ind}{\sim} Po(e^{\eta_i}) \\ \eta_i &= \beta_1 + \theta_i \\ \theta_i \mid P &\overset{i.i.d.}{\sim} P \\ P &\sim DP\,(M \cdot N(0,1)) \\ M &= 1 \end{aligned} \tag{22}$$

where $\beta_1 = 2$ is known. The algorithms "no gaps" and auxiliary Gibbs with $m = 1$, 2 and 30 were compared with the algorithm of Section 2. The criteria used to evaluate the algorithms were the computational cost per iteration (in microseconds) and the *autocorrelation times* for the following variables: number of clusters, k, and the random effects $\theta_1, \ldots, \theta_9$. The autocorrelation time of an MCMC chain (refer to Ripley, 1987, section 6.3) is defined as one plus twice the sum of the correlations from lag one upwards. It is interpreted as the factor by which the MCMC sample size is effectively reduced, relative to an i.i.d. posterior sample, for the computation of empirical average estimates of posterior expectations.

Using the normal proposals algorithm of Section 2, 1000 initial iterations were run to obtain reasonable parameter values. With this vector as the initial iterate of all the samplers, 20000 additional draws were generated using each sampler and its performance was evaluated using the sample of generated values and the above criteria for comparisons. The results for $\beta_1 = 2$ are displayed in Table 3. We find the normal proposals algorithm outperforms the "no gaps" algorithm for these data. In particular, the autocorrelation time for the number of clusters, k, is smaller implying that the normal proposals algorithm has a significantly better ability to begin new clusters. The auxiliary Gibbs algorithm with $m = 30$ involves much higher computational costs. However, the differences between the normal proposals algorithm and the auxiliary Gibbs algorithms with $m = 1$ and $m = 2$ is less clear. This suggests that there are data sets and models (depending on the assumed priors for the fixed effects and hyperparameters) for which each one of these three algorithms would outperform the others. However, the normal proposals algorithm is found to perform reliably in a broad range of problems.

Data sets with larger counts tend to favor the normal proposals algorithm, and its performance approaches that of conjugate MDP models because of the greater validity of the Laplace approximation. For further comparisons between the normal proposals and auxiliary Gibbs algorithms, I added the constant $\beta_1 = 4$ to the standard normal sample in Neal's paper, exponentiating the numbers and generating Poisson variables with these means to obtain $(Y_1, \ldots, Y_9) =$ (10, 18, 22, 20, 26, 68, 96, 89, 110). The data were analyzed using the model (22) with β_1 set equal to 4. Table 3 displays the simulation results. In this example, the normal proposals algorithm is clearly more effective than auxiliary Gibbs with either $m = 1$ or $m = 2$.

	Normal proposals	**“No gaps”**	**AG,** $m = 1$	**AG,** $m = 2$	**AG,** $m = 30$
Cost per iteration (10^{-3} s)	62.5	53.9	62.3	69.3	153.6
k	2.5	9.0	2.3	2.0	1.8
θ_1	8.5	12.5	6.6	5.9	5.1
θ_2	8.2	12.0	6.7	5.9	5.1
θ_3	6.0	8.5	4.8	4.3	3.6
θ_4	3.0	3.4	2.4	2.1	1.9
θ_5	8.2	12.3	6.6	6.0	5.2
θ_6	4.0	6.7	3.6	3.4	3.5
θ_7	3.4	9.0	4.9	3.7	3.3
θ_8	4.0	7.4	4.5	3.9	3.8
θ_9	4.0	6.8	3.6	3.3	3.3

Table 2: A comparison of the normal proposals, “no gaps” and auxiliary Gibbs (AG) algorithms for $\beta_1 = 2$. See the text for an explanation.

	Normal proposals	**AG,** $m = 1$	**AG,** $m = 2$
Cost per iteration (10^{-3} s)	77.7	70.6	77.8
k	2.8	6.1	4.4
θ_1	6.7	10.1	6.5
θ_2	3.9	5.0	4.1
θ_3	5.3	7.0	5.4
θ_4	4.9	6.1	4.9
θ_5	4.2	6.3	4.2
θ_6	3.4	12.2	5.3
θ_7	4.3	7.4	6.9
θ_8	3	5.8	4.7
θ_9	3.4	9.0	6.8

Table 3: A comparison of the normal proposals and auxiliary Gibbs (AG) algorithms for $\beta_1 = 4$. See the text for an explanation.

4 CONCLUSIONS

The paper proposes a new Metropolis-Hastings algorithm for generalized linear mixed models having non-conjugate mixture of Dirichlet process random effects with normal base measure. The sampler exploits special properties of the exponential family to make good proposals for new cluster structures using the Laplace approximation. The method is remarkably effective in a large number of problems and significantly reduces autocorrelation times, relative to other methods, for various posterior quantities of interest. The gains are found to be substantial for even data sets with modest asymptotic effects.

In addition to the examples discussed in this paper, the algorithm was found to be equally effective with multivariate random effects. Moreover, the technique can be extended to include multivariate outcomes. These and other extensions will be the focus of my future work.

5 APPENDIX

5.1 Proof of the Step (A) procedure for generating new clusters

The current and proposed values of the allocation variables can be categorized into four cases:

(i) $c_{i,0} \leq k^-$ and $c_{i,1} \leq k^-$ *(ii)* $c_{i,0} \leq k^-$ and $c_{i,1} = k^- + 1$

(iii) $c_{i,0} = k^- + 1$ and $c_{i,1} \leq k^-$ *(iv)* $c_{i,0} = k^- + 1$ and $c_{i,1} = k^- + 1$

Case (i): The Metropolis-Hastings acceptance probability equals $\min\{1, \psi_1\}$, where

$$\psi_1 = \frac{n^-_{c_{i,1}} \cdot [Y_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}^*_{c_{i,1}}, \varsigma]}{\sum_{j=1}^{k^-} n^-_j \cdot [Y_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}^*_j, \varsigma] + M \cdot E_i} \cdot \frac{\sum_{j=1}^{k^-} n^-_j \cdot [Y_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}^*_j, \varsigma] + M \cdot \hat{E}_i}{n^-_{c_{i,1}} \cdot [Y_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}^*_{c_{i,1}}, \varsigma]}$$

$$\times \frac{\sum_{j=1}^{k^-} n^-_j \cdot [Y_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}^*_j, \varsigma] + M \cdot E_i}{n^-_{c_{i,0}} \cdot [Y_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}^*_{c_{i,0}}, \varsigma]} \cdot \frac{n^-_{c_{i,0}} \cdot [Y_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}^*_{c_{i,0}}, \varsigma]}{\sum_{j=1}^{k^-} n^-_j \cdot [Y_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}^*_j, \varsigma] + M \cdot \hat{E}_i}$$

which is equal to 1.

Case (ii): The Metropolis-Hastings acceptance probability equals $\min\{1, \psi_2\}$, where

$$\begin{aligned}\psi_2 = {} & \frac{M \cdot E_i \cdot [\boldsymbol{\theta}^*_{(k^-+1),1} \mid Y_i, \boldsymbol{\varphi}]}{\sum_{j=1}^{k^-} n_j^- \cdot [Y_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}_j^*, \varsigma] + M \cdot E_i} \cdot \frac{\sum_{j=1}^{k^-} n_j^- \cdot [Y_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}_j^*, \varsigma] + M \cdot \hat{E}_i}{M \cdot \hat{E}_i \cdot N_q\left(\boldsymbol{\theta}^*_{(k^-+1),1} \mid \boldsymbol{\mu}^*_{(k^-+1),0}, \, \Omega_{(k^-+1),0}\right)} \\ & \times \frac{\sum_{j=1}^{k^-} n_j^- \cdot [Y_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}_j^*, \varsigma] + M \cdot E_i}{n^-_{c_{i,0}} \cdot [Y_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}^*_{c_{i,0}}, \varsigma]} \cdot \frac{n^-_{c_{i,0}} \cdot [Y_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}^*_{c_{i,0}}, \varsigma]}{\sum_{j=1}^{k^-} n_j^- \cdot [Y_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}_j^*, \varsigma] + M \cdot \hat{E}_i} \\ = {} & \frac{E_i \cdot [\boldsymbol{\theta}^*_{(k^-+1),1} \mid Y_i, \boldsymbol{\varphi}]}{\hat{E}_i \cdot N_q\left(\boldsymbol{\theta}^*_{(k^-+1),1} \mid \boldsymbol{\mu}^*_{(k^-+1),0}, \, \Omega_{(k^-+1),0}\right)} = \rho_1 \triangleq \frac{[Y_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}^*_{(k^-+1),1}, \varsigma] \cdot N_q(\boldsymbol{\theta}^*_{(k^-+1),1} \mid \mathbf{0}, D)}{\hat{E}_i \cdot N_q\left(\boldsymbol{\theta}^*_{(k^-+1),1} \mid \boldsymbol{\mu}^*_{(k^-+1),0}, \, \Omega_{(k^-+1),0}\right)}\end{aligned}$$

as claimed, because of the following identity:

$$\begin{aligned} E_i \cdot [\boldsymbol{\theta} \mid Y_i, \boldsymbol{\varphi}] & \equiv [Y_i, \mid \boldsymbol{\varphi}] \cdot [\boldsymbol{\theta} \mid Y_i, \boldsymbol{\varphi}] \\ & = [Y_i, \boldsymbol{\theta} \mid \boldsymbol{\varphi}] \\ & = [Y_i \mid \boldsymbol{\theta}, \boldsymbol{\varphi}] \cdot [\boldsymbol{\theta} \mid \boldsymbol{\varphi}] \\ & \equiv [Y_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}, \varsigma] \cdot N_q(\boldsymbol{\theta} \mid \mathbf{0}, D). \end{aligned}$$

The acceptance probabilities for cases *(iii)* and *(iv)* can be similarly proven.

5.2 Computation of the matrix Ω_β^{-1} in (19) for large n

Definition (19) states that $\Omega_\beta^{-1} = \Sigma_\beta^{-1} + X'T^{-1}X$, where the n by n non-diagonal matrix $T = W^{-1} + Q$. Matrix Q has typical element $q_{ij} = \boldsymbol{z}_i' D \boldsymbol{z}_j$ if $c(i) = c(j)$, and $q_{ij} = 0$ if $c(i) \neq c(j)$. When the number of cases n is large, the main computational burden is the evaluation of T^{-1}.

Let P be the orthogonal 0-1 matrix that permutes the indices of the cases in such a manner that indices $1, \ldots, n_1$ correspond to cluster 1, indices $(n_1 + 1), \ldots, n_2$ correspond to cluster 2, and so on. For the rearranged data set, let T^* be defined analogously to the matrix T . We have $T^* = PTP'$. It is easy to show that T^* is block diagonal with blocks of order $n_1, \ldots, n_k$. Because of this, inversion of T^* typically involves a substantially lower cost than that of T . So T^{-1} can be efficiently computed as

$$T^{-1} = P(T^*)^{-1}P'$$

where we simply permute the rows and columns of $(T^*)^{-1}$ rather than actually pre- and post-multiplying by P.

REFERENCES

Antoniak, C. E. (1974), "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *Annals of Statistics*, 2, 1152–1174.

Banerjee, S., Carlin, B. and Gelfand, A. (2004). *Hierarchical Modeling and Analysis for Spatial Data*, Chapman & Hall/CRC, Florida.

Besag, J., Mollié, A. and York, J. (1991), "Bayesian image restoration, with two applications in spatial statistics," *Annals of the Institute of Statistical Mathematics*, 43, 1–20.

Blackwell, D. and MacQueen, J. B. (1973), "Ferguson distributions via Pólya urn schemes," *The Annals of Statistics*, 1, 353–355.

Breslow, N. and Day, N. E. (1987), *Statistical Methods in Cancer Research, Volume 2: The Design and Analysis of Cohort Studies*, International Agency for Research on Cancer, Lyon.

Burden S., Guha, S. Morgan, G., Ryan, L. and Young L. (2005), "Spatio-temporal Analysis of Ischemic Heart Disease in NSW, Australia," *Environmental and Ecological Statistics*, 12, 427–448.

Bush, C. A. and MacEachern S. N. (1996), "A semi-parametric Bayesian model for randomized block designs," *Biometrika*, 83, 275–285.

Dey, D., Müller, P. and Sinha, D. (1998), *Practical Nonparametric and Semiparametric Bayesian Statistics*, Springer-Verlag, New York.

Escobar, M. D. (1994), "Estimating normal means with a Dirichlet process prior," *Journal of the American Statistical Association*, 89, 268–277.

Escobar, M. D. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Ferguson, T. S. (1973), "Estimating normal means with a Dirichlet process prior," *Annals of Statistics*, 1, 209–230.

Guha, S. and Ryan, L. M. (2006), “Computationally Efficient Estimation of Generalized Linear Mixed Models for Large Data Sets using the Penalized Quasi-likelihood Approach with Application to Spatially Varying Disease Rates,” in preparation.

Guha, S., Ryan, L. M., Morgan, G., Beard, J. (2006), “Age and gender modify the association between socio-economic factors and heart disease in New South Wales, Australia,” submitted.

Ishwaran, H. and James, L. (2001), “Gibbs Sampling Methods for Stick-Breaking Priors,” *Journal of the American Statistical Association*, 96, 161–173.

Ishwaran, H. and Zarepour, M. (2002), “Dirichlet prior sieves in finite normal mixtures,” *Statistica Sinica*, 12, 941–963.

Kleinman, K. P. and Ibrahim, J. G. (1998a), “A semiparametric Bayesian approach to the random effects model,” *Biometrics*, 54, 921–938.

Kleinman, K. P. and Ibrahim, J. G. (1998b), “A semi-parametric Bayesian approach to generalized linear mixed models,” *Statistics in Medicine*, 17, 2579–2596.

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

MacEachern, S. N. (1994), “Estimating Normal Means with a Conjugate Style Dirichlet Process Prior,” *Communications in Statistics: Simulation and Computation*, 23, 727–741.

MacEachern, S. N. (1998), “Computational Methods for Mixture of Dirichlet Process Models,” in *Practical Nonparametric and Semiparametric Bayesian Statistics*, (eds. Dey, D., Müller, P. and Sinha, D.), pp. 23–43, Springer-Verlag, New York.

MacEachern, S. N. and Müller, P. (1994), “Estimating Mixture of Dirichlet Process Models,” *Journal of Computational and Graphical Statistics*, 7, 223–238.

Marmot M. G., Bosma H., Hemingway H., et al. (1997). “Contribution of job control and other risk factors to social variations in coronary heart disease incidence,” *Lancet*, 350, 235–239.

McCullagh, P. and Nelder, J. A. (1999), *Generalized Linear Models* (2nd ed.), CRC Press LLC, Boca Raton, Florida.

Neal, R. M. (2000), "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 9, 283–297.

Papaspiliopoulos, O. and Roberts, G. (2006), "Retrospective MCMC for Dirichlet process hierarchical models," *Biometrika*, to appear.

Ripley, B. D. (1987), *Stochastic simulation*, New York: Wiley.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–650.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," *Annals of Statistics*, 22, 1701–1728.

Walker, S. G., Damien, P., Laud, P. W. and Smith, A. F. M. (1999), "Bayesian nonparametric inference for random distributions and related functions," *Journal of the Royal Statistical Society: Series B*, 61, 485–527.

West, M., Müller, P. and Escobar, M. D. (1994), "Hierarchical priors and mixture models, with application in regression and density estimation," in *Aspects of Uncertainty: A tribute to D. V. Lindley*, (eds. A. F. M. Smith and P. Freeman), pp. 363–368, New York: Wiley.

Zeger, S. L. and Karim, M. R. (1991), "Generalized linear models with random effects: A Gibbs sampling approach," *Journal of the American Statistical Association*, 86, 79–86.