Journal of Data Science **8**(2010), 43-59

A Nonparametric Approach Using Dirichlet Process for Hierarchical Generalized Linear Mixed Models

Jing Wang
Louisiana State University

Abstract: In this paper, we propose a nonparametric approach using the Dirichlet processes (DP) as a class of prior distributions for the distribution G of the random effects in the hierarchical generalized linear mixed model (GLMM). The support of the prior distribution (and the posterior distribution) is large, allowing for a wide range of shapes for G. This provides great flexibility in estimating G and therefore produces a more flexible estimator than does the parametric analysis. We present some computation strategies for posterior computations involved in DP modeling. The proposed method is illustrated with real examples as well as simulations.

Key words: Dirichlet process, generalized linear mixed model, Gibbs sampler, Metropolis–Hastings algorithm.

1. Introduction

Generalized linear models have been very useful for a wide variety of discrete, continuous, and censored responses in many research areas. Estimation of the GLMM is discussed by Zeger and Karim (1991) and Breslow and Clayton (1993), among others. In these models, the distribution G of random effects typically is assumed to have a parametric distribution form, such as a normal distribution.

Clearly, it is not always the case that random effects come from a known parametric family of distributions. There are possibilities that the distribution of the random effects is nonnormal, multimodal, or skewed. Assuming a parametric distribution or misspecifying the distribution would impose unreasonably constraints on the distribution and as a consequence produce poor estimates of parameters. It is therefore important to use nonparametric approaches to allow random effects to come from a sufficiently large class. Currently, there are some references about nonparametric estimation in hierarchical mixed models. Lavine (1992) proposes Bayesian nonparametric mixing using Polya trees as a prior for the distribution of random effects in the hierarchical GLMM. This approach is limited to the univariate random effects. Walker and Wakefield (1998) propose a Bayesian nonparametric approach using Dirichlet processes (Ferguson, 1973; Antoniak, 1974) as a class of prior distributions for the distribution of random effects in the context of the hierarchical nonlinear mixed model (NLMM) where the responses are continuous only. The DP has been very popular in modern nonparametric statistics. There is an extensive discussion on DP models in the literature. Excellent references include (Escobar and West, 1992, 1995; West, Muller, and Escobar, 1994; Muller and Rosner, 1997; MacEachern, 1992; Bush and MacEachern, 1996). In this article, we implement the Dirichlet process in the hierarchical GLMM by placing a DP prior on the distribution, G, of the random effects. In this model framework, G itself is assumed uncertain, drawn from a Dirichlet process on a family of all possible distribution functions on the real line. The support of the prior distribution is large, allowing for a wide range of shapes for G. This provides great flexibility in estimating G and therefore produces a more flexible estimator than does the parametric analysis.

This article is organized as follows. In Section 2 we review the hierarchical generalized linear model and the Dirichlet process. In Section 3 we present some computation strategies for posterior computations involved in DP modeling. In sections 4 and 5 we illustrate the application of the proposed method in real examples and simulated data. We conclude this article with discussions.

2. Hierarchical GLMMs Based on DP Priors

The general form of the hierarchical generalized linear mixed model can be described as follows. Let y_{ij} be the response for the ith ($i = 1, \cdots, n$) individual and the jth observation ($j = 1, \cdots, n_i$). The hierarchy of the GLMM model basically comprises three stages. At the first stage, y_{ij} follows an exponential family of the distribution of the form:

$$f(y_{ij}|\mathbf{b}_i, \beta, \tau) = \exp\{\tau[y_{ij}h(\eta_{ij}) - g(h(\eta_{ij}))] + d(y_{ij}, \tau)\}, \tag{2.1}$$

where $h(\cdot)$ is the link function, $g(\cdot)$ is the variance function, $\eta_{ij} = \mathbf{x}_{ij}^T\beta + \mathbf{z}_{ij}^T\mathbf{b}_i$ is the linear predictor, β is a p-vector of fixed parameters, $\mathbf{b}_i$ is a q-vector of random effects associated with the ith subject, $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are p-dimensional and q-dimensional covariates associated with the fixed parameters β and random effects $\mathbf{b} = (\mathbf{b}_i)^T$, and τ is a scalar dispersion parameter. In logistic and Poisson regression models, τ is intrinsically equal to 1, as shown in all the examples given in this paper. For normal random effects models, $\tau = 1/\sigma^2$ is assumed unknown. At the second stage of the model, the random effects $\mathbf{b}_i$ typically are assumed to have a parametric distribution, G, such as a normal distribution. The final stage of the model specifies the priors for fixed parameters β, the covariance matrix of G, and the dispersion parameter τ. For example, β are usually assumed to have a normal prior with mean μ and variance $\mathbf{\Sigma}$, the inverse of $\mathbf{D}$ has a Wishart distribution, say, $\mathbf{D}^{-1} \sim W(\rho, (\rho\mathbf{R})^{-1})$, and for normal random effects models, τ usually has a Gamma prior as $\tau \sim Ga(\nu_0/2, \nu_0\omega_0/2)$. Here, Ga, W, U denote respectively a Gamma distribution, a Wishart distribution, and a Uniform distribution; the hyperparameters $\mu, \mathbf{\Sigma}, \nu_0, \omega_0, \rho$, $\mathbf{R}$, and d are known. For example, we can use the following set of priors: $\mu = \mathbf{0}, \mathbf{\Sigma}^{-1} = \mathbf{0}, \nu_0 = 0, \rho = q$ (the dimension of $\mathbf{b}_i$); $\mathbf{R}$ is chosen to be an approximate prior estimate of $\mathbf{D}$ and d is a positive number in a reasonable range.

The parametric form of the distribution of the random effects $\mathbf{b}$ may be too restrictive to account for model features such as multimodality, nonnormality, and skewness. It is therefore important to use nonparametric approaches for modeling in order to capture these possibilities. Specifically, we consider here the Dirichlet process for estimation in the hierarchical GLMM. The idea is that, instead of specifying a parametric form for G, we assume G to be uncertain and modeled as a DP with parameters G_0 and α, where G_0 is a probability measure and α is a positive real constant. The parameter G_0 is a location parameter for the DP prior. It is the best guess at what G is believed to be and is the prior expectation of G so that $E(G(\mathbf{b}|\cdot)) = G_0(\mathbf{b}|\cdot)$. The parameter $\alpha > 0$ is a measure of the concentration of the prior for G about the guess prior G_0. When G is integrated over its prior distribution G_0, a sequence of $\mathbf{b}_1, \cdots, \mathbf{b}_n$ can be generated using a Polya urn scheme (Blackwell and MacQueen, 1973), described as follows. The first parameter, $\mathbf{b}_1$, is chosen from G_0. The second parameter, $\mathbf{b}_2$, is chosen from G_0 with probability $\alpha/(\alpha + i - 1)$ and is equal to $\mathbf{b}_1$ with probability $1/(\alpha + i - 1)$. The generation rule for $\mathbf{b}_i|\mathbf{b}_1, \cdots, \mathbf{b}_{i-1}$ is to set $\mathbf{b}_i$ equal to $\mathbf{b}_j, j < i$ with probability $1/(\alpha + i - 1)$ and to choose $\mathbf{b}_i$ from G_0 with probability $\alpha/(\alpha + i - 1)$; that is,

$$\mathbf{b}_i \sim \begin{cases} G_0 & \text{w. p. } \dfrac{\alpha}{\alpha + i - 1} \\ \mathbf{b}_j (j \neq i, i = 2, \cdots, n) & \text{w. p. } \dfrac{1}{\alpha + i - 1}. \end{cases} \tag{2.2}$$

In other words, the sequence $\mathbf{b}_1, \cdots, \mathbf{b}_n$ is actually drawn from a mixture distribution with mixing probabilities determined by the DP. From this it is easy to see that, when α is large a sampled G is likely to be close to G_0, and that, when α is small a sampled G is likely to come from just a few existing realized values of $\mathbf{b}_1, \cdots, \mathbf{b}_n$. Therefore, the parameter α is a type of dispersion parameter for the DP prior. As described by MacEachern (1992), the DP procedure as demonstrated in (2.2) generates a cluster structure for $\mathbf{b}_i$'s. This cluster structure partitions $\mathbf{b}_1 \cdots, \mathbf{b}_n$ into $c \leq n$ sets. All of the $\mathbf{b}_i$'s within a particular set are identical; those in different sets differ. This produces a rich class of prior distributions, allowing for a wide range of shapes for G. The base prior G_0 may be viewed as a "baseline" prior in the context of an analysis of sensitivity to the assumptions of the baseline parametric model. The closed form of the joint prior density of $\mathbf{b}_1, \cdots, \mathbf{b}_n$ is

$$f(\mathbf{b}_1, \cdots, \mathbf{b}_n) = \prod_i^n \frac{\alpha G_0(d\mathbf{b}_i) + \sum_{j=1}^{i-1} \delta(\mathbf{b}_j, d\mathbf{b}_i)}{\alpha + i - 1},$$

where $\delta(\mathbf{b}_j, d\mathbf{b}_i)$ denotes a unit point mass at $\mathbf{b}_i = \mathbf{b}_j$:

$$\delta(\mathbf{b}_j, d\mathbf{b}_i) = \begin{cases} 1 & \text{when } \mathbf{b}_i = \mathbf{b}_j \\ 0 & \text{when } \mathbf{b}_i \neq \mathbf{b}_j. \end{cases}$$

A further stage, therefore, can be added to the hierarchical GLMM model (2.1) based on the DP prior:

$$\mathbf{b}_i \sim G, \quad G \sim DP(G_0(\mathbf{b}|\cdot)). \tag{2.3}$$

This further stage adds uncertainty about G modeled as a DP, allowing for the modeling of deviation away from the specific distribution G_0. An important instance of model (2.3) is the normal DP model, in which $G_0(\mathbf{b}_i|\mathbf{D}) = N(\mathbf{0}, \mathbf{D})$.

3. Posterior Computations

Estimation of the DP model has been efficiently implemented by the Gibbs sampling scheme from the full conditionals of parameters (Escobar, 1988; Escobar and West, 1992, 1995; West, Muller, and Escobar, 1994; Muller and Rosner, 1997; MacEachern, 1992; Bush and MacEachern, 1996). In this section, we describe the implementation of the Gibbs sampler in our model framework.

Estimation based on the DP for the normal random effects model has been investigated by several authors, Escobar and West (1992), West, Muller, and Escobar (1994), Walker and Wakefield (1998), among others. In this article we focus on DP estimation in logistic and Poisson random effects models in which $\tau = 1$. For notational convenience, in model (2.1) we simplify $d(y_{ij}, \tau)$ into $d(y_{ij})$ and denote $s(y_{ij}, \mathbf{b}_i, \beta) = y_{ij} h(\eta_{ij}) - g(h(\eta_{ij})) + d(y_{ij})$.

For simplicity and ease of implementation, we consider throughout a normal baseline prior with mean $\mathbf{0}$ and covariance matrix $\mathbf{D}$, i.e., $G_0(\mathbf{b}|\mathbf{D}) = N(\mathbf{0}, \mathbf{D})$.

3.1 Posterior distribution of $\mathbf{b}_i$

Recall in Section 2, the sequence $\mathbf{b}_1, \cdots, \mathbf{b}_n$ produced by the DP procedure, as described in (2.2), can be partitioned into $c \leq n$ clusters. Let $(\tilde{\mathbf{b}}_1, \cdots, \tilde{\mathbf{b}}_c)$

denote the set of distinct values. Let $\mathbf{b}^{(i)}$ denote the vector with $\mathbf{b}_i$ removed, that is, $\mathbf{b}^{(i)} = (\mathbf{b}_1, \cdots, \mathbf{b}_{i-1}, \mathbf{b}_{i+1}, \cdots, \mathbf{b}_n)$. Let $c^{(i)}$ denote the number of different clusters formed by $\mathbf{b}^{(i)}$ and $c_k^{(i)}(k = 1, \cdots, c^{(i)})$ denote the number of observations sharing the common parameter value $\tilde{\mathbf{b}}_k^{(i)}$ in the kth cluster.

Conditional on $\mathbf{b}^{(i)}$, the prior of $\mathbf{b}_i$ can be given by:

$$(\mathbf{b}_i|\mathbf{b}^{(i)}, \beta, \mathbf{D}, \alpha) \propto \frac{\alpha}{\alpha+n-1} G_0(\mathbf{b}_i|\mathbf{D}) + \frac{1}{\alpha+n-1} \sum_{k=1}^{c^{(i)}} c_k^{(i)} \delta(\mathbf{b}_i|\tilde{\mathbf{b}}_k^{(i)}), \qquad (3.1)$$

where $\delta(\mathbf{b}_i|\tilde{\mathbf{b}}_k^{(i)})$ denotes a unit point mass at $\mathbf{b}_i = \tilde{\mathbf{b}}_k^{(i)}$. This shows that $\mathbf{b}_i$ is distinct from the other parameters and drawn from $G_0(\mathbf{b}_i|\mathbf{D})$ with probability $\alpha/(\alpha+n-1)$, otherwise, it is chosen from the existing values $\tilde{\mathbf{b}}_k^{(i)}$ with probabilities $c_k^{(i)}/(\alpha+n-1)$. It's clear that when $\alpha \to \infty$, $G \to G_0$ and when $\alpha \to 0$, $G \to \tilde{\mathbf{b}}_k^{(i)}$. Therefore in this process the parameter α governs the precision of the guess prior G_0.

With such a cluster feature as shown in (3.1), several algorithms have been proposed for implementing the Gibbs sampler to resample $\mathbf{b}_i$'s in the DP model (Escobar and West, 1992; MacEachern, 1992; West, Muller, and Escobar, 1994; Escobar and West, 1995; Muller and Rosner, 1997; Bush and MacEachern, 1996). Currently MacEachern's algorithm, outlined as follows, is recommended for use.

Escobar andWest (1995) show that, based on the prior (3.1), the conditional posterior distribution of $\mathbf{b}_i$ has the following form:

$$(\mathbf{b}_i|\mathbf{b}^{(i)}, \mathbf{y}_i, \beta, \mathbf{D}, \alpha) \propto q_0 f(\mathbf{y}_i|\mathbf{b}_i, \beta) G_0(\mathbf{b}_i|\mathbf{D}) + \sum_{k=1}^{c^{(i)}} q_k \delta(\mathbf{b}_i|\tilde{\mathbf{b}}_k^{(i)}),$$

where $f(\mathbf{y}_i|\mathbf{b}_i, \beta)$ is the density function of $\mathbf{y}_i$ conditional on $\mathbf{b}_i$ and β, and

$$\begin{aligned} q_0 &= \alpha \int f(\mathbf{y}_i|\mathbf{b}_i, \beta) dG_0(\mathbf{b}_i|\mathbf{D}), \\ q_k &= c_k^{(i)} f(\mathbf{y}_i|\tilde{\mathbf{b}}_k^{(i)}, \beta), \qquad 1 = q_0 + \sum_k q_k. \end{aligned}$$

The proportion q_0 can be computed

$$q_0 \propto \alpha |\mathbf{D}|^{-1/2} \int \exp\left\{ -\frac{1}{2} \left[\sum_{j=1}^{n_i} s(y_{ij}, \mathbf{b}_i, \beta) + \mathbf{b}_i{}^T \mathbf{D}^{-1} \mathbf{b}_i \right] \right\} d\mathbf{b}_i.$$

The above integral generally does not have a closed form solution. We apply the Laplace's method for integral approximation:

$$|\mathbf{H}_i|^{-1/2} \exp\left\{ -\frac{1}{2} \left[\sum_{j=1}^{n_i} s(y_{ij}, \hat{\mathbf{b}}_i, \beta) + \hat{\mathbf{b}}_i^T \mathbf{D}^{-1} \hat{\mathbf{b}}_i \right] \right\},$$

where $\hat{\mathbf{b}}_i$ and $\mathbf{H}_i$ are the mode and Hessian of the function $\sum_{j=1}^{n_i} s(y_{ij}, \mathbf{b}_i, \beta) + {\mathbf{b}_i}^T \mathbf{D}^{-1} \mathbf{b}_i$.

The proportion q_k can be computed directly

$$q_k = c_k^{(i)} f(\mathbf{y}_i | \tilde{\mathbf{b}}_k^{(i)}, \beta) \propto c_k^{(i)} \exp\left[-\frac{1}{2} \sum_{j=1}^{n_i} s(y_{ij}, \tilde{\mathbf{b}}_k^{(i)}, \beta) \right].$$

3.2 Posterior distribution of β

The conditional density of β can be computed

$$\begin{aligned} f(\beta | \mathbf{b}, \mathbf{y}, \mathbf{D}, \alpha) &\propto f(\beta) \prod_{i=1}^{n} f(\mathbf{y}_i | \mathbf{b}_i, \beta, \mathbf{D}, \alpha) \\ &\propto f(\beta) \exp\left[-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n_i} s(y_{ij}, \mathbf{b}_i, \beta) \right], \end{aligned} \tag{3.2}$$

where $f(\beta)$ is the density of β. Posterior density (3.2) typically cannot be written into a closed form and therefore direct sampling β is not available. Alternatively, we can use the Metropolis–Hastings (M–H) algorithm to obtain samples without knowing the analytical form of the posterior distribution. We choose the candidate distribution, $I(\beta)$, as follows. Because (3.2) is known up to a normalization constant, we can compute its mode $\hat{\beta}$ and Hessian $\mathbf{V}$ using numerical optimization techniques. This yields a natural choice of the candidate distribution, a normal distribution with mean $\hat{\beta}$ and variance $\mathbf{V}$. Then we can implement the Metropolis–Hastings algorithm as follows. Denote β^t the current value of β at the tth iteration. A new value β^* is drawn from the candidate distribution $I(\beta)$. The acceptance probability is computed

$$\min\left\{ 1, \frac{f(\beta^*) f(\mathbf{y} | \mathbf{b}, \beta^*, \mathbf{D}, \alpha) I(\beta^t)}{f(\beta^t) f(\mathbf{y} | \mathbf{b}, \beta^t, \mathbf{D}, \alpha) I(\beta^*)} \right\}.$$

Note that there is no need to compute the normalization constant because it cancels out in the acceptance probability.

3.3 Posterior Distribution of D

The posterior distribution depends on $\mathbf{b}$ only. Using a convenient prior for $\mathbf{D}^{-1}$, a Wishart distribution $W(\rho, (\rho\mathbf{R})^{-1})$ as specified in Section 2, it can be shown that the posterior of $\mathbf{D}^{-1}$ is the following Wishart distribution (Wakefield, Gelfand, Racine-Poon and Smith, 1995):

$$(\mathbf{D}^{-1}|\mathbf{b}, \mathbf{y}, \beta, \alpha) \sim W\left(\rho + n, \left[\rho\mathbf{R} + \sum_{i=1}^{n} \mathbf{b}_i \mathbf{b}_i{}^T\right]^{-1}\right), \qquad (3.3)$$

where W denotes a Wishart distribution, ρ is the Wishart degrees of freedom index which may be chosen to be the dimension of the random effects, and the matrix $\mathbf{R}$ is an approximate estimate of $\mathbf{D}$.

3.4 Posterior distribution of α

Suppose $\alpha \sim Ga(\tau_1, \tau_2)$, a Gamma distribution with a shape parameter $\tau_1 > 0$ and a scale parameter $\tau_2 > 0$. Given this Gamma prior, α can be resampled from a mixture of two gammas:

$$(\alpha|\xi, c) \sim \pi_\xi Ga(\tau_1 + c, \tau_2 - \log(\xi)) + (1 - \pi_\xi) Ga(\tau_1 + c - 1, \tau_2 - \log(\xi)), \quad (3.4)$$

where ξ is a latent variable sampled from a Beta distribution, $(\xi|\alpha, c) \sim Beta(\alpha + 1, c)$, and π_ξ is the weight computed by $\pi_\xi/(1-\pi_\xi) = (\tau_1 + c - 1)/[c(\tau_2 - log(\xi))]$. When both τ_1, τ_2 are small, the Gamma prior puts weight on both high and low values of α; when both τ_1, τ_2 are large, the Gamma prior favors low values of α.

3.5 Prediction and density estimation

Predictions of future values of $\mathbf{b}_{n+1}$ can be obtained by extending n to $n+1$:

$$(\mathbf{b}_{n+1}|\mathbf{b}, \beta, \mathbf{D}, \alpha) \sim \frac{\alpha}{\alpha + n} G_0(\mathbf{b}_{n+1}|\mathbf{D}) + \frac{1}{\alpha + n} \sum_{i=1}^{c} c_i \delta(\mathbf{b}_{n+1}|\tilde{\mathbf{b}}_i), \qquad (3.5)$$

where $\mathbf{b}_{n+1}$ is a new independent draw from G_0.

Estimation of the densities of $\mathbf{b}$ can be obtained using Monte Carlo approximations (Escobar and West, 1995), described as follows. First, we run the Gibbs sampler for M iterations, where M is a sufficiently large number, so that convergence of the the Gibbs sampler has achieved. With the obtained samples $\mathbf{b}^M, \beta^M, \mathbf{D}^M$, and α^M, we compute Monte Carlo approximations to $f(\mathbf{b}_{n+1}|\mathbf{y})$ as follows:

$$f(\mathbf{b}_{n+1}|\mathbf{y}) \approx \frac{1}{M_0} \sum_{m=M}^{M+M_0} f(\mathbf{b}_{n+1}|\mathbf{b}^m, \beta^m, \mathbf{D}^m, \alpha^m),$$

where M_0 is a large number required for Monte Carlo approximation, and $f(\mathbf{b}_{n+1}|\mathbf{b}^m, \beta^m, \mathbf{D}^m, \alpha^m)$ is the density for (3.5) computed at the mth iteration.

The posterior distribution of $\mathbf{D}^{-1}$ and α have a closed form, and therefore estimates of their marginal densities can be obtained by using the empirical distributions (Gelfand and Smith, 1990) of the sampled values $\mathbf{D}$ and α:

$$\begin{aligned} f(\mathbf{D}^{-1}|\mathbf{y}) &\approx \frac{1}{M_0}\sum_{m=M}^{M+M_0} f(\mathbf{D}^{-1}|\mathbf{b}^m, \mathbf{y}, \beta^m, \alpha^m), \\ f(\alpha|\mathbf{y}) &\approx \frac{1}{M_0}\sum_{m=M}^{M+M_0} f(\alpha|\xi^m, c^m), \end{aligned}$$

where $f(\mathbf{D}^{-1}|\mathbf{b}^m, \mathbf{y}, \beta^m, \alpha^m)$ and $f(\alpha|\xi^m, c^m)$ are the densities given in (3.3) and (3.4). As seen from (3.2), the posterior density $f(\beta|\mathbf{b}, \mathbf{y}, \mathbf{D}, \alpha)$ is not analytically available. With large data sets, it may be reasonable to use a Gaussian distribution as an approximation to $\beta|\mathbf{b}, \mathbf{y}, \mathbf{D}, \alpha$.

4. Illustrative Examples

4.1 Pups data

Ochi and Prentice (1984) analyzed the data, previously presented by Weil (1970), collected from an experiment comprising two treatments. One group of 16 pregnant female rats was fed a control diet while a second group of 16 pregnant females was fed a chemically treated diet. For each litter, the number m_i ($i = 1, \cdots, 32$) of pups alive after 4 days and the number z_i of pups that survived to 21 days were recorded. Denote the two levels of response y_i by success and failure, that is, $y_i = 1$ if the survival time exceeds 0 and 0 otherwise. Hence, z_i is the number of successes among the m_i observations in the ith litter. The number of survivors varies among litters in a manner that substantially exceeds that consistent with a standard binomial error structure. This extra-binomial variation, or overdispersion, reflects the fact that the binary responses of rats from the same litter tend to be more alike than are the responses from distinct litters at the same diet. To account for this overdispersion, a general probit model has been considered for the counts z_i, allowing the binary responses y_i's to have separate probabilities for control and treated groups:

$$\begin{aligned} &(z_i|b_i) \sim Binomial(m_i, p_i), \text{ and} \\ &p_i = P(y_i = 1) = \Phi(\beta_1 x_{1i} + \beta_2 x_{2i} + b_i), \end{aligned} \tag{4.1}$$

where x_{1i} and x_{2i} are binary indicators defined as

$$x_{1i} = \begin{cases} 1 & \text{if the litter receives the control diet} \\ 0 & \text{otherwise,} \end{cases}$$

$$x_{2i} = \begin{cases} 1 & \text{if the litter receives the treated diet} \\ 0 & \text{otherwise,} \end{cases}$$

β_1 and β_2 represent the treatment means for control and treated groups, p_i is the probability of the binary observation in the ith litter, and Φ is the standard normal cumulative distribution function. The $q = 1$ dimensional random effects are $b_i \sim N(0, D_1 x_{1i} + D_2 x_{2i})$, allowing for separate variations, say, D_1 and D_2, for control and treated groups respectively. Here, the b_i's represent the litter effects, used to describe the greater alikeness of the binary response within a litter as compared to that between litters for control and treated groups.

To illustrate the DP analysis for these data, we make a slight modification to model (4.1) by assuming a common covariance matrix D for the control and treated groups: $b_i \sim N(0, D)$. In other words, we assume homoscedasticity of variance for random effects among the control and treated groups. Basically, we want to see if the DP model might be useful to detect potential heterogeneity in random effects among the two groups.

To implement the Gibbs sampler, we choose the prior for β to be $N((0,0)^T, 10000\mathbf{I}_2)$, where $\mathbf{I}_2$ is the 2×2 identity matrix. The prior for D^{-1} is a Wishart distribution with $\rho = 1, R = 1$. The parameter α is modeled with a Gamma prior $Ga(0.1, 1)$, inducing a range of small values of α. We run the Gibbs sampler for 20,000 iterations, discarding the first 1,000 iterations and using the remaining 19,900 iterations to compute parameter estimates. Convergence is assessed graphically. Figure 1 displays the posterior densities by applying a Gaussian kernel estimate to the sampled values. The top row of Figure 1 is from $Ga(0.1, 1)$ and the bottom row is from the fully parametric analysis. These are the predictive distributions computed without knowledge of group membership. It's evident that the density under the Gamma prior is bimodal, suggesting separate variances for control and treated groups respectively. Following this, we fit the two groups separately using model (4.1). The parameter estimates for the two groups are computed as $\beta_1 = 1.33, \beta_2 = -0.47$, $D_1 = 0.24$, and $D_2 = 1.32$, indicating a small litter effect among the control litters and a strong effect a mong the treated litters. This can also be seen from the predicted distribution under DP analysis as shown in Figure 1, in which the mode on the left, due to the control diet, appears to be smaller than that on the right, due to the treated diet. In summary, analyses of these data indicate that the DP prior is useful in detecting some particular model features of interest.

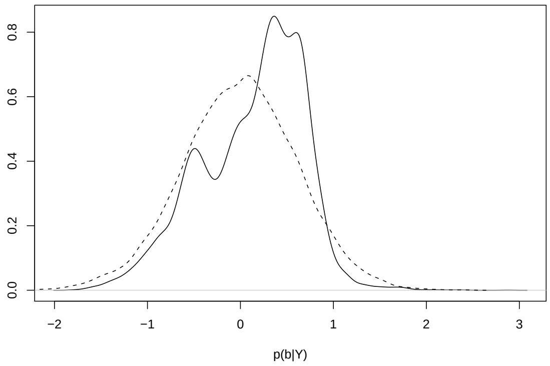

Figure 1: Estimated posterior density of b under parametric (dotted line) analysis and DP analysis (solid line) for the pups data.

4.2 Seed data

Breslow and Clayton (1993) analyzed the data, previously presented by Crowder (1978), on the seeds that germinated on each of 21 plates. Two factors, seed variety (S) and type of root extract (R), were examined, yielding a 2×2 factorial structure. The binary response indicator y_i was 1 if the ith seed germinated and 0 otherwise. The within-group variation was observed to exceed that predicted by binomial sampling theory. To account for this extraneous plat-to-plate variability, a logistic modeling analysis of treatment and interaction effects was used to model the germination rate:

$$logit\ P(y_i = 1|b_i) = \beta_0 + \beta_1 S_i + \beta_2 R_i + \beta_3 (S_i \times R_i) + b_i,$$

where $q = 1$ and $b_i \sim N(0, D)$, $i = 1, \cdots, 21$, represented random effects associated with each plate.

We choose the priors for β and D in a similar fashion to the pups data. The precision parameter α is modeled with each of the following three Gamma priors: $Ga(10, 10)$, $Ga(100, 0.1)$, and $Ga(1000, 0.1)$. We find that estimation results from these four Dirichlet processes are similar. For brevity, our presentation is limited to the analysis under the $Ga(10, 10)$ prior only. Table 1 summarizes posterior sample means and standard errors (SE) from the $Ga(10, 10)$ prior and the parametric analysis. It can be seen that parameter estimates from the two methods are very close. In general, the parametric method produces smaller standard errors, indicating that it tends to underestimate parameters. This is not surprising in the sense that under parametric analysis G is restricted to be normal while under DP analysis G varies in a larger range of distributions. Figure 2 displays posterior densities under the four Gamma priors and the parametric method. The four rows in Figure 2, from top to bottom, are from $Ga(10, 10), Ga(100, 0.1), Ga(1000, 0.1)$, and the fully parametric analysis. We see that there are various shapes including bimodal and skewed densities under the DP analysis. The $Ga(10, 10)$ prior seems to induce a more severe clustering of the b_i's, while each of the other two priors $Ga(100, 0.1)$ and $Ga(1000, 0.1)$ seems to induce more of a blend between the baseline predictive distribution and the predictive distribution from a standard normal model with each of the $Ga(100, 0.1)$ and $Ga(1000, 0.1)$ priors. This is expected because the $Ga(10, 10)$ prior induces much smaller α values than the other two. These DP predicted distributions show us flexibility provided by the DP prior in modeling the distribution of random effects. The DP prior specification for these data allows for an arbitrary distribution of the plate effects, and results in effective estimation of the treatment effects across a wide range of distributions for the plate effects.

Table 1: Posterior means and standard errors (SE) for the seed data

	DP method		Parametric method	
	Mean	SE	Mean	SE
β_0	-.56	.045	-.55	.013
β_1	.15	.026	.13	.023
β_2	1.32	.022	1.32	.019
β_3	-.78	.036	-.78	.032
D	.012	.006	.012	.004

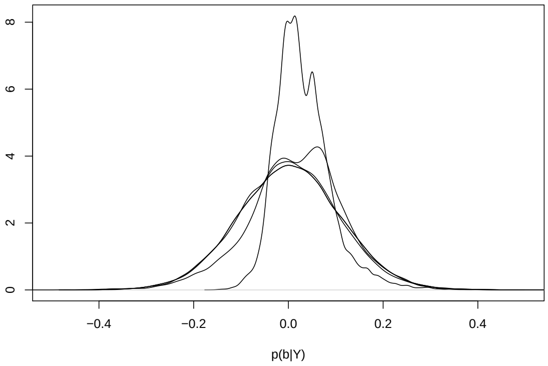

Figure 2: Estimated posterior density of b under DP and parametric analyses for the seed data

4.3 Longitudinal data

Thall and Vail (1990) presented the data arising from a clinical trial of 59 epileptics who were randomized to receive either a new drug or a placebo as an adjustment to the standard chemotherapy. The number of epileptic seizures, the response variable, occurring over the previous two weeks was recorded at each of four successive clinic visits. There were five covariates: the binary indicators for treatments (T) (T=1 if the epileptic received the new drug and 0 otherwise), the logarithm of 1/4 the 8-week baseline seizure counts (B), the logarithm of age in years (A), the visit (V), coded $V_1 = -3, V_2 = -1, V_3 = 1, V_4 = 3$, and the interaction between baseline seizure counts and treatment (BT). Four covariance models, each in a log-linear form, were considered for these data. We focus on the following two. The logarithm of the response variable y_{ij}, the seizure count for patient i on the jth visit $(i = 1, \cdots, 59, j = 1, \cdots, 4)$, is assumed to be Poisson-distributed with mean computed

$$\log \mu_{ij} = \beta_0 + \beta_1 V_4 + \beta_2 T_i + \beta_3 B_i + \beta_4 (BT)_i + \beta_5 A_i + b_{i0}, \tag{4.2}$$

$$\log \mu_{ij} = \beta_0 + \beta_1 V_4 + \beta_2 T_i + \beta_3 B_i + \beta_4 (BT)_i + \beta_5 A_i + b_{i0} + b_{i1} V_j/10, \tag{4.3}$$

where $q = 2$ and $\mathbf{b} = (b_{i0}, b_{i1})$ are bivariate normal random effects from $N(\mathbf{0}, \mathbf{D})$, which represent the residual level and rate of change in the event rate for the ith subject.

We choose the priors for $\mathbf{D}$ as follows: $\rho = 2$, $\mathbf{R} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ for (4.2) and $\mathbf{R} = \mathbf{I}_2$ for (4.3). The Gamma prior for α is $Ga(1000, 0.001)$, inducing quite large α values and therefore indicating no anticipation of departures from normality. Table 2 gives posterior sample means and standard errors from each model. In general, the DP and parametric analyses agree reasonably well in each case. This indicates that, when random effects are likely to come from a normal distribution, the DP and parametric models are equivalent.

Table 2: Results for the longitudinal data

	Model (4.2)				Model (4.3)			
	DP method		Parametric method		DP method		Parametric method	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
β_0	-1.35	.30	-1.18	.14	-1.33	.30	-1.37	.14
β_1	-.16	.005	-.16	.005	-.16	.10	-.27	.03
β_2	-1.06	.24	-.93	.05	-.81	.25	-.92	.04
β_3	.90	.04	.87	.02	.89	.04	.88	.01
β_4	.41	.14	.34	.03	.28	.14	.34	.02
β_5	.49	.09	.45	.04	.48	.09	.49	.04
$\mathbf{D}_{11}$	.42	.08	.41	.08	.18	.04	.18	.04
$\mathbf{D}_{12}$	—	—	—	—	-.03	.06	-.0008	.04
$\mathbf{D}_{22}$	—	—	—	—	.30	.07	.27	.06

5. Simulation Studies

5.1 Simulation 1

Zeger and Karim (1991) presented a simulation study involving $n = 100$ subjects of size $n_i = 7$. Two models, each in a logistic regression form, were considered for these data:

$$P(y_{ij} = 1|\mathbf{b}_i) = \beta_0 + \beta_1 t_j + \beta_2 x_i + \beta_3(t_j x_i) + b_{i0}, \tag{5.1}$$

$$P(y_{ij} = 1|\mathbf{b}_i) = \beta_0 + \beta_1 t_j + \beta_2 x_i + \beta_3(t_j x_i) + b_{i0} + b_{i1} t_j, \tag{5.2}$$

where $x_i = 0$ for half the sample and $x_i = 1$ for the other half and $t_j = j - 4$, $i = 1, \cdots, 100, j = 1, \cdots, 7$. The regression coefficients were fixed at $\beta_0 = -2.5, \beta_1 = 1.0, \beta_2 = -1.0$, and $\beta_3 = -0.5$. Zeger and Karim generated $q = 2$ dimensional random effects $\mathbf{b}_i = (b_{i0}, b_{i1})$ from a bivariate normal distribution. Here, to illustrate the DP method, for model (5.1) we generate b_{i0} from a mixture of two normals, with means equal to -0.5 and 0.5; variance 1. For model (5.2) we generate b_{i0} from a mixture of two normals, with means equal to -0.5 and 0.5; variance 0.49; b_{i1} are generated from $N(0, 0.25)$. The mixing probability for each model is 0.5. Under this mixture of normals, it can be easily verified that the mean and variance of b_{i0} are 0, 1.25 for model (5.1) and 0, 0.74 for model (5.2).

Table 3: Results for the simulated data fitted by models (5.1) and (5.2)

	Model (5.1)				Model (5.2)			
	DP method		Parametric method		DP method		Parametric method	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
β_0	-2.60	.30	-2.53	.29	-2.31	.19	-2.39	.034
β_1	1.03	.13	1.03	.13	1.19	.17	1.08	.02
β_2	-.75	.43	-.68	.43	-.85	.14	-.86	.05
β_3	-.73	.20	-.73	.20	-.64	.11	-.66	.03
$\mathbf{D}_{11}$	1.15	.18	1.09	.17	.73	.10	.72	.10
$\mathbf{D}_{12}$	—	—	—	—	.42	.14	.38	.07
$\mathbf{D}_{22}$	—	—	—	—	.47	.24	.31	.06

For each model, we analyze the data using both DP and parametric methods. The Gamma priors for α are $Ga(1, 1)$ and $Ga(1, 10)$ for models (5.1) and (5.2), suggesting that low values of α are favored. Table 3 presents sample means and standard errors obtained from each model. Overall, estimates from the DP method are in close agreement with those from the parametric method. Figures 3 and 4 display the posterior densities. The posterior densities under the DP analysis appear to be less diffuse over the data range than those under the parametric analysis. Bimodality of the densities under the DP analysis is evident, which is consistent with the distribution of b_0. However, this fact is not suggested by the parametric method. Results for these data show that, when the random effects are not likely to come from a normal distribution, the DP analysis performs better than the parametric analysis in that it detects some particular feature of the model while the parametric analysis can not.

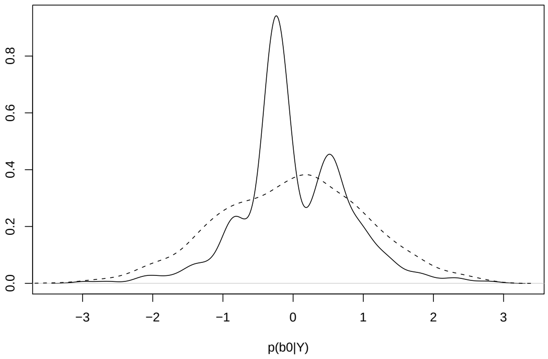

Figure 3: Estimated densities of b_0 for model (5.1). The solid line is the DP density, and the dotted line is the parametric density.

5.2 Simulation 2

In this simulation, we generate 100 independent Poisson counts, $y_i, i = 1, \cdots, 100$, with means

$$\mu_i = \beta_0 + \beta_1 x_i + b_{i0} + b_{i1} x_i,$$

where $x_i = i/100, \beta_0 = 10, \beta_1 = 1$, $q = 2$, b_{i0} are generated from a mixture of three normal distributions with means equal to -1, 0, 1; variance 4; and corresponding weights 0.33, 0.34, and 0.33; b_{i1} are generated from $N(0, .25)$. It can be verified that the mean and variance of b_{i0} resulting from this mixture of normals are 0 and 4.66. The Gamma prior for the parameter α is $Ga(1, 50)$. Table 4 summarizes results from both parametric and DP analyses. In general, sample means and standard errors of parameters produced by the DP method are larger than those produced by the parametric method. Figure 5 displays posterior densities of b_0 and b_1 under both methods. From Figure 5, three spikes are indicated in the posterior density of b_0 under the DP prior. The spike on the left is due to $N(-1, 4)$, the spike in the middle is due to $N(0, 4)$, and the spike on the right is due to $N(1, 4)$. However, the parametric analysis fails to capture this feature. In summary, our conclusion on the two methods for these simulated counts data is similar to that for the first simulation study.

Table 4: Results for the simulated counts data

	DP method		Parametric method	
	Mean	SE	Mean	SE
β_0	10.16	.34	9.89	.13
β_1	2.31	.45	2.22	.09
$\mathbf{D}_{11}$	5.09	1.02	4.21	.67
$\mathbf{D}_{22}$	.11	.02	.10	.01

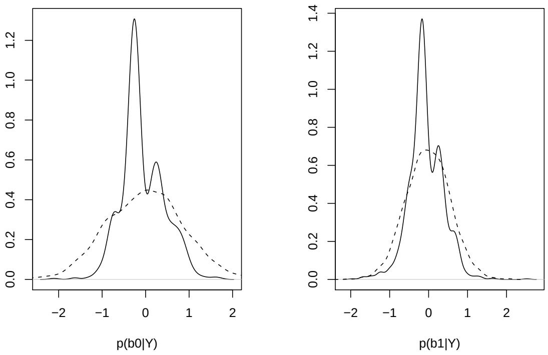

Figure 4: Estimated densities of b_0 and b_1 for model (5.2). The solid line is the DP density, and the dotted line is the parametric density.

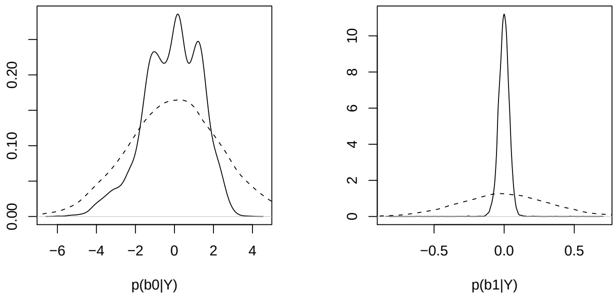

Figure 5: Estimated posterior densities of b_0 and b_1 for the simulated counts data. The solid line is the DP density, and the dotted line is the parametric density.

6. Conclusion

In this paper, we have presented a Bayesian nonparametric analysis using the DP prior for the hierarchical generalized linear model. The use of the DP prior provides great flexibility in estimating the distribution of random effects, and therefore allows us to explore the particular model features of interest. We have applied this method to real examples as well as simulated data. We find that when the distribution of the random effects is likely to be normal the DP and parametric methods give similar estimation results. However, when the random effects come from a distribution other than the normal distribution, as shown in the pups example and the two simulation studies, the DP analysis clearly is more useful than the parametric analysis in detecting some particular model features of interest.

Acknowledgements

The author thanks the referee for providing valuable suggestions that lead to an improvement of the earlier version of the manuscript.

References

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

Blackwell, D. and MacQueen, J. (1973). Ferguson distributions via Polya urn schemes. *Annals of Statistics* **1**, 353-355.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9-25.

Bush, C. A. and MacEachern, S. N. (1996). A semiparametric Bayesian model for randomized block designs. *Biometrika* **83**, 275-286.

Crowder, M. J. (1978). Beta-binomial anova for proportions. *Applied Statistics* **27**, 34-37.

Escobar, M. D. (1988). Estimating the means of several normal populations by nonparametric estimation of the distribution of the means. Ph.D. Dissertation. Yale University.

Escobar, M. D. and West, M. (1992). Computing nonparametric hierarchical models. Discussion paper 92-A20, Duke University.

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). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209-230.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398-409.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modeling. *The Annals of Statistics* **20**, 1222-1235.

MacEachern, S. (1992). Estimating normal means with a conjugate style Dirichlet process prior. Technical report, Duke University.

Muller, P. and Rosner, G. L. (1997). A Bayesian population model with hierarchical mixture priors applied to blood count data. *Journal of the American Statistical Association* **92**, 1279-1292.

Ochi, Y. and Prentice, R. L. (1984). Likelihood inference in a correlated probit regression model. *Biometrika* **71**, 531-543.

Thall, P. F. and Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics* **46**, 657-671.

Wakefield, J.C., Gelfand, A. E., Racine-Poon, A. and Smith, A. F. M. (1994). Bayesian analysis of linear and nonlinear population models using the Gibbs sampler. *Applied Statistics* **43**, 201-221.

Walker, S. G. and Wakefield, J. C. (1998). Population models with a nonparametric random coefficient distribution. *Sankhya* **60**, 196-212.

Weil, C. S. (1970). Selection of the valid number of sampling units and consideration of their combination in toxicological studies involving reproduction, teratogenesis, or carcinogenesis. *Food and Cosmetic Toxicology* **8**, 177-182.

West, M., Muller, P. and Escobar, M. D. (1994). *Hierarchical priors and mixture models, with application in regression and density estimation. Aspects of Uncertainty: A tribute to D. V. Lindley, A.F.M. Smith and P. Freeman.* 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.

Received Feberuary 18, 2008; accepted March 25, 2008.

Jing Wang
Department of Experimental Statistics
Lousiana State University
Baton Rouge, LA 70808, USA
jwang@lsu.edu