Posterior Computation for Hierarchical Dirichlet Process Mixture Models:
Application to Genetic Association Studies of Quantitative Traits in the Presence of Population Stratification

Nicholas M. Pajewski\textsuperscript{1} and Purushottam W. Laud
Division of Biostatistics
Department of Population Health
Medical College of Wisconsin
Milwaukee, WI 53226 USA

Introduction

In \cite{pajewski2008}, we introduced a unified hierarchical Bayesian semiparametric model for genetic association studies of quantitative traits in the presence of population stratification. The model uses a Dirichlet Process Mixture (DPM) construction to account for stratification in making association inference. It also involves a nonparametric sparsity prior to accommodate the expectation that most genetic markers are unrelated to the phenotype in a large association screen. In this technical report, we describe the necessary computational details for implementing the DPM model (C code available from http://www.biostat.mcw.edu/software/SoftMenu.html). We begin with a short description of the DPM model, and then discuss its implementation through Markov chain Monte Carlo (MCMC) sampling.

Consider a continuous phenotype $Y_i$ observed on a sample of $N$ unrelated individuals. Suppose each individual is then genotyped at $L$ Single Nucleotide Polymorphism (SNP) markers. The extension to more polymorphic markers is straightforward, although the available C code does not currently implement such a case. Define $V_{li} = 1$ (0 otherwise) if the $i$th individual is homozygous for the reference (or minor) allele at the $l$th SNP, and $W_{li} = 1$ (0 otherwise) if the individual is heterozygous at that SNP. Then let $\beta_{l1}$ and $\beta_{l2}$ represent the regression effects for individuals heterozygous and homozygous respectively at the $l$th SNP. Finally, let $X_{li} = [W_{li} V_{li}]$ and $\beta_l = [\beta_{l1}, \beta_{l2}]$. The hierarchical DPM model can then be

\textsuperscript{1}Correspondence to: Nicholas M. Pajewski, Division of Biostatistics, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226, (414) 456-8674, npajewsk@mcw.edu
defined as follows.

\[
L(Y_i|\mu_i, \tau_e) = \frac{\tau_e^{1/2}}{\sqrt{2\pi}} \exp\left[\frac{-\tau_e}{2} (Y_i - \mu_i)^2\right]
\]

\[
\mu_i = \beta_{0i} + \sum_{l=1}^{L} X_{li} \beta_l
\]

\[
L(W_{li}, V_{li}|\theta_l) = \frac{2^{W_{li}}e^{\theta_l(2V_{li}+W_{li})}}{(1+e^{\theta_l})^2}
\]

\[
i = 1, \ldots, N\quad l = 1, \ldots, L
\]

\[
\beta_{0i}, \theta_{l1}, \ldots, \theta_{lL}|G \sim \text{DP} (\alpha_G, G_0)
\]

\[
G|\alpha_G, G_0 \sim \text{DP} (\alpha_G, G_0)
\]

\[
G_0 = N (\beta_{0i}, \mu_0, \tau_0) \prod_{l=1}^{L} N (\theta_{li}; \mu_{\theta}, \tau_{\theta})
\]

\[
\beta_l|H \sim \text{DP} (\alpha_H, H_0)
\]

\[
H|\alpha_H, H_0 \sim \text{DP} (\alpha_H, H_0)
\]

\[
H_0 = \pi \delta_{(0,0)}(\cdot) + (1-\pi)MVN_2 (M_\beta, T_\beta)
\]

\[
\pi \sim \text{Beta} (c_1, d_1)
\]

\[
\tau_\epsilon \sim \text{Gamma} (\eta_1, \lambda_1)
\]

\[
\alpha_G \sim \text{Gamma} (\eta_2, \lambda_2) \quad \text{and} \quad \alpha_H \sim \text{Gamma} (\eta_3, \lambda_3)
\]

Note: Throughout the document, we use the following parametrization of gamma density, \(X \sim \text{Gamma} (\alpha, \lambda)\),

\[
f(x) \propto x^{\alpha-1}e^{-\lambda x}
\]

In the above formulation, \(\theta_l = \text{logit} (\pi_{li})\) where \(\pi_{li}\) presents the reference allele frequency for the \(i^{th}\) individual at the \(l^{th}\) SNP. \(\delta_{(0,0)}(\cdot)\) represents a Dirac delta function indicating a point mass at \((0,0)\). In addition, \(N(x; \mu, \tau)\) denotes a normal density with mean \(\mu\) and precision \(\tau\) and \(MVN_p(x; M, T)\) represents a p-dimensional multivariate normal with mean vector \(M\) and precision matrix \(T\). For each of the Dirichlet Processes, we have assumed gamma priors for the scalar mass parameters \(\alpha_G\) and \(\alpha_H\) following \(\gamma\); alternatively they could be taken as to be fixed constants. Figure 1 displays the model as a directed acyclic graph (DAG).
Posterior Computations

We now describe in full detail the necessary steps to implement posterior inference using MCMC sampling. Given an initial state \( \Theta_0 = \left[ \theta_i^{(0)} \text{ for all } i, \beta_l^{(0)} \text{ for all } l, \tau_\epsilon^{(0)}, \alpha_G^{(0)}, \alpha_H^{(0)} \right] \), iterate through the following steps.

**STEP 1: Update for \( \theta_i \)**

In order to update \( \theta_i = [\beta_0, \beta_1, \ldots, \beta_L] \) we employed a Metropolis-Hastings based algorithm described in \( ? \) (algorithm 5). The algorithm of Neal utilizes the notion of a configuration in updating each \( \theta_i \). At a given MCMC iteration, the \( \theta_i \) will have clustered to a set of \( K_\theta < N \) distinct values denoted as \( \theta^* = [\theta_1^*, \ldots, \theta_{K_\theta}^*] \). Note that each element of \( \theta^* \) represents an \( L + 1 \) dimensional vector containing the regression parameter \( \beta_0 \) and the logit of allele frequencies at each SNP. We then define the configuration indicators \( s_i \) where \( s_i = j \) if and only if \( \theta_i = \theta_j^* \). Finally let \( n_j \) represent the number of \( s_i \) currently equal to \( j \).

Figure 1: DAG for Hierarchical DPM Model of Quantitative Traits
**Step 1a:** Perform the following proposal step for $R$ iterations. For $i = 1, 2, ..., N$; propose a new distinct atom membership ($s_i^*$) for the $i^{th}$ observation. The approach of ? uses the conditional prior as a proposal distribution for $s_i^*$. Let $s_{(-i)}$ denote the set of all configuration indicators minus $s_i$, and let $n_j^{(-i)}$ denote the number of $s_c = j$ for $c = 1, 2, ..., i-1, i+1, ..., N$.

\[
P (s_i^* = j | s_{(-i)}) = \frac{n_j^{(-i)}}{\alpha_G + N - 1} \text{ for } j = 1, 2, ..., K\theta \text{ and}
\]

\[
P (s_i^* = K\theta + 1 | s_{(-i)}) = \frac{\alpha}{\alpha_G + N - 1}
\]

Note that if $s_i^* = K\theta + 1$ is proposed then a new value $\theta_{K\theta+1}$ needs to be sampled from $G_0$. Accept the move to $s_i^*$ with the following probability.

\[
P (s_i, s_i^*) = \min [1, R] \text{ where } R = \frac{L(Y_i, W_i, V_i | \theta_{s_i}^*)}{L(Y_i, W_i, V_i | \theta_{s_i}^*)} \text{ and }
\]

\[
L(Y_i, W_i, V_i | \theta_{s_i}^*) = L(Y_i | W_{li}, V_{li}, \beta^0_{ij}, \tau, \beta_l \forall l) \times \prod_{l=1}^{L} L(W_{li}, V_{li} | \theta_{s_i}^*)
\]

When updating the configuration indicators $s_i$, there are two potential moves which would alter the number of distinct points in $\theta^*$. If $n_{s_i}^{(-i)} = 0$ (i.e. the $i^{th}$ observation is currently a singleton), unless a proposal of $s_i^* = K\theta + 1$ is accepted, there is now one less distinct point in $\theta^*$. Therefore, $K\theta = K\theta - 1$. Similarly, if $n_{s_i}^{(-i)} > 0$ and a proposal of $s_i^* = K\theta + 1$ is accepted, then $K\theta = K\theta + 1$.

**Step 1b:** After updating each $s_i$, let $K\theta$ denote the current atoms in $\theta^*$ where $n_j > 0$. For $j=1,2,...,K\theta$, update $\theta_{s_i}^*$. This entails a series of independent updates for each element of $\theta_{s_i}^*$. Begin by sampling $\beta_{0j}$ from a Normal ($\mu^*, \tau^*$) distribution, where

\[
\tau^* = n_j \tau_\epsilon + \tau_0
\]

\[
\mu^* = \frac{1}{\tau^*} \left[ \tau_\epsilon \sum_{i:s_i = j} Y_i - \sum_{l=1}^{L} X_{li} \beta_l \right] + \tau_0 \mu_0
\]

Then, for $l = 1, 2, ..., L$, the unnormalized log full conditional density for $\theta_{s_i}^*$ takes the following form.

\[
\log [\theta_{s_i}^* | s, W, V] = \theta_{s_i}^* \sum_{i: s_i = j} (2V_{li} + W_{li}) - 2n_j \log (1 + e^{\theta_0}) - \frac{\tau_0}{2} (\theta_{ij} - \mu)^2
\]
Although the above log target density does not take a standard distributional form, the
density is log-concave, and so a new value for $\theta^*_j$ can be sampled using Adaptive-Rejection
sampling (?).

**STEP 2: Update for $\beta_l$**

In order to update each $\beta_l$, we employed the Blocked Gibbs Sampler of ?. The Blocked Gibbs
Sampler is based on the stick-breaking representation of the Dirichlet Process, discussed in
the work of ?. Although the stick-breaking representation of the DP involves an infinite
sum of discrete points, in actual implementation, the Blocked Gibbs Sampler utilizes a finite
approximation, imposing a limit $F_L$ to the number of distinct atoms amongst the $\beta_l$. Denote
this collection of distinct points as $\beta^* = [\beta^*_1, ..., \beta^*_F_L]$. ? show that even for large sample
sizes, a limit of $F_L = 150$ provides a suitable approximation to the Dirichlet Process. Because
of the point mass mixture construction in $H_0$, without a loss of generality, we can include
the additional distinct point $\beta^*_0$ to represent the cluster denoting no effect (i.e. $\beta_1 = 0$
and $\beta_{i2} = 0$) with associated model weight $\pi$. Similar to the configuration representation for $\theta_i$,
define the pointers $z_l$ where $z_l = j$ if and only if $\beta_l = \beta^*_j$ for $j = 0, 1, 2, ..., F_L$. Then define
$m_j$ as the number of $z_l$ currently equal to $j$.

**Step 2a:** For $j = 1, 2, ..., F_L$; update $\beta^*_j$. Note, because $\beta^*_0$ represents the null effect clus-
ter, its value need not be updated. If $m_j = 0$, then $\beta^*_j \sim H_0$. Else draw $\beta^*_j \sim MVN_2(M^*, T^*)$
where

$$
T^* = \tau_t G'_j G_j + T_{\beta},
$$

$$
M^* = (T^*)^{-1} \left[ \tau_t G'_j (Y - B_0 - X\beta^{(-j)}) + T_{\beta} M_{\beta} \right] 
$$

$Y$ denotes a $n \times 1$ column vector of the quantitative traits $Y_i$. Similarly, $B_0$ represents a
$n \times 1$ column vector where the $i^{th}$ element is $\beta_{0s}$. $G_j$ is a $n \times 2$ matrix whose $i^{th}$ row equals

$$
\left[ \sum_{l:z_l=j} W_{li} \sum_{l:z_l=j} V_{li} \right].
$$

Finally, $X\beta^{(-j)}$ is a $n \times 1$ column vector whose $i^{th}$ element is $X_{ci}\beta^*_c$. 


**Step 2b:** For $l = 1, 2, ..., L$; independently sample $z_l$ where,

$$P(z_l = 0) \propto \pi L(Y | s, \beta_0^*, \tau_e)$$

$$P(z_l = j) \propto (1 - \pi) p_j L(Y | s, \beta_j^*, \tau_e) \text{ for } j = 1, 2, ..., F_L$$

where

$$L(Y | s, \beta_j^*, \tau_e) \propto \exp \left[ \frac{-\tau_e}{2} \sum_{i=1}^{N} \left( Y_i - \beta_{0s_i} - X_{hi} \beta_j^* - \sum_{c \neq l} L \left( X_{ci} \beta_{zc_i} \right) \right) ^2 \right]$$

**Step 2c:** Update $\pi$ and the stick-breaking weights ($p_j$). Sample $\pi \sim \text{Beta}(c_1 + m_0, d_1 + (L - m_0))$. Then for $j = 1, 2, ..., F_L$; set

$$p_1 = V_1$$

$$p_k = (1 - V_1)(1 - V_2) \cdots (1 - V_{k-1})V_k \text{ for } k = 2, 3, ..., F_L - 1$$

where

$$V_k \sim \text{Beta} \left( \frac{\alpha_H}{F_L} + m_k, \frac{\alpha_H(F_L - k)}{F_L} + \sum_{c=k+1}^{F_L} m_c \right) \text{ for } k = 1, 2, ..., F_L - 1$$

Then because the $p_j$ must sum to 1, $p_{F_L} = 1 - \sum_{j=1}^{F_L-1} p_j$.

**STEP 3:** Updating the scalar mass parameters of the Dirichlet Process ($\alpha_G, \alpha_H$)

If $\alpha_G$ and $\alpha_H$ are given Gamma priors, then they can be updated using the following procedure described in $\text{?}$. Assume there are $K_G$ and $K_H$ distinct atoms in the configuration representations for both G and H at the current MCMC iteration.

**STEP 3a:** Update for $\alpha_G$

1. Sample $x_G | \alpha_G \sim \text{Beta}(\alpha_G, N)$

2. Let $\pi_G$ equal

$$\pi_G = \frac{\eta_2 + K_G - 1}{\eta_2 + K_G - 1 + N(\lambda_2 - \log(X_G))}$$

3. Sample $\alpha_G | x_G, K_G \sim$

$$\pi_G \text{ Gamma} (\eta_2 + K_G, \lambda_2 - \log(x_G)) + (1 - \pi_G) \text{ Gamma} (\eta_2 + K_G - 1, \lambda_2 - \log(x_G))$$
**STEP 3b: Update for $\alpha_H$**

1. Sample $x_H|\alpha_H \sim \text{Beta}(\alpha_H, L)$

2. Let $\pi_H$ equal

$$\pi_G = \frac{\eta_3 + K_H - 1}{\eta_3 + K_H - 1 + L(\lambda_3 - \log(X_H))}$$

3. Sample $\alpha_G|x_G, K_G \sim$

$$\pi_H \text{ Gamma (} \eta_3 + K_H, \lambda_3 - \log(x_H)) \ + \ (1 - \pi_G) \text{ Gamma (} \eta_3 + K_H - 1, \lambda_3 - \log(x_H))$$

**STEP 4: Update error precision $\tau_\epsilon$**

Sample $\tau_\epsilon \sim \text{Gamma}(\alpha^*, \lambda^*)$ where

$$\alpha^* = \frac{N}{2} + \eta_1$$

$$\lambda^* = \lambda_1 + \frac{1}{2} \sum_{i=1}^{N} \left( Y_i - \beta_{0s_i} - \sum_{l=1}^{L} X_{li}\beta_{z_l}^* \right)^2$$