TECHNICAL REPORT 55
MARCH 2008

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

Nicholas M. Pajewski[1] and Purushottam W. Laud
Division of Biostatistics
Department of Population Health
Medical College of Wisconsin

de ned as follows.

$$
\begin{aligned}
L(Y_i | \ _i; \ ) &= \frac{^{1=2}}{\sqrt{2}} \exp \ \frac{-}{2} (Y_i - \ _i)^2 \\
_i &= \ _{0i} + \sum_{l=1}^{} X_{li} \ _l \\
L(W_{li}; V_{li} | \ _{li}) &= \frac{2^{W_{li}} e^{\ _{li}(2V_{li} + W_{li})}}{(1 + e^{\ _{li}})^2} \quad i = 1; :::; N \quad l = 1; :::; L \\
_{0i}; \ _{1i}; ::::; \ _{Li} | G &\overset{i:i:d}{\sim} G \quad i = 1; :::; N \\
G | \ _G; G_0 &\sim \mathrm{DP}(\ _G; G_o) \\
G_0 &= \mathrm{N}(\ _0; \ _0; \ _0) \prod_{l=1}^{} \mathrm{N}(\ _l; \ ; \ ) \\
_l | H &\overset{i:i:d}{\sim} H \quad l = 1; :::; L \\
H | \ _H; H_0 &\sim \mathrm{DP}(\ _H; H_0) \\
H_0 &= \ _{(0;0)}(\cdot) + (1 - \ ) MVN_2(M \ ; T \ ) \\
&\sim \mathrm{Beta}(c_1; d_1) \\
&\sim \mathrm{Gamma}(\ _1; \ _1) \\
_G \sim \mathrm{Gamma}(\ _2; \ _2) &\quad \text{and} \quad \ _H \sim \mathrm{Gamma}(\ _3; \ _3)
\end{aligned}
$$

Note: Throughout the document, we use the following parametrization of gamma density, $X \sim$ Gamma (;),

$$f(x) \propto x^{\ -1} e^{- \ x}$$

In the above formulation, $_{li}$ = logit ($_{li}$) where $_{li}$ presents the reference allele frequency for the i^{th} individual at the l^{th} SNP. $_{(0;0)}(\cdot)$ represents a Dirac delta function indicating a point mass at $(0; 0)$. In addition, $N(x; \ ; \ )$ denotes a normal density with mean and precision 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 $_G$ and $_H$ following ?; alternatively they could be taken as to be xed constants. Figure 1 displays the model as a directed acyclic graph (DAG).

$_{0i}$

Y_i

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 con guration indicators minus s_i, and let $n^{(-i)}$

Although the above log target density does not take a standard distributional form, the density is log-concave, and so a new value for $^*_{jl}$ can be sampled using Adaptive-Rejection sampling (?).

STEP 2: Update for $_l$

In order to update each $_l$, 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 in nite sum of discrete points, in actual implementation, the Blocked Gibbs Sampler utilizes a nite approximation, imposing a limit F_L to the number of distinct atoms amongst the $_l$. Denote this collection of distinct points as $^* = \; ^*_1; ::::; \; ^*_{F_L}$. ? 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 $^*_0$ to represent the cluster denoting no e ect (i.e. $_{l1} = 0$ and $_{l2} = 0$) with associated model weight . Similar to the con guration representation for $_i$, de ne the pointers z_l where $z_l = j$ if and only if $_l = \; ^*_j$ for $j = 0; 1; 2; ::::; F_L$. Then de ne m_j as the number of z_l currently equal to j.

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

$$
\begin{aligned}
T^* &= \quad G_j' G_j + T \\
M^* &= (T^*)^{-1} \quad G_j' \;\; Y - B_0 - X \;\;^{(-j)} \;\; + T \;\; M
\end{aligned}
$$

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 j^{th}]TJ/F22 11.9552 Tf 19.1289 0 Td [(element ...

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

$$
\begin{aligned}
P(z_l = 0) &\propto \quad L(Y|s;\ {}^*_0;\ \ ) \\
P(z_l = j) &\propto (1 - \ )p_j L\ \ Y|s;\ {}^*_j; \qquad \text{for } j = 1, 2, \ldots, F_L
\end{aligned}
$$

where

$$
L(Y|s;\ {}^*_j;\ \ ) \propto \exp\left[-\frac{}{2}\sum_{i=1}^{n}\left(Y_i - {}_{0s_i} - X_{li}\ {}^*_j - \sum_{c\neq l}(X_{ci}\ {}_{z_c})\right)^2\right]
$$

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

$$
\begin{aligned}
p_1 &= V_1 \\
p_k &= (1 - V_1)(1 - V_2)
\end{aligned}
$$

STEP 3b: Update for $_H$

1. Sample $x_H | \ _H \sim \text{Beta}(\ _H; L)$
2. Let $_H$ equal

$$_G = \frac{_3 + K_H - 1}{_3 + K_H - 1 + L(\ _3 - \log(X_H))}$$

3. Sample $_G | x_G; K_G \sim$

$$_H \ \text{Gamma}(\ _3 + K_H; \ _3 - \log(x_H)) + (1 - \ _G)\ \text{Gamma}(\ _3 + K_H - 1; \ _3 - \log(x_H))$$

STEP 4: Update error precision

Sample $\sim \text{Gamma}(\ ^*; \ ^*)$ where

$$^* = \frac{N}{2} + \ _1$$

$$^* = \ _1 + \frac{1}{2}\sum_{i=1}^{N}\left(Y_i - \ _{0s_i} - \sum_{l=1}^{k} X_{li} \ ^*_{z_l}\right)^2$$