**Documentation for STRUCTDPM.c**

Author: Nicholas M. Pajewski
Updated: 3/26/2008
Questions or bug reports can be sent npajewsk@mcw.edu

# Description

Implements Bayesian Dirichlet Process Mixture Model for a quantitative trait genetic association study of unrelated individuals in the presence of population stratification. Methods described in Pajewski and Laud (2008a). For a description of the details of the MCMC sampling algorithm, see Pajewski and Laud (2008b) available at
http://www.biostat.mcw.edu/tech/Tech.html.

# Necessary Hard Coding

Unfortunately, the current implementation is a little cumbersome, although we are working on making it more user-friendly. However, until that happens, there is some necessary hard-coding. There is a struct definition (*genetic_data*) in the beginning of the program with a number of arrays where the array sizes need to be hard coded.

| | | |
|---|---|---|
| double q[$F_L$]; | double a_q[$F_L$]; | double b_q[$F_L$]; |
| double theta[$C_L$][L] | double clust_allele_freq[$C_L$][L][2]; | int genotypes[N][L][2]; |
| double gen_profile[N]; | double phenotype[N]; | double beta_ge[$C_L$][2]; |
| double beta0[$C_L$]; | int distinctH[$F_L$]; | |

$F_L$ denotes the finite limit of the Dirichlet Process (DP) on the regression effects (See parameter setup file description for a discussion of suitable values). N denotes the number of sampled individuals, and L equals the number of genotyped SNPs. Finally, $C_L$ is a suitable cap to the number of distinct atoms in the DP on the allele frequencies. In theory, because we employ a computational algorithm described in Neal (2000), this limit could be equal to N. However, in practice, the DP produces substantial clustering amongst the $\theta_i$, and so a value such as 25 or 50 should be more than adequate.

# Input File Format

The program looks for input files (and places output files) in the directory where it is being run. However, this can be adjusted by changing the paths beginning at line 1170.

The program currently takes three files as input, a parameter setup file *(default name: parm_setup.txt)*, a file containing the observed genotype data *(default name: genome.txt)*, and a separate file with the observed phenotypes *(default name: phenotype.txt)*. In addition, the program utilizes a number of GSL subroutines and so it assumes GSL has been installed on your system (GSL freely available from http://www.gnu.org/software/gsl/).

1. **arms.c** and **arms.h:** These are two files, available at Wally Gilks website, http://www.maths.leeds.ac.uk/∼wally.gilks/adaptive.rejection/web_page/Welcome.html, for implementing Adaptive Rejection Sampling (Gilks and Wild, 1992) for the full conditional distribution of $\theta^*_{jl}$. They should be saved in the same directory as STRUCT-DPM.c. The files can be saved in a different directory by adjusting the include statements at the beginning of the program.

2. **Setup file** *parm_setup.txt***:** This is a setup file with the following structure. Note: All fields must be included

| Line | Description | Example |
|---|---|---|
| 1 | Number of Individuals | 2100 |
| 2 | Number of SNPs | 703 |
| 3 | Number of MCMC iterations | 6500 |
| 4 | Burn-in | 1500 |
| 5 | Repeat | 2 |
| 6 | Finite Limit | 50 |
| 7 | Initial Clusters | 10 |
| 8 | Seed | 7564383 |
| 9 | $a_{0G}$ $b_{0G}$ $a_{0H}$ $b_{0H}$ | 1.0 1.0 1.0 1.0 |
| 10 | $\mu_{\beta_0}$ $\tau_{\beta_0}$ | 100.0 0.001 |
| 11 | $\alpha_\tau$ $\lambda_\tau$ | 0.01 0.01 |
| 12 | $a_\pi$ $b_\pi$ | 0.95 0.05 |
| 13 | $\mu_{\beta_1}$ $\tau_{\beta_1}$ $\mu_{\beta_2}$ $\tau_{\beta_2}$ | 0.00 0.01 0.00 0.01 |
| 14 | Output Iteration | 25 |

**Notes:** Line 3 denotes the number of desired MCMC iterations + burn-in, so to get 1000 iterations following a burn-in of 1000, lines 3 and 4 should 2000 and 1000. Line 5 denotes the number of Metropolis-Hastings Proposal sub-iterations (R) to perform in the sampling algorithm of Neal (2000). Neal suggests using R=4, although R=2 seems to perform just as well in this situation. Line 6 denotes the number of initial clusters in the configuration representation of the Dirichlet Process on the allele frequencies.

The program then randomly allocates each individual amongst Initial Clusters distinct atoms. In our experience, a random configuration of roughly 8 to 10 initial distinct points seems to perform adequately well. Line 7 represents the finite limit approximation to the Dirichlet Process used in the Blocked Gibbs Sampler of Ishwaran and James (2001). Ishwaran and James show that even for large sample sizes a suitable approximation is a limit of 150. However, we have found adequate performance in our limited simulations using a limit of 50 to 75. Line 8 represents the gamma prior parameters for the mass parameters of the two Dirichlet Processes. G refers to the process on the allele frequencies, and H refers to the sparsity process on the regression effects. Note that we use the following parametrization of a Gamma density,

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

Line 9 is a random number seed. Line 10 is the prior parameters (mean, precision) for the normal prior on $\beta_0$ within $G_0$. Line 11 is the prior parameters for the gamma prior on $\tau_\epsilon$, the residual error precision in the linear regression likelihood. Line 12 represents the prior parameters for beta prior on $\pi$, the common weight parameter that controls how many of the SNP effects are clustered to zero. We have seen reasonable performance for vague priors that have prior expectation $> 0.90$, such as a Beta(0.95,0.05). Line 13 contains the prior parameters within $H_0$ for the regression effects. As a default, the bivariate expectation for the distribution on $\beta_l$ is constructed as the product of two independent normal densities. However, the program can handle dependence via the precision matrix T, simply by editing the matrix *PRE* on line 1300. Line 14 denotes the iteration interval with which to output progress of the MCMC chain to the screen. For example, a value of 25 outputs the progress of the sampler every 25 iterations. Currently, the program outputs the current iteration, the number of distinct atoms in the configuration representations for both G and H, and the current values of the mass parameters $(\alpha_G, \alpha_H)$ for the two DPs. Finally, in terms of the configuration on the regression effects, the program currently initializes each effect to be zero, i.e. $z_l = 0$ for all $l$. This can be changed by making appropriate changes to the *initial_samp* function.

3. **Genotype Data:** Each row contains SNP genotype information for each individual. At the moment, the code is not able to handle more polymorphic markers. Each row contains two indicators for each SNP genotype, one denoting heterozygotes for the disease allele, and the other representing disease allele homozygotes. The specification of the disease allele is arbitrary, so it can be set to represent either the minor or reference (from HapMap for example) allele. The following table illustrates the necessary layout of the genotype data file. Note that the individual ID column and the SNP headers are merely for illustration, and so should not be placed in the actual file. The program

is not currently setup to handle missing genotype data, so one solution would be to use a program such as fastPHASE (Scheet and Stephens, 2006) to impute necessary genotypes before using STRUCTDPM.c

| *individual* | *SNP1* | | *SNP 2* | | *SNP 3* | | $\cdots$ |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 0 | 0 | 0 | $\cdots$ |
| 2 | 0 | 1 | 0 | 0 | 0 | 1 | $\cdots$ |
| 3 | 1 | 0 | 0 | 0 | 1 | 0 | $\cdots$ |

For example, individual 1 is homozygous for the reference allele at the first SNP, heterozygous at the second, and homozygous for the other allele at the third.

4. **Phenotype Data:** The quantitative trait data is just a single column, which each row denoting the phenotype for the $i^{th}$ individual, i.e.

*phenotype 1*
*phenotype 2*
*phenotype 3*
.
.

**Output Files**

- cluster_probs.txt: Contains posterior probability estimates of two individuals being in the same cluster of the Dirichlet Process (See Huelsenbeck and Andolfatto (2007)). Useful for tracking whether the DP detects the underlying composition of a stratified population sample. The output format is the $N \times N$ matrix where the $(i, j)$ element denotes the posterior probability $P(s_i = s_j)$, i.e. the probability that the allele frequencies for the $i^{th}$ and $j^{th}$ individuals originate from the same distinct atom in the Dirichlet Process. **Note:** Because the matrix is symmetric, the program outputs only the upper triangular portion of the matrix, with all elements for $i < j$ set to 0.
- DPMtimer.txt: Contains run time statistics *(in seconds)*
- post_ge.txt: Contains posterior samples (past burn-in) of the genetic effects $\beta_{1l}$ and $\beta_{2l}$ for each SNP. Output format is

| *iteration* | *SNP* | *Effect* | *Sample* |
|---|---|---|---|
| 1000 | 1 | 1 | 1.23 |
| 1000 | 1 | 2 | 3.34 |
| 1000 | 2 | 1 | 0.00 |
| 1000 | 2 | 2 | 0.00 |
| 1000 | 3 | 1 | 4.56 |

- post_gezero.txt: Contains posterior samples (past burn-in) of the indicator whether each SNP's genetic effects were clustered outside of the no effect (0,0) cluster. The indicator is 1 if $z_l \neq 0$ at the current iteration.

| *iteration* | *SNP* | $I(z_l \neq 0)$ |
|---|---|---|
| 1000 | 1 | 0 |
| 1000 | 2 | 1 |
| 1000 | 3 | 0 |
| 1000 | 4 | 0 |
| 1000 | 5 | 0 |

  So, in the above example, only the $2^{nd}$ SNP had a non-zero genetic effect at the $1000^{th}$ iteration.

- out_allele.txt: Contains samples from the predictive distribution for $\theta_{l,N+1}$. Useful for looking at the predictive distribution for the allele frequency at each SNP.

| *iteration* | *SNP* | $\theta_{l,N+1}$ |
|---|---|---|
| 1000 | 1 | 0.3456 |
| 1000 | 2 | -1.3234 |
| 1000 | 3 | 1.4568 |
| 1000 | 4 | -2.3455 |
| 1000 | 5 | 0.7645 |

# References

W. R. Gilks and P. Wild. Adaptive rejection sampling for Gibbs sampling. *JRSS C: Applied Statistics*, 41:337–348, 1992.

J.P. Huelsenbeck and P. Andolfatto. Inference of population structure under a Dirichlet process model. *Genetics*, 175:1787–1802, Apr 2007.

H. Ishwaran and L.F. James. Gibbs sampling methods for stick breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

R.M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

N.M. Pajewski and P.W. Laud. A flexible Bayesian semiparametric approach to genetic association studies of quantitative traits in the presence of population stratification. *submitted*, 2008a.

N.M. Pajewski and P.W. Laud. Posterior computation for hierarchical Dirichlet Process Mixture models: An application to genetic association studies of quantitative traits in the presence of population stratification. Technical report # 55, Medical College of Wisconsin, 2008b.

P. Scheet and M. Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.*, 78:629–644, Apr 2006.