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_probile[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 $_i$, 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 les as input, a parameter setup le *(default name: parm_setup.txt)*, a le containing the observed genotype data

The program then randomly allocates each individual amongst Initial Clusters distinct atoms. In our experience, a random con guration of roughly 8 to 10 initial distinct points seems to perform adequately well. Line 7 represents the nite limit approxi-

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*		···
1	0	1	1	0	0	0	···
2	0	1	0	0	0	1	···
3	1	0	0	0	1	0	···

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

iteration	*SNP*	*E ect*	*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 e ects were clustered outside of the no e ect (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 e ect at the 1000^{th} iteration.

- out_allele.txt: Contains samples from the predictive distribution for $_{l;N+1}$

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 exible Bayesian semiparametric approach to genetic association studies of quantitative traits in the presence of population strati cation. *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 strati cation. Technical report # 55, Medical College of Wisconsin, 2008b.

P. Scheet and M. Stephens. A fast and exible 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.