STA613/CBB540 HOMEWORK 5

DUE TUESDAY, MARCH 5^{TH}

(1) *Gibbs Sampling.* We return to question 3 from the previous homework, now in order to consider a Gibbs sampling algorithm to estimate parameters for a Gaussian mixture model.
   (a) Write out the conditional probability of the soft assignment variables Z, conditioned on all of the other variables in the model (for each of the K components: mixture proportions π_j, cluster specific means μ_j and covariance matrices Σ_j, and data X). Note that the equation for the expected value of variable Z for a new data point X^* might be useful in this context.
   (b) Write out the conditional probability of the mixture proportions π_j, with prior $\pi \sim Dir(\alpha)$ with $\alpha = 1$.
   (c) Write out the conditional probability of the parameters μ_j, with prior $\mu_j \sim \mathbb{N}(0, 1)$.
   (d) Let's say we have another way to estimate parameters Σ_j. Write out pseudocode for how you would estimate the model parameters using a Gibbs sampler based on these conditional probabilities, including what parameter(s) you would initialize to start with and how you choose to aggregate the Gibbs sample estimates.

(2) *Admixture models.* Let's say I have a collection of mutts: dogs that are admixed from purebreds. Let's also say I have genetic data from a purebred labrador and from a purebred poodle. Using the admixture model from the [Pritchard et al.] paper, how can I find the following information (for a given number of ancestral populations in the model)? If you can't get this information from the model, say why; otherwise, describe the parameters from which you can extract the information.
   (a) What proportion of each dogs' genome come from each of my K ancestral populations?
   (b) What are the allele frequencies for each of the ancestral populations?
   (c) What specific alleles distinguish ancestral population A from B?
   (d) What specific allele frequencies are shared between ancestral population A and B?
   (e) Let's include the genetic data for the labrador and the poodle in the model, but fix their Q^i parameters so that they both are 100% descended from populations 1 and 2, respectively. How can I determine the proportion of dog C's ancestry from poodle?
   (f) Using the same procedure, how can I determine the proportion of dog C's ancestry from dachshund?

(g) Using the same proceedure, at a specific locus, what proportion of my sample appears to have poodle ancestry?

(3) *PCA and factor analysis.* Let's consider two different types of population structure, and how well PCA and factor analysis perform on these models. Download population_a.txt and population_b.txt from the course website. Population A has 210 individuals; population B has 225.
   (a) Normalize each site of each of the files as in [Patterson et al.]. Compute the covariance matrix for the individuals and then the principal components (see: eigen). For both data sets, plot the first two PCs (they should be vectors of length 210, 225)
   (b) Without normalizing, compute the factors for the two populations (see: factanal). Use two factors for each and plot the two factor loadings against each other (they should be vectors of length 210, 225). Use three factors each, and plot each factor loading separately (i.e., x-axis is individual number, y-axis is factor loading value). What is the difference between the different numbers of factors?
   (c) Compare the structure found in the PC-based analysis versus the FA-based analysis.

(4) *Markov Chains and Wright-Fisher model.* Say you are studying a set of 1000 individuals, and you find that there are 400 instances of the minor allele in your sample.
   (a) Build a Markov chain simulator in R, so that, starting at this point, you can simulate possible trajectories of the allele frequency in this stable, non-migrating, diploid population with no mutation or selection (i.e., given the assumptions of the Wright-Fisher model).
   (b) Run this chain 1000 times until either the minor allele or the major allele is fixed. Plot these trajectories (x-axis: time steps; y-axis: number of copies of the minor allele).
   (c) What is the probability, from your simulations, that the minor allele will ultimate be fixed?
   (d) What is the average amount of time for either allele to be fixed based on your simulations? What is the average amount of time for the minor allele to be fixed?