2 EM Implementation and data analysis [35 pts]

You are given two datasets each of which contains genotype data from 1000 individuals at 5000 SNPs. Dataset 1 contains the values of the hidden values and parameters that you can use to debug.

For this question, you will implement an EM algorithm and obtain MLE of the parameters of the mixture model for each dataset as well as infer the hidden variables. The first dataset (dataset 1) includes the true values of the θ and Z and can be used to test your implementation. The second dataset (dataset 2) does not include the true parameter values.

Use the following specifications for the EM.

- *Initialization*: To initialize, draw each $f_{j,k}$ from a uniform random variable. Draw each π_k from an exponential random variable and normalize (remember that π_k needs to sum to 1). Alternately, draw π from a Dirichlet distribution.
- *Stopping criterion*: Run for at most 100 iterations or till the change in log likelihood $< 10^{-8}$.
- *Number of random restarts*: Since EM only reaches a local optimum, we often run the algorithm with multiple random initializations and pick the solution with the highest log likelihood. We use 3 restarts.
- *The number of mixture components or populations K*: Set $K = 2$.

(a) For dataset 1, plot the log likelihood as a function of iteration for a single run of EM (*i.e.* for a single initialization of EM).

(b) For dataset 1, what is the MLE of π?

(c) For dataset 1, fixing π and f at its MLE from the previous question, we can compute the posterior probabilty of Z_i for each indvidual. Note that these posterior probabilities are computed in the last E-step of the EM algorithm. We would like to compute the accuracy of the inferred Z_i by comparing the posterior probability of Z_i to the true Z_i available in mixture1.ganc file (each row of this file indicates which population the corresponding individual belongs to). The inferred population label for individual i is set to the MAP (maximum a posteriori) estimate, *i.e.*, choose the population with the maximum value of the posterior probability. Our measure of accuracy is simply the fraction of individuals for which the MAP estimate matches the true population label.

On dataset 1, What is the accuracy of the inferred population label?

(d) How similar are the solutions on dataset 1 across the different initializations. Report the log likelihoods of the optimum found as well as the accuracy of the population label.

(e) To see how accuracy of predicting population label 1 changes as we get more SNPs, redo the EM using only the first 10, 100, 1000 SNPs and 5000 SNPs. Plot accuracy as a function of number of SNPs.

(f) For dataset 2, run EM and report the MLE of π.

(g) For dataset 2, plot how the log likelihood varies with K (try $K = 1, \ldots, 4$). Which value of K would you choose based on this plot ?

Implementation details: The fixed point to which EM algorithm converges is not guaranteed to be the global optimum. To obtain a better solution, it is recommended to start the algorithm at multiple random points in the parameter space and pick the theta with the maximum likelihood across the random restarts.

The log likelihood is guaranteed to be non-decreasing in each iteration and is a useful check of your implementation.

You will need to compute the posterior probabilities $r_{i,k}$ by applying Bayes theorem. One difficulty is that the likelihood, $P(\boldsymbol{x}_{i,1:m}|z_i = k, \theta)$ can become very small when the number of SNPs is large leading to underflow errors. One solution is to work with the log likelihood and to notice that the posterior probability is invariant if we re-scale each likelihood by the same amount. Let $l_i^* = max_{k'} l_{i,k'}$. We can then write

$$\begin{aligned} r_{i,k} &= \frac{\exp(l_{i,k})}{\sum_{k'} \exp(l_{i,k'})} \\ &= \frac{\exp(l_{i,k} - l_i^*)}{\sum_{k'} \exp(l_{i,k'} - l_i^*)} \end{aligned}$$

The second line is better behaved in practice.

3 Principal Component Analysis [15 pts]

We will apply Principal Component Analysis on the genetic data of 1,092 (real) individuals from the 1000 Genomes Project.

You are given the genotype data containing $M = 13,237$ SNPs for $N = 1,092$ individuals of African, East Asian, (Admixed) American and European descent. Both the SNP data and the true population labels are given.

If using R, please set the seed for the random number generator to 0 in the beginning of your script (set.seed(0)). If using sklearn, use np.random.seed(0).

(a) Use the prcomp function in R to run PCA on this dataset and plot PC1 against PC2 in a scatter plot. Color your points based on the population the individual comes from.

(b) Briefly comment on why you think the first two components of PCA exhibit clustering by population. Why can PCA successfully capture population membership?

(c) Use the kmeans function in R to run the k-means algorithm on the SNP data (*not* the PCs) with K=4 and nstart=5 (*i.e.* use 5 different initializations since kmeans is not guaranteed to converge to a global optimum). Now rename all the clusters you have obtained by size: the largest cluster by number of individuals should be named Cluster1 and the smallest cluster should be named Cluster4. Generate the same plot as in part (a) (PC1 vs. PC2), but this time color your points based on cluster assignments (*i.e.* was an individual assigned to cluster 1, 2, 3 or 4) instead of the population labels.

(The location of the points on the PC1 vs. PC2 plot should not change, but their colors should now indicate cluster membership instead of true population labels.)

(d) Match clusters to population labels by inspection (*e.g.* "Cluster 1 most closely resembles the ASN population.") What fraction of the cluster assignments agree with the true population labels?