CM124/224, Fall 2023

Problem Set 4: Graphical models and EM implementation

Due Dec 8, 11:59pm (First two questions)

Due Dec 15, 11:59pm (EM implementation)

1 Correcting for population stratification in GWAS with PCA [15 pts]

In class, we have talked about how population structure can lead to false discoveries in genome-wide association studies. One approach to correct for population stratification in GWAS is to "explain away" population stratification with principal component analysis.

You are given a genotype matrix containing $M = 2000$ SNPs and $N = 1000$ individuals and the binary phenotype associated with each individual.

(a) Run linear regressions on each SNP and the phenotype and report your p-values in a QQ plot. Do you observe inflation in your p-values? How many SNPs are significantly associated with the phenotype at $\alpha = 0.05$? Control FWER using Bonferroni correction.
In R, you can use the lm function:
(https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/lm)
and in Python, you can use the statsmodel.api.OLS function:
(https://www.statsmodels.org/dev/examples/notebooks/generated/ols.html)
to compute linear regression estimates and their associated p-values. **Solution:**

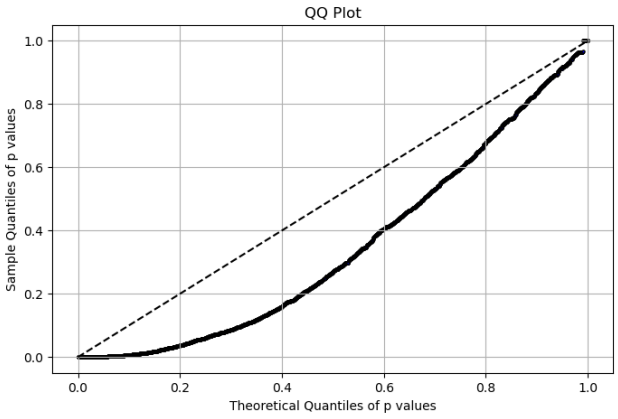

Figure 1: QQ Plot

We observe inflation in p-values. 14 SNPs are significantly associated with the phenotype at $\alpha = 0.05$ after Bonferroni correction procedure.

(b) Use the prcomp function in R (Documentation: prcomp) or the sklearn.decomposition.PCA function in Python (Documentation: sklearn PCA) to perform PCA on the genotype matrix and plot PC1 against PC2. How many populations do you think are represented in the dataset, under the assumption that each individual can only belong to one population (i.e. not considering admixture)? **Solution:**

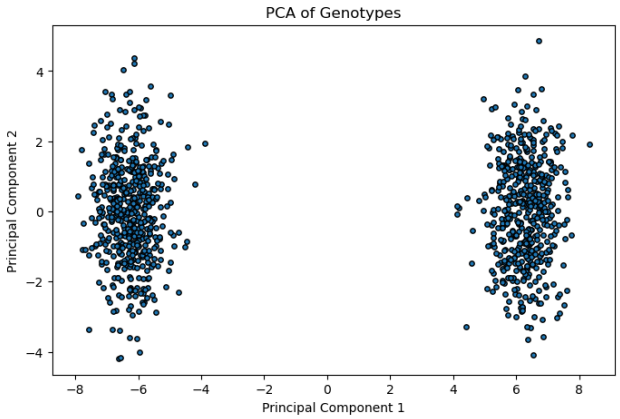

Figure 2: PC1 vs PCv2

There are two populations which are separated by the PC1.

(c) Run linear regressions on each one of your SNPs again, but this time include a term for PC1 in your regression. What does the QQ plot indicate? How many SNPs are significantly associated with the phenotype at $\alpha = 0.05$? Again, control FWER using Bonferroni correction. **Solution:**

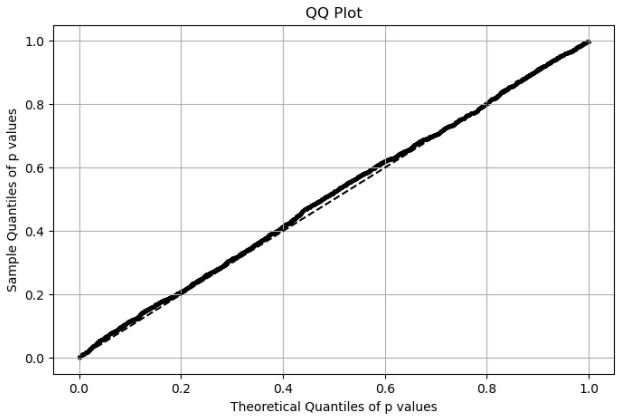

Figure 3: QQ plot after including population parameter

We do not observe inflation after including the PC1 term. There are no SNPs significantly associated with the phenotype at $\alpha = 0.05$ after Bonferroni correction procedure.

2 Graphical models and phylogenetic trees [10 pts]

In this problem, we will study graphical models structured as trees. Recall that we can label the two alleles at a SNP as 0 and 1. We observe the frequency of the 1 allele in a population, *i.e.*, the fraction of individuals in the population that carry the 1 allele at the SNP. We assume that we observe the frequencies of an allele in different populations.

Tree-structured graphical models are a natural model to represent populations relationships. Each node corresponds to a population. The leaves of the tree are the observed values of allele frequencies in present-day populations while the internal nodes are the allele frequencies in an ancestral population.

Let $X_1 \in [0, 1]$ denote the frequency of the 1 allele at a single SNP in population 1. Population 1 splits into 2 populations, 2 and 3. The allele frequency in 2 and 3 are independent given X_1. Population 2 then splits into 4 and 5. We can represent the joint distribution of allele frequencies in the five populations $(X_1, X_2, X_3, X_4, X_5)$ as a graphical model (Figure 4) where each $X_i \in [0, 1]$.

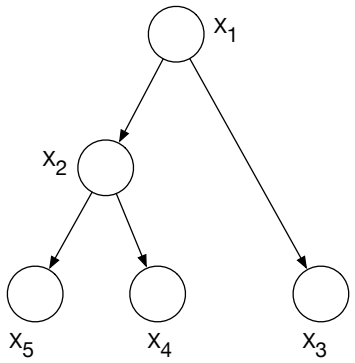

Figure 4:

(a) [4 pts] Write down the joint distribution of $P(x_1, x_2, x_3, x_4, x_5)$ as a product of conditional probabilities. **Solution:**

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)\, P(x_2 \mid x_1)\, P(x_3 \mid x_1)\, P(x_5 \mid x_2)\, P(x_4 \mid x_2)$$

(b) [6 pts] For the graphical model in Figure 4, which of the following conditional independence statements holds ?

i. $X_4 \perp\!\!\!\perp X_5$ **Solution:** False

$$p(x_4|x_5) \stackrel{?}{=} p(x_4)$$

$$RHS : p(x_4) = \sum_{x_1}\sum_{x_2}\sum_{x_3}\sum_{x_5} p(x_1, x_2, x_3, x_4, x_5)$$

$$= \sum_{x_1}\sum_{x_2}\sum_{x_3}\sum_{x_5} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_5|x_2)p(x_4|x_2)$$

$$= \sum_{x_1} p(x_1) \sum_{x_2} p(x_2|x_1)p(x_4|x_2) \sum_{x_3} p(x_3|x_1) \sum_{x_5} p(x_5|x_2)$$

$$= \sum_{x_1} p(x_1) \sum_{x_2} p(x_2|x_1)p(x_4|x_2)$$

$$LHS : p(x_4|x_5) = \frac{p(x_4, x_5)}{p(x_5)}$$

$$p(x_4, x_5) = \sum_{x_1} \sum_{x_2} \sum_{x_3} p(x_1, x_2, x_3, x_4, x_5)$$

$$= \sum_{x_1} \sum_{x_2} \sum_{x_3} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_5|x_2)p(x_4|x_2)$$

$$= \sum_{x_1} p(x_1) \sum_{x_2} p(x_2|x_1)p(x_5|x_2)p(x_4|x_2) \sum_{x_3} p(x_3|x_1)$$

$$= \sum_{x_1} p(x_1) \sum_{x_2} p(x_2|x_1)p(x_5|x_2)p(x_4|x_2)$$

$$p(x_5) = \sum_{x_1} \sum_{x_2} \sum_{x_3} \sum_{x_4} p(x_1, x_2, x_3, x_4, x_5)$$

$$= \sum_{x_1} \sum_{x_2} \sum_{x_3} \sum_{x_4} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_5|x_2)p(x_4|x_2)$$

$$= \sum_{x_1} p(x_1) \sum_{x_2} p(x_2|x_1)p(x_5|x_2) \sum_{x_3} p(x_3|x_1) \sum_{x_4} p(x_4|x_2)$$

$$= \sum_{x_1} p(x_1) \sum_{x_2} p(x_2|x_1)p(x_5|x_2)$$

$$LHS : \frac{p(x_4, x_5)}{p(x_5)} = \frac{\sum_{x_1} p(x_1) \sum_{x_2} p(x_2|x_1)p(x_5|x_2)p(x_4|x_2)}{\sum_{x_1} p(x_1) \sum_{x_2} p(x_2|x_1)p(x_5|x_2)}$$

$$LHS : \frac{\sum_{x_1} p(x_1) \sum_{x_2} p(x_2|x_1)p(x_5|x_2)p(x_4|x_2)}{\sum_{x_1} p(x_1) \sum_{x_2} p(x_2|x_1)p(x_5|x_2)} \neq \sum_{x_1} p(x_1) \sum_{x_2} p(x_2|x_1)p(x_4|x_2) : RHS$$

Therefore, LHS $\neq$ RHS which means, $X_4 \perp\!\!\!\perp X_5$ is FALSE.

ii. $X_4 \perp\!\!\!\perp X_5|X_1$ **Solution:** False

$$p(x_4|x_1) \stackrel{?}{=} p(x_4|x_5, x_1)$$

$$LHS : p(x_4|x_1) = \frac{p(x_4, x_1)}{p(x_1)}$$

$$p(x_4, x_1) = \sum_{x_2} \sum_{x_3} \sum_{x_5} p(x_1, x_2, x_3, x_4, x_5)$$

$$= \sum_{x_2} \sum_{x_3} \sum_{x_5} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_5|x_2)p(x_4|x_2)$$

$$= p(x_1) \sum_{x_2} p(x_2|x_1)p(x_4|x_2) \sum_{x_3} p(x_3|x_1) \sum_{x_5} p(x_5|x_2)$$

$$= p(x_1) \sum_{x_2} p(x_2|x_1)p(x_4|x_2)$$

$$LHS : \frac{p(x_1) \sum_{x_2} p(x_2|x_1)p(x_4|x_2)}{p(x_1)} = \sum_{x_2} p(x_2|x_1)p(x_4|x_2)$$

$$RHS : p(x_4|x_5, x_1) = \frac{p(x_4, x_5, x_1)}{p(x_5, x_1)}$$

$$p(x_5, x_1) = \sum_{x_2} \sum_{x_3} \sum_{x_4} p(x_1, x_2, x_3, x_4, x_5)$$

$$= \sum_{x_2} \sum_{x_3} \sum_{x_4} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_5|x_2)p(x_4|x_2)$$

$$= p(x_1) \sum_{x_2} p(x_2|x_1)p(x_5|x_2) \sum_{x_3} p(x_3|x_1) \sum_{x_4} p(x_4|x_2)$$

$$= p(x_1) \sum_{x_2} p(x_2|x_1)p(x_5|x_2)$$

$$p(x_4, x_5, x_1) = \sum_{x_2} \sum_{x_3} p(x_1, x_2, x_3, x_4, x_5)$$

$$= \sum_{x_2} \sum_{x_3} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_5|x_2)p(x_4|x_2)$$

$$= p(x_1) \sum_{x_2} p(x_2|x_1)p(x_5|x_2)p(x_4|x_2) \sum_{x_3} p(x_3|x_1)$$

$$= p(x_1) \sum_{x_2} p(x_2|x_1)p(x_5|x_2)p(x_4|x_2)$$

$$RHS : \frac{p(x_4, x_5, x_1)}{p(x_5, x_1)} = \frac{p(x_1) \sum_{x_2} p(x_2|x_1)p(x_5|x_2)p(x_4|x_2)}{p(x_1) \sum_{x_2} p(x_2|x_1)p(x_5|x_2)} = \frac{\sum_{x_2} p(x_2|x_1)p(x_5|x_2)p(x_4|x_2)}{\sum_{x_2} p(x_2|x_1)p(x_5|x_2)}$$

$$LHS : \sum_{x_2} p(x_2|x_1)p(x_4|x_2) \neq \frac{\sum_{x_2} p(x_2|x_1)p(x_5|x_2)p(x_4|x_2)}{\sum_{x_2} p(x_2|x_1)p(x_5|x_2)} : RHS$$

Therefore, LHS $\neq$ RHS which means, $X_4 \perp\!\!\!\perp X_5|X_1$ is FALSE.

iii. $X_4, X_5 \perp\!\!\!\perp X_3|X_1, X_2$ **Solution:** True

$$p(x_4, x_5|x_1, x_2) \stackrel{?}{=} p(x_4, x_5|x_1, x_2, x_3)$$

or(easier)

$$p(x_3|x_1, x_2) \stackrel{?}{=} p(x_3|x_1, x_2, x_4, x_5)$$

Continuing with the second equation:

$$LHS : p(x_3|x_1, x_2) = \frac{p(x_1, x_2, x_3)}{p(x_1, x_2)}$$

$$RHS : p(x_3|x_1, x_2, x_4, x_5) = \frac{p(x_1, x_2, x_3, x_4, x_5)}{p(x_1, x_2, x_4, x_5)}$$

Then continuing with RHS,

$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_5|x_2)p(x_4|x_2)$$

$$p(x_1, x_2, x_4, x_5) = \sum_{x_3} p(x_1, x_2, x_3, x_4, x_5) = \sum_{x_3} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_5|x_2)p(x_4|x_2)$$

$$= p(x_1)p(x_2|x_1)p(x_5|x_2)p(x_4|x_2) \sum_{x_3} p(x_3|x_1)$$

$$= p(x_1)p(x_2|x_1)p(x_5|x_2)p(x_4|x_2)$$

$$RHS : \frac{p(x_1, x_2, x_3, x_4, x_5)}{p(x_1, x_2, x_4, x_5)} = \frac{p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_5|x_2)p(x_4|x_2)}{p(x_1)p(x_2|x_1)p(x_5|x_2)p(x_4|x_2)} = p(x_3|x_1)$$

Continuing with LHS,

$$p(x_1, x_2, x_3) = \sum_{x_4} \sum_{x_5} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_5|x_2)p(x_4|x_2)$$

$$= p(x_1)p(x_2|x_1)p(x_3|x_1) \sum_{x_4} p(x_4|x_2) \sum_{x_5} p(x_5|x_2)$$

$$= p(x_1)p(x_2|x_1)p(x_3|x_1)$$

$$p(x_1, x_2) = \sum_{x_3} \sum_{x_4} \sum x_5 p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_5|x_2)p(x_4|x_2)$$

$$= p(x_1)p(x_2|x_1) \sum_{x_3} (x_3|x_1) \sum_{x_4} p(x_4|x_2) \sum_{x_5} p(x_5|x_2)$$

$$= p(x_1)p(x_2|x_1)$$

$$LHS : \frac{p(x_1, x_2, x_3)}{p(x_1, x_2)} = \frac{p(x_1)p(x_2|x_1)p(x_3|x_1)}{p(x_1)p(x_2|x_1)} = p(x_3|x_1) : RHS$$

Therefore, LHS $\triangleq$ RHS which means, $X_4, X_5 \perp\!\!\!\perp X_3|X_1, X_2$ is TRUE.

3 EM Implementation and data analysis [35 pts]

In problem set 3, we defined a mixture model for genetic data and used it to infer an individual's genetic ancestry. For simplicity, we made the assumption that a given individual belongs to one of $K = 2$ populations. Here, we will relax this assumption and allow the random variable Z_i, denoting the population of individual i, to take any discrete value for arbitrary K populations. In other words, now we have hidden assignment $Z_i \in \{1, 2, \ldots, K\}$ and soft ancestry assignment matrix $r \in \mathbf{R}^{N \times K}$, where $r_{i,k}$ denotes $P(Z_i = k | x_i, \theta)$. In a similar fashion, we will expand our model parameters. Specifically, we column-wise append our allele frequency vectors $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_K$ to form an allele frequency matrix $F \in \mathbf{R}^{M \times K}$, where $F_{j,k}$ denotes how frequently the allele encoded as 1 is seen at SNP/feature j for individuals belonging to population k. We also expand our proportion variable to a vector $\boldsymbol{\pi} \in \mathbf{R}^K$, where π_k denotes the proportion of individuals coming from population k.

You will implement the full EM algorithm, which alternates between obtaining the MLE of the parameters of the mixture model and inferring the hidden variables until convergence. We have provided two datasets, each of which contains genotype data from $N = 1000$ individuals at $M = 250$ SNPs/features. The first dataset (dataset 1) also includes the true values of F (mixture1.freq) and Z (mixture1.ganc) and can be used to test your implementation. The second dataset (dataset 2) does not include the true parameter values. In addition, we have also provided skeleton codes (both `python` and `R` version), in which the main structure of the model has already been implemented and only requires you to fill in critical EM updates and convergence related pieces.

(a) *M step* [5 pts]: Implement the following MLE updates on the model parameters

Implementation details:

- When calculating the frequency matrix, if certain cluster has no partial assignment across all samples, directly set the frequency terms associated with that cluster to 0. Otherwise, you might run into a divide-by-zero error.

$$\pi_k^{(t+1)} = \frac{\sum_{i=1}^N r_{i,k}^{(t)}}{N} \qquad f_{j,k}^{(t+1)} = \frac{\sum_{i=1}^N r_{i,k}^{(t)} x_{i,j}}{\sum_{i=1}^N r_{i,k}^{(t)}}$$

What is the MLE of $\boldsymbol{\pi}$ when setting $r^{(t)}$ to be the ground truth genetic ancestry matrix provided? Does your estimation of F look similar to the ground truth frequency matrix?
Solution:

$$\boldsymbol{\pi} = [0.394, 0.606]$$

Yes, Although they are not exactly same, order of magnitudes and first digits are commonly same.

(b) *E step* [5 pts]: Implement the following posterior update on the soft assignment $r_{i,k}^{(t)}$ and calculate the probability of data $P(X_{1:N,1:M} = x_{1:N,1:M}|\theta^{(t)})$

Implementation details:

- You will need to compute the posterior probabilities $r_{i,k}^{(t)}$ by applying Bayes theorem. One difficulty is that the likelihood, $P(\boldsymbol{x}_{i,1:M}|Z_i = k, \theta^{(t)})$ can become very small when the number of SNPs/features is large leading to underflow errors. One solution is to work with the log likelihood.
- Sometimes we need to go back to the original scale when calculating the sum of the exponential of a few log probabilities. Instead of directly exponentiating possibly large negative log-scale values, it is numerically more stable to use the function `logsumexp`.
- If certain entries in your current estimated allele frequency matrix F (the probability of observing certain allele at certain SNP/feature for a particular population) are 0, substitute them with ϵ that are small but still non-zero (e.g. $\epsilon = 10^{-8}$). This would circumvent the issue of calculating the log of 0.

$$\begin{aligned} r_{i,k}^{(t)} &= P(Z_i = k|\boldsymbol{x}_{i,1:M}, \boldsymbol{\theta}^{(t)}) \\ &= \frac{P(\boldsymbol{x}_{i,1:M}, Z_i = k|\boldsymbol{\theta}^{(t)})}{P(\boldsymbol{x}_{i,1:M}|\boldsymbol{\theta}^{(t)})} \end{aligned}$$

$$\begin{aligned} \log\left(P(\boldsymbol{x}_{i,1:M}|Z_i = k, \boldsymbol{\theta}^{(t)})\right) &= \log\left(P(\boldsymbol{x}_{i,1:M}|Z_i = k, \boldsymbol{\theta}^{(t)}) \cdot P(Z_i = k|\boldsymbol{\theta}^{(t)})\right) \\ &= \log\left(\prod_{j=1}^{M}\left(f_{j,k}^{(t)}\right)^{x_{i,j}}\left(1 - f_{j,k}^{(t)}\right)^{(1-x_{i,j})}\right) + \log\left(\pi_k^{(t)}\right) \end{aligned}$$

$$\begin{aligned} \log\left(P(\boldsymbol{x}_{i,1:M}|\boldsymbol{\theta}^{(t)})\right) &= \log\left(\sum_k P(\boldsymbol{x}_{i,1:M}, Z_i = k|\boldsymbol{\theta}^{(t)})\right) \\ &= \log\left(\sum_k \exp\left(\log P(\boldsymbol{x}_{i,1:M}, Z_i = k|\boldsymbol{\theta}^{(t)})\right)\right) \\ &= \text{logsumexp}_k\left(\log P(\boldsymbol{x}_{i,1:M}|Z_i = k, \boldsymbol{\theta}^{(t)})\right) \end{aligned}$$

$$\begin{aligned} \log\left(P(\boldsymbol{x}_{1:N,1:M}|\boldsymbol{\theta}^{(t)})\right) &= \log\left(\prod_{i=1}^{N} P(\boldsymbol{x}_{i,1:M}|\boldsymbol{\theta}^{(t)})\right) \\ &= \sum_{i=1}^{N} \log P(\boldsymbol{x}_{i,1:M}|\boldsymbol{\theta}^{(t)}) \end{aligned}$$

What is the accuracy of the inferred posterior probability r_i when initializing the E step with the true allele frequency matrix F in the mixture1.freq file and $\boldsymbol{\pi}$ set to the MLE obtained from the previous question? Our measurement of accuracy is calculated by comparing the MAP (maximum a posteriori) estimate of the population label for individual i (*i.e.*, choose

the population k with the maximum value of $r_{i,k}$) against the true Z_i available in the mixture1.ganc file. We use `adjusted_rand_score` as our metric. **Solution:** Accuracy is 1 when inferred labels used against the true Z_i which means we obtain the ground truth without any wrong classification at clustering.

(c) *Full ancestry model* [5 pts]: Implement the full model by repeatedly performing E step and M step until convergence. We will use the following specifications.

- *Initialization*: To initialize, we draw each $r_{i,k}$ from an uniform random variable (between 0.01 and 0.99) and normalize the soft assignments across ancestry groups for each sample such that $\sum_k r_{i,k} = 1$. Alternately, one can draw $\boldsymbol{r_i}$ from a Dirichlet distribution.
- *Stopping criterion*: The EM algorithm should by default run for at most 100 iterations or until the change in the average log likelihood per sample is smaller than $< 10^{-4}$, whichever condition is met first.
- *The number of mixture components or populations* K: By default we will set $K = 2$. Yet, your implementation should be flexible enough to allow exploring different values of K.
- *Number of random restarts*: The fixed point to which the 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. We will use 3 restarts by default and keep track of all the local optimums EM converges to. At the end, we will pick the θ with the maximum likelihood across the random restarts as the final estimator of the model.

The log likelihood is guaranteed to be non-decreasing in each iteration and is a useful check of your implementation. Fit the entire EM ancestry model on **just the first 10 SNPs/features in dataset 1**, while leaving everything else to default. Plot the log likelihood (*i.e.* $\log P(X_{1:N,1:M} = x_{1:N,1:M}|\theta^{(t)})$) as a function of iteration for all three internal random initializations of EM.

Solution:

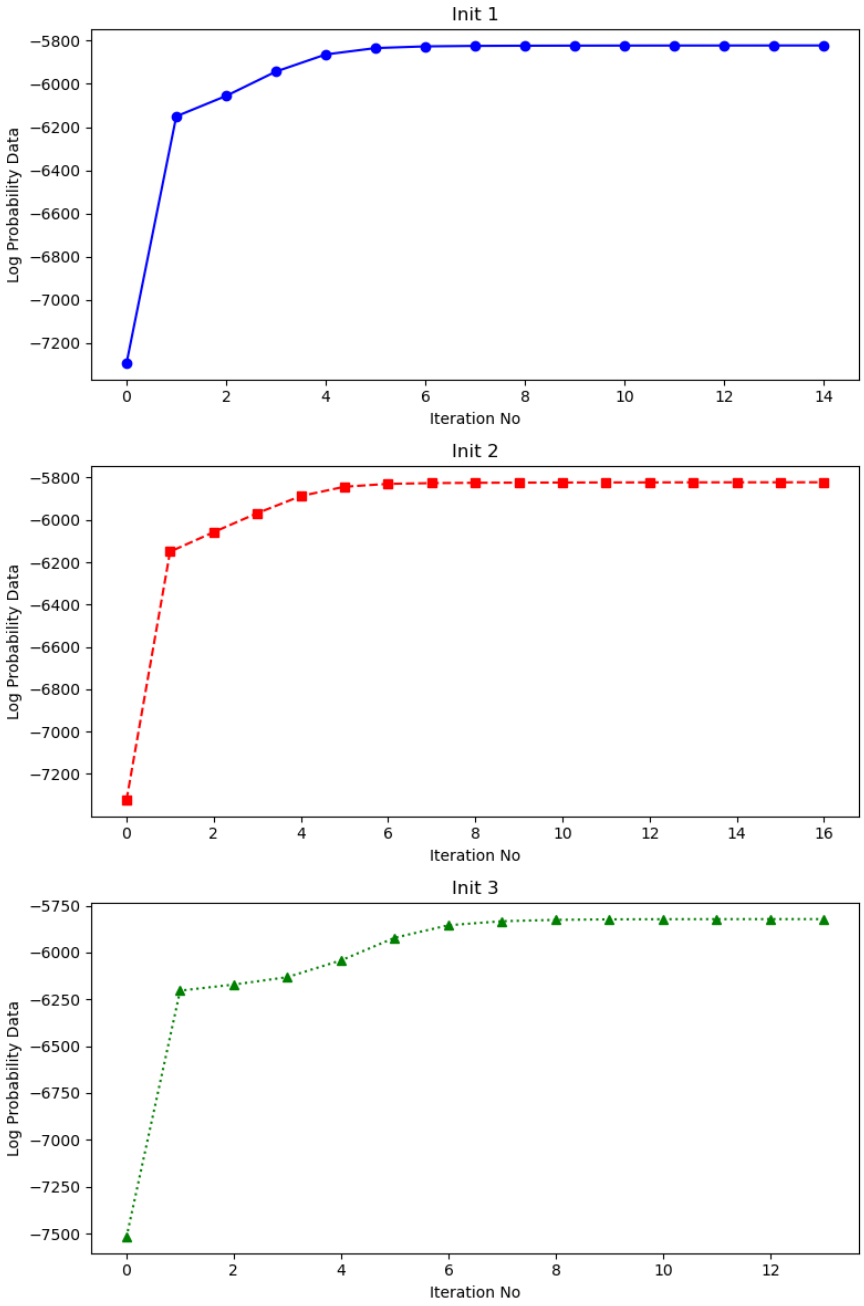

Figure 5: Log Likelihood of Data through iterations

(d) [2 pts] Next, we will explore how similar are the EM solutions across different initializations. Fit the entire EM ancestry model again on **just the first 10 SNPs/features in dataset 1**. **We will in addition, set** $K = 3$ while leaving everything else to default. Report the log likelihoods of the optimum found across all three different initializations.
Solution: Log likelihoods of the optimum:

$$[-5819.384247780635, -5813.984426089208, -5813.458229577478]$$

(e) [3 pts] Let's further visually check the concordance between the population labels obtained from each random initialization from part (d) by projecting samples to the first two principal components, **calculated on the full set of SNPs/features in dataset 1**, and coloring each individual with his or her inferred ancestry label. Visualize all 3 sets of inferred ancestry labels, one per optima/random initialization, and describe any discrepancy you observe. **Solution:**

The main difference is that population labels are different in each initialization, hence the color of points are different in each Figure. However, it is easy to observe that two visible clusters are colored by a single color. Remaining additional color is spread among the two clusters. However, accuracy does not change since we adjusted rand score. The score can easily handle switch at labels. Also, note that log likelihoods are very close to each other.

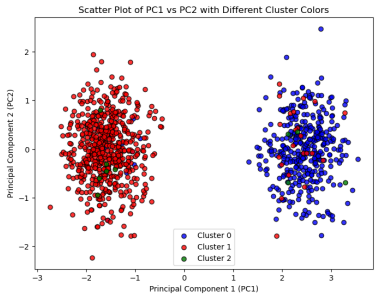

Figure 6: Init 1

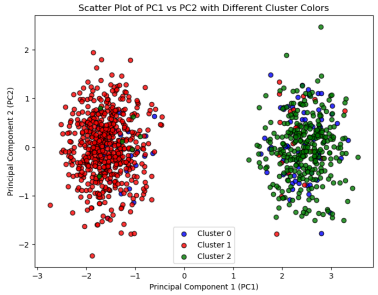

Figure 7: Init 2

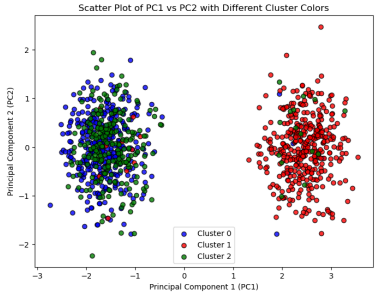

Figure 8: Init 3

(f) [5 pts] To see how the adjusted rand score of the prediction changes as we include more SNPs/features, we now perform EM on dataset 1 **with K = 2**, using the first 10, 25, 50, 100, 150, 200, and 250 SNPs/features while leaving everything else to default. Under each configuration, calculate the adjusted rand score between the ground truth genetic ancestry and the estimated label corresponding to the best random initialization. Plot the adjusted rand score as a function of number of SNPs/features. **Solution:**

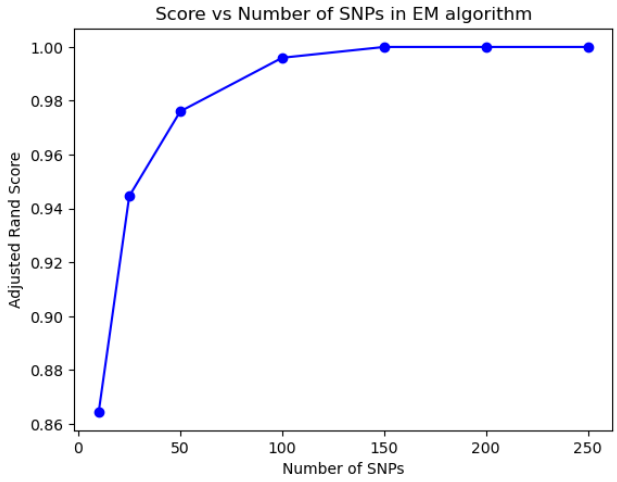

Figure 9: Adjusted Rand Score vs Number of Features

(g) [5 pts] For dataset 2, plot how the log likelihood (corresponding to the best random initialization out of 3) varies under different K ($K = 1, \ldots, 5$). For computational reasons, **try fitting the model on just the first 100 SNPs/features for each K**. Plot the log likelihood as a function of number of components. Which value of K would you choose based on this plot?

Solution:

I would have chosen K = 2 since there is a sharp increase from K = 1 to K = 2, and there is not much positive change between subsequent increases.

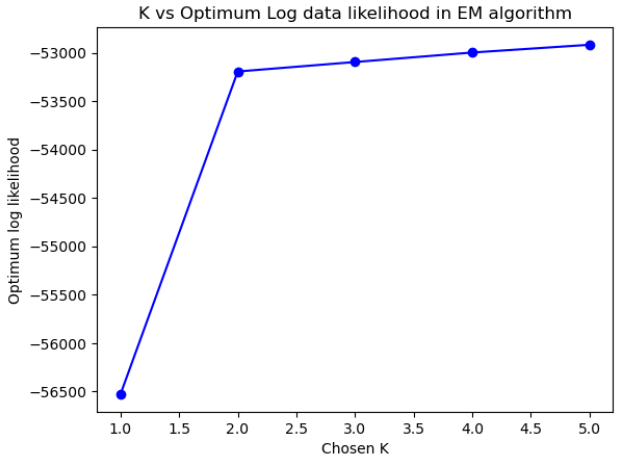

Figure 10: K vs Optimum Log Likelihood

(h) [5 pts] For dataset 2, run EM on the full dataset with $K = 2$ and report the MLE of π.
Solution:

$$\boldsymbol{\pi} = [0.776, 0.224]$$