CMZ26. PS3.

Mingjia Yao

Q1

(a) $$Q(\theta;\theta^{(t)}) = E_{\theta^{(t)}}\left(\sum_{i=1}^{n}\sum_{k=1}^{k} z_{ik}\left(\log \pi_k + \sum_{j=1}^{m}\log P(X_{ij}\mid f_{jk})\right)\right)$$

$$= \sum_{k=1}^{k}\sum_{i=1}^{n} r_{ik}\log\pi_k + \sum_{k=1}^{k}\left(\sum_{i=1}^{n}\sum_{j=1}^{m} r_{ik}\log P(X_{ij}\mid f_{jk})\right)$$

(b) $$r_{ik}^{(t)} = \frac{P(X_i\mid Z_i=k)P(Z_i=k)}{P(X_i)} = \frac{P(X_i\mid Z_i=k)P(Z_i=k)}{\sum_{k'=1}^{k} P(X_i\mid Z_i=k')P(Z_i=k')}$$

$$= \frac{P(X_{i,1:m}\mid f_k^{(t)})\pi_k^{(t)}}{\sum_{k'=1}^{k} P(X_{i,1:m}\mid f_{k'}^{(t)})\pi_{k'}^{(t)}}$$

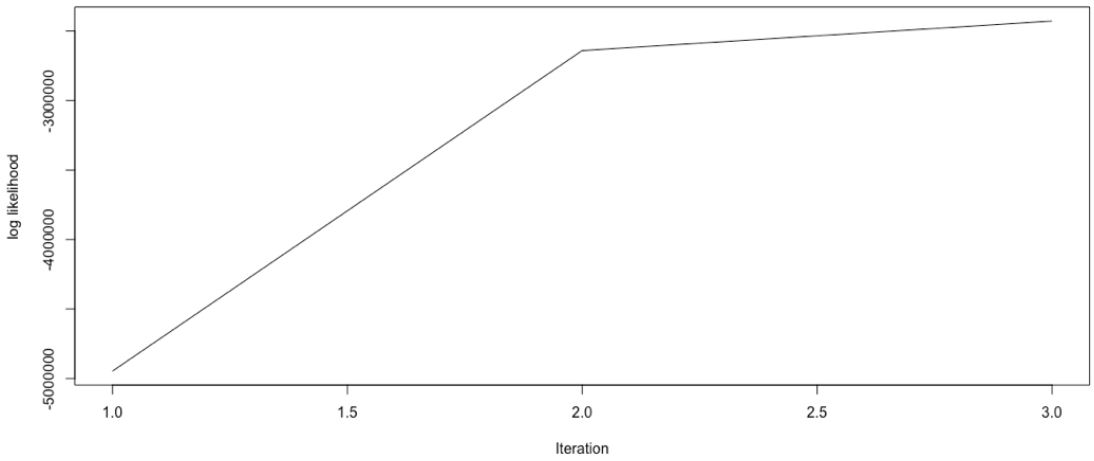
(c)

$$M\text{-step}: \pi_k = \frac{\sum_{i=1}^{n} r_{ik}^{(t)}}{\sum_{k'}\sum_{i=1}^{n} r_{ik'}^{(t)}}$$

$$f_{jk} = \frac{1}{\sum_{i=1}^{n} r_{ik}^{(t)}}\sum_{i=1}^{n} r_{ik}^{(t)} X_{ij}$$

Q2.

(a)



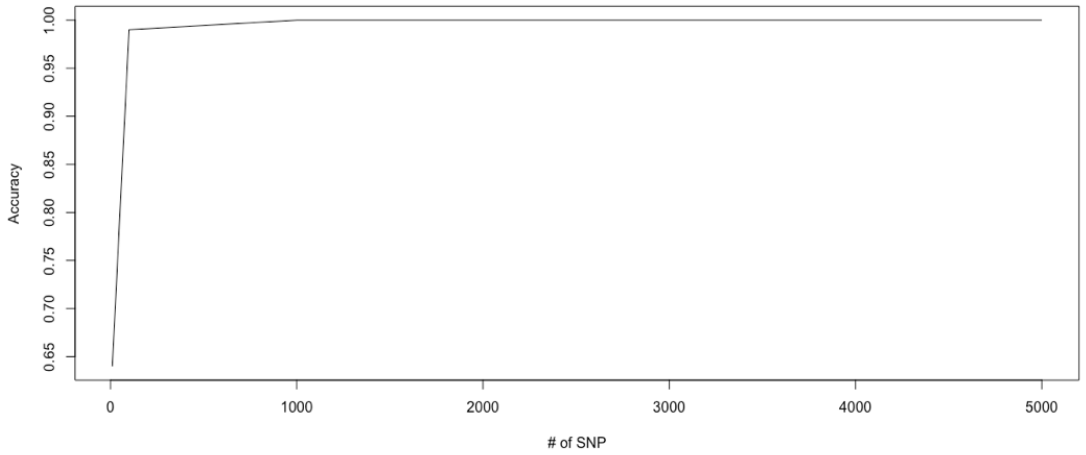
At iteration 1, log likelihood = -4946557.660070642
At iteration 2, log likelihood = -2640862.0329081262
At iteration 3, log likelihood = -2429180.164042054, and converge

(b)
Pi_1 = 0.394
Pi_2 = 0.606

(c)
The accuracy is 100%

(d)
The solutions are very similar, almost totally same across different initializations. The optimum log likelihood is -2429180.164042054, and the accuracy is 100%.

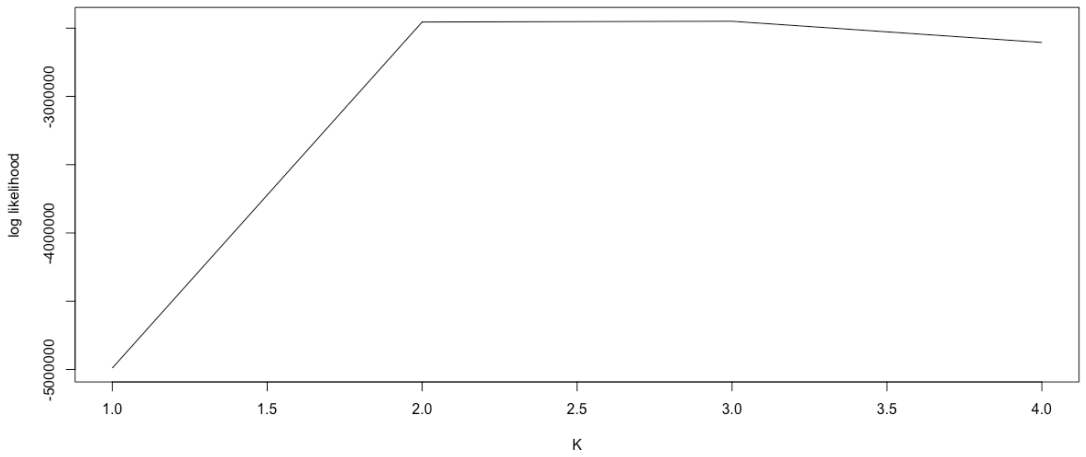
(e)



With 10 SNPs, accuracy = 0.64
With 100 SNPs, accuracy = 0.99
With 1000 SNPs, accuracy = 1.0
With 5000 SNPs, accuracy = 1.0

(f)
Pi_1 = 0.224
Pi_2 = 0.776

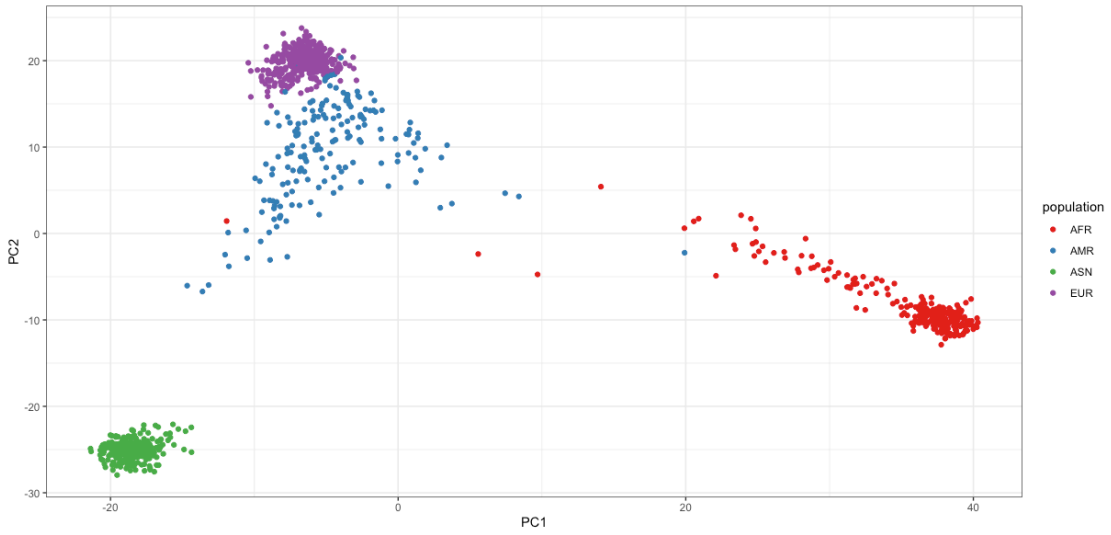
(g)



K=1, log likelihood = -4990114.384554308
K=2, log likelihood = -2454791.9576821947
K=3, log likelihood = -2449315.873989708
K=4, log likelihood = -2605016.469979796.
Based on the plot, I would choose k=3

Q3.

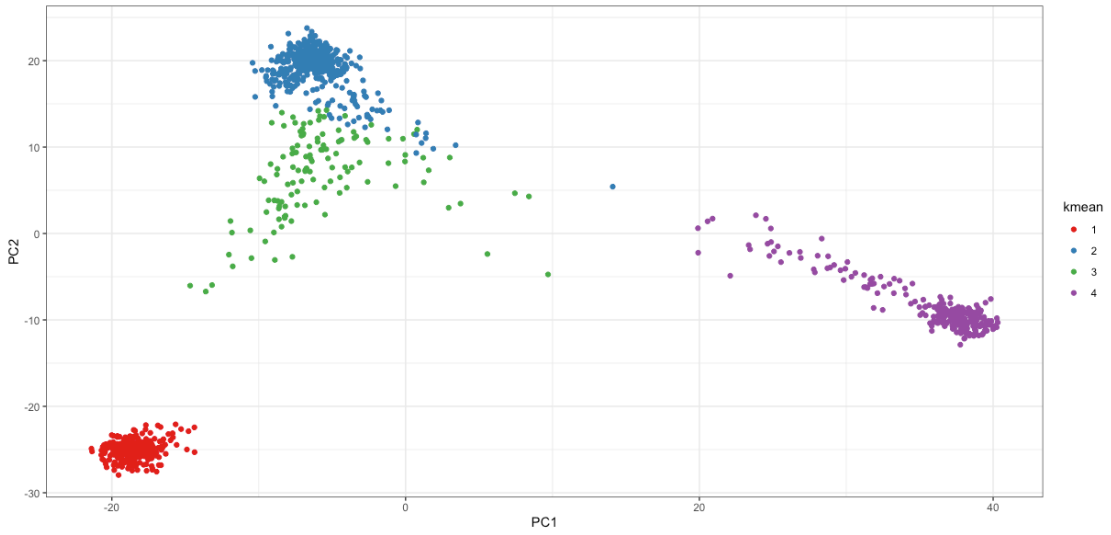
(a)



(b)

Because the PCA method find a direction that maximizes variance of projected data. The PC1 and PC2 are the PC's associated with the largest eigenvalues, thus maximum of variance. With capturing max of variance between individuals, we can cluster them into different populations.

(c)



(d)

0.9413919 of the cluster assignments agree with the true population labels.