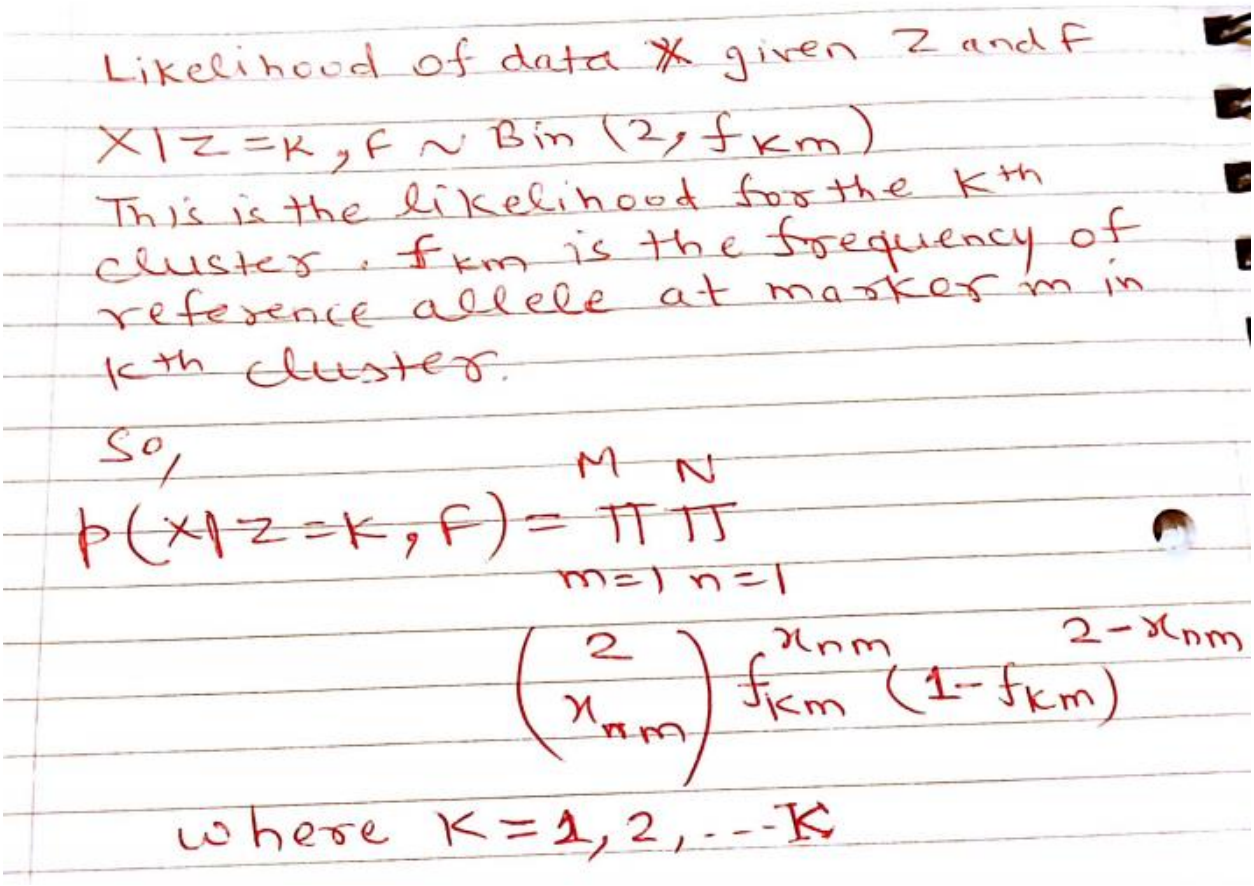
Name: Prashant Pandey
Student ID: 16200112

Bayesian Analysis Project Report

1.



2.

Z is initialized randomly and F is calculated based on cluster membership of individuals in Z.

3.

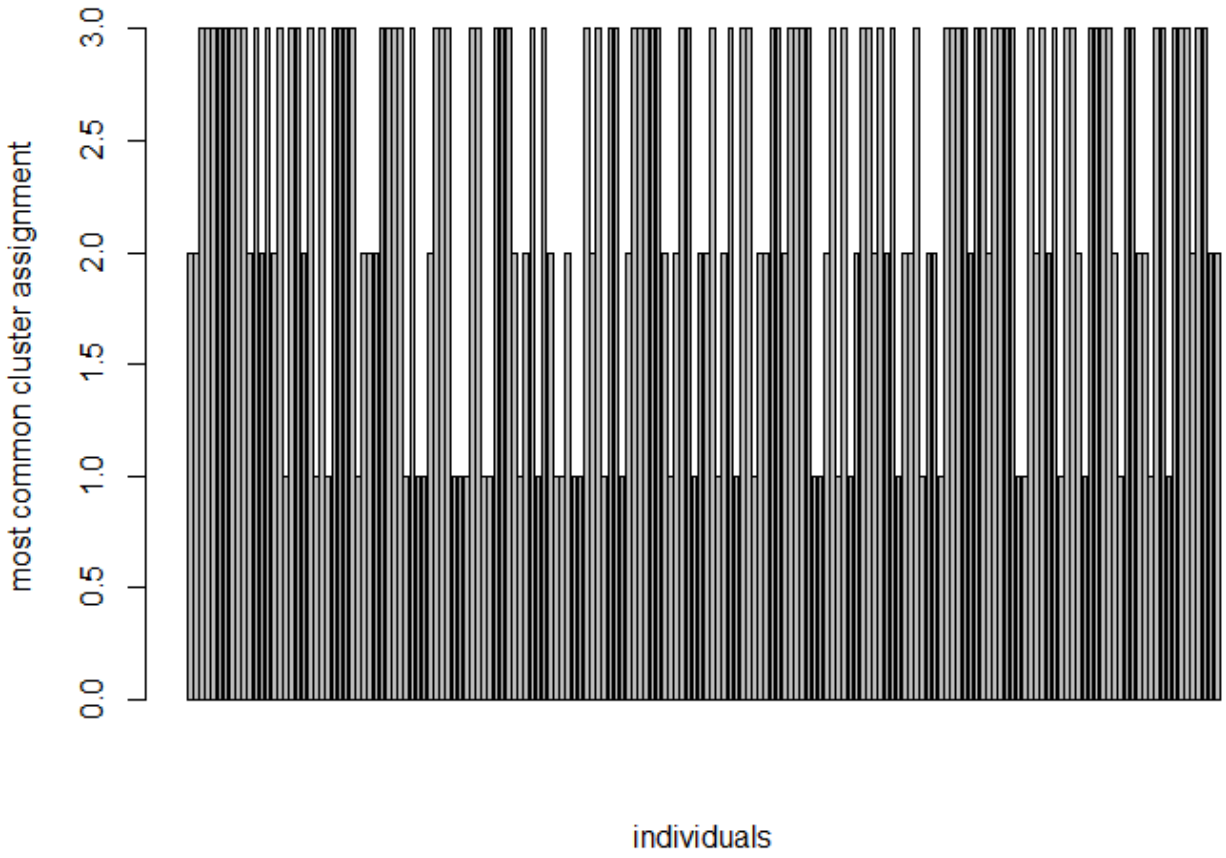
A JAGS model is created with Binomial likelihood. The priors on parameters Z and F are flat.
Z is assigned categorical with equal probability for each category (or cluster here) prior and F is assigned flat beta prior with parameters alpha and beta as one.

4.

Models is run using given data and initialized parameters to sample posterior of parameters Z and F. The number of iterations are 10000 using two chains.

5.

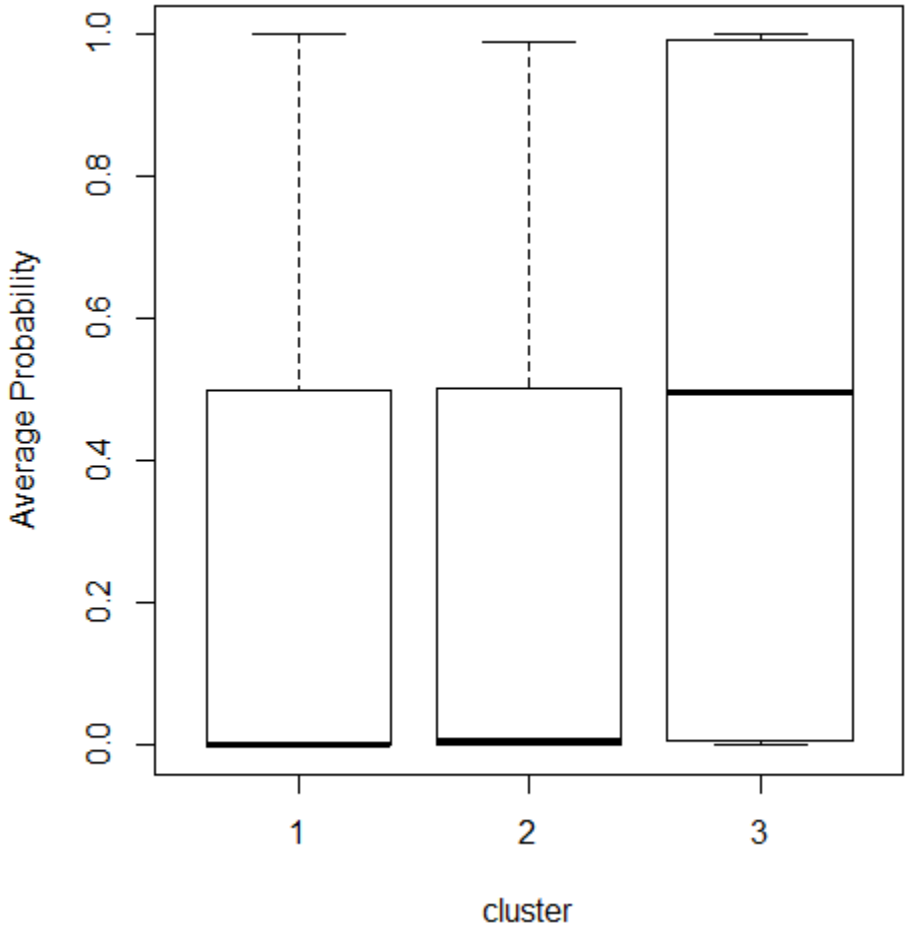
Most common cluster assignment for each individual in (1:172):



	cluster		
population	**1**	**2**	**3**
Dai	10	0	0
French	1	0	27
Ireland	0	0	7
Japanese	28	0	0
Mandenka	0	22	0
Moroccan	0	0	22
Spanish	0	0	34
Yoruba	0	21	0

The inferred clusters do correspond to the population labels. There is a very neat clustering. Populations are clustered like this: {dai, japanese} in cluster 1 , {yoruba, mandenka} in cluster 2, {french,ireland,moroccan,spanish} in cluster 3

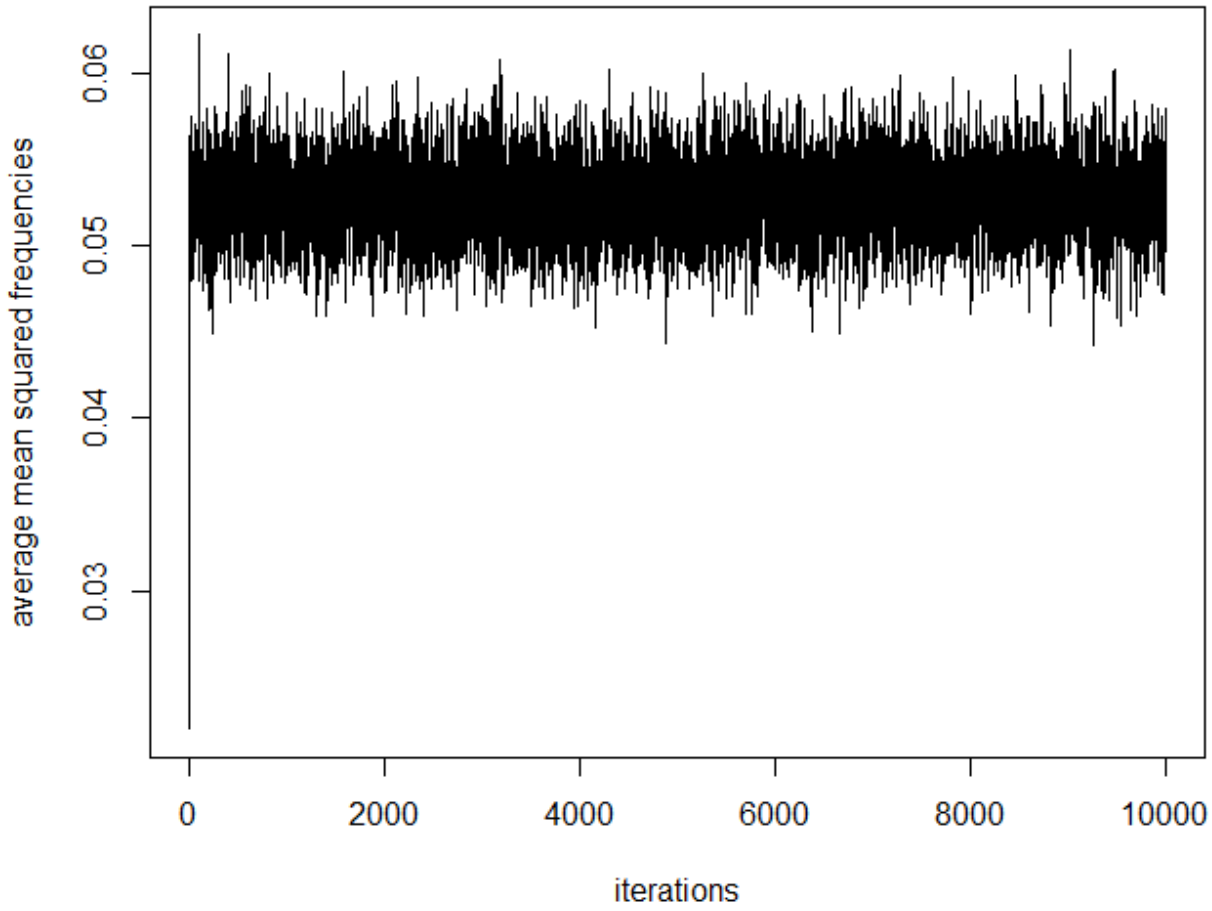
6.



Box-plot showing average probability of individuals in each population assigned to the clusters

Cluster 1 and 2 are spread over smaller range of average probability while cluster 3 is quite spread. This shows that cluster 1 and 2 are tight clusters while cluster three is loose and individuals from this cluster can move in other clusters. In fact, it is the case, when we increase the size of K to 4 and 5, it's this cluster which breaks and individuals from this cluster reshuffles to other clusters. So, there is high randomness in cluster 3.

7.



Traceplot of average mean squared difference in frequencies among all clusters

In the initial number of the iterations (like up to 10) of Gibbs sampling, the average mean squared difference in frequencies among the three clusters was low. This shows that the clusters were highly similar. But in the later number of iterations, when samples started mixing well, the difference increases and clusters become more dissimilar to each other.

8.

Log-posterior

$$p(z, f \mid x) \propto \prod_{m=1}^{M} \prod_{n=1}^{N} f_{km}^{x_{nm}} (1-f_{km})^{2-x_{nm}} \cdot p(f) \cdot p(z)$$

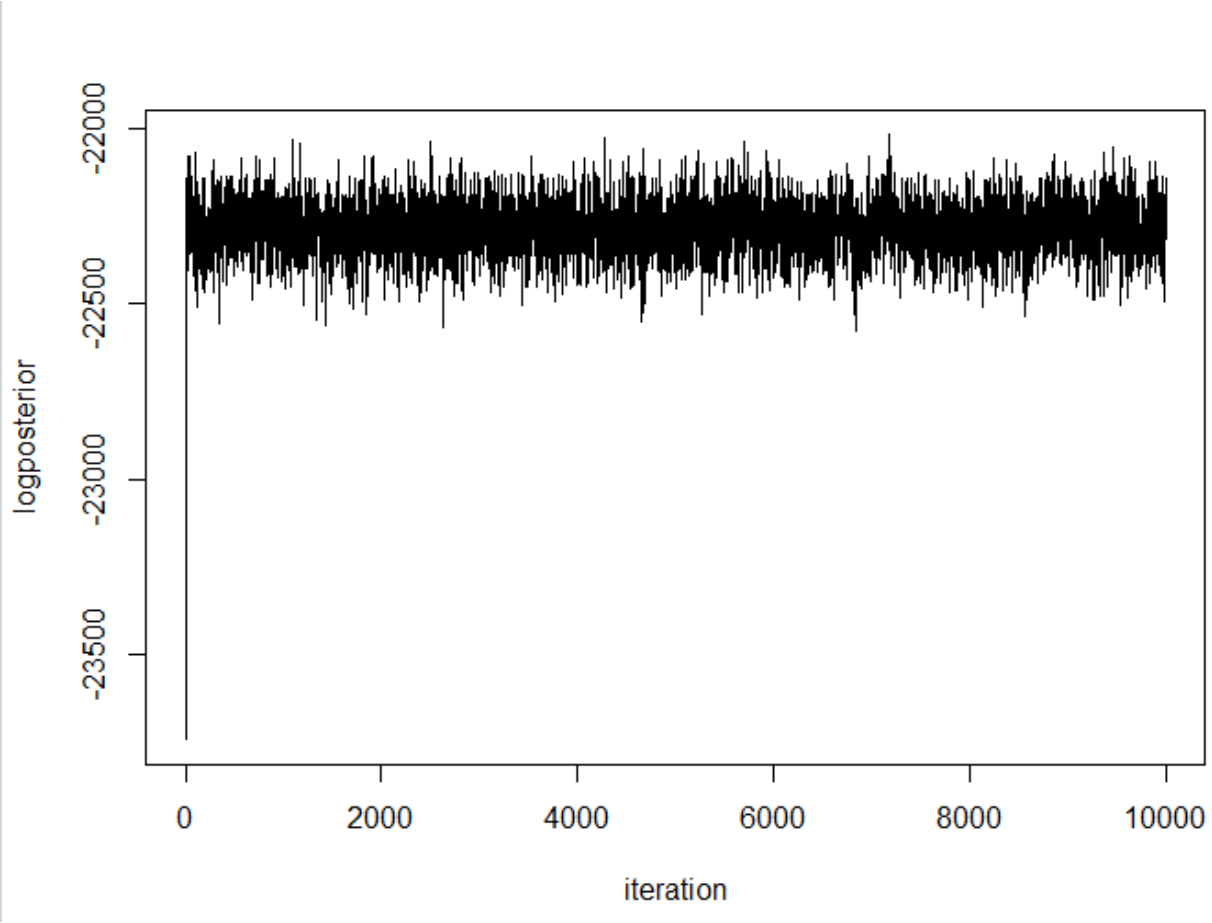
If we take flat prior for $p(f) \propto 1$

Also $p(z) = p_i(z_i = k)$

$$p(z, f \mid x) \propto \prod_{m=1}^{M} \prod_{n=1}^{N} f_{km}^{x_{nm}} (1-f_{km})^{2-x_{nm}} \; p_i(z_i = k)$$

$$\log\left(p(z, f \mid x)\right) \propto \sum_{m=1}^{N} \sum_{n=1}^{N} \left(x_{nm} \log f_{km} + (2 - x_{nm}) \log (1 - f_{km}) + \log p_n(z_n = k) \right)$$

up to a constant of proportionality



Traceplot of log-posterior

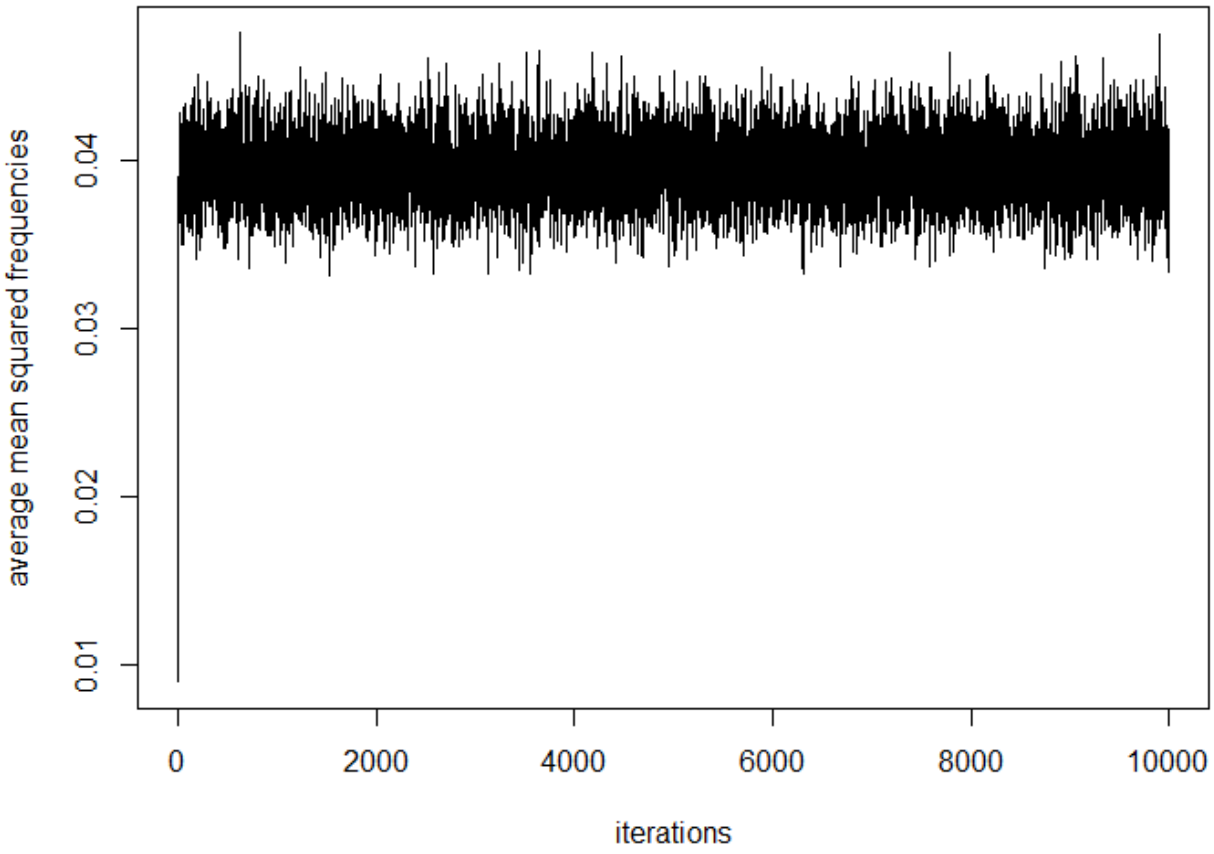
The mixing is quite well for the joint logposterior. The chain seems to have converged after few number of iterations and we are indeed sampling from a stationary distribution which is the posterior distribution.

9.

The sampler **is** sensitive to the initialization. If I initialize Z to be all 1, then the clustering results changes completes. Other clusters are **not even reported** even though a flat categorical prior is assigned to Z giving equal probability to all clusters. Maybe the sampler gets stuck in some local minima/maxima for this kind of initialization.

10.

With strong prior on frequencies (k=3)



Traceplot of average mean squared difference in frequencies among all clusters with a strong Beta prior Beta(9,8) for F.

	cluster		
population	**1**	**2**	**3**
Dai	0	10	0
French	26	2	0
Ireland	7	0	0
Japanese	0	28	0
Mandenka	0	0	22
Moroccan	20	1	1
Spanish	34	0	0
Yoruba	0	0	21

Individuals in each population assigned to different clusters

Here, we can see that if we know from prior beliefs that reference alleles are higher, then the clusters now look more similar because their average mean squared difference in frequencies among all clusters has dropped as compared to that with a flat prior on F. There is less mutation as we are assuming then populations will be more similar. If we see the table above, some individuals have changed clusters as the clusters are now more identical.

Range of average mean squared difference with a strong prior on F
0.01490942 0.04733069

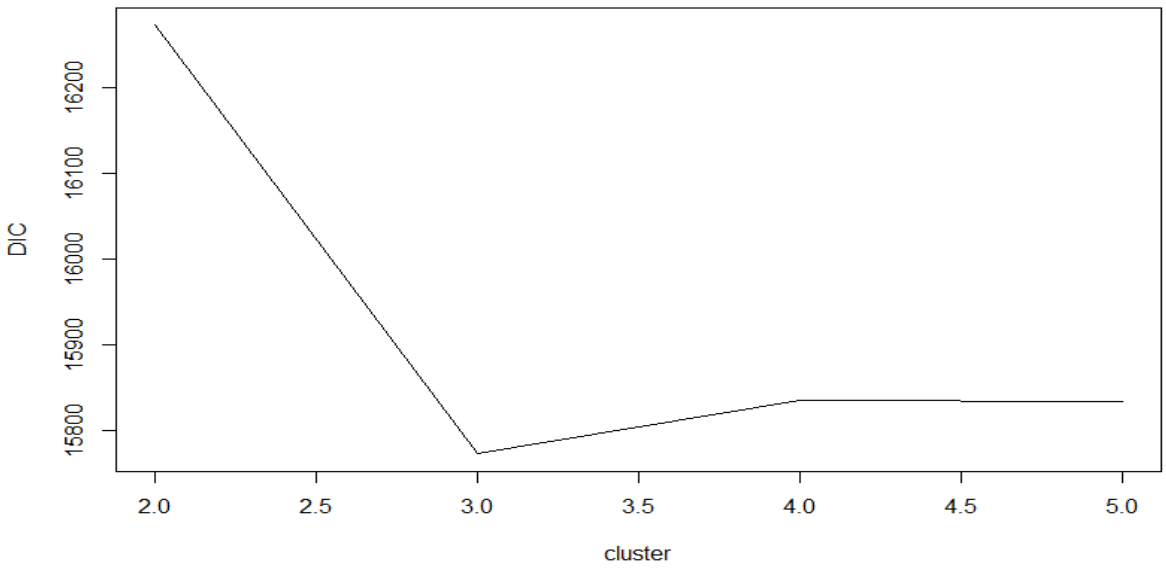
Range of average mean squared difference with a flat prior
0.021967930.06100491

This clearly shows that clusters are more dissimilar when we don't assume that reference alleles have higher frequency.

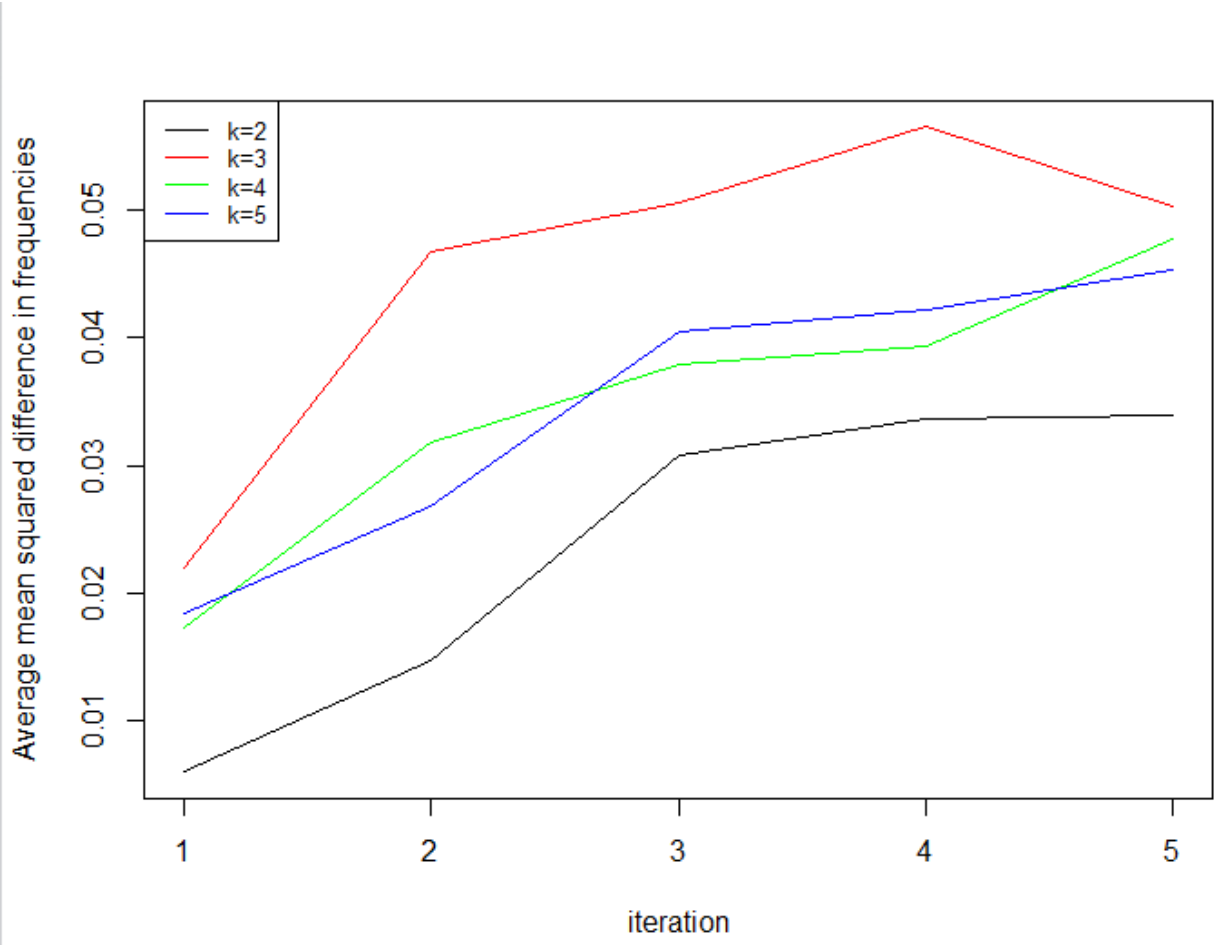
11.

The DIC values for different clusters:

K=2
Penalized deviance: 16273
K=3
Penalized deviance: 15773
K=4
Penalized deviance: 15835
K=5
Penalized deviance: 15834



The plot shows that the DIC is least for K=3 among all values of K. So, the model with K=3 is preferred.



It can also be shown from average mean squared difference in frequencies that with k=3, the clusters were more dissimilar at the start (say for first five iterations). We can see from the above plot the average mean squared difference in frequencies is highest for K=3 among all clusters. So, with K=3, we get a better clustering with more dissimilar clusters.