HighDimensionCluster

Jinchang Fan

3/20/2019

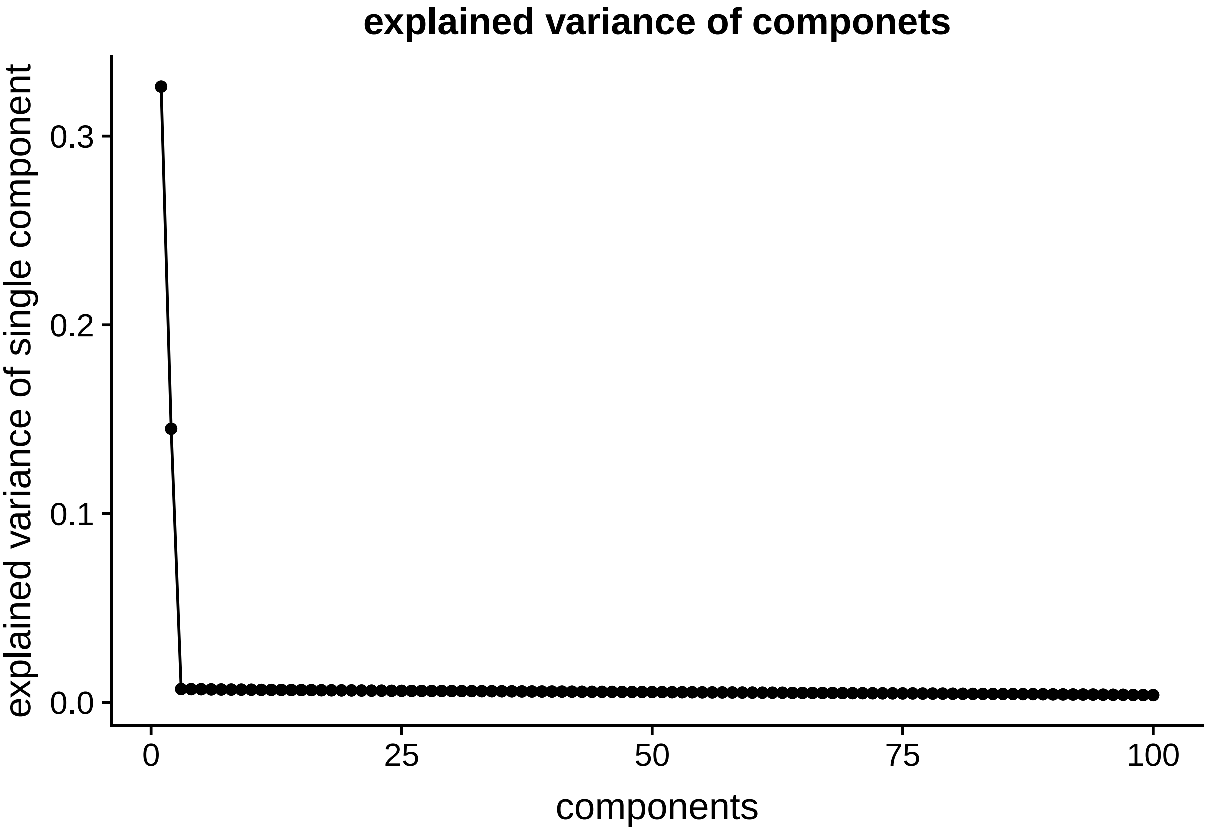
1. Data simulation

Simulate high-dimensional data (p=1000) with three groups of observations where the number of observations is n=100

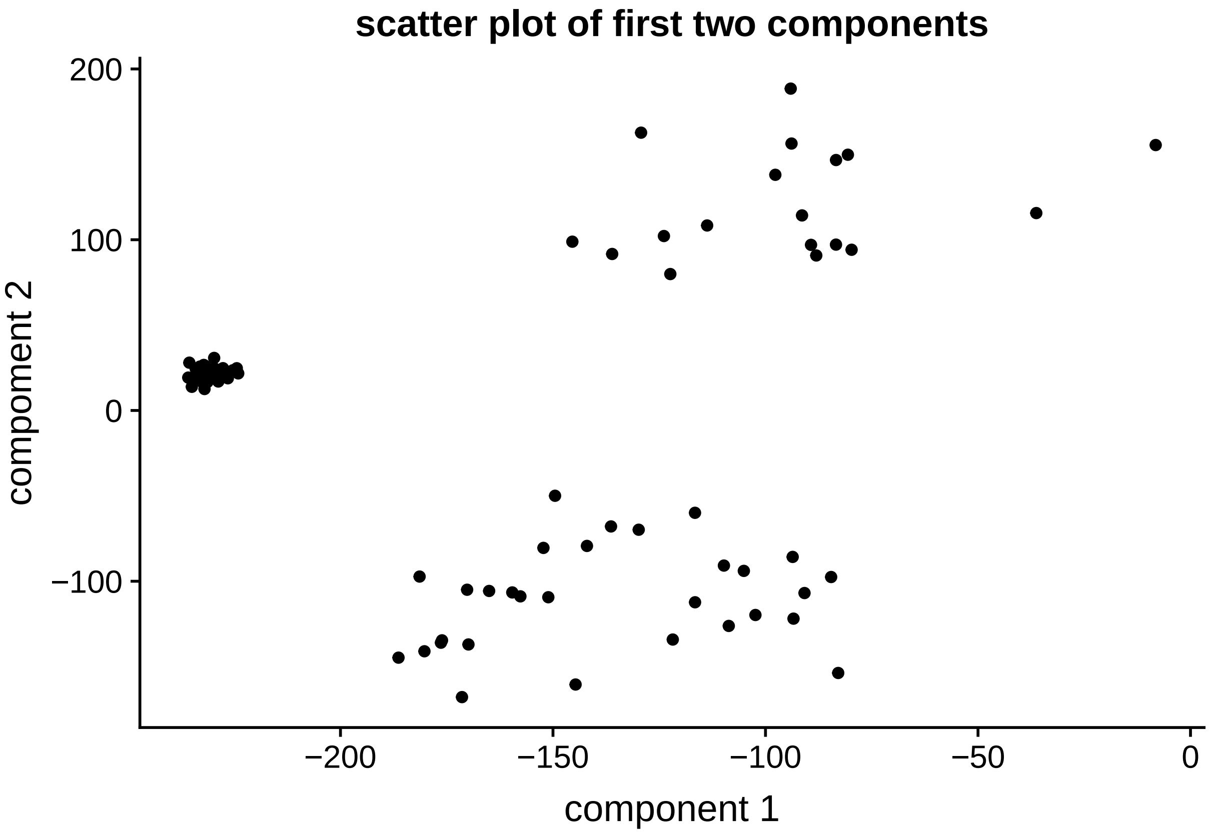
2. Perform k-means to identify the number of clusters in the data.

kmeans after PCA

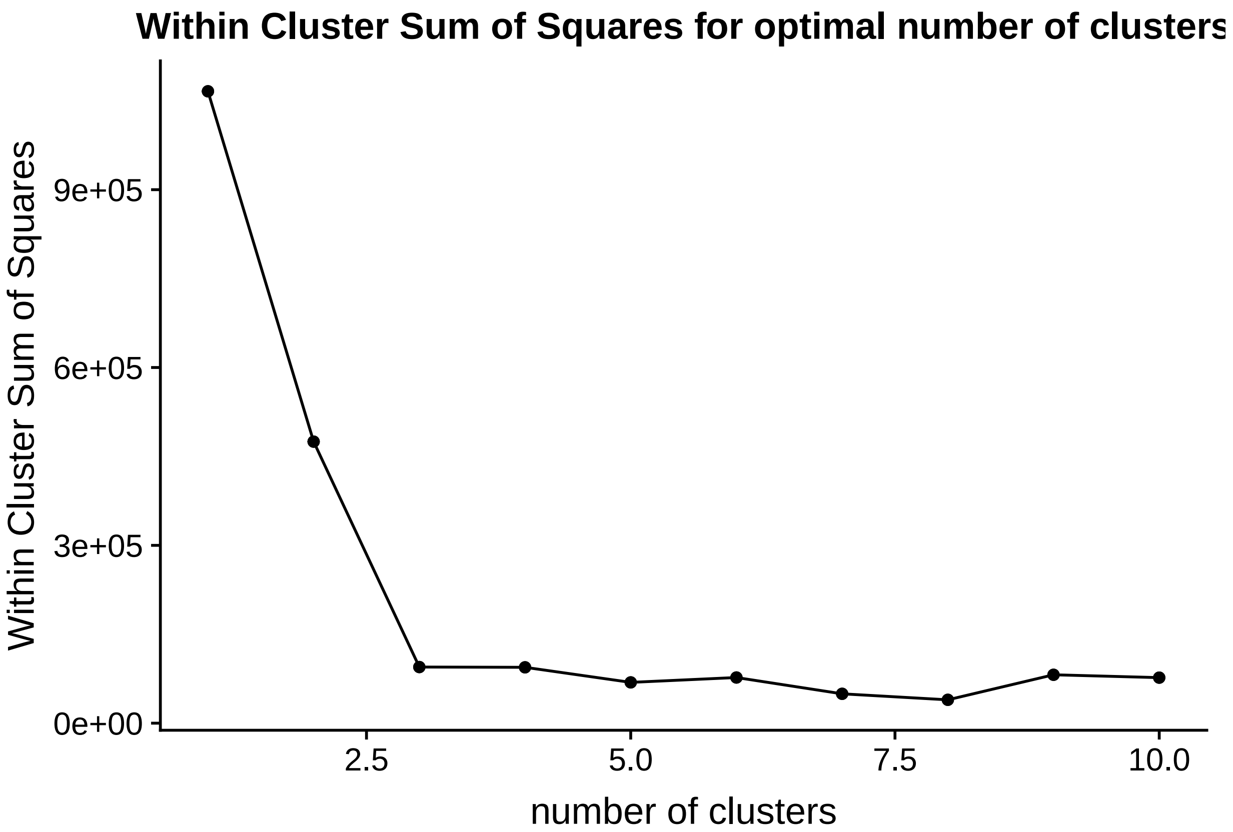
The structure of high dimension data is more complicated, so try dimension reduction at first. Perfrom principal component analysis first. Considering the large size of correlation matrix, it is more efficient to perfrom singular value decomposition as the singular value is the square root of eigne value. The below figure shows the explained variance of every single component, it is clear the the first two components contain much more information than others, and the remaining are all at the same level. In that case we can pick the first two components to perform clustering.



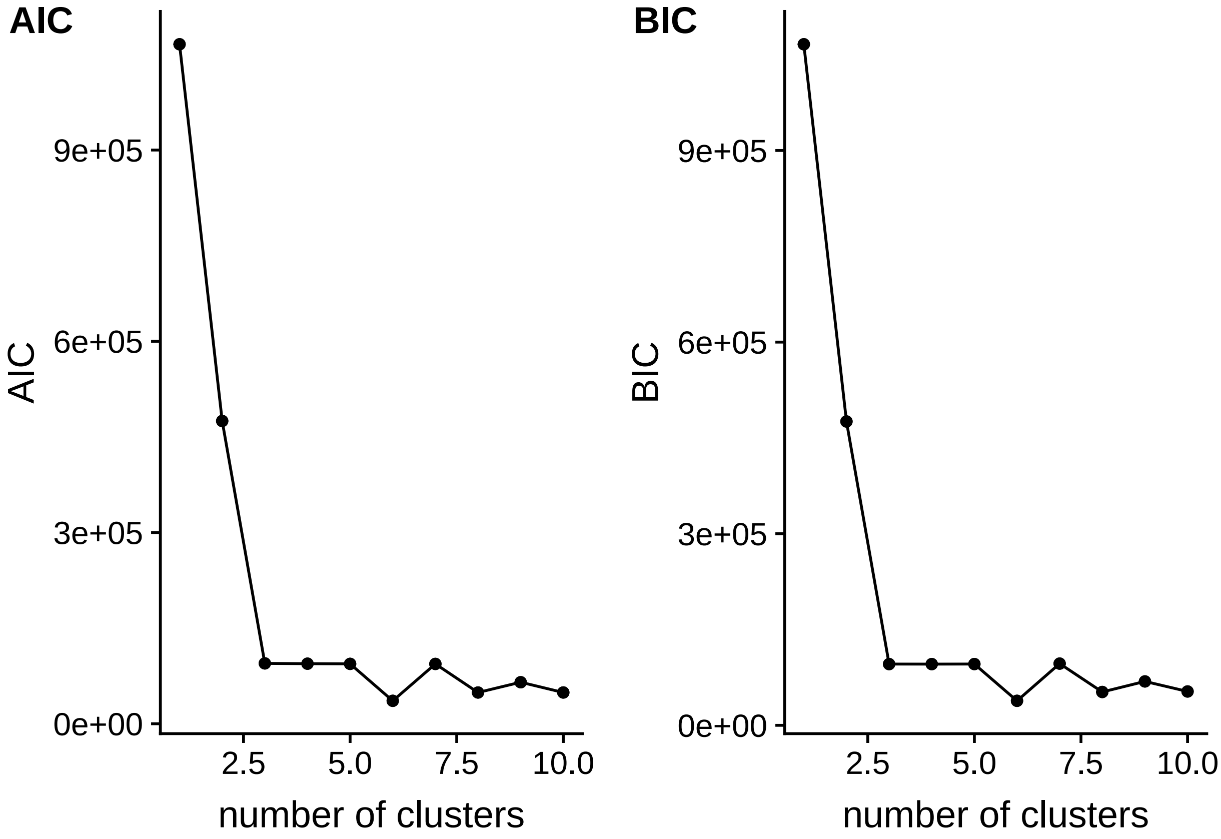
Based on the scatter plot we can intuitively guess there are three groups, but the pattern is not clear.



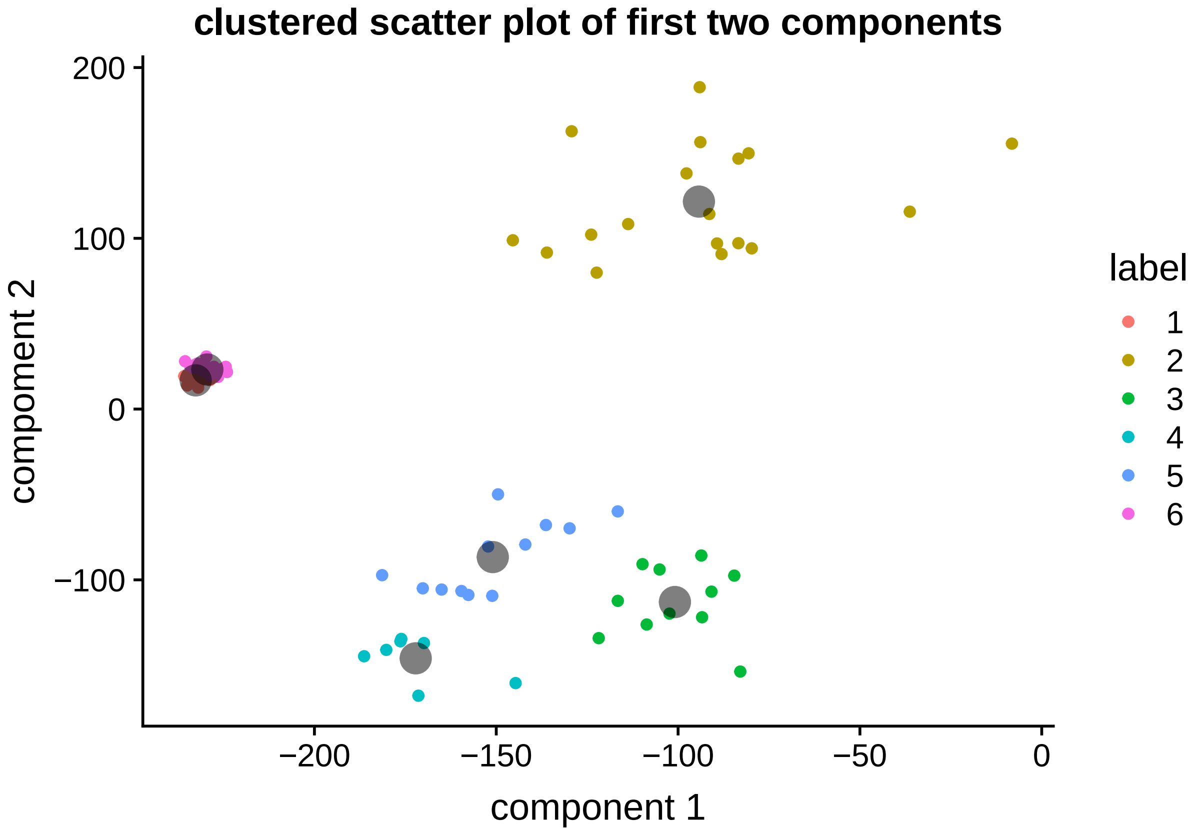
Concerning the Within Cluster Sum of Squares, the information could not explained by clusters decreases rapidly when k increases from 1 to 3, and goes flat with further increasing. This conclusion is more clear than scatter plot.



For a clear optimal choice of k with defined decison role, AIC and BIC are calculated. Figure below shows the AIC and BIC or different k values. there is a clear minimal value for BIC but not for AIC, and the optimal k based on BIC is 3 which is coordinated with our intuitively guess. So $argmin_k(BIC)$ is taken as the strategy for optimal k.

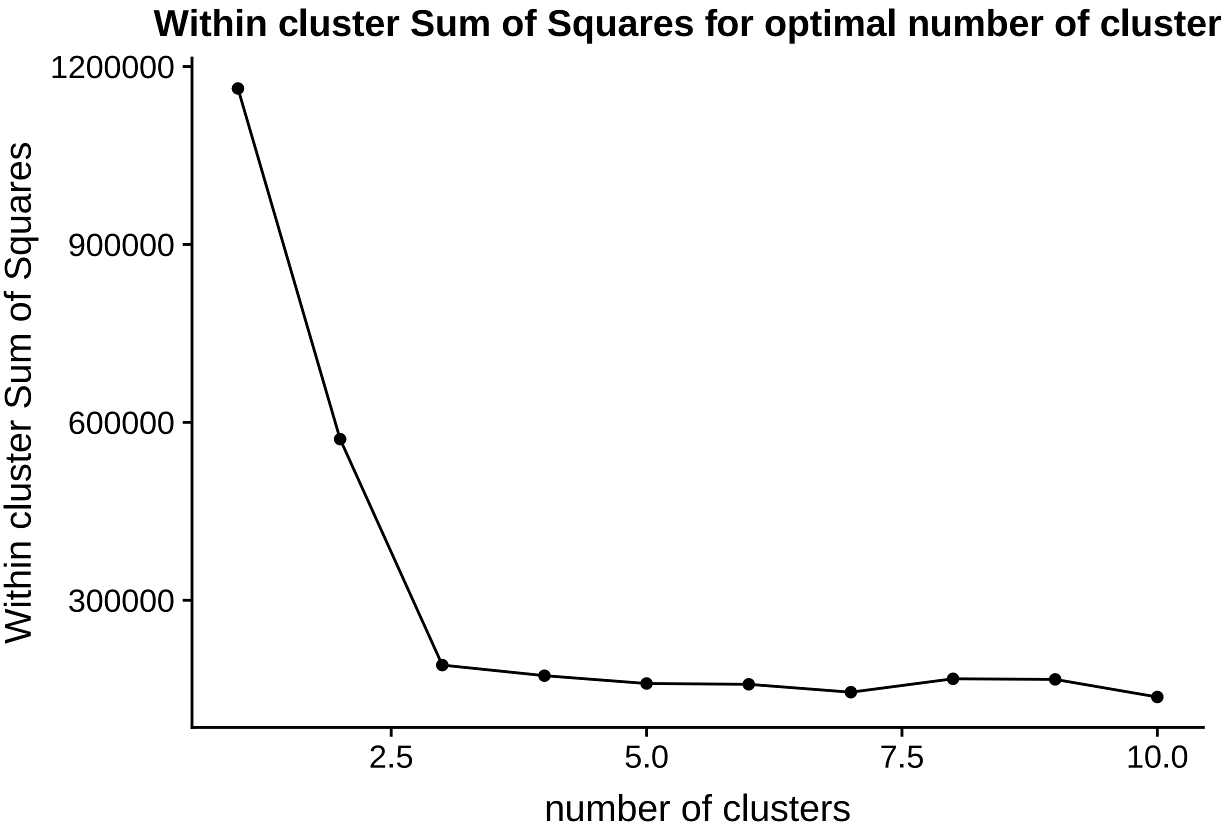


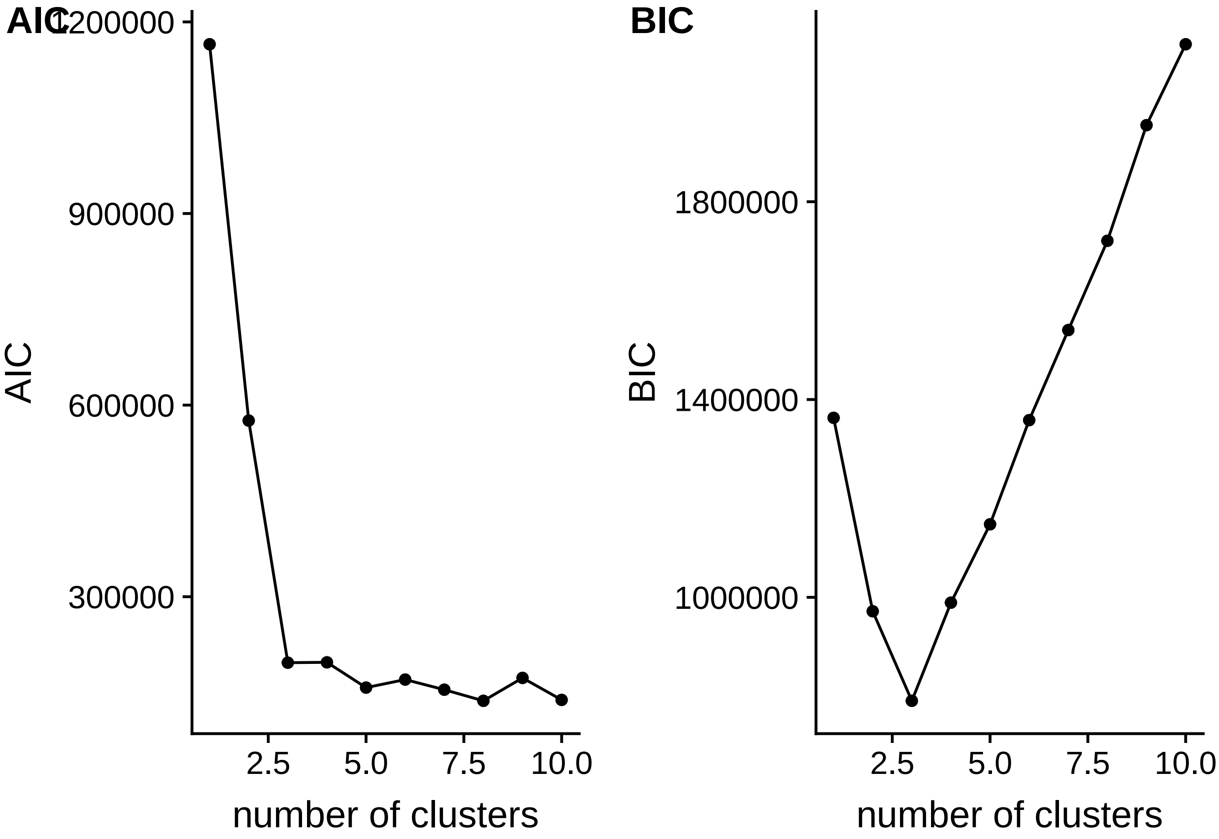
The cluster results:



kmeans for original data

For compariation we can alse perfom similar clustering for original data without pca. This time we can't make clear data visualization so just Within Cluster Sum of Square, AIC, BIC are calculated.





3. To assess the accuracy, calculate the adjusted rand index and then calculate the within clusters sum of squares.

assess the accuracy:

```
## [1] 0.6442095
```

```
##            SSW     SST     Ratio
## [1,] 146256.3 1163010 0.1257568
```

```
##            SSW     SST     Ratio
## [1,] 667437.8 1163010 0.5738884
```

```
## [1] 0.3888138
```

Perform 100 times:

4. Record both metrics.

Then, create a data visualization summarizing the results.