HighDimensionCluster

Jinchang Fan

3/24/2019

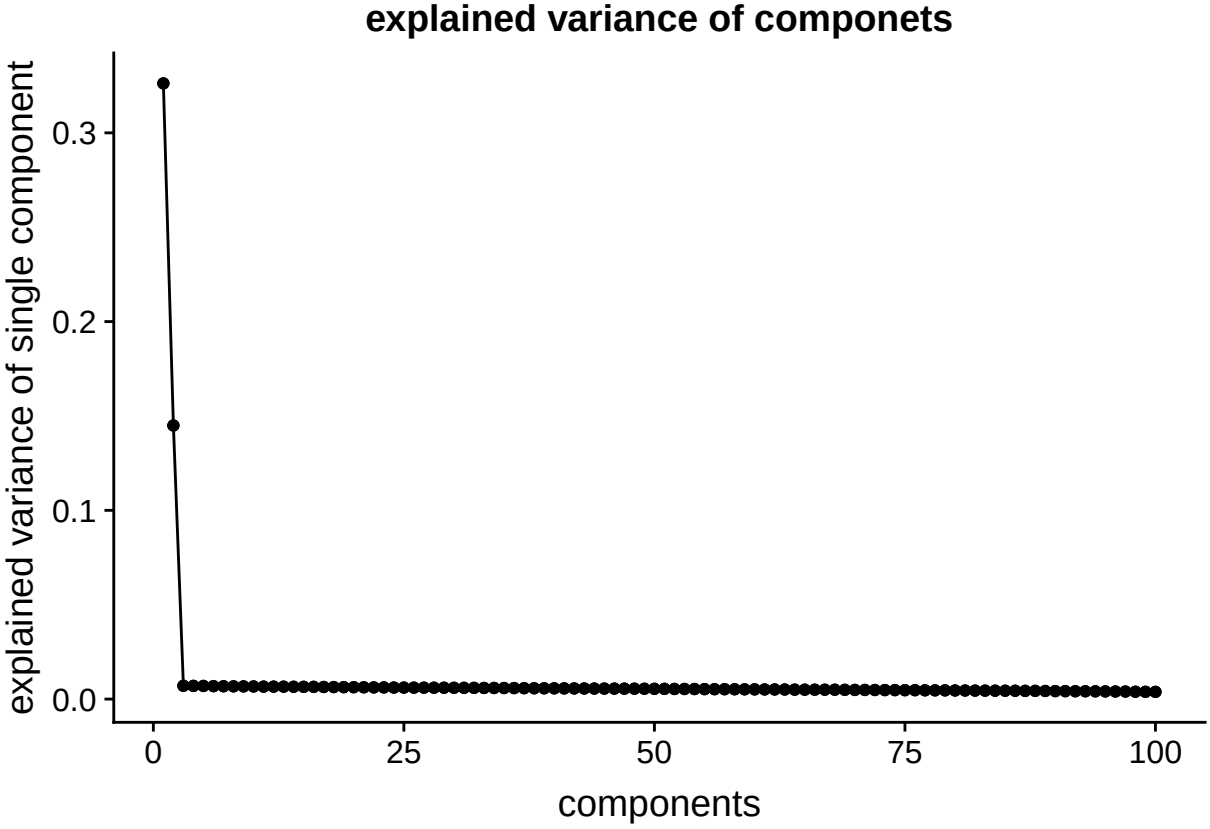
1. Data simulation

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

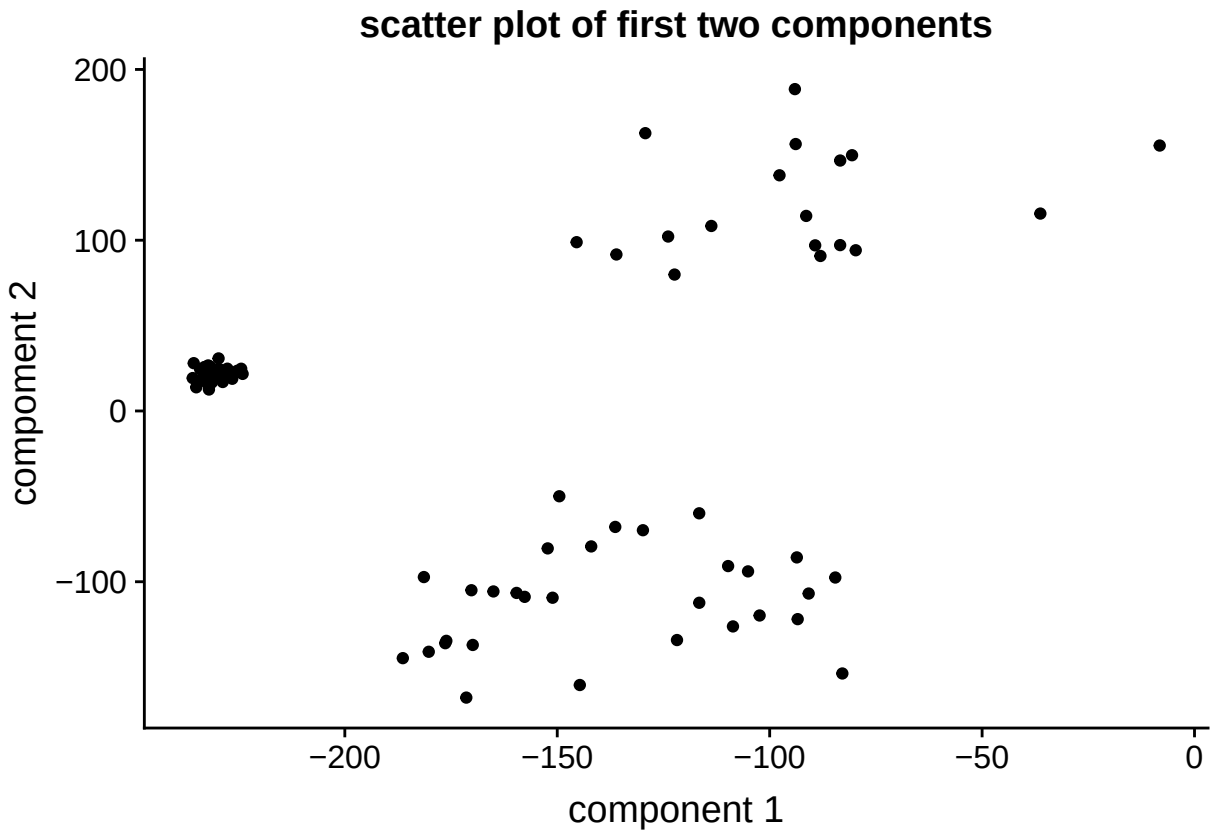
2. Perform k-means to identify the number of clusters in the data.

kmeans after PCA

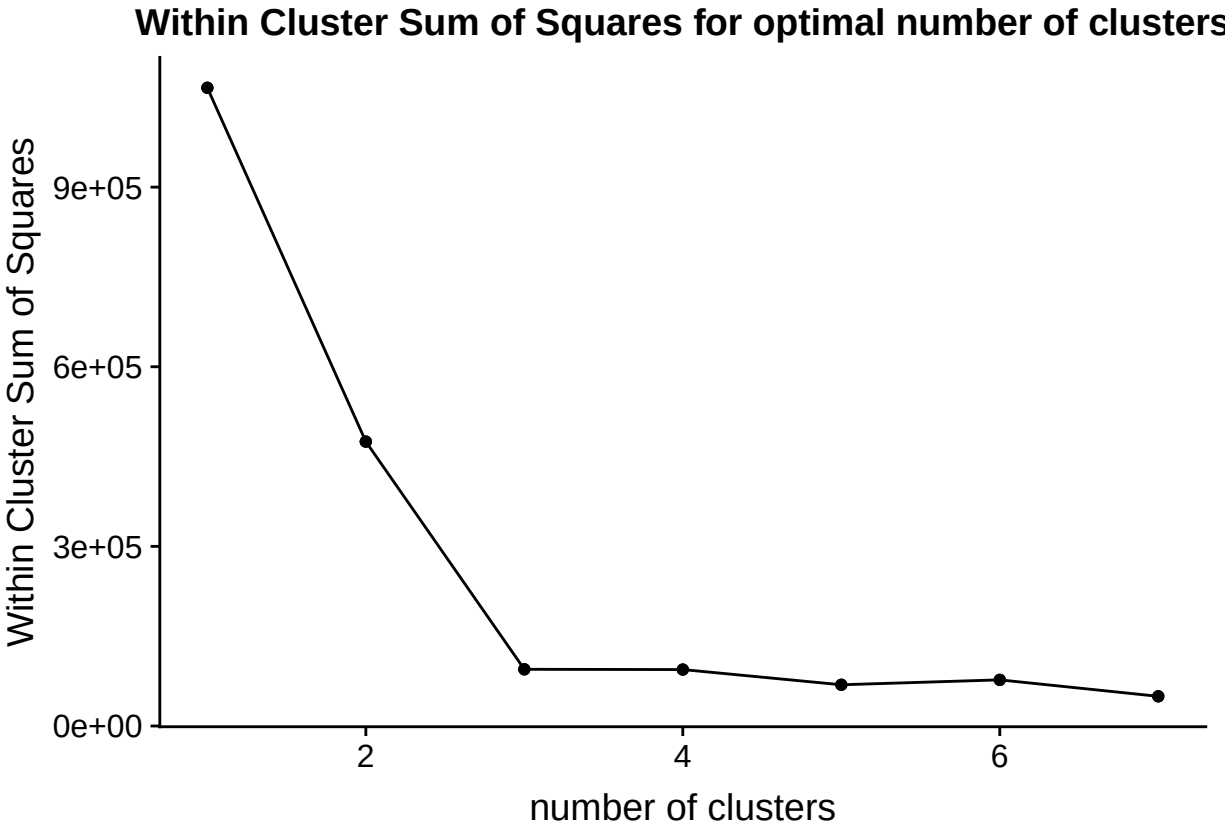
The structure of high dimension data is more complicated, so try dimension reduction at first. Perfrom principal component analysis first. Considering the large size of correlation matrix, it is more efficient to perfrom singular value decomposition as the singular value is the square root of eigne value. 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.



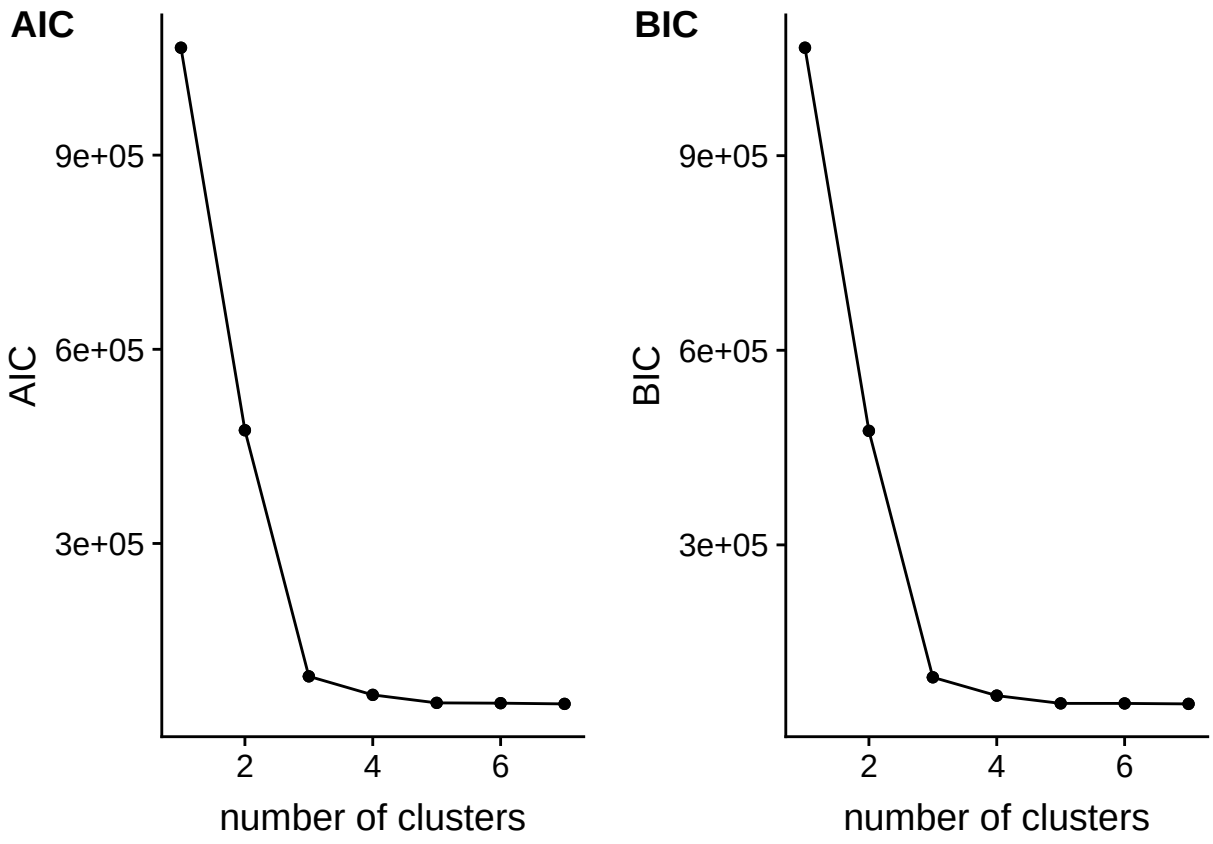
Based on the scatter plot we can intuitively guess there are three groups, but the pattern is not clear.



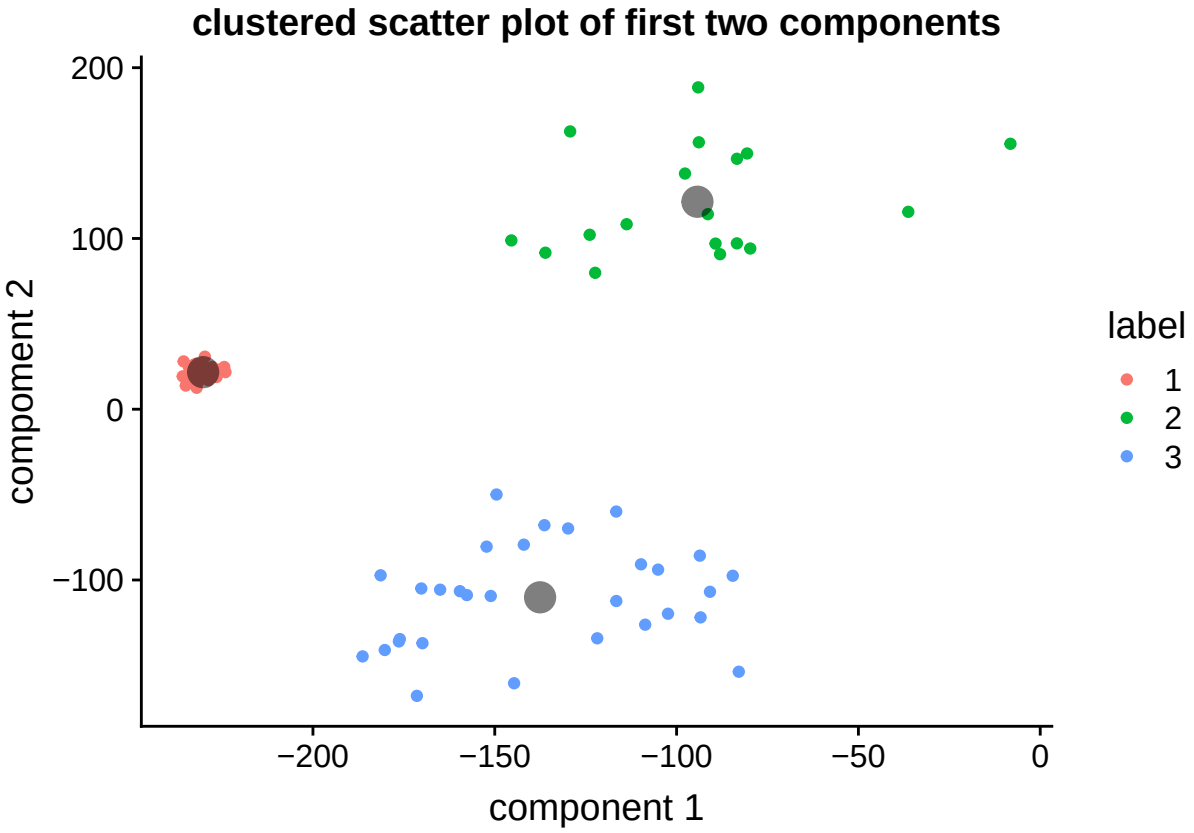
Concerning the Within Cluster Sum of Squares, the information could not explained by clusters decreases rapidly when k increases from 1 to 3, and goes flat with further increasing. This conclusion is more clear than scatter plot.



Try to penish parameter by AIC and BIC. Figure below shows the AIC and BIC or different k values. It is quite similar compated with the result without panalty, as the scale of sum of square error is great larger than panalty on parameters. So minimize AIC or BIC is not a good idea to determine k.



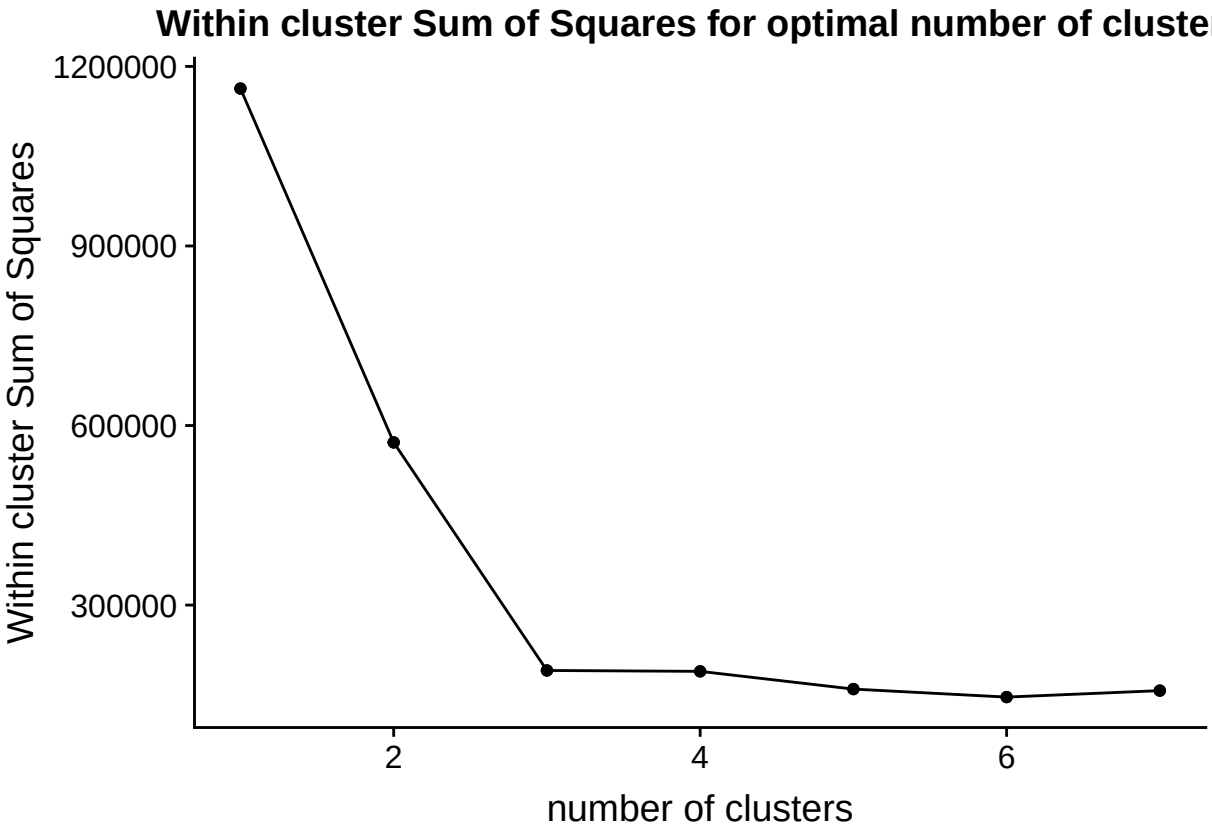
Since AIC nor BIC is a good fit, elbow method is applied. Result is k=3 and the cluster result is as following:



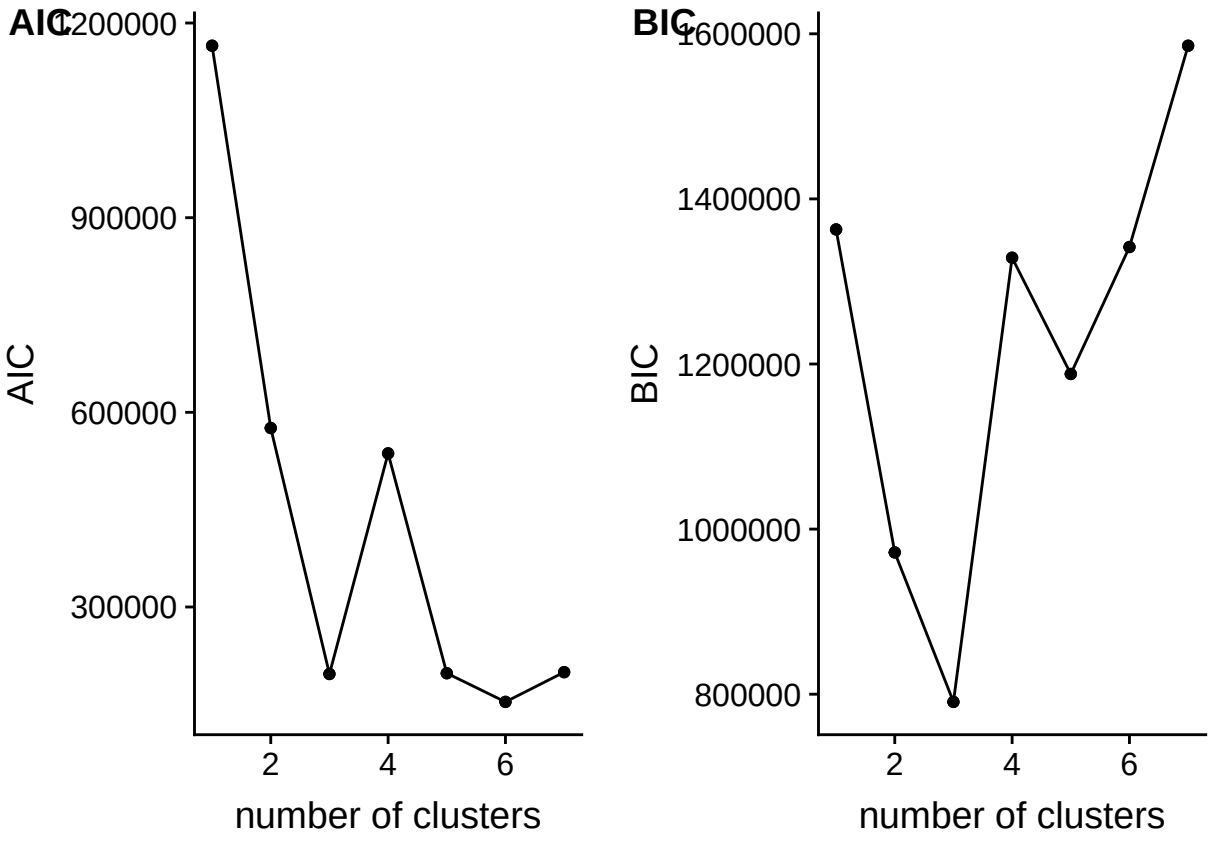
kmeans for original data

For compariation similar clustering for original data without pca is performed as well. It is hard to make clear data visualization on high dimension, so just Within Cluster Sum of Square, AIC, BIC are calculated.

Concerning the Within Cluster Sum of Squares, the result is quite familiar with clustering after pca. Information could not explained by clusters decreases rapidly when k increases from 1 to 3, and goes flat with further increasing.



Penish parameter by AIC and BIC. BIC is found to have a sharp minimal value, as after multiplied by dimension of variables, penalty is at the same scale as sum of squared error.



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

adjusted rand index for kmeans after pca:

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

within clusters sum of squares:

```
##           SSW     SST     Ratio
## [1,] 190713.1 1163010 0.1639824
```

adjusted rand index for kmeans after pca:

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

within clusters sum of squares:

```
##           SSW     SST     Ratio
## [1,] 190713.1 1163010 0.1639824
```

Perform 100 times then compare and summary.:

Compare the adjusted rand index for two methods:

adjusted rand index for kmeans after pca:

```
## mean:  0.7827099 standard diviation:  0.2322727
```

adjusted rand index for simple kmeans:

```
## mean:  0.8621053 standard diviation:  0.2120299
```

within clusters sum of squares for kmeans after pca:

```
## mean:  0.3271211 standard diviation:  0.1779515
```

within clusters sum of squares for simple kmeans:

```
## mean:  0.2564183 standard diviation:  0.1486152
```

time efficiency of two method:

```
## kmeans after pca:  0.02881178 simple kmeans:  0.04304173
```

It shows that simple kmeans with determinating k by BIC shows better performance on adjusted rand index than kmeans after pca and determinante k by elbow (0.86 over 0.78), with similar standard diviation. While kmeans after pca is better on within clusterss sum of squares (0.33 over 0.26). And keams after pca has better time efficiency.

Further discussion

- Different methods to determinate k are applied to simple kmeans and kmeans after pca, As personally I think BIC would be a better method than elbow method. It is reasonable to use elbow for both if necessary of pca is considered.
- Based on the scatter plot of first two pinciple components, the clusters have different variance, which violates the assumption of kmeans that clusters have constant variance. Method accepting different variance such as Gaussian mixture models could be considered for imporvement.