626 midterm 2

Yuxiang Gao

2023-04-27

```
#setwd("/Users/yuxianggao/Desktop/626 Midterm2")
df=read.table('hgdp.txt')
```

```
d<-df[,-c(1:3)]
my<-setupSNP(d,1:ncol(d),sep = "")
```

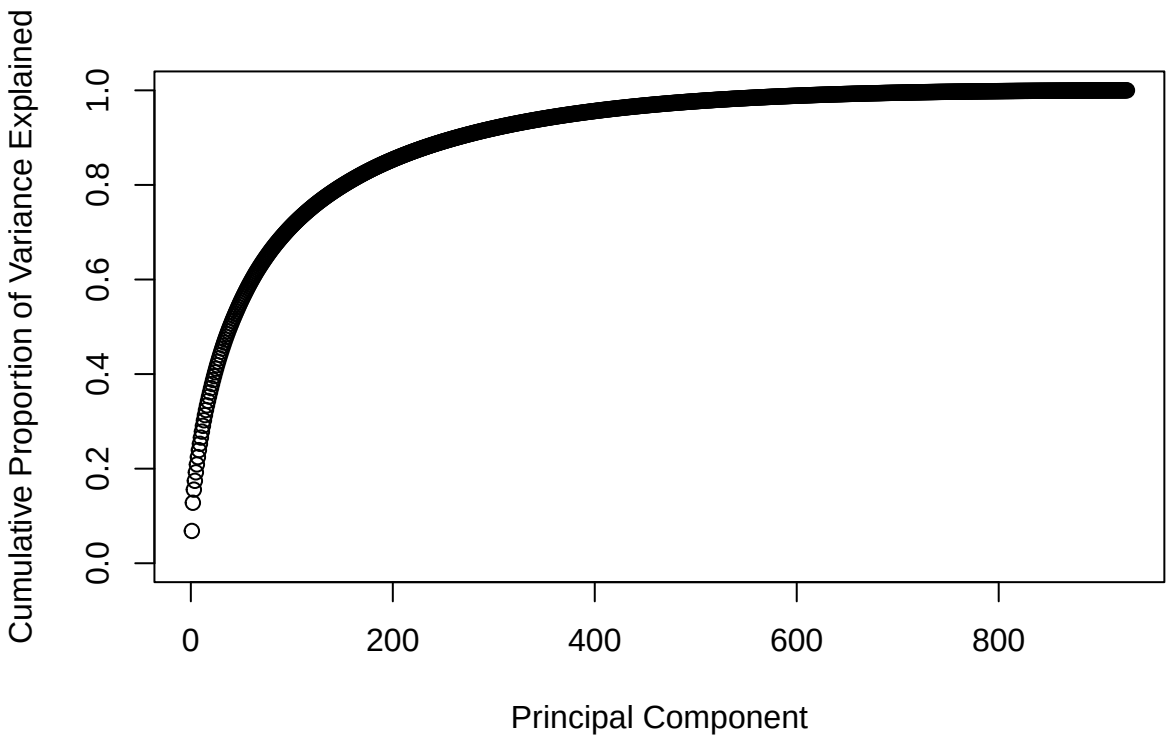
```
myda<- apply(my,2,additive)
```

1.

```
pca=prcomp(myda)
#plot(cumsum(pca$sdev^2)/sum(pca$sdev^2),type='l')
```

```
pr.var <- pca$sdev^2
pve <- pr.var / sum (pr.var)
```

```
plot ( cumsum (pve), xlab = " Principal Component ",
ylab = " Cumulative Proportion of Variance Explained ",
ylim = c(0, 1), type = "b")
```




```
m=which(cumsum(pca$sdev^2)/sum(pca$sdev^2)>.8)[1]
x.pca=pca$x[,1:m]
```

2.

```
set.seed(1)
tot.withinss=c()
for(k in 2:10){
  km=kmeans(x.pca,centers=k)
  tot.withinss=c(tot.withinss,km$tot.withinss)
}
tot.withinss
```

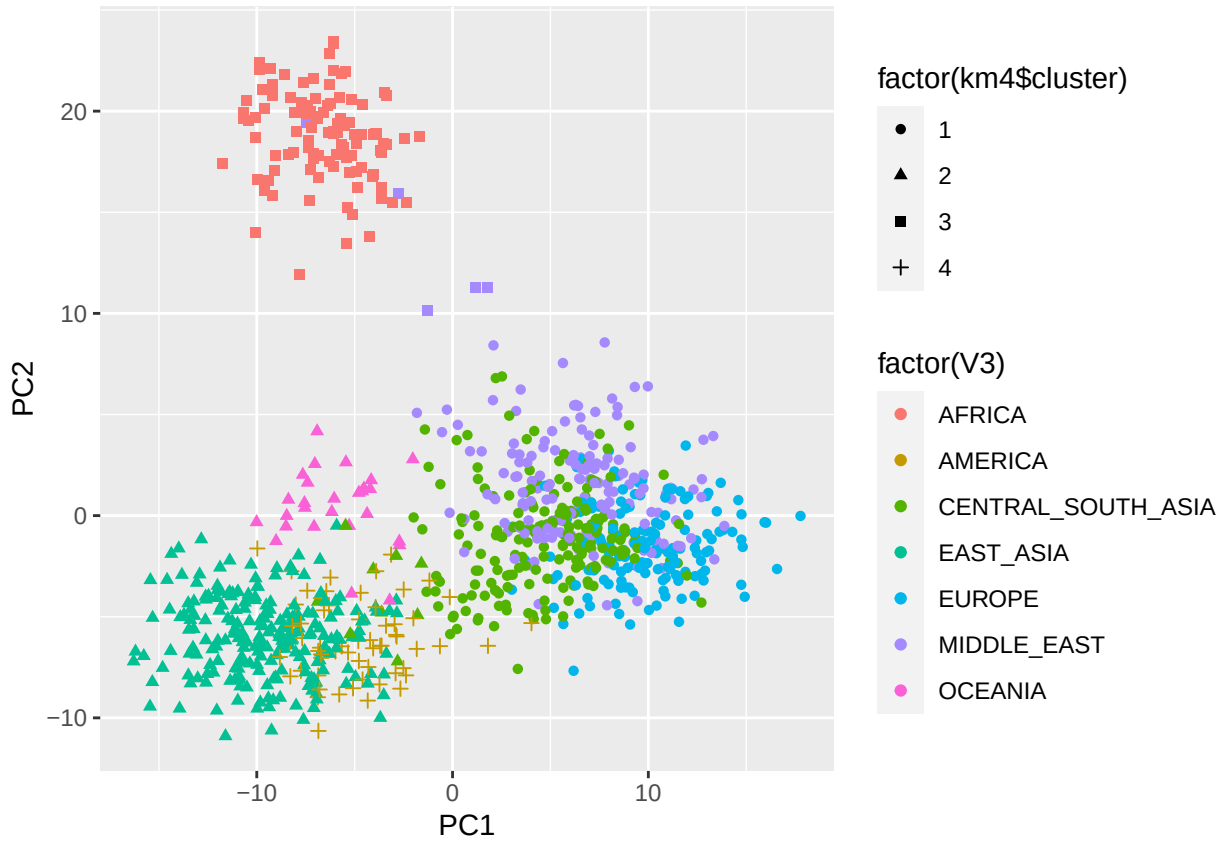
```
## [1] 666956.5 616436.9 598162.2 587848.7 583811.5 574407.4 568417.3 564246.6
## [9] 571890.5
```

```
km4=kmeans(x.pca,centers=4)
```

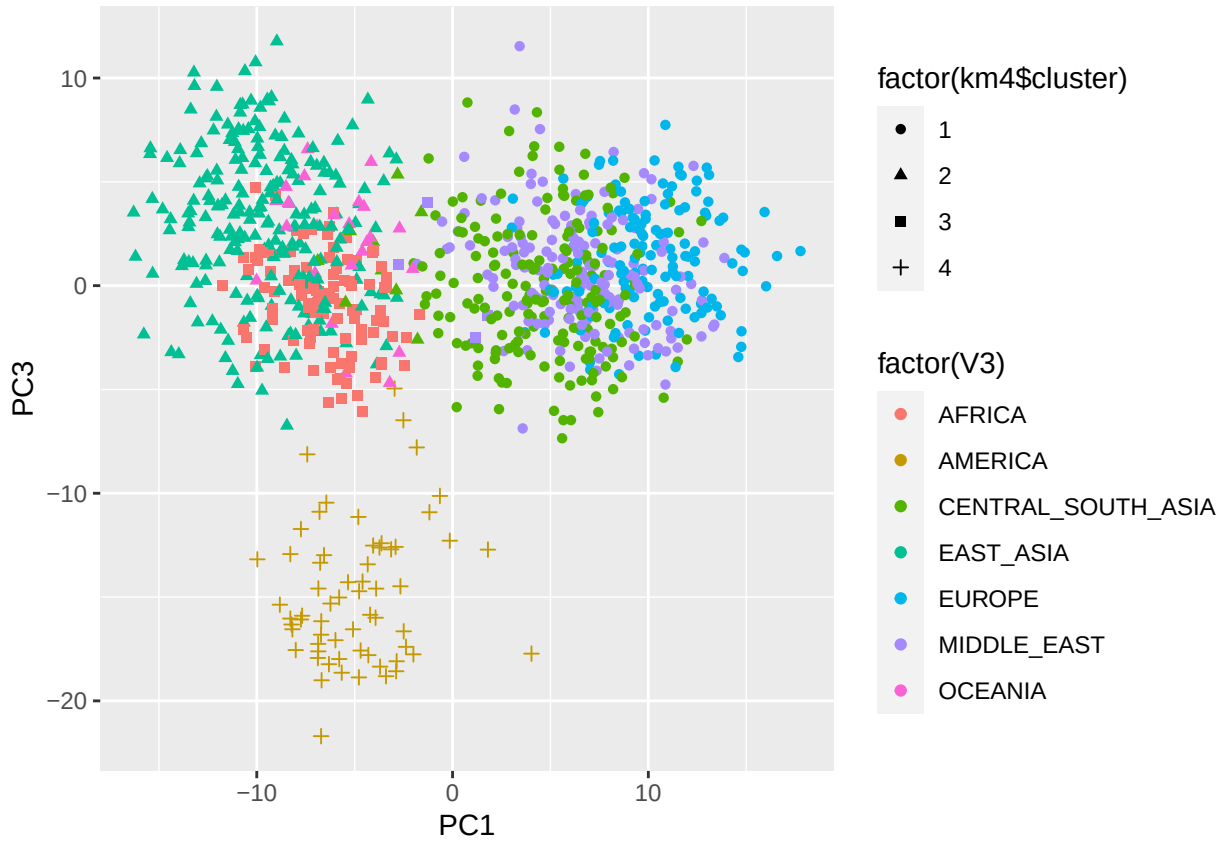
```
#plot
library(ggplot2)
data=cbind(df[,1:3],x.pca)

p=ggplot(data, aes(PC1,PC2))

p + geom_point(aes(colour=factor(V3),shape=factor(km4$cluster)))
```




```
p=ggplot(data, aes(PC1,PC3))
p + geom_point(aes(colour=factor(V3),shape=factor(km4$cluster)))
```



3. From the graph above, we can see that a centers=4 works well in clustering. Each of the clustering is decentralized. Different clustering without much similarities are far from each other. We can also see that within one clustering, distribution of data points are compact. In general, we would say a k=4 kmeans is a good clustering choice in splitting up data.

4.

```
ts_result <- tsne(myda, k=4)
```

```
## sigma summary: Min. : 0.410924985855584 |1st Qu. : 0.500894573842357 |Median : 0.538033636166382 |Mea

## Epoch: Iteration #100 error is: 19.2961963552742

## Epoch: Iteration #200 error is: 1.48210669536236

## Epoch: Iteration #300 error is: 1.42445915654439

## Epoch: Iteration #400 error is: 1.40471961844012

## Epoch: Iteration #500 error is: 1.3978247237512

## Epoch: Iteration #600 error is: 1.39356973865347

## Epoch: Iteration #700 error is: 1.39131984265338
```

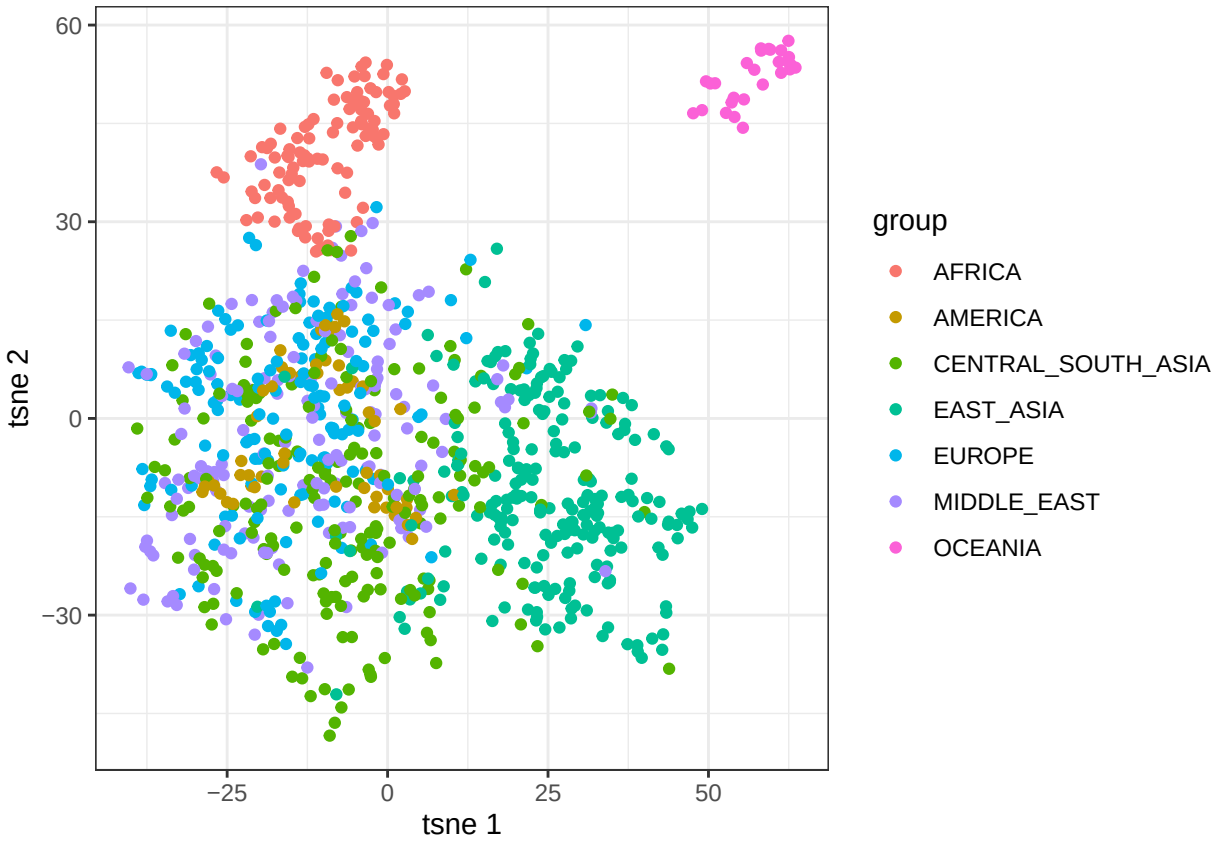
```
## Epoch: Iteration #800 error is: 1.38959679477327

## Epoch: Iteration #900 error is: 1.38765067778548

## Epoch: Iteration #1000 error is: 1.38594049522618
```

```
my_res <- as.data.frame(ts_result)
my_res$group <- data$V3
```

```
#graph from 2 dimension
ggplot(my_res)+geom_point(aes(V1,V2,color=group))+
  theme_bw()+
  labs(x="tsne 1",y="tsne 2")
```



Comparing T-SNE result to PCA result in 2 dimension, we can see that the clustering structure is similar, but the T-SNE doesn't include factor labels in the graph, which makes it easier for us to read and interpret. For me, I think T-SNE is better than PCA in visualization.