STA567 homework 8

Lina Lee

11/26/2019

Problem 1) What do the factor loadings, scree plot and biplot tell you about the eight dimensions of numeric information from the olive oils?

Load data for accessing

```
data("oliveoil")


sub_olive<-oliveoil[,3:10]
olive_scaled <- scale(sub_olive)
olive_scaled<-as.data.frame(olive_scaled)
```

(1) Loadings (eigenvectors)

```
olive_pca <- princomp(~ . , data=sub_olive, cor=TRUE)
olive_pca$loadings
```

```
## Loadings:
##             Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## palmitic     0.461         0.114  0.280  0.535         0.525  0.354
## palmitoleic  0.450  0.241  0.143  0.212  0.138  0.167 -0.787
## stearic             -0.258  0.802 -0.471  0.213
## oleic       -0.494 -0.159         0.200         0.113 -0.181  0.799
## linoleic     0.366  0.343        -0.512 -0.401 -0.305         0.467
## linolenic    0.219 -0.605 -0.191         0.125 -0.698 -0.191
## arachidic    0.228 -0.447 -0.427 -0.482  0.147  0.554
## eicosenoic   0.312 -0.405  0.301  0.332 -0.672  0.257  0.140
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.125  0.125  0.125  0.125  0.125  0.125  0.125  0.125
## Cumulative Var  0.125  0.250  0.375  0.500  0.625  0.750  0.875  1.000
```

-The loading vectors for the first principal component defines a direction in feature space along which the data vary the most. The direction of the vectors in biplot is determined by the sign of the associated eigenvector(loadings) sign, and the magnitude is based on the eigenvector value.(from ISL text book)

-From the loadings table, We see the loadings of the first component is comprised of a linear combination of all the acids except stearic acid. palmitic,palmitoleic, and olic weighted about 0.45, linoleic,linolenic, arachidic, and eicosenoic weighted less. oleic has negative sign, which we can see the vector of oleic headed negative in the biplot below.

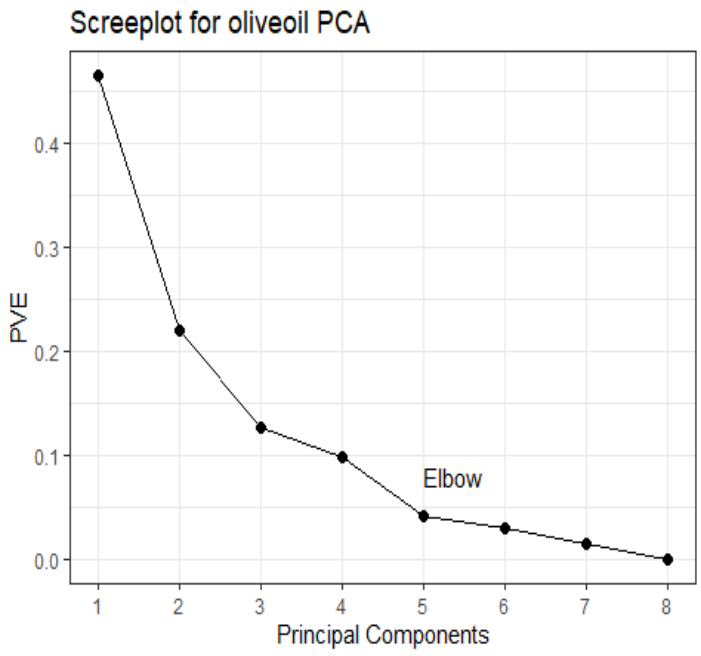
-The loadings of the second component is a linear combination of all the acids that has maximal variance out of all linear combinations that are uncorrelated with the first principal component. The second principal component is comprise of all the acids except palmitic. Palmitoleic and linoleic have positive signs, and stearic,oleic,linolenic,arachidic,and eicosenoic acids have negative signs. linolenic acid weighted most as about -0.6. linoleic, arachidic, and eicosenoic acids weighted more than palmitoleic,stearic,oleic. Oleic acid weighted least.

get data into data frames

```
component_loadings <- as.data.frame(olive_pca$loadings[1:8,1:8])
component_loadings$var <- row.names(component_loadings)
olive_scores <- as.data.frame(olive_pca$scores)
olive_scores$oliveoil <- row.names(olive_scores)
```

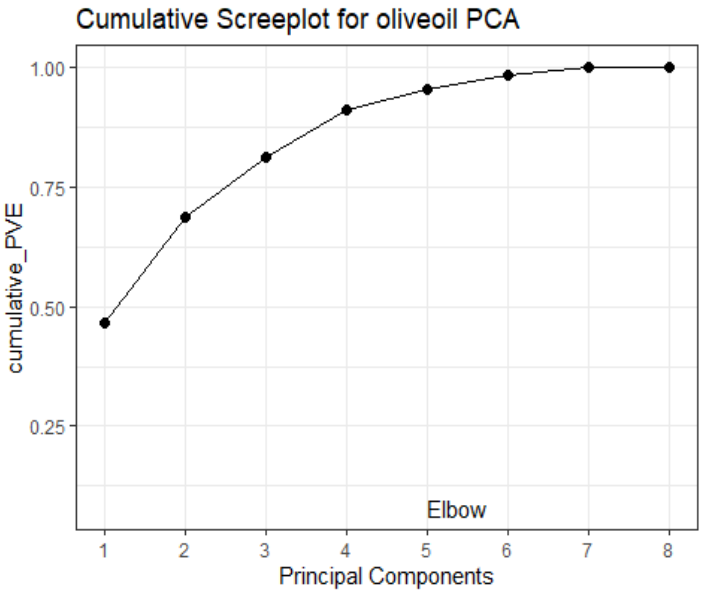
(2) Screeplot

```
# pull off eigenvalues to plot PVE
olive_pca$sdev

##     Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6
## 1.92909565 1.32883314 1.00814455 0.89044867 0.57776956 0.49881727
##     Comp.7     Comp.8
## 0.34470293 0.04562633

scree_data <- data.frame(comp=1:8,
                         PVE=olive_pca$sdev^2/sum(olive_pca$sdev^2),
                         cumulative_PVE = cumsum(olive_pca$sdev^2/sum(olive_pca$sdev^2)))
# build plot
ggplot()+
  geom_point(aes(x=comp,y=PVE),size=2, data=scree_data)+
  geom_line(aes(x=comp,y=PVE), data=scree_data) +
  annotate(geom="text",x=5,y=.08,label="Elbow",hjust=0)+
  labs(x="Principal Components",
       title="Screeplot for oliveoil PCA")+
  scale_x_continuous(breaks=1:8)+
  theme_bw() + theme(panel.grid.minor.x = element_blank())
```




```
ggplot()+
  geom_point(aes(x=comp,y=cumulative_PVE),size=2, data=scree_data)+
  geom_line(aes(x=comp,y=cumulative_PVE), data=scree_data) +
  annotate(geom="text",x=5,y=.08,label="Elbow",hjust=0)+
  labs(x="Principal Components",
       title="Cumulative Screeplot for oliveoil PCA")+
```

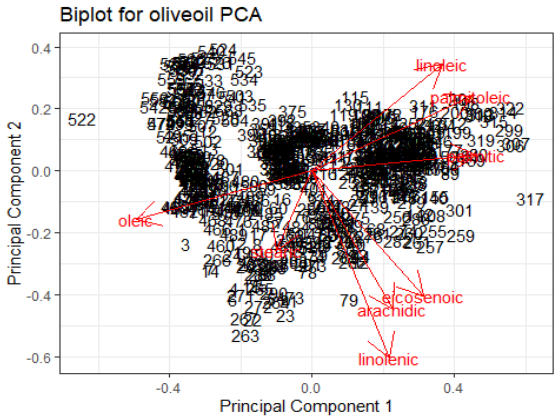
```
  scale_x_continuous(breaks=1:8)+
  theme_bw() + theme(panel.grid.minor.x = element_blank())
```



-From the scree plot, we see that the first component explains about 50 percent of variability. The first, second, third principal components explains about 80 percent of variability. The principal components 4-8 does not appear to explain variability much.

(3) Biplots

```
ggplot()+
  geom_text(aes(x=Comp.1/8,y=Comp.2/8, label=oliveoil), color="black",
            data=olive_scores) +
  geom_segment(x=0,y=0,aes(xend=Comp.1,yend=Comp.2),color="red",
               data=component_loadings, arrow=arrow(angle=30))+
  geom_text(aes(x=Comp.1,y=Comp.2, label=var), color="red",
            data=component_loadings)+
  labs(x="Principal Component 1",
       y="Principal Component 2",
       title="Biplot for oliveoil PCA")+
  theme_bw()
```

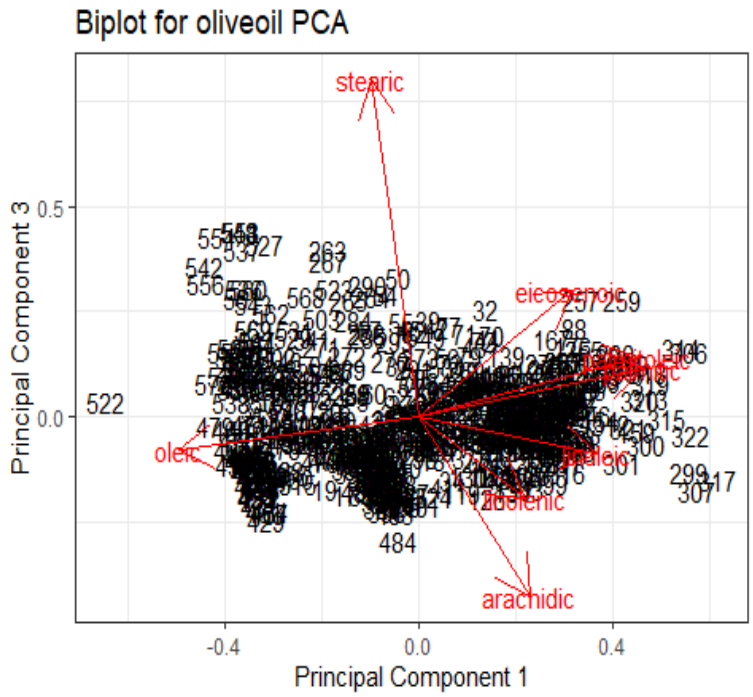


-From the biplot, we can observe that for the Principal component1, linoleic,palmitoleic, palmitic, linolenic,arachidic, and eicosenoic acids headed positive. On the otherhand, oleic headed negative.

-For the Principal component2, direction of linoleic and palmitoleic acids are positive, and Stearic, Oleic, linolenic, arachidic, and eicosenoic acids headed negative.

-Considering direction and magnitude of vectors of variables, linoleic, palmitoleic, and palmitic acids are correlated each other, and linolenic, arachidic, and eicosenoic acids are correlated with each other. Stearic and Oleic seems not correlated with other elements. Oleic acid looks negatively correlated with other variables.

```
ggplot()+
  geom_text(aes(x=Comp.1/8,y=Comp.3/8, label=oliveoil), color="black",
            data=olive_scores) +
  geom_segment(x=0,y=0,aes(xend=Comp.1,yend=Comp.3),color="red",
               data=component_loadings, arrow=arrow(angle=30))+
  geom_text(aes(x=Comp.1,y=Comp.3, label=var), color="red",
            data=component_loadings)+
  labs(x="Principal Component 1",
       y="Principal Component 3",
       title="Biplot for oliveoil PCA")+
  theme_bw()
```



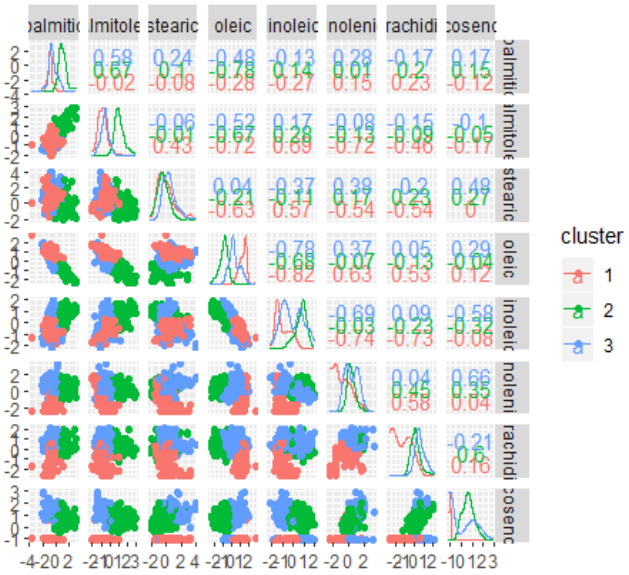
For the third principal component, Stearic headed positive and arachidic headed negative. Two elements are extremely apart from the cluster of elements in the middle. eicosenoic, palmitoleic, palmitic are located close each other. and linoleic, linolenic are located close each other.

Problem 2

1) 3 macro-areas.

(1) k-means clustering

```
km.out <- kmeans(olive_scaled, 3)
plotdat1 <- data.frame(cluster=as.factor(km.out$cluster),
                       olive_scaled)



ggscatmat(plotdat1, columns=2:9, color="cluster")
```



from the above matrix, we can observe some evidences of three clusters

*Evaluation of k-means clustering with the 3 macro-areas.

```
table(plotdat1$cluster,oliveoil$macro.area)
```

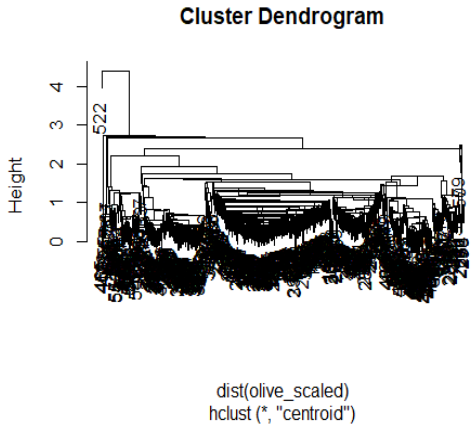
```
##
##     South Sardinia Centre.North
##   1     2        0          121
##   2   219        0            0
##   3   102       98           30
```

Clusters based on the k-means clustering, the cluster 1 lines up with Centre.North and cluster 2 lines up with South. Cluster 3 seems to have trouble identifying clusters that would correspond with three areas. Overally, k-means clustering does not identify three macro areas within the data well.

(2) hierarchical clustering

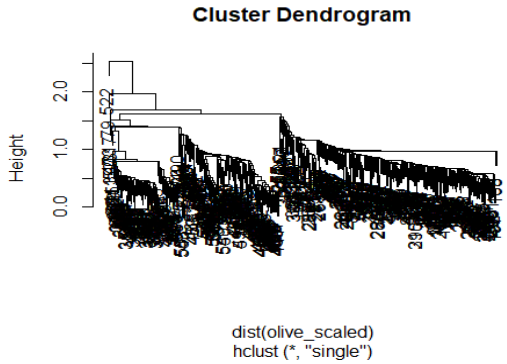
centroid method

```
hc.out <- hclust(dist(olive_scaled), method="centroid")
plot(hc.out)
```



Single method

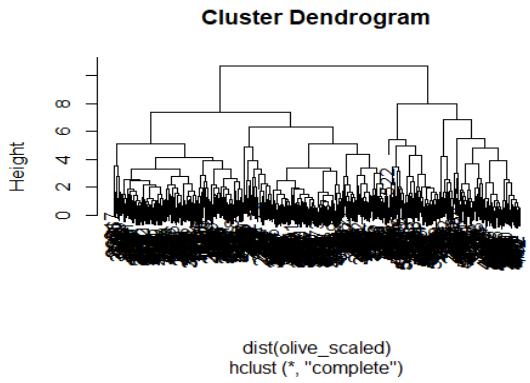
```
hc.out <- hclust(dist(olive_scaled), method="single")
plot(hc.out)
```



From single and centroid hierarchical clusterings, it is difficutl to find three obvious clusters.

Complete method

```
hc.out_fi <- hclust(dist(olive_scaled), method="complete")
plot(hc.out_fi)
```



*Evaluation of hierarchical clustering with the 3 macro-areas.

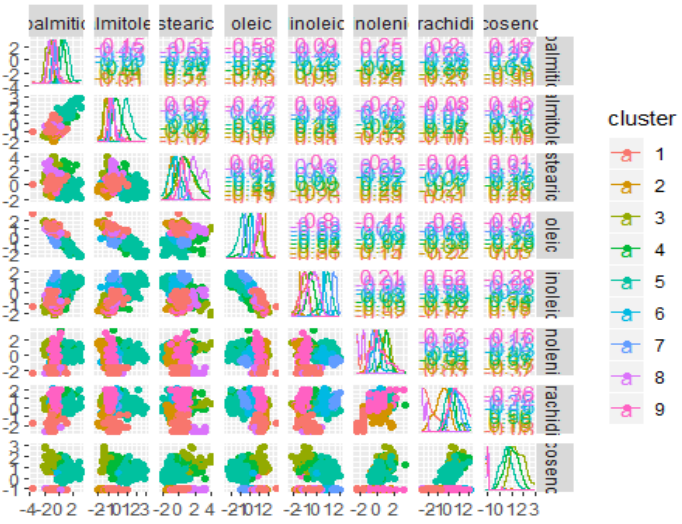
```
plotdat2 <- data.frame(cluster=as.factor(cutree(hc.out_fi,3)),
                       olive_scaled)
table(plotdat2$cluster,oliveoil$macro.area)

##
##     South Sardinia Centre.North
##   1    34        0           80
##   2   289       98            0
##   3     0        0           71
```

-Clusters based on the hierarchical clustering with complete method, cluster 3 perfectly lines up with Centre.North. Cluster 1 mostly lines up with Centre. North, but it seems to have trouble indentifying clusters that correspond with South or Centre. North some extent. cluster 2 mostly lines up with South but it also seems to have trouble indentifying clusters that correspond with South and Sardinia some extent. The hierarchical clustering also does not align with three macro areas perfectly although it identifies clusters that would correspond to three areas some extent. The hierarchical clustering seems to work better than k-means clustering in this case.

2) 9 regions

(1) k-means clustering

```
km.out3 <- kmeans(olive_scaled, 9)
plotdat3 <- data.frame(cluster=as.factor(km.out3$cluster),
                       olive_scaled)



ggscatmat(plotdat3, columns=2:9, color="cluster")
```



since there were 9 clusters, it is difficult to see clear pattern about 9 clusters easily from the above chart.

*Evaluation of k-means clustering with the 9 regions.

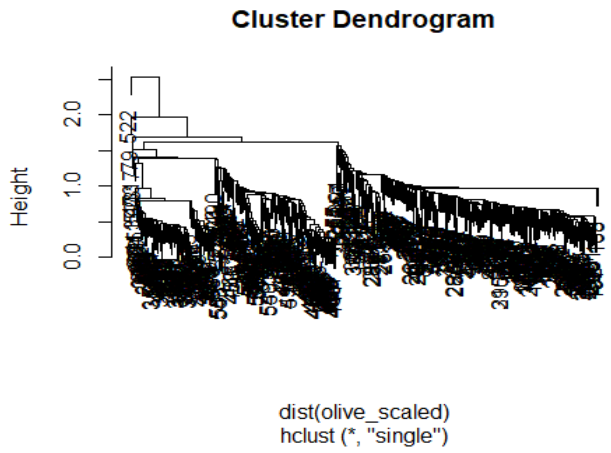
```
table(plotdat3$cluster,oliveoil$region)

##
##     Apulia.north Calabria Apulia.south Sicily Sardinia.inland
##   1            0        0            0      0               0
```

```
##   2           2        0           0      0              0
##   3          21        0           0     17              0
##   4           2       54           8     15              0
##   5           0        1         197      4              0
##   6           0        0           1      0             65
##   7           0        0           0      0              0
##   8           0        0           0      0              0
##   9           0        1           0      0              0
##
##     Sardinia.coast Liguria.east Liguria.west Umbria
##   1              0           13           35      0
##   2              0            3            0     50
##   3              0            0            0      0
##   4              0            0            0      0
##   5              0            0            0      0
##   6              3            0            0      0
##   7             30            0            0      0
##   8              0            0           15      0
##   9              0           34            0      1
```

-Cluster 1 mostly lines up with Liguria.west although it seems to have a little trouble identifying clusters that correspond to Liguria.east and Liguria.west, cluster2 lines up with Umbria. Cluster3 seems to have trouble identifying clusters that correspond to Apulia.north and Sicily. Cluster4 lines up with Calabria. cluster5 lines up mostly with Apulia.south, cluster6 lines up with Sardinia.inland. Cluster7 lines up with Sardinia.coast. Cluster8 lines up with Liguria.west. Cluster9 lines up with Liguria.east. Overally, k-means clustering seems to align with 9 regions well.

(2) hierarchical clustering

```
hc.out4 <- hclust(dist(olive_scaled), method="complete")
plot(hc.out)
```



*Evaluation of hierarchical clustering with 9 regions.

```
aligndat2 <- data.frame(cluster=as.factor(cutree(hc.out4,9)),
                        olive_scaled)
table(aligndat2$cluster,oliveoil$region)
```

```
##
##     Apulia.north Calabria Apulia.south Sicily Sardinia.inland
##   1           13        0            0      0               0
##   2           10        0            0     11               0
##   3            2       35           20      8              65
##   4            0       21            4     16               0
##   5            0        0          170      1               0
##   6            0        0           12      0               0
##   7            0        0            0      0               0
##   8            0        0            0      0               0
##   9            0        0            0      0               0
##
```

```
##     Sardinia.coast Liguria.east Liguria.west Umbria
##   1              0           22            0     51
##   2              0            0            0      0
##   3             33            0            0      0
##   4              0            0            0      0
##   5              0            0            0      0
##   6              0            0            0      0
##   7              0            7            0      0
##   8              0           20           50      0
##   9              0            1            0      0
```

-Cluster1 seems to have trouble identifying clusters that would correspond with Apulia.north,Liguria.east and Umbria. Cluster2 seems to have trouble identifying clusters that would correspond with Apulia.north and Sicily. Cluster3 also seems to have trouble identifying clusters that would correspond with Calabria, Apulia.south, Sardinia.inland, and Sardinia.coast Cluster4 seems to have trouble identifying clusters that would correspond with Calabria and Sicily. Cluster5 and Cluster6 lines up with Apulia.south. Cluster7 lines up with Liguria.east. Cluster 8 seems to have trouble identifying clusters that would corresponds with Liguria.east and Liguria.west. Cluster 9 only has one observation that corresponds with Liguria.east.

-Overally, hierchical clustering seems not to align with 9 regions well.