HW06

Objective:

1. Perform a PCA-based clustering analysis for the samples based on the data set “tobacco_clr$data”.
   - Color the sample points using “tobacco_clr$sample.color” and shape the sample points using “tobacco_clr$sample.pch”.
   - Interpret your results.
2. The data matrix “tobacco_clr$H” is calculated from a phylogenetic tree, a branching diagram that characterizes the evolutionary relationship between OTUs (microbes). Perform the CoIA on “tobacco_clr$data”and “tobacco_clr$H”and construct a CoIA-plot to show the sample clustering of the observations.
   - Color the sample points using "sample.color" and shape the sample points using "sample.pch".
   - Interpret your results.
3. Compare the plots constructed in Steps 1 and 2. Discuss their similarities and differences.

1. PCA-Based Clustering Analysis

```
library(CCA)
```

```
Loading required package: fda

Loading required package: splines

Loading required package: fds

Loading required package: rainbow

Loading required package: MASS

Loading required package: pcaPP

Loading required package: RCurl

Loading required package: deSolve


Attaching package: 'fda'
```

```
The following object is masked from 'package:graphics':

    matplot

Loading required package: fields

Loading required package: spam

Spam version 2.10-0 (2023-10-23) is loaded.
Type 'help( Spam)' or 'demo( spam)' for a short introduction
and overview of this package.
Help for individual functions is also obtained by adding the
suffix '.spam' to the function name, e.g. 'help( chol.spam)'.


Attaching package: 'spam'

The following objects are masked from 'package:base':

    backsolve, forwardsolve

Loading required package: viridisLite


Try help(fields) to get started.
```

```
library(dplyr)
```

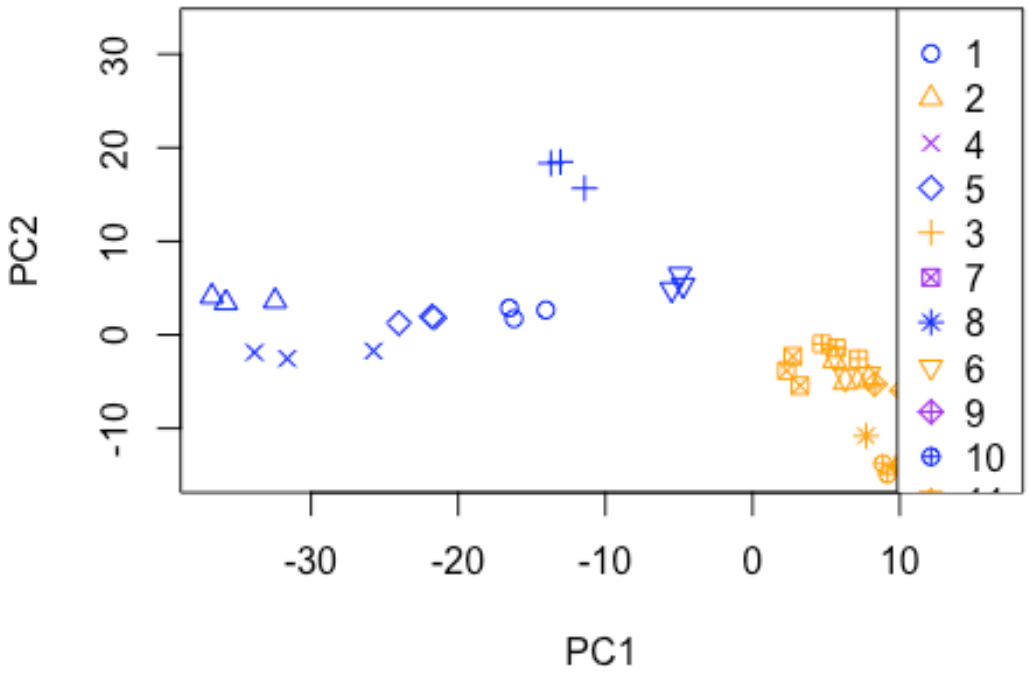
```
Attaching package: 'dplyr'

The following object is masked from 'package:MASS':

    select

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(plyr)
```

```
------------------------------------------------------------------------------
-

You have loaded plyr after dplyr - this is likely to cause problems.
If you need functions from both plyr and dplyr, please load plyr first, then
dplyr:
library(plyr); library(dplyr)
```

```
------------------------------------------------------------------------------
-


Attaching package: 'plyr'

The following objects are masked from 'package:dplyr':

    arrange, count, desc, failwith, id, mutate, rename, summarise,
    summarize
```

```
load("/Users/eschuler/Downloads/tobacco_clr.Rdata")
tobacco_data1 <- tobacco_clr$data
tobacco_h <- tobacco_clr$H
tobacco_color <- tobacco_clr$sample.color
tobacco_pch <- tobacco_clr$sample.pch
tobacco_otu_names <-tobacco_clr$otu.names

#PCA Clustering Analysis
tobacco_c = scale(tobacco_data1, center = T, scale = F)
sigmahat = 1/150*t(tobacco_c)%*%tobacco_c
sigmahat.pca = eigen(sigmahat)
tobacco_c_eigenv = sigmahat.pca$vectors
PCA.cord1 = as.matrix(tobacco_data1)%*%tobacco_c_eigenv[,1]
PCA.cord2 = as.matrix(tobacco_data1)%*%tobacco_c_eigenv[,2]
plot(PCA.cord1,PCA.cord2, xlab="PC1", ylab= "PC2", col=tobacco_color,
pch=tobacco_pch)
legend("topright", legend = unique(tobacco_pch), col = unique(tobacco_color),
pch=unique(tobacco_pch))
```

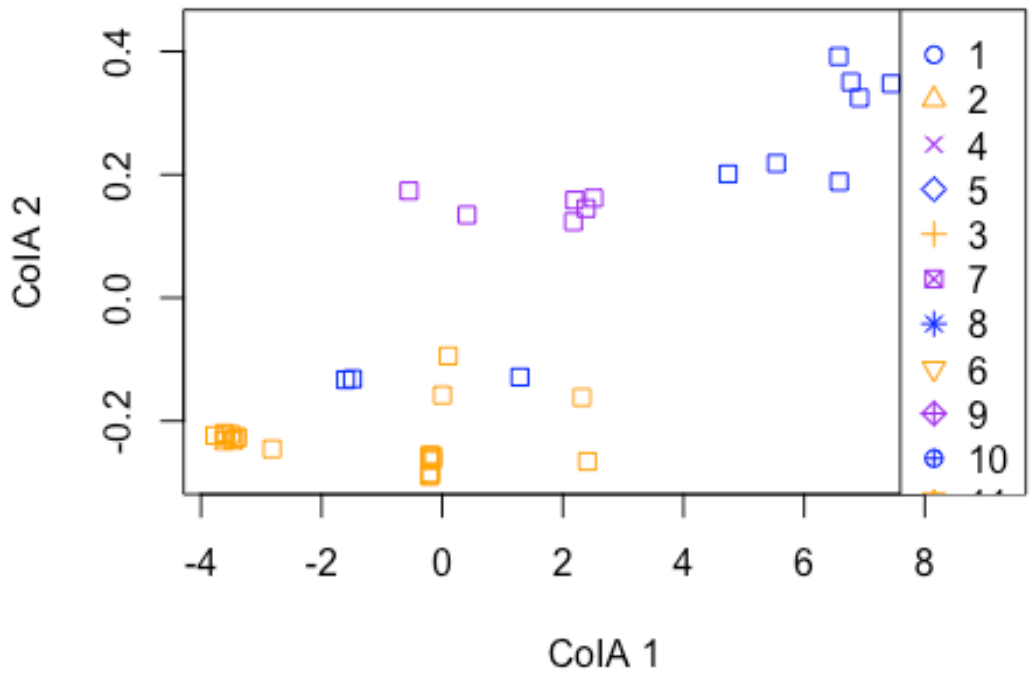



```
#Interpretation
pca_result <-prcomp(tobacco_data1[,1:10], scale. = TRUE)
summary(pca_result)

Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     1.900 1.3461 1.0612 0.99138 0.94433 0.79616 0.58147
Proportion of Variance 0.361 0.1812 0.1126 0.09828 0.08918 0.06339 0.03381
Cumulative Proportion  0.361 0.5422 0.6548 0.75309 0.84226 0.90565 0.93946
                           PC8     PC9    PC10
Standard deviation     0.54724 0.45509 0.31432
Proportion of Variance 0.02995 0.02071 0.00988
Cumulative Proportion  0.96941 0.99012 1.00000
```

The “Standard Deviation” column denotes the variance captured by each PC, with higher values indicating greater data variance explanation. The “Proportion of Variance” reveals the percentage of total dataset variance explained by each PC, with PC1 being most influential at 36.1%, followed by decreasing contributions. The “Cumulative Proportion” depicts the accumulation of explained variance as more PCs are considered, demonstrating how much of the total variance is captured cumulatively. For instance, the first four components explain approximately 75.3% of the variance, and the first six encapsulate around 90.56%. PC1 is pivotal, holding the most variance, aiding in dimensionality reduction or feature selection.

2. CCA and Co-Inertia Analysis

```
x = tobacco_data1 [,1:10]
y = tobacco_h [,1:10]
cca.tobacco = cc(x,y)
CCA.x1 = cca.tobacco$scores$xscores[,1]
CCA.y1 = cca.tobacco$scores$yscores[,2]
CoIA= svd(t(x)%*%y)
CoIA.x1 = x%*%CoIA$u[,1]
CoIA.y1 = y%*%CoIA$v[,1]

plot(CoIA.x1,CoIA.y1, xlab= "CoIA 1", ylab="CoIA 2", col=tobacco_color, pch=
tobacco_h)
legend("topright", legend = unique(tobacco_pch), col = unique(tobacco_color),
pch=unique(tobacco_pch))
```




```
#Interpretation
CoIA_summary <- summary(svd(t(x) %*% y))
print(CoIA_summary)

  Length Class  Mode
d  10    -none- numeric
u 100    -none- numeric
v 100    -none- numeric
```

In the CoIA plot, the orange and purple points seem to be in closer proximity and cluster more than the blue points. This indicates that purple and orange points have strong siimilarites and blue points have weak assosciations between corresponding variables from the two datasets.

3. **Comparing PCA and CoIA:**

   Both plots show that the blue points are more dispersed and do not cluster. They also show that the orange and purple points in both graphs cluster and are in closer proximity than the blue points. This indicates that the blue points are not as closely correlated in CoIA and are more variable within the single data set. The purple and orange points have lower variance and are more closely correlated.

   PCA is used for a single dataset, while CoIA analyzes relationships between two separate datasets. PCA focuses on capturing variance within a single dataset, whereas CoIA identifies shared structures between datasets. PCA's principal components are uncorrelated, while CoIA's components capture shared covariance between datasets. PCA emphasizes variance within a dataset, while CoIA assesses relationships and commonalities between different datasets. PCA is for exploring internal structure within a single dataset, while CoIA is for assessing relationships between distinct datasets.