Homework 5

Elena Volpi

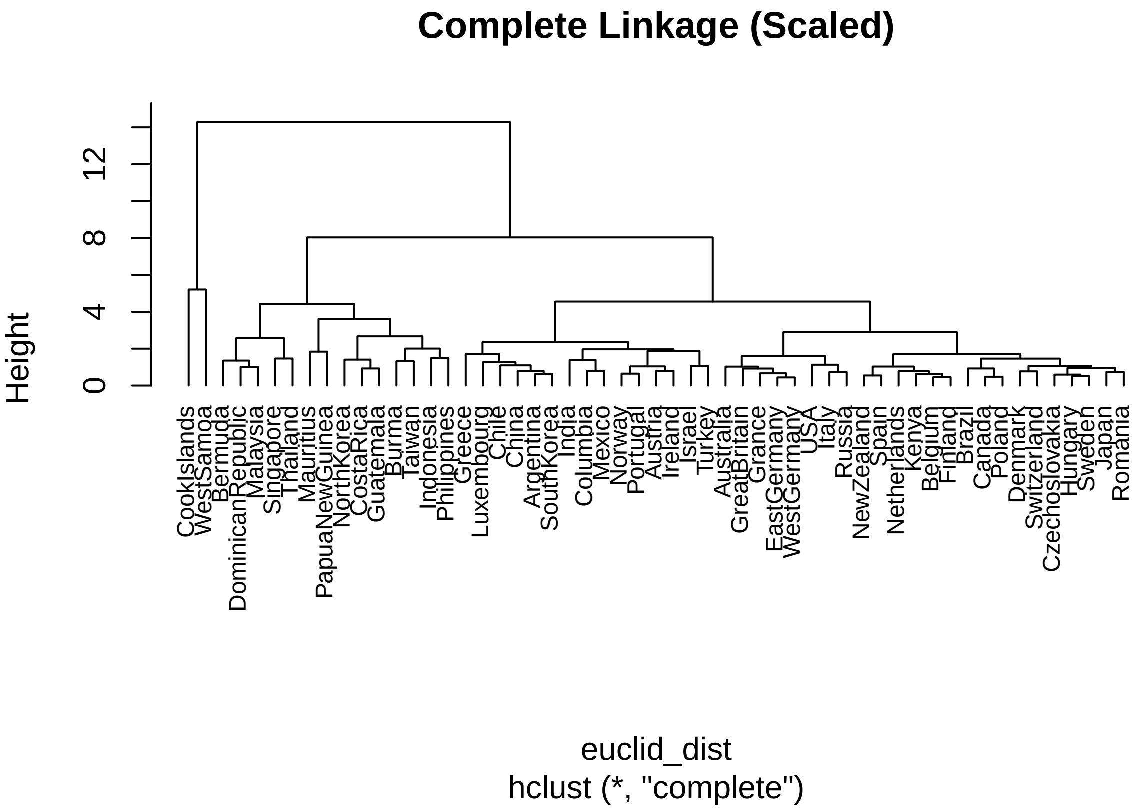
12/6/2021

#Problem 1

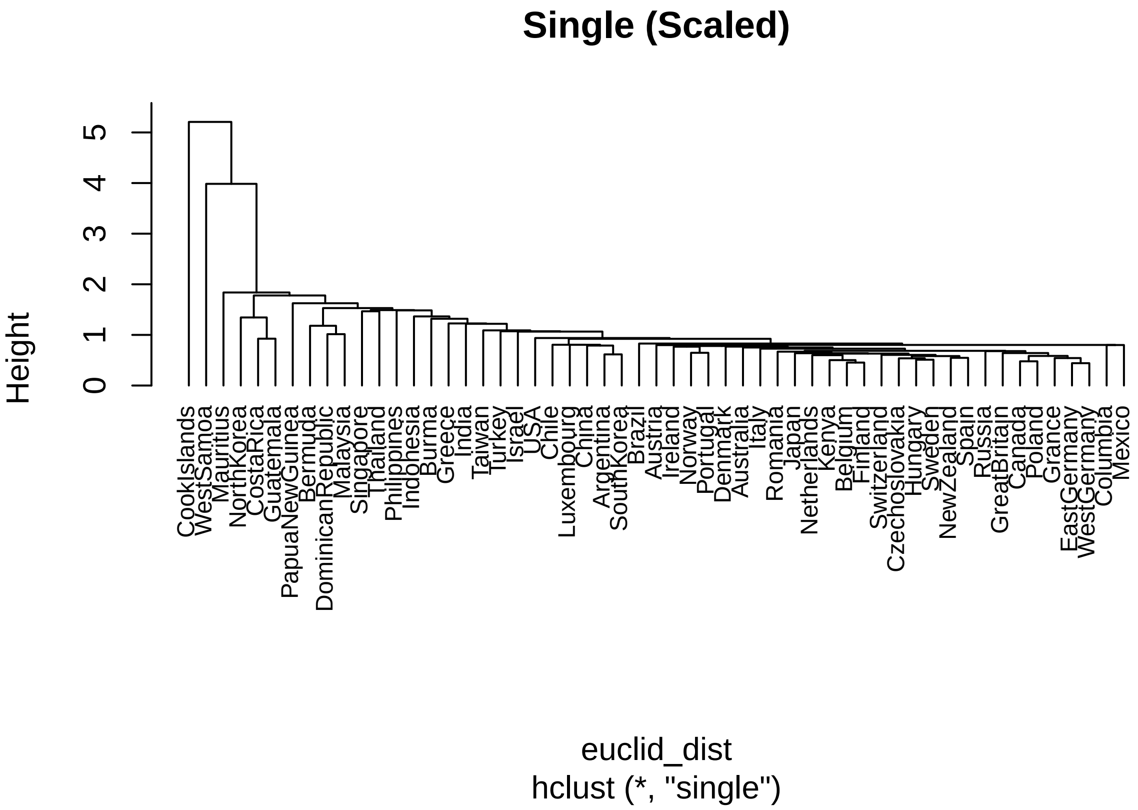
```
TrackData <- read_csv("~/Desktop/TrackData.csv", show_col_types = FALSE)
```

a)Calculate the Euclidean distances between all of the pairs of countries. Using these distances as a measure of dissimilarity, cluster the countries using single linkage and complete linkage hierarchical clustering procedures. Plot the dendrograms and compare the results. Which linkage produces a 'better' clustering of this data, in your opinion? Explain your answer.

```
#scale data and find distance
track_scaled <- scale(TrackData[,3:10])
euclid_dist <- dist(track_scaled)

#cluster

track.hcEucSc <- hclust(euclid_dist, method="complete")
track.hsEucSc <- hclust(euclid_dist, method="single")

plot(track.hcEucSc, labels=TrackData$Country, hang=-1, cex = 0.75, main = "Complete Linkage (Scaled)")
```



```
plot(track.hsEucSc, labels=TrackData$Country, hang=-1, cex = 0.75, main = "Single (Scaled)")
```



The linkage that produced the better clustering of the data would be the complete clustering. The single clustering doesn't give enough information on which countries are similar in terms of track performance and produces many more one-off clusters.

b) K-means for k=2,3,4.

```
#Cluster
track.km2 <- kmeans(track_scaled[,-c(1,2)], centers=2, nstart=10)
track.km3 <- kmeans(track_scaled[,-c(1,2)], centers=3, nstart=10)
track.km4 <- kmeans(track_scaled[,-c(1,2)], centers=4, nstart=10)
```

```
#k=2
TrackData[track.km2$clus==1,1]
```

```
## # A tibble: 41 x 1
##    Country
##    <chr>
##  1 Argentina
##  2 Australia
##  3 Austria
##  4 Belgium
##  5 Brazil
##  6 Canada
##  7 Chile
##  8 China
##  9 Columbia
## 10 Czechoslovakia
## # ... with 31 more rows
```

```
TrackData[track.km2$clus==2,1]
```

```
## # A tibble: 14 x 1
##    Country
##    <chr>
##  1 Bermuda
##  2 Burma
##  3 CookIslands
##  4 CostaRica
##  5 DominicanRepublic
##  6 Guatemala
##  7 Indonesia
##  8 Malaysia
##  9 Mauritius
## 10 PapuaNewGuinea
## 11 Philippines
## 12 Singapore
## 13 Thailand
## 14 WestSamoa
```

```
#k=3 clusters
TrackData[track.km3$clus==1,1]
```

```
## # A tibble: 34 x 1
##    Country
##    <chr>
##  1 Australia
##  2 Austria
##  3 Belgium
##  4 Brazil
##  5 Canada
##  6 Chile
##  7 Columbia
##  8 Czechoslovakia
##  9 Denmark
## 10 Finland
## # ... with 24 more rows
```

```
TrackData[track.km3$clus==2,1]
```

```
## # A tibble: 19 x 1
##    Country
##    <chr>
##  1 Argentina
##  2 Bermuda
##  3 Burma
##  4 China
##  5 CostaRica
##  6 DominicanRepublic
##  7 Guatemala
##  8 Indonesia
##  9 Israel
## 10 SouthKorea
## 11 NorthKorea
## 12 Luxembourg
## 13 Malaysia
## 14 Mauritius
## 15 PapuaNewGuinea
## 16 Philippines
## 17 Singapore
## 18 Taiwan
## 19 Thailand
```

```
TrackData[track.km3$clus==3,1]
```

```
## # A tibble: 2 x 1
##   Country
##   <chr>
## 1 CookIslands
## 2 WestSamoa
```

```
#K=4 clusters
TrackData[track.km4$clus==1,1]
```

```
## # A tibble: 30 x 1
##    Country
##    <chr>
##  1 Australia
##  2 Austria
##  3 Belgium
##  4 Brazil
##  5 Canada
##  6 Czechoslovakia
##  7 Denmark
##  8 Finland
##  9 Grance
## 10 EastGermany
## # ... with 20 more rows
```

```
TrackData[track.km4$clus==2,1]
```

```
## # A tibble: 2 x 1
##   Country
##   <chr>
## 1 CookIslands
## 2 WestSamoa
```

```
TrackData[track.km4$clus==3,1]
```

```
## # A tibble: 14 x 1
##    Country
##    <chr>
##  1 Argentina
##  2 Burma
##  3 Chile
##  4 China
##  5 Columbia
##  6 CostaRica
##  7 Greece
##  8 Guatemala
##  9 Israel
## 10 SouthKorea
## 11 NorthKorea
## 12 Luxembourg
## 13 Taiwan
## 14 Turkey
```

```
TrackData[track.km4$clus==4,1]
```

```
## # A tibble: 9 x 1
##   Country
##   <chr>
## 1 Bermuda
## 2 DominicanRepublic
## 3 Indonesia
## 4 Malaysia
## 5 Mauritius
## 6 PapuaNewGuinea
## 7 Philippines
## 8 Singapore
## 9 Thailand
```

c) Compare to complete linkage

```
table(track.km2$clus, cutree(track.hcEucSc, k=2))
```

```
##
##      1  2
##   1 41  0
##   2 12  2
```

```
table(track.km3$clus, cutree(track.hcEucSc, k=3))
```

```
##
##      1  2  3
##   1 34  0  0
##   2  5 14  0
##   3  0  0  2
```

For k=2, hierarchical clustering groups 53 countries in cluster 1 and 2 countries in group 2. Whereas, k-means groups 41 countries in cluster 1 and 14 in cluster 2.

For k=3, k-means has 2,19,and 34 countries in clusters 1,2,3, respectively. Hierarchical clustering has 39 countries in cluster 1, 14 countries in cluster 2, and 2 countries in cluster 3.

The k-means clustering seems to give more even clusters and the clusters of countries seem more sensible, although my track knowledge is limited.

#Problem 2

```
Archaeo <- read_csv("ArchaeoData.csv")
```

```
## Rows: 9 Columns: 10
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): X
## dbl (9): Site1, Site2, Site3, Site4, Site5, Site6, Site7, Site8, Site9
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

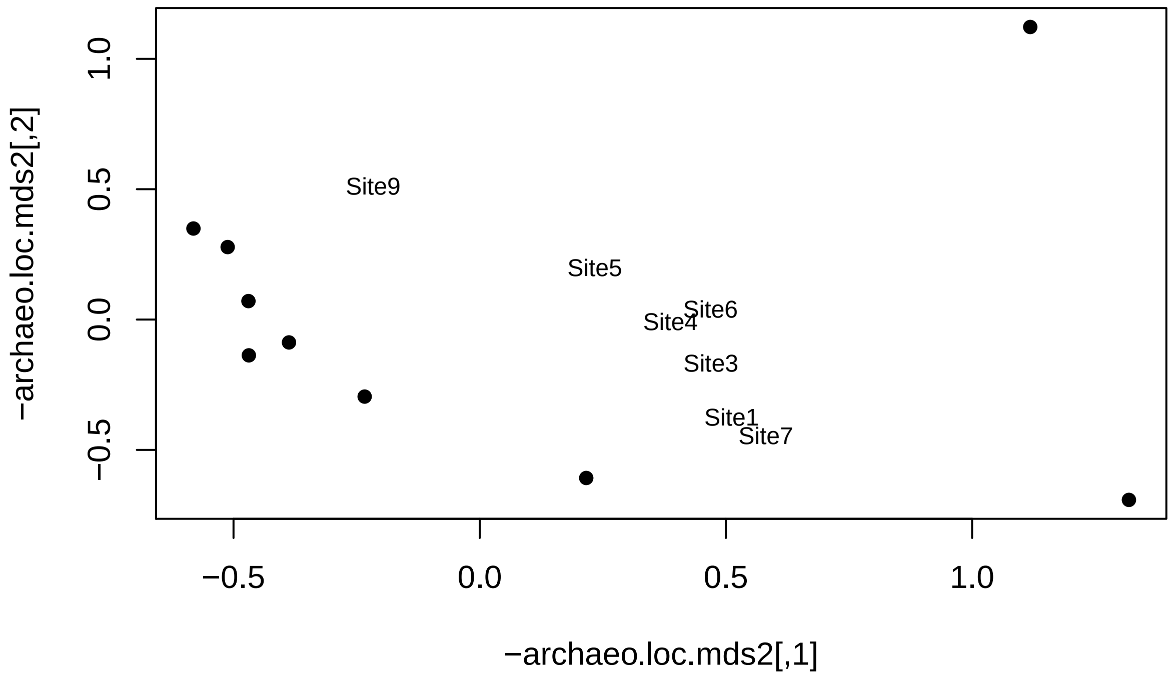
```
#Determine coordinates of q=2, q=3 dimensions using multidimensional scaling.

archaeo.loc.mds2 <- cmdscale(Archaeo[,-1], k=2)
archaeo.loc.mds3 <- cmdscale(Archaeo[,-1], k=3)

-archaeo.loc.mds2
```

```
##             [,1]        [,2]
##  [1,] -0.5119010  0.27797661
##  [2,]  1.3184960 -0.69177869
##  [3,] -0.4696574  0.07075632
##  [4,] -0.3874028 -0.08774518
##  [5,] -0.2336943 -0.29550962
##  [6,] -0.4688497 -0.13734912
##  [7,] -0.5814134  0.34919001
##  [8,]  1.1180751  1.12218941
##  [9,]  0.2163475 -0.60772973
```

```
-archaeo.loc.mds3
```

```
##             [,1]        [,2]        [,3]
##  [1,] -0.5119010  0.27797661 -0.24210462
##  [2,]  1.3184960 -0.69177869 -0.62299269
##  [3,] -0.4696574  0.07075632 -0.18553022
##  [4,] -0.3874028 -0.08774518 -0.04893247
##  [5,] -0.2336943 -0.29550962  0.32518484
##  [6,] -0.4688497 -0.13734912  0.21876261
##  [7,] -0.5814134  0.34919001 -0.45732159
##  [8,]  1.1180751  1.12218941  0.31595964
##  [9,]  0.2163475 -0.60772973  0.69697450
```

```
plot(-archaeo.loc.mds2,pch=16)
text(archaeo.loc.mds2, labels = Archaeo$X, pos=1, cex=0.75)
```



#Problem 3 For each of the following scenarios, indicate which of the following test procedures you would use to answer the specified questions.

a) Twenty subjects were given each of three diets (in random order) and the subjects' blood pressures were measured at the end of each diet, so there were three blood pressure measurement associated with each subject. Question: Did the different treatments affect the subjects' blood pressure differently?

Answer a) :Hotelling's Tˆ(2) repeated measures.

b) Two varieties of chickweed are difficult to distinguish. Measurements on four variables were obtained for chickweed plants whose variety was known.

Question: Use these observations to establish a rule for classifying a new candidate plant into one of the two varieties.

Answer b): Discriminant Function Analysis/Linear Discriminant Analysis

c) Each of 50 eight-year-old girls and 50 eight-year-old boys were given a total of 10 tests. Five of these tests had to do with language and five had to do with mathematical reasoning.

Question: Do scores differ between boys and girls?

Answer: Hotelling's two-sample T 2test

Question: Combining the boys and girls, what combination of the language tests is most associated with some combination of the math tests?

Answer: Canonical Correlation Analysis

d) Daily measurements of seven pollution-related variables were recorded over an extended period of time at a single location in Los Angeles.

Question: Find a low-dimensional representation for these variables that captures most of the variability.

Answer: Multidimensional scaling

Question: Test whether the pollution on weekends differed from that on weekdays.

Answer: Hotelling's two sample Tˆ2 test.

e)For each of a sample of 42 new microwaves made by a certain manufacturer, the amount of radiation emitted when the door of the microwave is closed and the amount of radiation emitted when the door of the microwave is opened are measured.

Question: Construct a confidence interval for the difference in amount of radiation emitted under these two conditions.

**Answer: Univariate t-test (paired differences)

f) A sample of 50 married couples was obtained. The wife and the husband each answered four questions regarding their relationship on a scale of 0 to 10.

Question: Do the wife's answers tend to be similar to the husband's answers, and in what way are they most similar? That is, what combination of the wife's answers is most similar to what combination of the husband's answers?

**Answer: Canonical Correlation Analysis

g)The standardized scores for each of the ten events in the decathlon were obtained for each of 50 entrants.

Question: Can the variation in the scores be explained by three underlying athletic abilities, and how might these abilities be described?

Answer: Factor analysis

h) For 15 different species of predator fish, data were gathered on several aspects of their diet.

Question: How can these species of fish be grouped based on similarities in their diet?

Answer: Clustering

i) Calcite content was measured at 25 equally-spaced locations along the leg bone for each of seven Tyrannosaurus Rex skeletons and also for each of five skeletons of a newly-discovered type of dinosaur.

Question: Do the calcite concentrations at these locations differ between the two dinosaur species? Answer: Hotelling's two-sample Tˆ2 test.

Question: Combining the dinosaur species, is calcite concentration the same at all of the measured locations in the leg bone? Answer: ANOVA

Question: Based on these measurements, construct a rule for classifying a new bone as coming from a Tyrannosaurus Rex or from the newly-discovered species. Answer: Discriminant Function Analysis/Linear Discriminant Analysis

j) Blood samples from 40 patients were obtained and each divided into six subsamples, which were sent to six different laboratories to have iron content measured

Question: Do the six different laboratory results have the same means?

**Answer: ANOVA

k) Measurements on six accounting and financial variables were obtained from a sample of insurance companies that were distressed (close to bankrupt) and an independent sample of insurance companies that were solvent.

Question: Establish a rule for classifying future insurance companies as solvent or distressedbased on these variables. Answer: LDA/discriminant analysis

l) DNA analysis was performed on hair specimens from each of 100 mummies taken from Egyptian pyramids. For each mummy, twenty variables concerning the DNA sequence were measured.

Question: Based on the measured variables, identify groups of mummies that are related to each other (have similar values of the variables). Answer: Clustering

Question: Based on the distances between these variables, construct a two-dimensional plot of the mummies to visualize the groupings. Answer: Multidimensional Scaling

m) SAT subject test scores are obtained for a random sample of 100 12th graders who took Math, Biology, Literature, and World History subject tests

Question: Test whether the average score for all four tests is 500. Answer: One sample Hotelling's T test

Question: Question: Test whether the average scores are equal for all four tests.

** Answer: Hotelling's one-sample T^2 test (repeated measures)

n) A wildlife ecologist measured tail length and wing length for a sample of 45 female hook-billed kites and 45 male hook-billed kites.

Question: Are average tail length and wing length the same for female and male hook-billed kites? Answer: Hotelling's two sample

o) Several measurements were obtained on chief executive officers (CEO) of companies, regarding the degree to which the officers took risks. Several additional measurements were available on the success of the company under their leadership.

Question: What aspects of risk-taking propensity of the CEO are associated with which aspects of company success? Answer: Canonical Correlation Analysis

Question: What combination of risk-taking propensities displays the greatest variation between CEOs? Answer: PCA

p) The age, diameter, and height were measured for a sample of trees that contained eagle roost sites and for an independent sample of trees that did not contain eagle roost sites.

Question: Construct confidence intervals for the difference in age, difference in diameter, and difference in height between roosting trees and non-roosting trees.

Answer: Bonferroni Simultaneous confidence intervals.

Question: Determine a rule for classifying a new tree as a likely roosting site or unlikely roosting site, based on these three variables.

Answer: LDA/discriminant function analysis

q) For all of the NBA rookies who started in 2000, data were collected on their free-throw percentages each year for the first five years of their NBA careers Question: Does average free-throw percentage change over these five years? Answer: Hotelling's one-sample repeated measures

r) Twelve measurements were taken on fossilized skull measurements from 20 kinds of squirrels. The goal of the analysis was to order the 20 squirrels chronologically, on the basis of the similarities between the skull measurements for different squirrels.

Question: Find a one-dimensional representation of the 20 squirrels that best captures the differences between the measured variables. Answer: Multi-dimensional scaling

s) The protein, fat content, calories, and Vitamin A content were measured for each of ten brands of hot dogs. Question: Group the brands of hot dogs based on their nutritional content. Answer: Clustering

Question: What combination of these nutritional measurements captures the greatest difference between the hot dog brands? Answer: PCA

t) Measurements were obtained on five pre-college predictor variables and four college performance variables for each of several hundred students.

Question: What combination of pre-college variables is most associated with a combination of college performance? Answer: Canonical Correlation Analysis

Question: Question: Combining the two variable sets, are there a few underlying abilities that explain the pre-college and college performance? Answer: Factor Analysis.