Homework 6

ST557

Nick Sun

December 4, 2019

Question 1

Here we are analyzing diabetic and non-diabetic patients across 3 different primary variables: glucose intolerance, insulin response, and insulin resistance. There are also 2 different secondary variables: relative weight and fasting plasma glucose.

For this question, we will be doing a canonical correlation analysis for the provided covariance matrix.

```
diabetic.sig11 <- matrix(c(1106, 396.7, 108.4,
                        396.7, 2382, 1143,
                        108.4, 1143, 2136), nrow = 3)
diabetic.sig22 <- matrix(c(0.02, 0.22,
                          0.22, 70.56), nrow = 2)
diabetic.sig12 <- matrix(c(0.79, 26.23,
                           -0.21, -23.96,
                           2.19, -20.84), byrow = T, nrow = 3)
```

We'll begin by computing S_{11} and S_{22} following the code from the R supplement:

```
diabetic.sig11.eig <- eigen(diabetic.sig11)
diabetic.sig11.5 <- diabetic.sig11.eig$vectors %*%
  diag(sqrt(diabetic.sig11.eig$values)) %*%
  t(diabetic.sig11.eig$vectors)
```

```
diabetic.sig22.eig <- eigen(diabetic.sig22)
diabetic.sig22.5 <- diabetic.sig22.eig$vectors %*%
  diag(sqrt(diabetic.sig22.eig$values)) %*%
  t(diabetic.sig22.eig$vectors)
```

Using S_{11} and S_{22}, we can compute A_1 and A_2 as well as the eigendecomposition of these matrices.

```
diabetic.A1 <- solve(diabetic.sig11.5) %*%
  diabetic.sig12 %*% solve(diabetic.sig22) %*%
  t(diabetic.sig12) %*% solve(diabetic.sig11.5)
diabetic.A1.eig <- eigen(diabetic.A1)

diabetic.A2 <- solve(diabetic.sig22.5) %*%
  t(diabetic.sig12) %*% solve(diabetic.sig11) %*%
  diabetic.sig12 %*% solve(diabetic.sig22.5)
diabetic.A2.eig <- eigen(diabetic.A2)
```

Finally, we get the follwing coefficent vectors:

Table 1: Canonical Coefficients of X(1)

	a1	a2
Glucose Intolerance	0.0131880	0.0247137
Insulin Response	-0.0144335	-0.0092853
Insulin resistance	0.0233716	-0.0087275

Table 2: Canonical Coefficients of X(2)

	b1	b2
Relative Weight	-7.1854840	0.3802300
Fasting Plasma Glucose	0.0161129	-0.1200668

For a_1 and b_1:

A possible interpretation of these canonical variables is that a weighted difference of relative weight and plasma is highly correlated with the weighted difference in insulin response and insulin intolerance/insulin resistance.

For a_2 and b_2:

The weighted difference in fasting weight and glucose is highly correlated with a weighted difference in glucose intolerance vs. insulin response/insulin resistance.

Question 2

For this question, we are analyzing crude oil drawn from two different zones of sandstone: sub-mulina and upper sandstone.

The samples were analyzed for three trace elements (vanadium, iron, beryllium) and two hydrocarbon metrics (saturated and aromatic).

Part a.

Let's perform a linear discriminant analysis of this data. This part is easy enough using the `lda()` function. We will need this in order to get Fisher's Discriminant Function.

	LD1
Vanadium	0.2108
Iron	-0.03707
Beryllium	2.96
SatHydroCarb	-0.8462
AroHydroCarb	-0.001727

We will classify x_0 as part of π_1 aka part of the sub-mulina if the following is true

$$\boldsymbol{a}^T x_0 \leq \frac{\boldsymbol{a}^T \bar{\boldsymbol{X}}_1 + \boldsymbol{a}^T \bar{\boldsymbol{X}}_2}{2}$$

In other words, if we get a value of $\boldsymbol{a}^T x_0$ that is less than -3.6596338 is coming from sub-mulina π_1.

Part b.

Let's construct a confusion matrix and calculate the APER.

```
confusion.matrix <- rep(1, nrow(crudeOil))
```

Table 3: Canonical Correlations

	x
r1	0.4611246
r2	0.1254845

Table 5: APER

x
0.0408163

```
  }
}

APER <- 1 - mean(confusion.matrix == crudeOil$Population)
kable(APER,
      caption = "APER")
```

The apparent error rate is actually pretty low, only around 0.0408163!

Part c.

Now we observe a new $\mathbf{x} = [4, 17, .5, 5.54, 3.51]^T$. How would Fisher's Discriminant Function classify it?

This is easy enough to do using the `predict()` function.

- **class**: *2*
- **posterior**:

1	2
0.03048	0.9695

- **x**:

LD1
-0.2701

The `$class` element in this list shows that Fisher's Discriminant Function classified $\mathbf{x}$ to population 2, the upper sandstone!

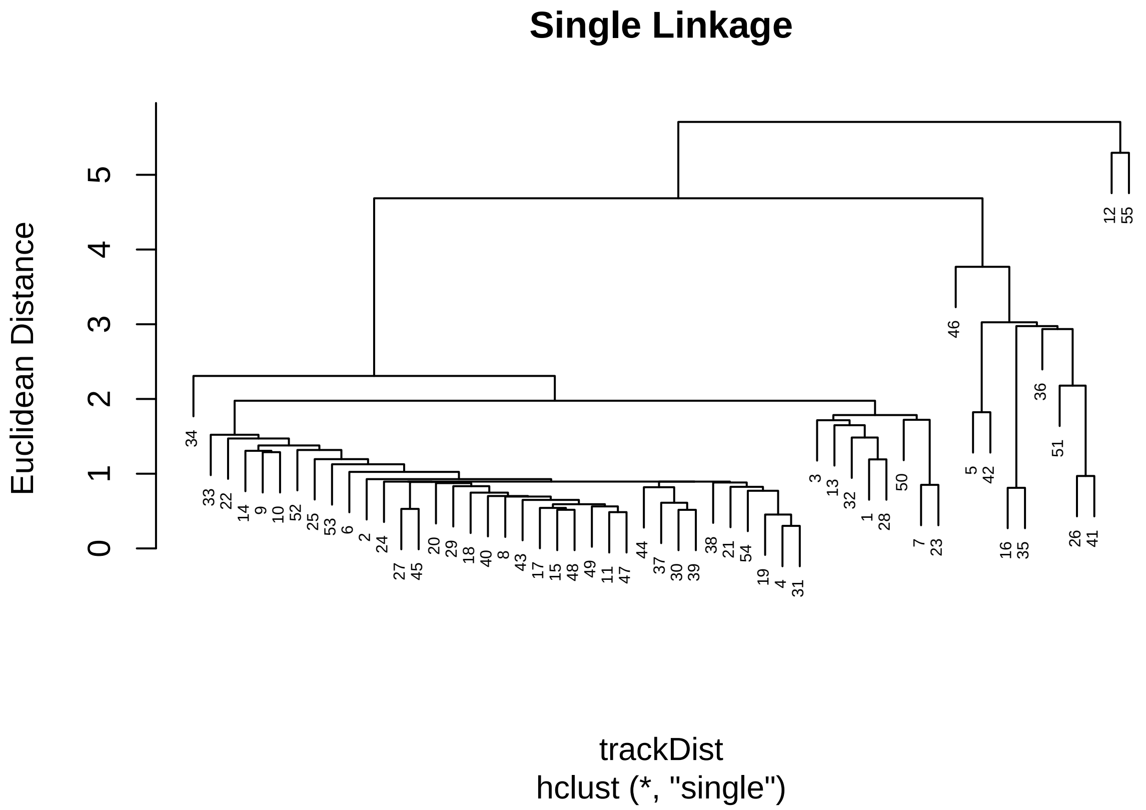
Question 3

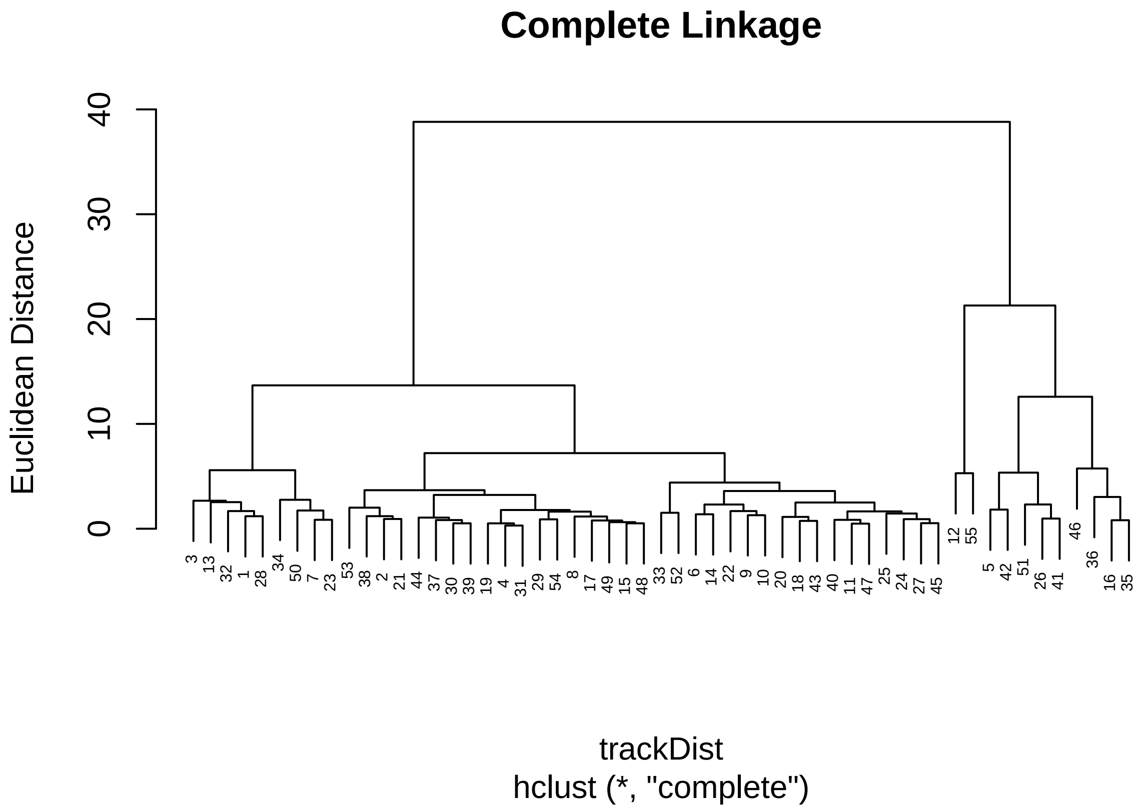
For this problem, we will analyze track and field data for 55 different countries.

Part a.

Let's cluster these countries by their track times using hierarchical clustering.

We will use Euclidean Distance as the distance metric and try both single and complete linkage.


Single Linkage
Euclidean Distance
0
1
2
3
4
5
34
33
22
14
9
10
52
25
53
6
2
24
27
45
20
29
18
40
8
43
17
15
48
49
11
47
44
37
30
39
38
21
54
19
4
31
3
13
32
1
28
50
7
23
46
5
42
16
35
36
51
26
41
12
55
trackDist
hclust (*, "single")



In my opinion, the complete linkage provides a more interesting clustering since it seems to partition the data into two distinct groups.

```
complete_indices <- cutree(trackComplete, h = 30)
track$Country[complete_indices == 2]
```

```
##  [1] Bermuda           CookIslands       DominicanRepublic
##  [4] Indonesia         Malaysia          Mauritius
##  [7] PapuaNewGuinea    Philippines       Singapore
## [10] Thailand          WestSamoa
## 55 Levels: Argentina Australia Austria Belgium Bermuda Brazil ... WestSamoa
```

The clustering produces a large cluster that seems to encompass most large countries and a second smaller cluster that consists mostly of smaller island nations or countries that are in southeast Asia. Pretty interesting!

Part b.

Now let's cluster using k-means when k = 2, 3, and 4.

k = 2

With 2 clusters, we get the following clustering of countries.

Bermuda, *CookIslands*, *DominicanRepublic*, *Indonesia*, *Malaysia*, *Mauritius*, *PapuaNewGuinea*, *Philippines*, *Singapore*, *Thailand* and _WestSamoa___Argentina_, *Australia*, *Austria*, *Belgium*, *Brazil*, *Burma*, *Canada*,

Table 8: Counts in Each cluster

Var1	Freq
1	11
2	44

Chile, *China*, *Columbia*, *CostaRica*, *Czechoslovakia*, *Denmark*, *Finland*, *Grance*, *EastGermany*, *WestGermany*, *GreatBritain*, *Greece*, *Guatemala*, *Hungary*, *India*, *Ireland*, *Israel*, *Italy*, *Japan*, *Kenya*, *SouthKorea*, *NorthKorea*, *Luxembourg*, *Mexico*, *Netherlands*, *NewZealand*, *Norway*, *Poland*, *Portugal*, *Romania*, *Spain*, *Sweden*, *Switzerland*, *Taiwan*, *Turkey*, *USA* and *Russia*

This is actually identical to the clustering produced by the hierarchical clustering with complete linkage!

k = 3

With 3 clusters, we get the following clustering:

Var1	Freq
1	36
2	10
3	9

```
## [1] "Cluster 1"
```

Australia, *Austria*, *Belgium*, *Brazil*, *Canada*, *Chile*, *China*, *Columbia*, *Czechoslovakia*, *Denmark*, *Finland*, *Grance*, *EastGermany*, *WestGermany*, *GreatBritain*, *Greece*, *Hungary*, *India*, *Ireland*, *Italy*, *Japan*, *Kenya*, *NorthKorea*, *Mexico*, *Netherlands*, *NewZealand*, *Norway*, *Poland*, *Portugal*, *Romania*, *Spain*, *Sweden*, *Switzerland*, *Turkey*, *USA* and *Russia*

```
## [1] "Cluster 2"
```

Argentina, *Bermuda*, *Burma*, *CostaRica*, *Guatemala*, *Israel*, *SouthKorea*, *Luxembourg*, *Philippines* and *Taiwan*

```
## [1] "Cluster 3"
```

CookIslands, *DominicanRepublic*, *Indonesia*, *Malaysia*, *Mauritius*, *PapuaNewGuinea*, *Singapore*, *Thailand* and *WestSamoa*

k = 4

Var1	Freq
1	9
2	18
3	9
4	19

```
## [1] "Cluster 1"
```

Argentina, *Bermuda*, *Burma*, *CostaRica*, *Guatemala*, *Israel*, *Luxembourg*, *Philippines* and *Taiwan*

```
## [1] "Cluster 2"
```

Australia, *Belgium*, *Canada*, *Denmark*, *Finland*, *EastGermany*, *GreatBritain*, *Italy*, *Japan*, *Kenya*, *Mexico*, *Netherlands*, *NewZealand*, *Portugal*, *Sweden*, *Switzerland*, *USA* and *Russia*

```
## [1] "Cluster 3"
```

CookIslands, *DominicanRepublic*, *Indonesia*, *Malaysia*, *Mauritius*, *PapuaNewGuinea*, *Singapore*, *Thailand* and *WestSamoa*

```
## [1] "Cluster 4"
```

Austria, *Brazil*, *Chile*, *China*, *Columbia*, *Czechoslovakia*, *Grance*, *WestGermany*, *Greece*, *Hungary*, *India*, *Ireland*, *SouthKorea*, *NorthKorea*, *Norway*, *Poland*, *Romania*, *Spain* and *Turkey*

Part c.

From a strategic standpoint, I prefer hierarchical clustering since we don't know (at least I don't) *a priori* how many clusters there will be in the data (one for each continent? one for each level of developed world? It's hard to know). During our exploration of k = 3,4, I don't immediately see any practical interpretation for the clusters like I did with just 2 clusters. In that case, k-means and complete linkage hierarchical clustering produced the same groups, and I prefer the hierarchical clustering dendrogram since we can easily visualize the "compactness" of the clusters.

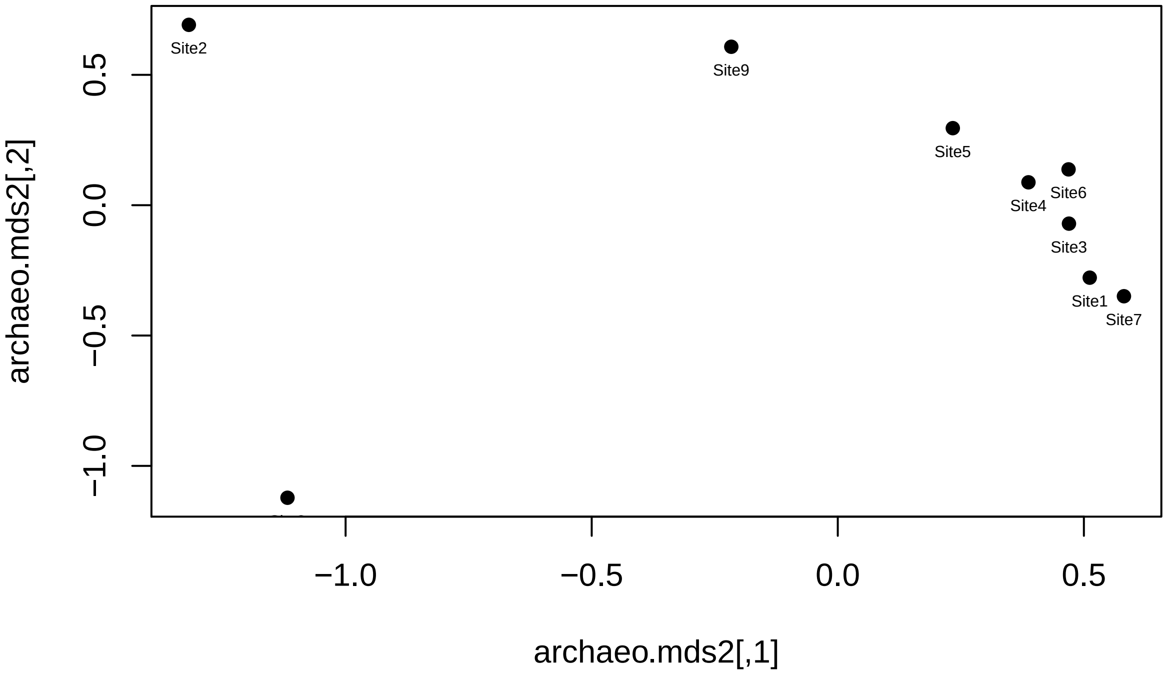
Question 4

Now let's do some multidimensional scaling with archaelogical data!

q = 2

Here is the results for q = 2:

Site1	0.5119	-0.278
Site2	-1.318	0.6918
Site3	0.4697	-0.07076
Site4	0.3874	0.08775
Site5	0.2337	0.2955
Site6	0.4688	0.1373
Site7	0.5814	-0.3492
Site8	-1.118	-1.122
Site9	-0.2163	0.6077



q = 3

Site1	0.5119	-0.278	0.2421
Site2	-1.318	0.6918	0.623
Site3	0.4697	-0.07076	0.1855
Site4	0.3874	0.08775	0.04893
Site5	0.2337	0.2955	-0.3252
Site6	0.4688	0.1373	-0.2188
Site7	0.5814	-0.3492	0.4573
Site8	-1.118	-1.122	-0.316
Site9	-0.2163	0.6077	-0.697

Question 5

Now it's time for a multivariate marathon!

Part a.

Twenty subjects were given three diets and blood pressure was measured with each diet. Did the different treatments affect the subjects' blood pressure differently?

- f. Hotelling's one-sample T^2 test (repeated measures)

Part b.

Two varieties of chickweed are difficult to distinguish and measurement on 4 variables are taken on chickweed whose variety is known. Let's establish a rule for classifying a new candidate plant.

- o. Discriminant Function Analysis/Linear Discriminant Analysis

Part c.

50 8-year old girls and 50 8-year old boys were given 10 tests. Fives tests had to do with language and fives tests had to do with math.

Do scores differ between boys and girls?

- g. Hotelling's two-sample T^2 test

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

- n. Canonical Correlation Analysis

Part d.

Daily measurments of seven air pollution variables were measured in a LA.

Find a low dimensional representation of these variables that captures the most variability

- q. Multidimensional Scaling

Test whether pollution on weekdays differs than on weekends.

- g. Hotelling's two-sample T^2 test (don't think this is repeated measures since air is an open system?)

Part e.

Microwave radiations are measured with the doors open and doors closed.

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

- b. Univariate one-sample t-test (paired differences)

Part f.

50 married couples were asked four questions regarding their relationship.

Do the wife's answers tend to be similar to the husband's answers and in what way or combination are they the most similar?

- n. Canonical Correlation Analysis

Part g.

The standardized scores for 10 events in the decathlon were obtained for each entrant.

Can the variation be explained by three underlying athletic abilities and how can these abilities be described?

- m. Factor Analysis

Part h.

For 15 different species of predator fish, data was gathered on their diet.

How can these species be grouped based on similarities in their diet?

- p. Clustering

Part i.

Calcite was measured at 25 locatiions on the leg bone for 7 T-rex skeletons and also 5 skeletons of a new dinosaur.

Do calcite concentrations at these locations differ between the dinosaur species?

- g. Hotelling's two-sample T^2 test

Combining the species, is calcite the same at all measured locations on the leg bone.

- f. Hotelling's one-sample T^2 test (repeated measures)

Construct a new rule for classifying a new bone as coming from the T-rex or the newly discovered species.

- o. Discriminant Function Analysis/Linear Discriminant Analysis

Part j.

Blood samples from 40 patients were split into 6 subsamples which were sent to 6 different labs.

Do the six different labs have the same means?

- f. Hotelling's one-sample T^2 test (repeated measures)

Part k.

Measurements on 6 accounting variables were obtained from a sample of insurance companies that were in financial trouble and in independent sample of insurance companies that were solvent.

Establish a rule for classifying future insurance companies as solvent or distressed.

- o. Discriminant Function Analysis/Linear Discriminant Analysis

Part l.

DNA analysis was performed hair from 100 mummies. 20 variables concerning DNA sequences were measured.

Identify groups of mummies that are lreated to each other.

- p. Clustering

Based on the distances between these variables, construct a 2 dimenstional plot of the mummies to visualize the groupings.

- q. Multidimensional Scaling

Part m.

SAT subject tests were obtained for 100 12th graders.

Test whether the average score for all four tests is 500.

- h. Bonferroni simultaneous tests

Test whether the average scores are equal for 100 four tests.

- i. ANOVA

Part n.

Tail length and wing length for male and female kites are measured.

Are average tail length and wing length the same for female and male kites?

- g. Hotelling's two sample T^2 test

Part o.

Several measurments were obtained on CEOs on risk-taking and the success of their companies.

What aspects of risk-taking in CEOs are associated with which aspects of success?

- n. Canonical Correlation Analysis

What combination of risk-taking propensities displays the greatest variation between CEOs?

- l. Principal Components Analysis

Part p.

The age, diameter, height were measured on trees that contained eagle roosts and an independent sample of sites that did not have eagle roosts.

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

- h. Bonferroni simultaneous tests

Determine a rule for classifiying a new tree as a likely roosting site based on these three variables?

- o. Linear Discriminant Analysis

Part q.

For all NBA rookies, data was collected on their free throw percentages over their first 5 years.

Does average free throw percentage change over these five years?

- f. Hotelling's one-sample T^2 test (repeated measures)

Part r.

Skull measurements are taken on 20 kinds of squirrels.

Find a one dimensional representation of the 20 squirrels that best captures the difference between the measured variables.

- q. Multidimensional Scaling

Part s.

Variables are measured on different hotdogs.

Group the brands of hot dogs based on their nutritional content.

- p. Clustering

What combination of these nutritional measurments captures the greatest difference between the hot dog brands?

- o. Linear Discriminant Analysis

Part t.

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

What combination of pre-college variables is most associated with a combination of college performance.

- n. Canonical Correlation Analysis

Combining the two variable sets, are there a few underlying abilities that explain the pre-college and college performance?

- m. Factor Analysis