CSC424 – Lab Day

1. Outline
    a. Multiple Regression
        i. Example: Bike Share
        ii. Assignment: Bike Share continued
    b. Principle Component Analysis
        i. Example: European Protein Consumption
        ii. Assignment: Rain Forest
    c. Linear Discriminant Analysis
        i. Example: Iris
        ii. Assignment: Thyroid
    d. Multi-dimensional Scaling
        i. Example: Eurodist
        ii. Assignment: Kinship

- This lab is meant to give you semi-supervised experience working with R.
- The document is **not** meant to be a tutorial, it is merely a guide for the lecture. If you are a DL student, please watch the lecture.
- For each topic, we will do a sample problem as a class.
- I will then give you time to work on a similar problem.
- These4 problems will be your Lab 2.
- Your completed Lab will be due next week (see schedule).
- Please take special note of the "**Submit**" for each assignment.

Task 1: Multiple Regression

Example: Bike Share

Examine the data fields. Notice that many of the variables are categorical even though they appear numerical. We will create some dummy variables, look at the correlations, build a model and evaluate the output. Here is some example code to get us started…

- BikeShareDay$S1 <- (BikeShareDay$season==1)*1
- BikeShareDay$W1 <- (BikeShareDay$weathersit==1)*1
- cor(BikeShareDay[3:22])
- CntModel<- lm(cnt ~ atemp + hum + windspeed + S1 + S2 + S3 + W1 + W2 + W3, data = BikeShareDay)
- summary(CntModel)

Assignment 1: Bike Share Continued

For your assignment:

- Where appropriate, create dummy variables for other variables in the dataset.
- Evaluate the correlation matrix and exclude variables from the models when appropriate.
- Create models to predict casual, registered and cnt.
- **Submit:**
  - A the summary for each model (i.e. summary(myModel))
  - A short 2-4 paragraph comparison of the models. You might explore questions like:
    - Is it easier to predict casual or registered?
    - Are certain variables more predictive for casual than for registered?

CNT Start Model

```
Call:
lm(formula = cnt ~ atemp + hum + windspeed + S1 + S2 + S3 + W1 + 
    W2 + W3 + WD0 + WD1 + WD2 + WD3 + WD4 + WD5 + Mnth1 + Mnth2 + 
    Mnth3 + Mnth4 + Mnth5 + Mnth6 + Mnth7 + Mnth8 + Mnth9 + Mnth10 + 
    Mnth11, data = BikeShareDay)

Residuals:
    Min      1Q  Median      3Q     Max 
-3587.5  -950.7  -236.5  1056.8  4028.1 

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2792.553    591.636   4.720 2.84e-06 ***
atemp        6731.908    705.502   9.542  < 2e-16 ***
hum         -3063.436    478.549  -6.402 2.81e-10 ***
windspeed   -3150.756    671.345  -4.693 3.23e-06 ***
S1          -1556.840    298.815  -5.210 2.48e-07 ***
S2           -622.524    350.886  -1.774   0.0765 .  
S3           -775.764    316.439  -2.452   0.0145 *  
W1           1909.799    326.272   5.853 7.38e-09 ***
W2           1669.802    305.265   5.470 6.25e-08 ***
W3                 NA         NA      NA       NA    
WD0          -411.068    176.513  -2.329   0.0202 *  
WD1          -303.722    176.365  -1.722   0.0855 .  
WD2          -141.313    176.975  -0.798   0.4249    
WD3           -46.367    177.156  -0.262   0.7936    
WD4          -108.008    177.026  -0.610   0.5420    
WD5           -30.253    176.961  -0.171   0.8643    
Mnth1         144.531    301.671   0.479   0.6320    
Mnth2         170.368    303.559   0.561   0.5748    
Mnth3         444.160    304.346   1.459   0.1449    
Mnth4         167.176    398.542   0.419   0.6750    
Mnth5         398.733    418.584   0.953   0.3411    
Mnth6          -6.996    418.845  -0.017   0.9867    
Mnth7        -523.312    447.625  -1.169   0.2428    
Mnth8          69.371    426.578   0.163   0.8709    
Mnth9         785.013    352.880   2.225   0.0264 *  
Mnth10        401.447    268.487   1.495   0.1353    
Mnth11       -135.126    255.744  -0.528   0.5974    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1274 on 705 degrees of freedom
Multiple R-squared:  0.582,	Adjusted R-squared:  0.5672 
F-statistic: 39.27 on 25 and 705 DF,  p-value: < 2.2e-16
```

CNT Final Model

```
Call:
lm(formula = cnt ~ atemp + hum + windspeed + S1 + S2 + S3 + W1 +
    W2, data = BikeShareDay)

Residuals:
    Min      1Q  Median      3Q     Max
-3919.2  -937.3  -247.1  1082.8  4167.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2579.8      563.1   4.581 5.45e-06 ***
atemp         6645.8      527.6  12.597  < 2e-16 ***
hum          -2670.9      464.1  -5.754 1.29e-08 ***
windspeed    -2996.6      678.7  -4.415 1.16e-05 ***
S1           -1501.3      153.5  -9.782  < 2e-16 ***
S2            -522.9      151.2  -3.459 0.000573 ***
S3            -864.0      186.4  -4.636 4.21e-06 ***
W1            1872.0      328.5   5.698 1.77e-08 ***
W2            1656.8      308.3   5.375 1.04e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1303 on 722 degrees of freedom
Multiple R-squared:  0.5523,    Adjusted R-squared:  0.5473
F-statistic: 111.3 on 8 and 722 DF,  p-value: < 2.2e-16
```

Casual Start Model

```
Call:
lm(formula = casual ~ atemp + hum + windspeed + S1 + S2 + S3 + 
    W1 + W2 + W3 + WD0 + WD1 + WD2 + WD3 + WD4 + WD5 + Mnth1 + 
    Mnth2 + Mnth3 + Mnth4 + Mnth5 + Mnth6 + Mnth7 + Mnth8 + Mnth9 + 
    Mnth10 + Mnth11, data = BikeShareDay)

Residuals:
     Min       1Q   Median       3Q      Max 
-1236.93  -230.60   -36.16   182.59  2068.66 

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   734.200    181.997   4.034 6.08e-05 ***
atemp        2032.786    217.024   9.367  < 2e-16 ***
hum          -758.883    147.210  -5.155 3.30e-07 ***
windspeed   -1073.742    206.517  -5.199 2.62e-07 ***
S1              5.481     91.921   0.060  0.95247    
S2            218.061    107.938   2.020  0.04374 *  
S3             53.330     97.342   0.548  0.58396    
W1            323.194    100.367   3.220  0.00134 ** 
W2            244.991     93.905   2.609  0.00927 ** 
W3                 NA         NA      NA       NA    
WD0          -147.460     54.298  -2.716  0.00677 ** 
WD1          -815.837     54.253 -15.038  < 2e-16 ***
WD2          -940.523     54.440 -17.276  < 2e-16 ***
WD3          -939.027     54.496 -17.231  < 2e-16 ***
WD4          -935.508     54.456 -17.179  < 2e-16 ***
WD5          -754.569     54.436 -13.862  < 2e-16 ***
Mnth1         -23.388     92.799  -0.252  0.80109    
Mnth2         -63.014     93.380  -0.675  0.50001    
Mnth3         186.063     93.622   1.987  0.04727 *  
Mnth4         177.975    122.598   1.452  0.14703    
Mnth5         200.810    128.763   1.560  0.11932    
Mnth6          16.324    128.844   0.127  0.89922    
Mnth7          -1.401    137.697  -0.010  0.99189    
Mnth8          91.257    131.222   0.695  0.48701    
Mnth9         300.980    108.552   2.773  0.00571 ** 
Mnth10        363.168     82.591   4.397 1.27e-05 ***
Mnth11        177.053     78.671   2.251  0.02472 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 392 on 705 degrees of freedom
Multiple R-squared:  0.6852,	Adjusted R-squared:  0.674 
F-statistic: 61.37 on 25 and 705 DF,  p-value: < 2.2e-16
```

Casual Final Model

```
Call:
lm(formula = casual ~ atemp + hum + windspeed + S1 + S2 + W1 + 
    W2 + WD0 + WD1 + WD2 + WD3 + WD4 + WD5, data = BikeShareDay)

Residuals:
     Min       1Q   Median       3Q      Max 
-1326.23  -230.48   -24.57   165.32  1963.25 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   830.42     177.37   4.682 3.40e-06 ***
atemp        1975.37     120.62  16.376  < 2e-16 ***
hum          -603.47     143.03  -4.219 2.77e-05 ***
windspeed    -948.13     210.30  -4.508 7.63e-06 ***
S1           -173.01      47.32  -3.656 0.000275 ***
S2            185.23      37.26   4.971 8.32e-07 ***
W1            320.19     102.31   3.130 0.001821 ** 
W2            236.75      96.11   2.463 0.013999 *  
WD0          -148.40      55.88  -2.656 0.008094 ** 
WD1          -817.64      55.81 -14.651  < 2e-16 ***
WD2          -941.91      55.94 -16.839  < 2e-16 ***
WD3          -941.84      56.01 -16.814  < 2e-16 ***
WD4          -931.88      55.92 -16.664  < 2e-16 ***
WD5          -752.00      56.03 -13.421  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 403.7 on 717 degrees of freedom
Multiple R-squared:  0.6605,	Adjusted R-squared:  0.6543 
F-statistic: 107.3 on 13 and 717 DF,  p-value: < 2.2e-16
```

Registered Start Model

```
Call:
lm(formula = registered ~ atemp + hum + windspeed + S1 + S2 +
    S3 + W1 + W2 + W3 + WD0 + WD1 + WD2 + WD3 + WD4 + WD5 + Mnth1 +
    Mnth2 + Mnth3 + Mnth4 + Mnth5 + Mnth6 + Mnth7 + Mnth8 + Mnth9 +
    Mnth10 + Mnth11, data = BikeShareDay)

Residuals:
    Min      1Q  Median      3Q     Max
-3362.5  -816.4  -208.5   910.6  2803.7

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2058.35     498.26   4.131 4.04e-05 ***
atemp        4699.12     594.15   7.909 1.00e-14 ***
hum         -2304.55     403.02  -5.718 1.59e-08 ***
windspeed   -2077.01     565.38  -3.674 0.000257 ***
S1          -1562.32     251.65  -6.208 9.14e-10 ***
S2           -840.58     295.50  -2.845 0.004576 **
S3           -829.09     266.49  -3.111 0.001939 **
W1           1586.60     274.77   5.774 1.16e-08 ***
W2           1424.81     257.08   5.542 4.22e-08 ***
W3                NA         NA      NA       NA
WD0          -263.61     148.65  -1.773 0.076609 .
WD1           512.12     148.53   3.448 0.000598 ***
WD2           799.21     149.04   5.362 1.11e-07 ***
WD3           892.66     149.19   5.983 3.48e-09 ***
WD4           827.50     149.09   5.551 4.03e-08 ***
WD5           724.32     149.03   4.860 1.45e-06 ***
Mnth1         167.92     254.06   0.661 0.508861
Mnth2         233.38     255.65   0.913 0.361603
Mnth3         258.10     256.31   1.007 0.314292
Mnth4         -10.80     335.64  -0.032 0.974341
Mnth5         197.92     352.52   0.561 0.574665
Mnth6         -23.32     352.74  -0.066 0.947308
Mnth7        -521.91     376.97  -1.384 0.166651
Mnth8         -21.89     359.25  -0.061 0.951439
Mnth9         484.03     297.18   1.629 0.103816
Mnth10         38.28     226.11   0.169 0.865615
Mnth11       -312.18     215.38  -1.449 0.147660
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1073 on 705 degrees of freedom
Multiple R-squared:  0.543,     Adjusted R-squared:  0.5268
F-statistic: 33.51 on 25 and 705 DF,  p-value: < 2.2e-16
```

Registered Final Model

```
Call:
lm(formula = registered ~ atemp + hum + windspeed + S1 + S2 + 
    S3 + W1 + W2 + WD0 + WD1 + WD2 + WD3 + WD4 + WD5, data = BikeShareDay)

Residuals:
    Min      1Q  Median      3Q     Max 
-3610.4  -792.3  -207.3   930.5  2785.0 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1782.6      479.7   3.716 0.000218 ***
atemp         4495.9      442.0  10.171  < 2e-16 ***
hum          -1927.2      390.3  -4.937 9.86e-07 ***
windspeed    -2022.4      567.7  -3.562 0.000392 ***
S1           -1321.8      128.3 -10.301  < 2e-16 ***
S2            -670.1      126.4  -5.302 1.52e-07 ***
S3            -789.7      155.9  -5.066 5.18e-07 ***
W1            1641.8      276.4   5.940 4.45e-09 ***
W2            1476.7      259.4   5.693 1.83e-08 ***
WD0           -270.6      150.8  -1.795 0.073109 .  
WD1            505.1      150.6   3.353 0.000841 ***
WD2            795.7      151.0   5.270 1.81e-07 ***
WD3            890.8      151.2   5.893 5.85e-09 ***
WD4            843.1      151.0   5.585 3.32e-08 ***
WD5            723.4      151.2   4.785 2.08e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1089 on 716 degrees of freedom
Multiple R-squared:  0.522,	Adjusted R-squared:  0.5126 
F-statistic: 55.85 on 14 and 716 DF,  p-value: < 2.2e-16
```

Multiple Regression Write Up

The Bike Share data has several different response variables that were investigated.

CNT- the count of total rental bikes including both casual and registered.

Casual- count of casual users.

Registered- count of registered users.

The Bike Share data has several explanatory variables that were investigated.

atemp- Normalized feeling temperature in Celsius.

hum - Normalized humidity.

weathersit - The data on the weather at that time. Mixture of cloudiness and participation.

season - season(1:spring, 2:summer, 3:fall, 4 winter)

Mnth - for which month of the year it is.

Weekday - for which weekday of the month it is.

The variable season(S[N]) , weathersit(W[N]) , mnth[N], weekday(WD[N]) were broken down into dummy variables for the model. To start testing the models all variables were included. Then, variables that have non-significant T-test of at least .01 were removed to produce the final models. When comparing the models to each other certain inferences can be made. It seems to be easier to predict a casual users verses a registered one. The explanatory variables for the casual users model account for about 65% of the number of casual users. Whereas the explanatory variables for the registered users model account for about 51%. The most important explanatory variables for how many casual users of a given day are those dependent on the weather(atemp,hum,windspeed) . Also, it seems that more users come on the weekend with a few more users on Saturday then Sunday.

Task 2: Principle Component Analysis

Example: European Protein Consumption

In this task, we will look at the protein consumption of 25 European countries. Some sample code is given below to get us started.

- cor(Protein[,-1])
- pca<- prcomp(Protein[,-1], scale=TRUE)
- pca
- plot(pca)
- summary(pca)
- pred<- predict(pca)
- pred
- plot(pred)
- plot(pred[,1:2])
- text(x=pred[,1], y=pred[,2], labels=Protein$Country)

Assignment 2: Rain Forest

Read RainForestReadMe.txt and familiarize yourself with the data. Perform a PCA analysis on the data:

- Look at the correlations
- Compute the PCA model and evaluate the scree plot
- Compute the new components
- Create plots to show the interactions between components
- **Submit:**
    - The correlation matrix
    - The transformation matrix
    - The new components
    - The scree plot
    - The component plots of with appropriate labels:
        - PCA1 vs. PCA2
        - PCA1 vs. PCA3
        - PCA2 vs PCA3
    - 2-4 paragraphs drawing conclusions of the results.

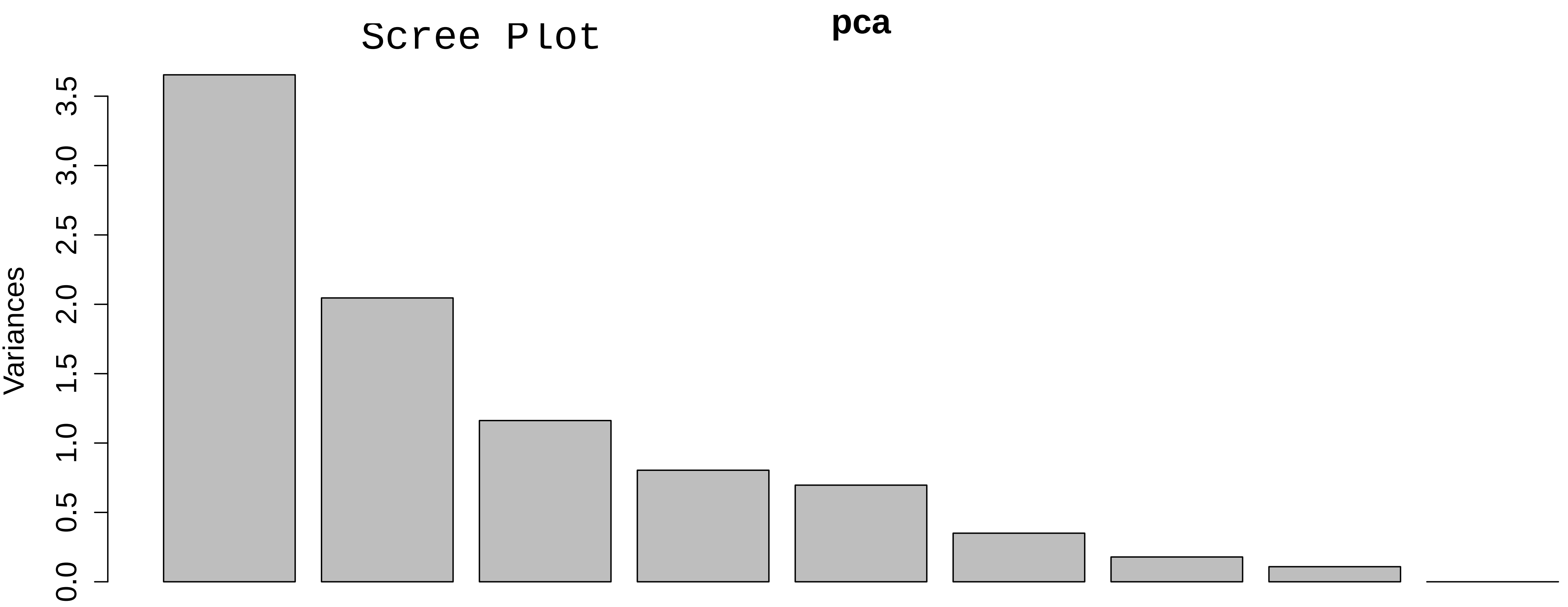
Correlation Matrix

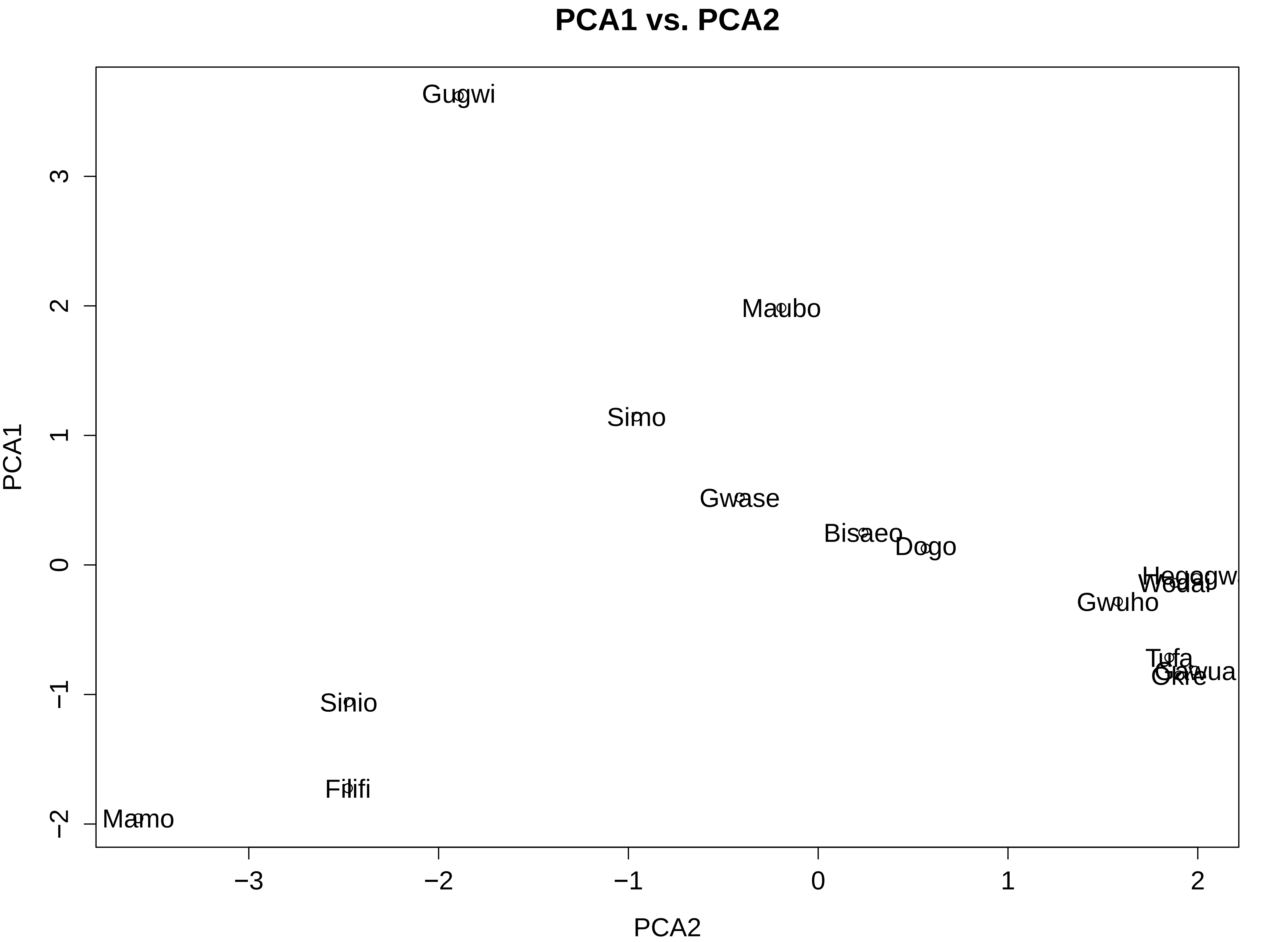
	Age	Nights	Pig	Cassowary	Fish.Spear
Age	1.00000000	-0.07844842	0.0797523	0.2418343	0.23310189
Nights	-0.07844842	1.00000000	0.2321227	0.1724949	0.37063911
Pig	0.07975230	0.23212274	1.0000000	0.5522945	0.45669415
Cassowary	0.24183426	0.17249494	0.5522945	1.0000000	0.21211009
Fish.Spear	0.23310189	0.37063911	0.4566941	0.2121101	1.00000000
Fish.Hook	0.09533439	0.08336283	-0.2111626	-0.1999805	0.07403181
Fish.Other	0.36966521	0.43743435	-0.1446823	-0.1719287	0.55445240
Other_Vertebrates	0.54050744	0.44780119	0.4775304	0.5631174	0.42488567
Total	0.21177404	0.32134246	0.9518226	0.7017020	0.59114815

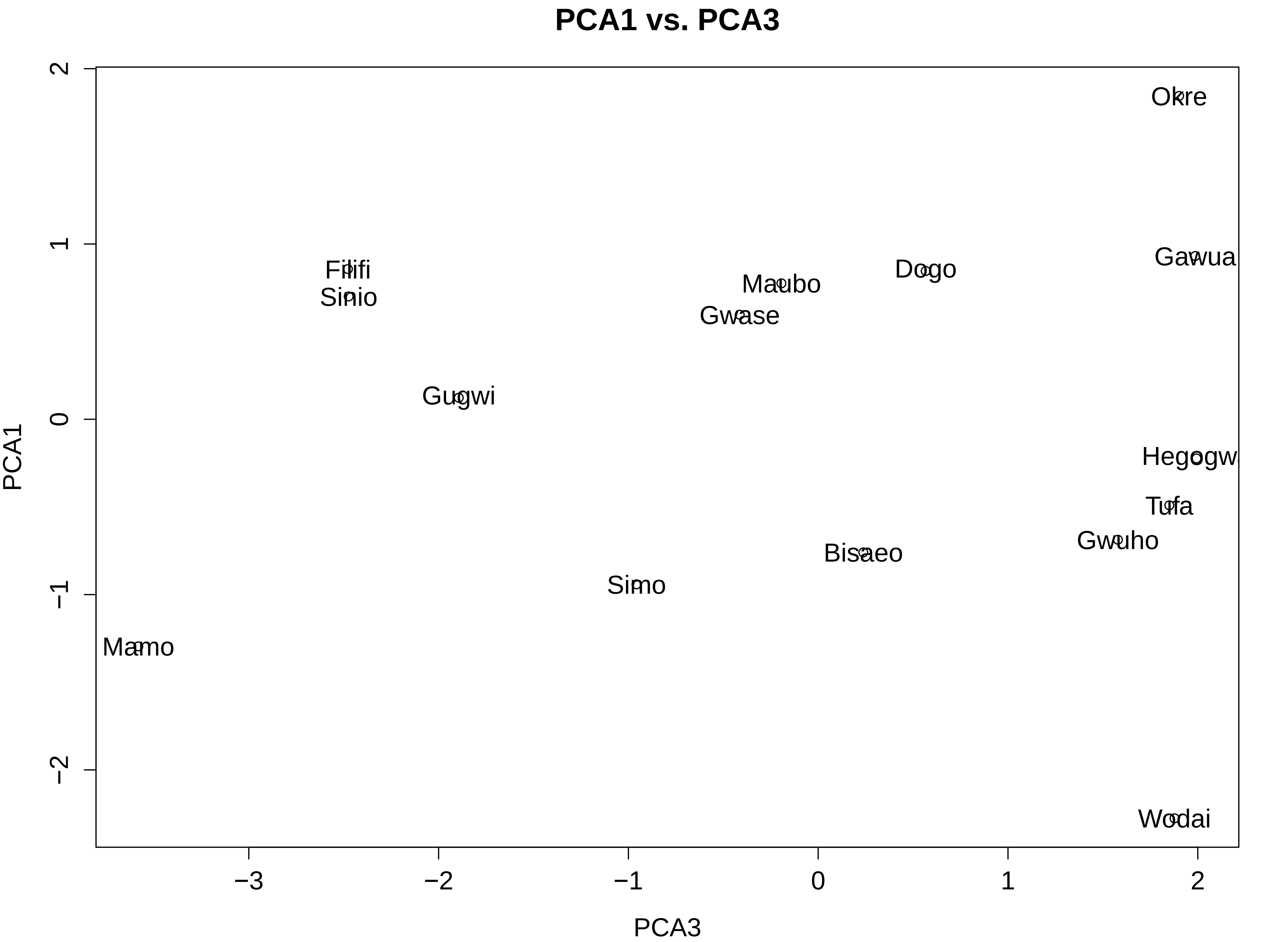
	Fish.Hook	Fish.Other	Other_Vertebrates	Total
Age	0.09533439	0.36966521	0.5405074	0.21177404
Nights	0.08336283	0.43743435	0.4478012	0.32134246
Pig	-0.21116257	-0.14468230	0.4775304	0.95182262
Cassowary	-0.19998054	-0.17192868	0.5631174	0.70170204
Fish.Spear	0.07403181	0.55445240	0.4248857	0.59114815
Fish.Hook	1.00000000	0.43267538	0.2477290	-0.08974128
Fish.Other	0.43267538	1.00000000	0.3113512	0.02112371
Other_Vertebrates	0.24772899	0.31135117	1.0000000	0.65371295
Total	-0.08974128	0.02112371	0.6537130	1.00000000

Transformation Matrix

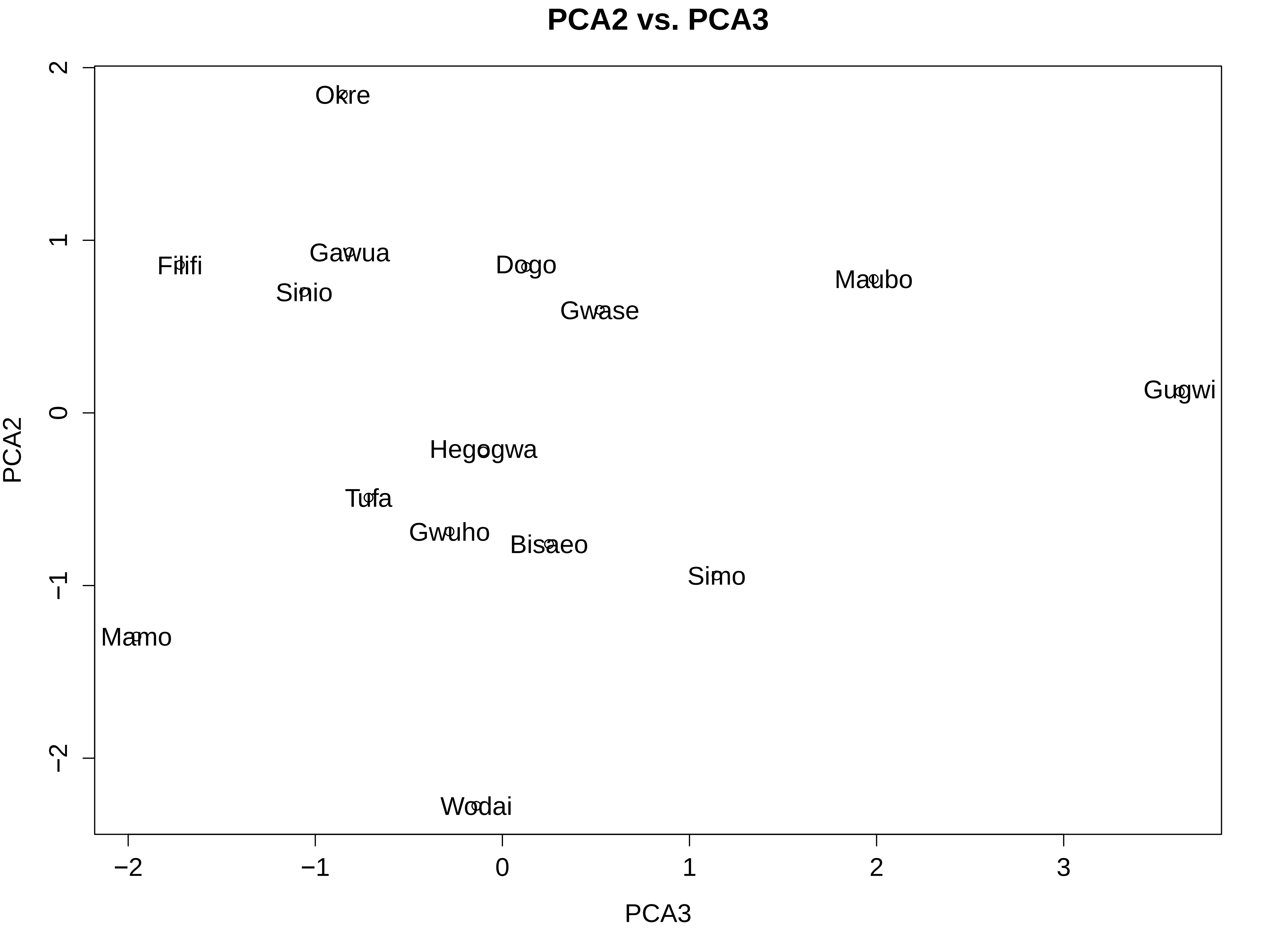
```
Standard deviations:
[1] 1.911532e+00 1.430325e+00 1.077958e+00 8.967105e-01 8.345574e-01
[6] 5.918372e-01 4.226686e-01 3.297368e-01 8.419756e-17

Rotation:
                          PC1        PC2         PC3         PC4         PC5
Age                -0.22048636  0.2175623 -0.70781103  0.28842735 -0.19767643
Nights             -0.26354978  0.2172908  0.54859195 -0.23212652 -0.56721368
Pig                -0.40886729 -0.3241692  0.13726364  0.03535543  0.32500030
Cassowary          -0.35988512 -0.3078468 -0.20790793 -0.23926950 -0.24588486
Fish.Spear         -0.36723983  0.2084725  0.24189571  0.49745997  0.30596673
Fish.Hook          -0.01798598  0.4777989 -0.09538552 -0.62473777  0.53033551
Fish.Other         -0.16336626  0.6069642  0.09251259  0.26145148 -0.07166637
Other_Vertebrates -0.43755685  0.1341179 -0.23585691 -0.31310648 -0.18176360
Total              -0.48156453 -0.2178307  0.06468703 -0.01645598  0.24507682
                          PC6         PC7         PC8           PC9
Age                 0.26850626 -0.13758483 -0.43561960  0.000000e+00
Nights              0.20130822 -0.05676739 -0.40374682 -2.443827e-16
Pig                 0.41199851 -0.31703792  0.05202315  5.736402e-01
Cassowary          -0.72945430 -0.20704137 -0.06485041  1.886870e-01
Fish.Spear         -0.32482174  0.48211384 -0.26499932  1.179448e-01
Fish.Hook          -0.07805349 -0.08093265 -0.27413231  5.424287e-02
Fish.Other         -0.16053409 -0.49725521  0.49923839  2.411327e-02
Other_Vertebrates  0.20333274  0.55981219  0.49536979  6.106655e-02
Total               0.08255986 -0.18663444  0.01757063 -7.836902e-01
```


Scree Plot
pca
Variances
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5


PCA1 vs. PCA2
PCA1
PCA2
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
Gugwi
Maubo
Simo
Gwase
Bisaeo
Dogo
Gwuho
Gawua
Sinio
Filifi
Mamo



PCA1 vs. PCA3
PCA1
PCA3
2
1
0
−1
−2
−3
−2
−1
0
1
2
Okre
Gawua
Fifi
Sinio
Maubo
Dogo
Gwase
Gugwi
Hegogw
Tufa
Gwuho
Bisaeo
Simo
Mamo
Wodai


PCA2 vs. PCA3
Okre
Gawua
Fiifi
Dogo
Maubo
Sinio
Gwase
Gugwi
Hegogwa
Tufa
Gwuho
Bisaeo
Simo
Mamo
Wodai
PCA2
PCA3
2
1
0
−1
−2
−2
−1
0
1
2
3

Principle Component Analysis (PCA)

It is often difficult to interpret the PCA model since the human brain does not deal with visualizing many different dimensions at once. What is important to notice is that PCA1 accounts for the largest amount of variance in the model with PCA2 accounting for about half of PCA1. Those two together account for more than all of the others together. So the most important chart to gather information on how the groups relate to each other.

Looking at the plot of PCA1 vs. PCA2 you can infer that Gugwi is the greatest out liner. Also these group of six are really clustered close to each other: Heaogwa, Wodai, Gwuho, Tufa, Gawua, and Okre. This cluster means that they have a lot in common with each other.

Task 3: Linear Discriminant Analysis

Example: Iris

For our classwork, we will look at the iris dataset. Note that this is a supervised learning approach. We will begin by creating a training and test partitions of the data. We will also need to install the package MASS.

- install.packages("MASS")
- require(MASS)
- set.seed(1234)
- ind<- sample(2, nrow(iris), replace=TRUE, prob=c(0.7, 0.3))
- ind
- trainData<- iris[ind==1,]
- testData<- iris[ind==2,]
- plot(iris[1:4], col=iris[,5])
- ldaModel<- lda(V5 ~ V1 + V2 + V3 + V4, data = trainData)
- ldaModel
- predictions <- predict(ldaModel, testData)
- ld<- predictions$x
- class <- predictions$class
- confusionMatrix<- table(testData$V5, class, dnn=c("V5", "pred"))
- ldaModel$scaling
- transformed <- as.matrix(trainData[1:4])%*%ldaModel$scaling
- plot(transformed, col=trainData[,5])

Assignment 3: Thyroid

Read ThyroidReadMe.txt and familiarize yourself with the data. Create an LDA model to predict the class attribute (1 = normal, 2 = hyper, 3 = hypo)

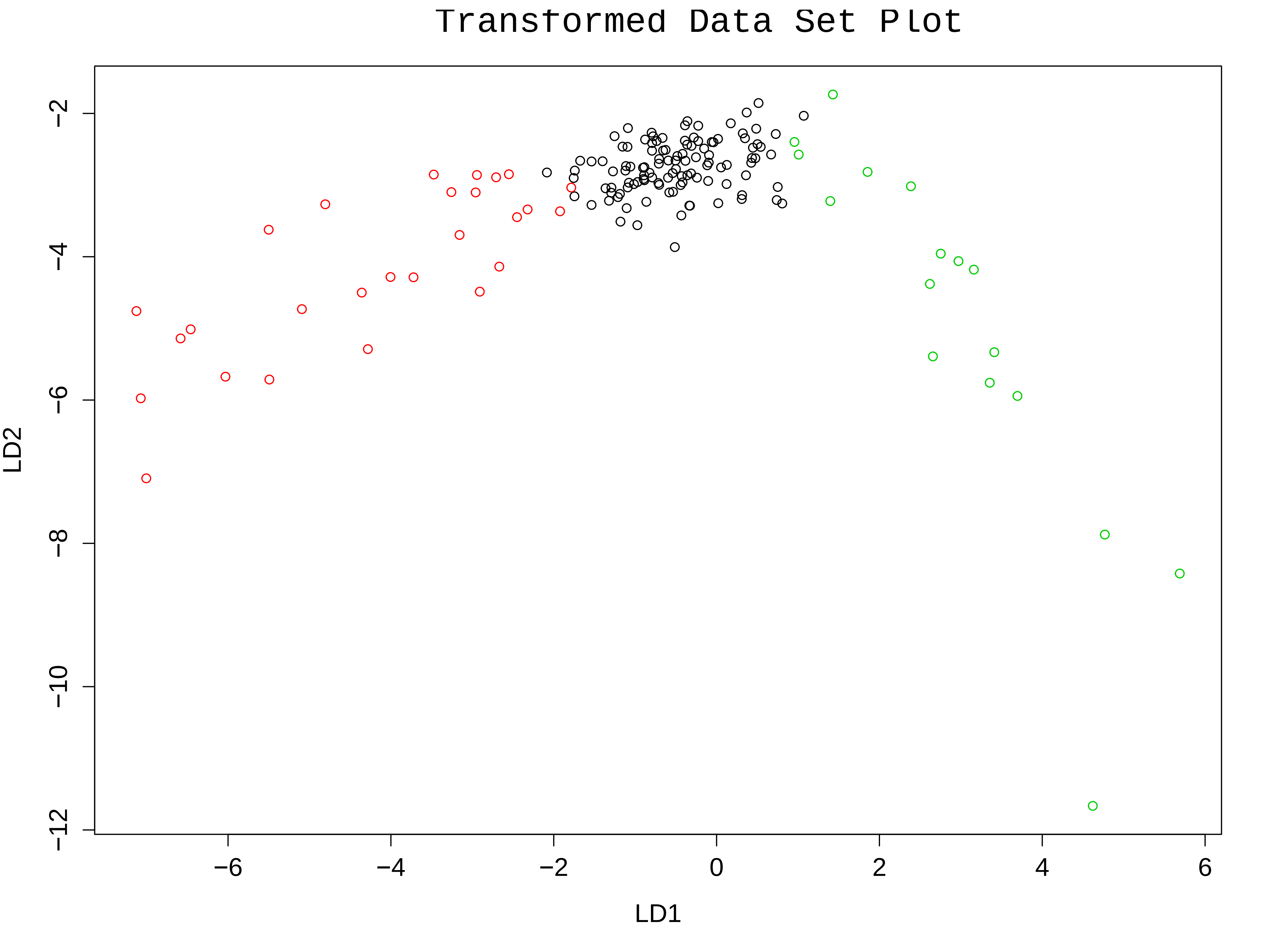
- Divide the data into training and testing sets.
- Create an lda model with the training data.
- Use the model to make predictions on the testing data
- **Submit:**
  - A description of the model (i.e. just type “myModel “and copy the priors, means, etc.)
  - A confusion matrix with appropriate labels
  - A plot showing the transformed data set, with appropriate labels and colored accordingly.
  - Include 2-4 paragraphs describing the results of the classifier. You might discuss:
    - How well the model performs.
    - It is strong/weak for particular class?
    - Which variables have the strongest influence?

Model Description

```
Call:
lda(V1 ~ V2 + V3 + V4 + V5 + V6, data = trainData)

Prior probabilities of groups:
        1         2         3 
0.7179487 0.1730769 0.1089744 

Group means:
         V2        V3       V4         V5          V6
1 110.48214  9.284821 1.763393  1.3000000  2.48392857
2  97.18519 17.888889 4.248148  0.9518519 -0.04444444
3 122.47059  3.600000 1.105882 10.2823529 12.52941176

Coefficients of linear discriminants:
           LD1          LD2
V2  0.02300923 -0.006545691
V3 -0.32894647 -0.076216235
V4 -0.13034640 -0.463199461
V5  0.03549785 -0.214009358
V6  0.08077231 -0.086010847

Proportion of trace:
   LD1    LD2 
0.8502 0.1498 
```

Confusion Matrix

```
> confusionMatrix
   pred
V1   1  2  3
  1 38  0  0
  2  2  6  0
  3  1  0 12
```


Transformed Data Set Plot
LD2
LD1
−2
−4
−6
−8
−10
−12
−6
−4
−2
0
2
4
6

Linear Discriminant Analysis(LDA)

The LDA model was used with the Thyroid data for classification predictions. The data has three different class that were trying to be predicted: normal, hyper and hypo. The data was not evenly distributed between the three classes. Normal class had disproportionately more than the other two classes. There was 5 input variables to help determine the classification. The fourth variable "Total Serum thyroxin as measured by the isotopic displacement method " seemed to be the greatest influence in the model determining the predicted class.

Looking at the confusion matrix to check out the performance the model itself appears to be doing very well on the testing data. The model predicts 56 out of the 59. When the class was normal the prediction was correct 38 out of 38. When the class was hyper it was correct only 6 out of 8. With the 2 miss classification predicting normal case. When the class was hypo the prediction was correct 12 out of 13. With the one miss classification belonging to being classified as normal.

Task 4: Multi-dimensional Scaling

Example: eurodist

In this example, we will download a dataset directly from CRAN.The dataset contains the pairwise distances between several European countries. In the end, we will have a plot showing the relative position of the cities in our new dimensions.

- data(eurodist)
- eurodist
- euro.mds<- cmdscale(eurodist)
- euro.mds
- eur.mds<- cmdscale(eurodist, eig = TRUE)
- eur.mds
- Dim1 <- euro.mds [,1]
- Dim2 <- euro.mds [,2]
- plot(Dim1, Dim2, type="n", xlab="", ylab="", main="cmdscale(eurodist)")
- segments(-1500, -0, 1500, 0, lty="dotted")
- segments(0, -1500, 0, 1500, lty="dotted")
- text(Dim1, Dim2, rownames(euro.mds), cex=0.8)

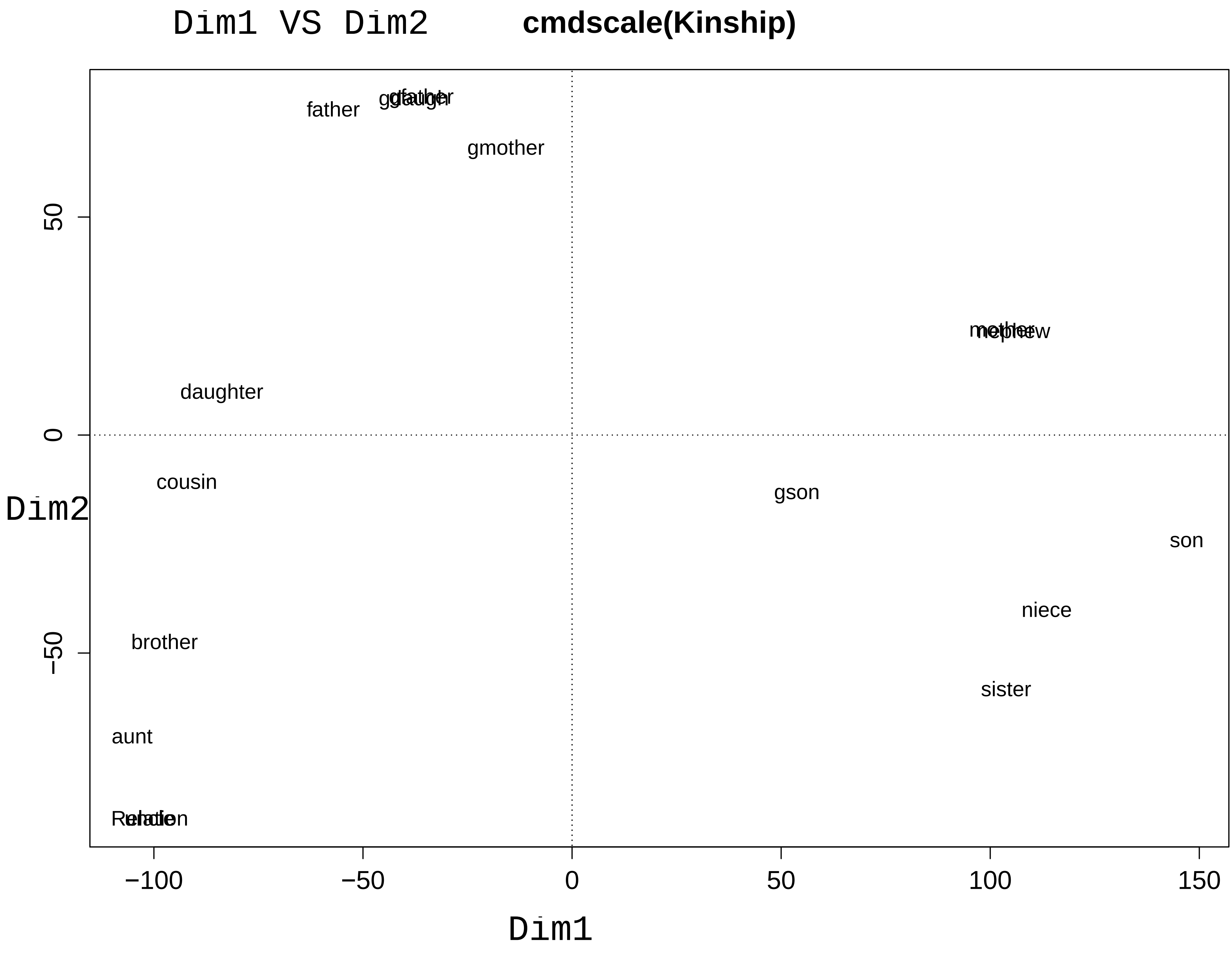
Assignment 4: Kinship

Rosenberg and Kim (Rosenberg and Kim, 1975) set out to analyze 15 kinship terms (aunt, brother, cousin, daughter, father, granddaughter, grandfather, grandmother, grandson, mother, nephew, niece, sister, son, uncle). They asked four groups of college students (two female, two male) to sort these terms on the basis of similarities. Two groups (one female, one male) were asked to sort twice, with the second sorting based on a different criterion from the first sort. Thus, a total of six "sources" were obtained.

- Perform a MDS analysis on the Kinship data.
- Note that you will need to create a "distance matrix".
    - This can be done using the command kindist<- dist(Kinship)
- Create a two-dimensional plot showing dimensions 1 and 2. Label the points
    - You may want to use colnames rather than rownames
- **Submit:**
    - The model
    - The goodness of fit measure
    - The plot showing Dim1 versus Dim, in which each point is labeled.
    - A 2-4 paragraph interpretation of the plot.

Model and Goodness of Fit

```
$points
             [,1]      [,2]
 [1,] -101.01599 -87.86581
 [2,] -105.21585 -69.14102
 [3,]  -97.45734 -47.38217
 [4,]  -92.13458 -10.64632
 [5,]  -83.76511   9.52581
 [6,]  -57.10127  74.78874
 [7,]  -37.86091  76.83293
 [8,]  -36.06059  77.22107
 [9,]  -15.82272  65.50095
[10,]   53.76065 -13.92503
[11,]  102.79302  24.27575
[12,]  105.58873  23.47661
[13,]  113.51569 -40.01442
[14,]  103.79358 -58.19943
[15,]  146.98268 -24.44765

$eig
 [1] 1.223911e+05 4.362055e+04 2.434949e+04 1.529304e+04 9.991875e+03
 [6] 4.071250e+03 2.838316e+03 2.798139e+03 2.281604e+03 1.034232e+03
[11] 7.372577e+02 2.376060e+02 7.309003e+01 1.789542e+01 2.773018e-12

$x
NULL

$ac
[1] 0

$GOF
[1] 0.7226209 0.7226209
```



Dim1 VS Dim2
cmdscale(Kinship)
father
gdaugh
gfather
gmother
mother
nephew
daughter
cousin
gson
son
niece
brother
sister
aunt
Relation
uncle
50
0
−50
Dim2
−100
−50
0
50
100
150
Dim1

Multi-Dimensional Scaling(MDS)

With MDS you can look at the plots of kinship and see which relationships are similar to each other by the distance between the each relationships on the plot. Looking at the plot a few relationships seems to be very close to each other when plotting the first dimension versus the second dimension. Father and grandfather are plotted very close to each other showing a similarly between the two. Also, grandfather and grandmother seem to have a lot in common with each other on the two dimensions. The next two closes pair seems to be the niece and the sister. An interesting point is while sister is to niece as brother is to nephew. The distance between brother and nephew is very far but not the distance between sister and niece. This fact seems counterintuitive.