Homework Assignment 4

1. (30 pts) Ex. 4.1 Consider 51 objects $O_1, \ldots, O_{51}$ assumed to be arranged along a straight line with the jth object being located at a point with coordinate j. Define the similarity s_{ij} between object i and object j as

$$s_{ij} = \begin{cases} 9 & \text{if } i = j \\ 8 & \text{if } 1 \le |i-j| \le 3 \\ 7 & \text{if } 4 \le |i-j| \le 6 \\ \quad \cdots & \\ 1 & \text{if } 22 \le |i-j| \le 24 \\ 0 & \text{if } |i-j| \ge 25. \end{cases}$$

Convert these similarities into dissimilarities δ_{ij} by using $\delta_{ij} = \sqrt{s_{ii} + s_{jj} - 2s_{ij}}$, and then apply classical multidimensional scaling to the resulting dissimilarity matrix. Explain the shape of the derived two-dimensional solution.

a. (10 pts) Define the similarity s_{ij} between object i and object j

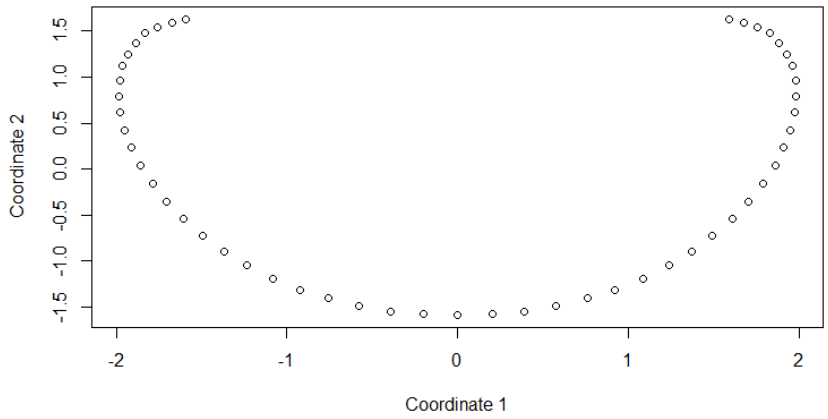
```
n<-51
s<- matrix(NA,nrow=n,ncol=n);s
for(i in 1:n){
 for(j in 1:n){
  if (i==j){
   s[i,j]<-9
  } else if (abs(i-j)>=25) {
   s[i,j]<-0
  } else
   for(k in 1:8){
    if (3*k-2<=abs(i-j) & abs(i-j)<=3*k) {
     s[i,j]=9-k
     break
    }
   }
 }
}
```

b. (8 pts) Convert these similarities into dissimilarities δ_{ij} by using $\delta_{ij} = \sqrt{s_{ii} + s_{jj} - 2s_{ij}}$

```
delta<-matrix(NA, nrow=n, ncol=n)
for(i in 1:n){
 for(j in 1:n){
  delta[i,j]=sqrt(s[i,i]+s[j,j]-2*s[i,j])
 }
}
```

c. (12 pts) Apply classical multidimensional scaling to the resulting dissimilarity matrix. Explain the shape of the derived two-dimensional solution.

```
delta_mds <- cmdscale(delta, k = 5, eig = TRUE)
delta_mds$eig
> (cumsum(abs(delta_mds$eig))/sum(abs(delta_mds$eig)))
 [1] 0.4428488 0.6744323 0.7382541 0.7657231 0.7924526 0.8183668 0.
> (cumsum((delta_mds$eig)^2)/sum((delta_mds$eig)^2))
 [1] 0.7608520 0.9689196 0.9847222 0.9876495 0.9904214 0.9930267 0.
 x <- delta_mds$points[,1]
y <- delta_mds$points[,2]
plot(x,y, xlab = "Coordinate 1", ylab = "Coordinate 2")
```



Two dimensional map of data cover 67% for $P^{(1)}$ AND 96.8% for $P^{(2)}$. The map is a heart shape.

2. (20 pts) The Table in the below summarizes data collected during a survey in which subjects were asked to compare a set of eight legal offenses, and to say for each one how unlike it was, in terms of seriousness, from the others. Each entry in the table shows the percentage of respondents who judged that the two offenses are very dissimilar. Find a two-dimensional scaling solution and try to interpret the dimensions underlying the subjects' judgements.

   Dissimilarity Matrix for a Set of Eight Legal Offenses organized in table

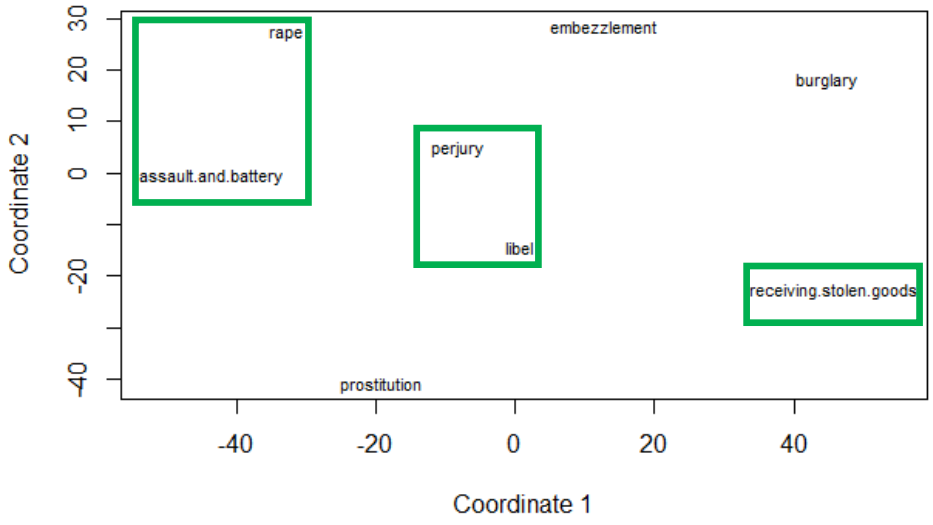
Offense	1	2	3	4	5	6	7	8
1	0							
2	21.1	0						
3	71.2	54.1	0					
4	36.4	36.4	36.4	0				
5	52.1	54.1	52.1	**0.7**	0			
6	89.9	75.2	36.4	54.1	53	0		
7	53	73	75.2	52.1	36.4	88.3	0	
8	90.1	93.2	71.2	63.4	52.1	36.4	73	0

   Offenses: (1) assault and battery, (2) rape, (3) embezzlement, (4) perjury, (5) libel, (6) burglary, (7) prostitution, (8) receiving stolen goods.

```
[1]  7.661446e+03  4.295776e+03  1.250223e+03  4.550082e+02  4.204946e+01  7.95
[8] -4.687390e+02
> (cumsum(abs(offenses_mds$eig))/sum(abs(offenses_mds$eig)))
[1] 0.5247079 0.8189118 0.9045356 0.9356976 0.9385775 0.9385775 0.9678976 1.000
> (cumsum((offenses_mds$eig)^2)/sum((offenses_mds$eig)^2))
[1] 0.7399532 0.9725834 0.9922876 0.9948975 0.9949198 0.9949198 0.9972302 1.000
> x <- offenses_mds$points[,1]
> y <- offenses_mds$points[,2]
> plot(x, y, xlab = "Coordinate 1", ylab = "Coordinate 2", xlim = range(x)*1.2,
> text(x, y, labels = colnames(offenses), cex = 0.7)
```

The two-dimension map covers at least 80% of variation of data structure, and the first characteristic covers at least 50% of variations of the data.

We can see like perjury and libel are judged as similar, because they have the similar values for the first coordinate and a slight different values for the second coordinator. The violent crimes like assault and battery and rape are judged as similar. The crime of receiving stolen goods appears to be quite different than the other crimes.



3. (20 pts) Ex. 4.3 In the data *garden flowers* data in Table 4.7 (from Kaufman and Rousseeuw 1990), the dissimilarity matrix of 18 species of garden flowers is shown. Use some form of multidimensional scaling to investigate which species share common properties.

```
gardenflowers<-read.csv("D:/STAT 4400/Data/gardenflowers.csv",head=TRUE,row.names=1)
 (garden_mds<-cmdscale(dist,k=nrow(gardenflower)-1,eig=T))
(lam<-garden_mds$eig)
(pm1<-cumsum(abs(lam))/sum(abs(lam)))
(pm2<-cumsum(lam^2)/sum(lam^2))
(df<-as.data.frame(garden_mds$points)
x<-df$V1
y<-df$V2
plot(x,y)
text(x,y,labels=rownames(df),pos=3)
```

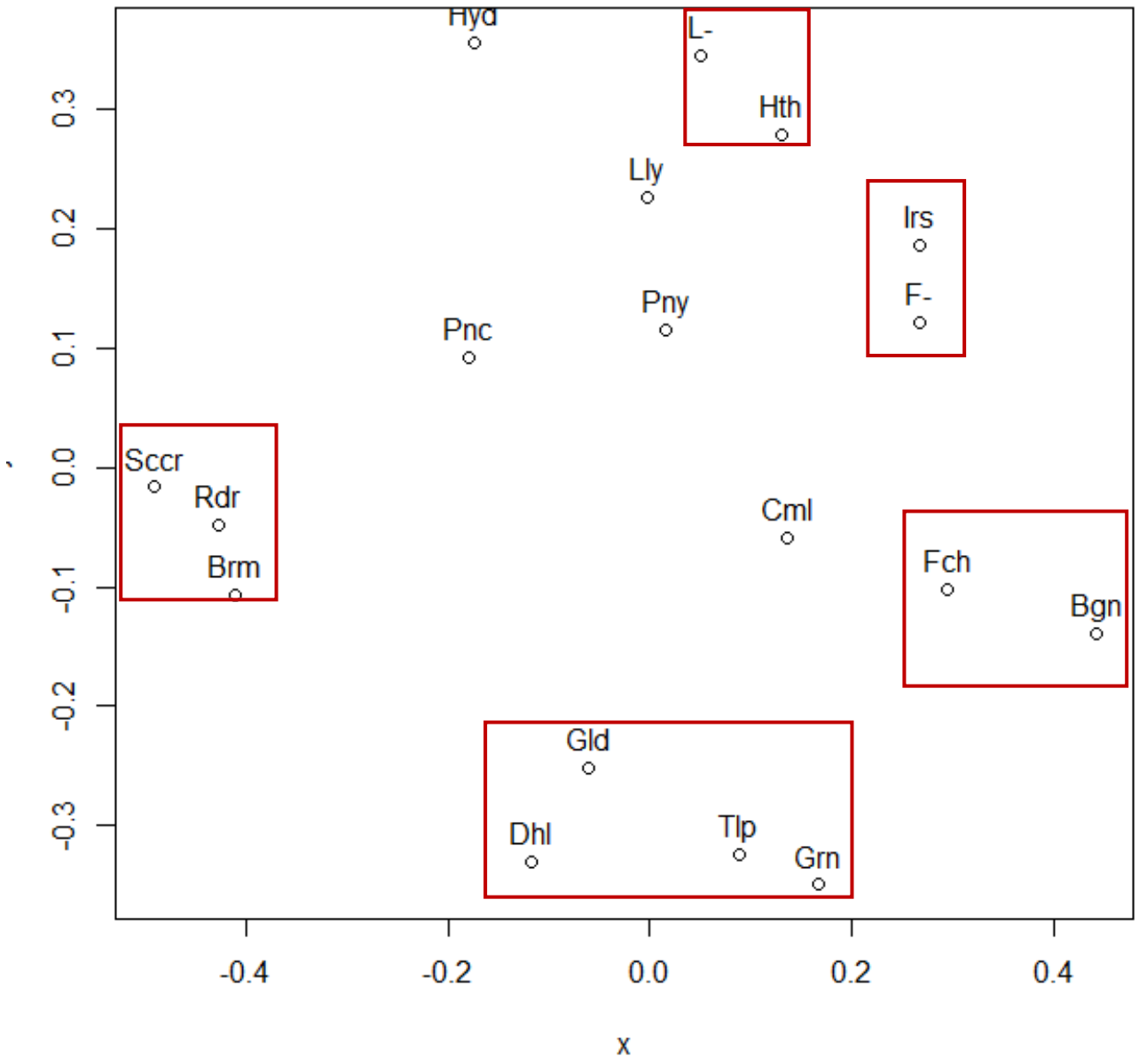
```
>(pm1<-cumsum(abs(lam))/sum(abs(lam)))

[1] 0.2549273 0.4493002 0.5738645 0.6791096 0.7364481 0.7863903 0.8046086 0.8190510 0.8255356
[10] 0.8255356 0.8302034 0.8391005 0.8485894 0.8678594 0.8905760 0.9183028 0.9555567 1.0000000

> (pm2<-cumsum(lam^2)/sum(lam^2))
 [1] 0.4611160 0.7291863 0.8392805 0.9178730 0.9412006 0.9588982 0.9612532 0.9627331 0.9630315
[10] 0.9630315 0.9631861 0.9637478 0.9643866 0.9670214 0.9706829 0.9761377 0.9859851 1.0000000
```

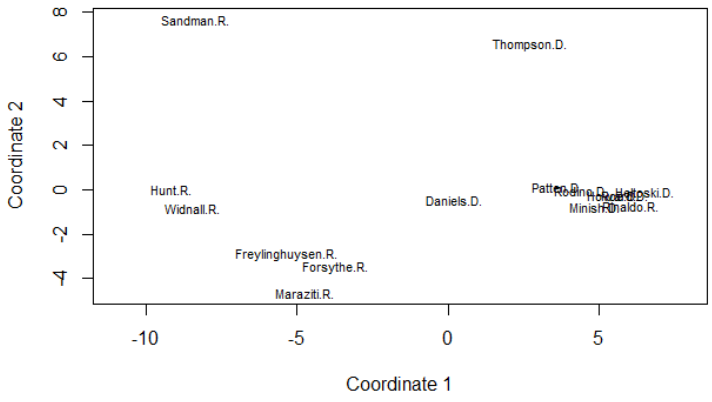
$P_m^{(2)}$ suggests that a three dimensional solution would be an adequate fit, but $P_m^{(1)}$ indicates that we may need seven dimensions. We may use two –dim map even though two shouldn't be enough dimensions – because it's easier to read.

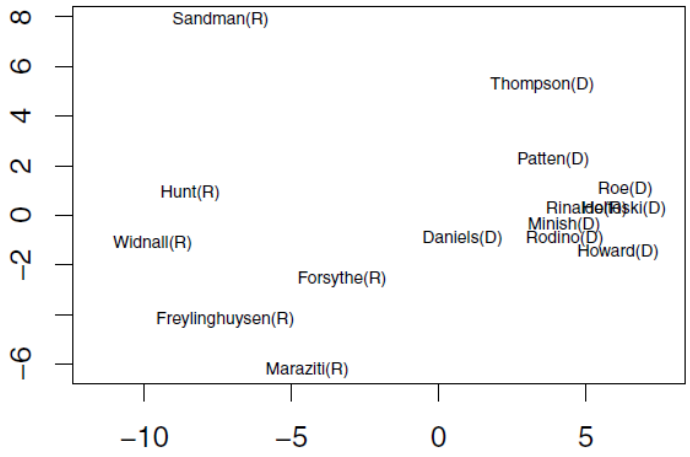
The flowers in the small boxes have the similar values of coordinates for both x and y. For instance, camellia and fuchsia appear to be similar; as well as scotch rose, red rose, and broom.


Hyd
L-
Hth
Lly
Irs
F-
Pny
Pnc
Sccr
Rdr
Brm
Cml
Fch
Bgn
Gld
Dhl
Tlp
Grn
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.2
0.0
0.2
0.4
x

4. (10 pts) Consider the *voting data* problem outlined in Section 4.5.1. Carry out a classical scaling of the data and show that the solution. Compare your solution to the nonmetric scaling solution given in Section 4.51.

```
> voting<-read.csv("D:/STAT 4400/Data/voting.csv", row.names = 1, header=TRUE)
> voting_mds <- cmdscale(voting, k = 7, eig = TRUE)
> voting_mds$eig
 [1]  4.977608e+02  1.461762e+02  1.029131e+02  7.687756e+01  5.511540e+01  2.474374e+01  8.
005009e+00
 [8]  6.171710e+00  2.358183e+00 -2.842171e-14 -2.026091e+00 -1.521409e+01 -1.869433e+01 -2.
040153e+01
[15] -3.398575e+01
> (cumsum(abs(voting_mds$eig))/sum(abs(voting_mds$eig)))
 [1] 0.4926161 0.6372815 0.7391310 0.8152140 0.8697597 0.8942477 0.9021700 0.9082779 0.91061
17 0.9106117
[11] 0.9126169 0.9276737 0.9461748 0.9663655 1.0000000
> (cumsum((voting_mds$eig)^2)/sum((voting_mds$eig)^2))
 [1] 0.8498269 0.9231165 0.9594436 0.9797152 0.9901344 0.9922344 0.9924542 0.9925849 0.99260
40 0.9926040
[11] 0.9926180 0.9934120 0.9946107 0.9960383 1.0000000
> x <- voting_mds$points[,1]
> y <- voting_mds$points[,2]
> plot(x, y, xlab = "Coordinate 1", ylab = "Coordinate 2", xlim = range(x)*1.2, type = "n")
> text(x, y, labels = colnames(voting), cex = 0.7)
```





Although no individual voter has moved very far, there are some distinct differences between the results from the two scaling methods. Classic scaling bunches the democrats much more tightly, making Thompson look like an outlier in the y-coordinate. Forsythe and Freylinghuysen have also moved closer together. In terms of similarities, Rinaldo is still the odd republican in the midst of the democrats, and most voters are close to where they were in the nonmetric solution.