HW3. for Multivariate Statistics ll

October 29, 2020

201611531/Department of Statistics/Jeong Hojae

Chapter 7. Multidimensional Scaling(MDS)

1. [Data 7.7.2] (color3.txt) is the result of an experiment that recognizes color with the human eye. For 91 cases with two color pairs each of 14 colors according to the wavelength of the color (434 - 674), 31 were scored on a scale of 0 (not quite alike) - 4 (identical) And the matrix data of 14 colors were made with the average score. For reference, you can look at several colors depending on the wavelength, such as 434 = Navy, 445 = Blue, 472 = Cyan, 504 = Green, 555 = Yellow, 600 = Yellow, 628 = Orange, 651 = Orange and 674 = Red.

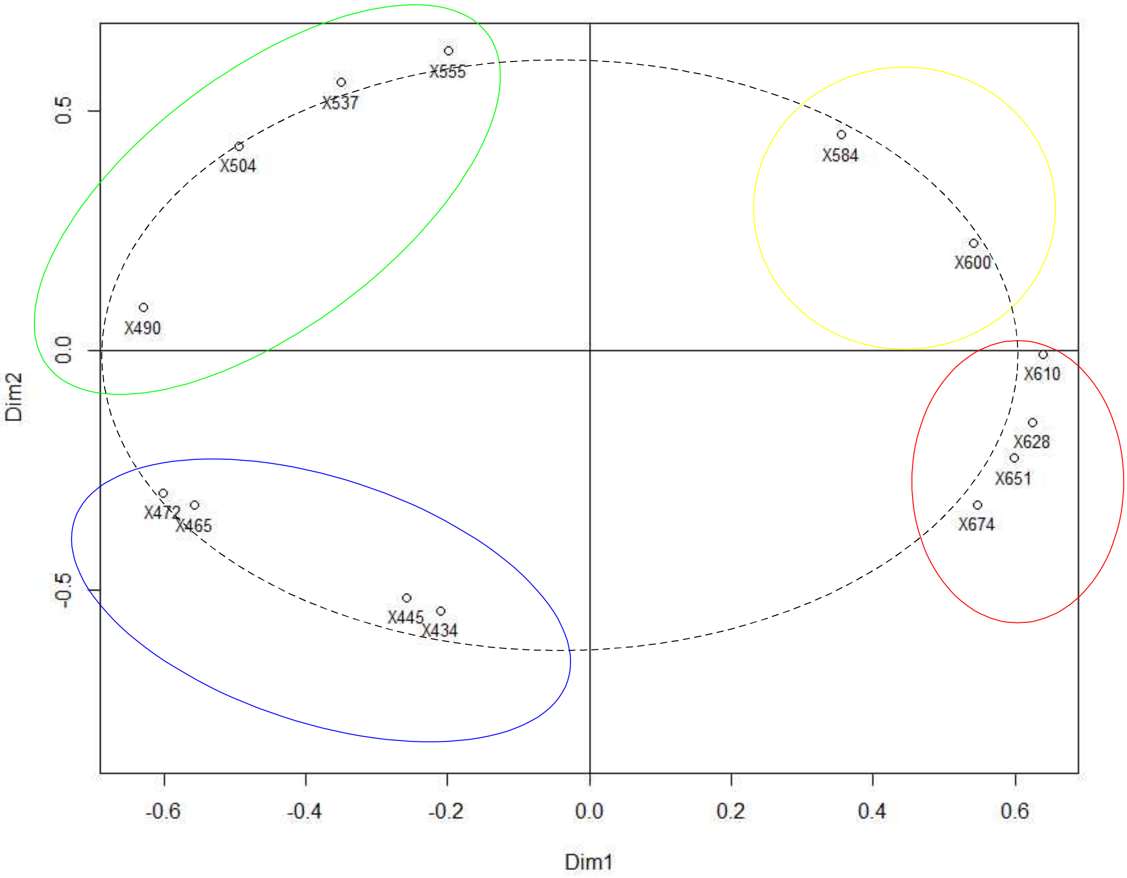
[Data 7.7.2] 14 color materials (color3.txt)

434	1													
445	0.86	1												
465	0.42	0.5	1											
472	0.42	0.44	0.81	1										
490	0.18	0.22	0.47	0.54	1									
504	0.06	0.09	0.17	0.25	0.61	1								
537	0.07	0.07	0.10	0.10	0.31	0.62	1							
555	0.04	0.07	0.08	0.09	0.26	0.45	0.73	1						
584	0.02	0.02	0.02	0.02	0.07	0.14	0.22	0.33	1					
600	0.07	0.04	0.01	0.01	0.02	0.08	0.14	0.19	0.58	1				
610	0.09	0.07	0.02	0.00	0.02	0.02	0.05	0.04	0.37	0.74	1			
628	0.12	0.11	0.01	0.01	0.01	0.02	0.02	0.03	0.27	0.5	0.76	1		
651	0.13	0.13	0.05	0.02	0.02	0.02	0.02	0.02	0.2	0.41	0.62	0.85	1	
674	0.16	0.14	0.03	0.04	0.00	0.01	0.00	0.02	0.23	0.28	0.55	0.68	0.76	1

(1) Explain whether [Data 7.7.2] is a similarity matrix or a dissimilarity matrix.

In general, dissimilarity represented by the distances between two objects r and s from the multivariate data matrix X. And similarity and dissimilarity are the relation of inverse proportionality. In clustering, dissimilarity is used for clustering observations, the measurements are used by some sort of distances. And similarity is used for clustering variables, they are usually grouped on the basis of correlation coefficients or like measurement of association. In general, the biggest difference between similarity and dissimilarity matrix is that the diagonal elements of the dissimilarity matrix are zero. Therefore, the above matrix is a similarity matrix.

(2) In the non-metric MDS, show that the MDS Map provides a color circle.



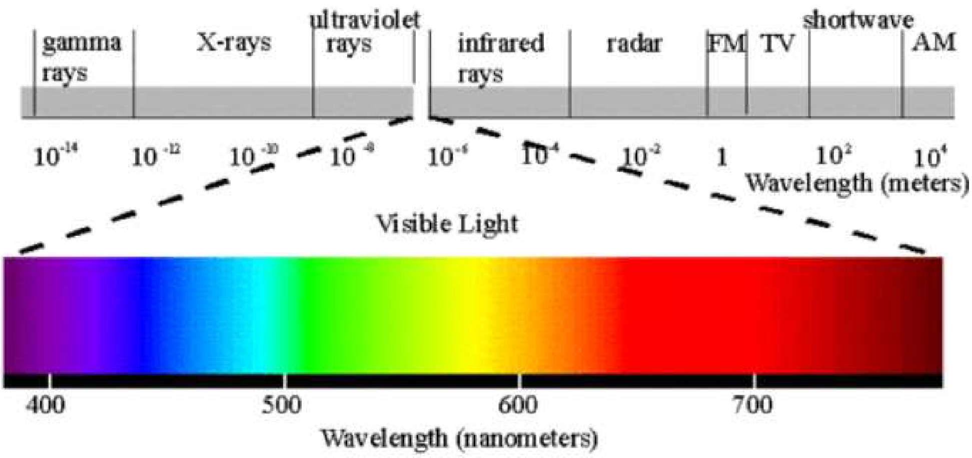


Table of Contents Visual Stimulus

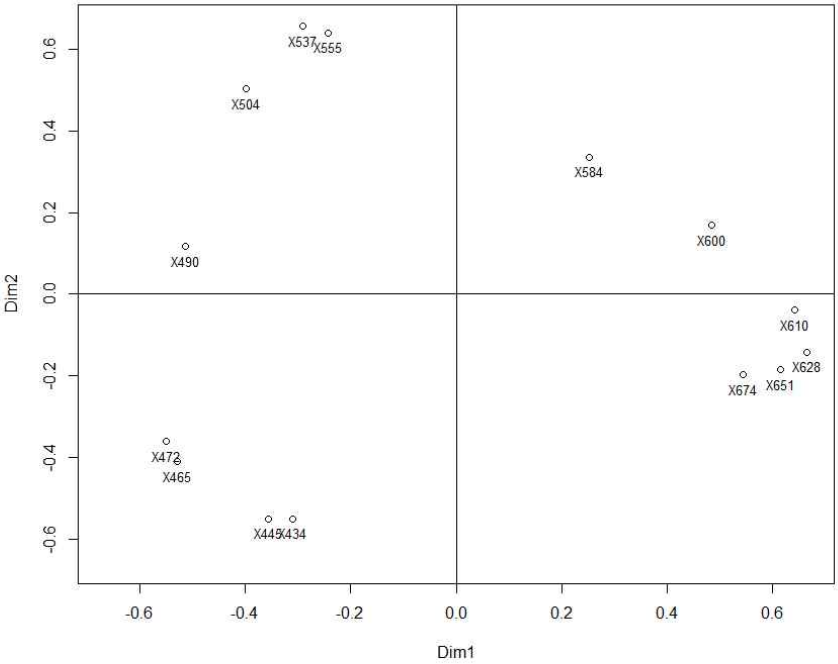
The MDS Map provides a color circle. Here, We can see that yellow group is in the first quadrant, green group is in the second quadrant, blue group is in the third quadrant, and red group is in the fourth quadrant.

(3) Compare the results of non-metric MDS and metric MDS.

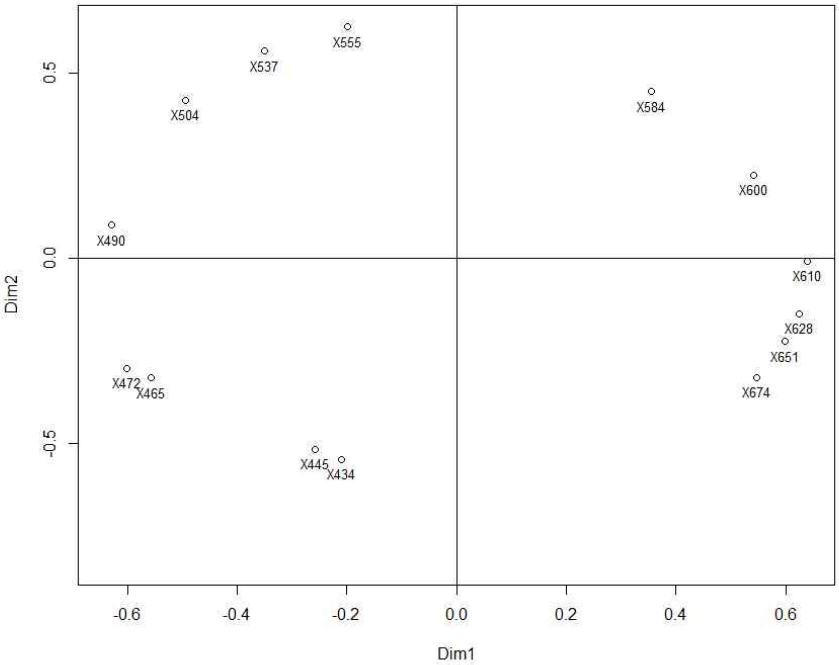
```
# non-metric MDS
> con
$points
            [,1]         [,2]
 [1,] -0.2105054 -0.543262570
 [2,] -0.2572935 -0.516221314
 [3,] -0.5573630 -0.323428434
 [4,] -0.6002278 -0.297529505
 [5,] -0.6291550  0.088391368
 [6,] -0.4933075  0.426426710
 [7,] -0.3502059  0.558884009
 [8,] -0.1992573  0.623464641
 [9,]  0.3553991  0.448696362
[10,]  0.5413890  0.224624229
[11,]  0.6381938 -0.008451818
[12,]  0.6248672 -0.150728044
[13,]  0.5993453 -0.224716023
[14,]  0.5477104 -0.321800498

$stress
[1] 2.92494
```

```
# metric MDS
> con
$points
            [,1]        [,2]
 [1,] -0.3101855 -0.55045611
 [2,] -0.3553803 -0.55037008
 [3,] -0.5279452 -0.41012571
 [4,] -0.5503163 -0.36052523
 [5,] -0.5128399  0.11725946
 [6,] -0.3988954  0.50303889
 [7,] -0.2905171  0.65580899
 [8,] -0.2426751  0.63996589
 [9,]  0.2520936  0.33544473
[10,]  0.4833546  0.16889700
[11,]  0.6419235 -0.03882865
[12,]  0.6633622 -0.14233568
[13,]  0.6150147 -0.18571739
[14,]  0.5425954 -0.19770698

$eig
 [1] 3.191448955 2.246925843 1.140005643
1.042208426
 [5] 0.631524254 0.533523075 0.317146887
0.267548219
 [9] 0.215838029 0.198364540 0.189887045
0.117545980
[13] 0.086827996 0.001919393

$x
NULL

$ac
[1] 0

$GOF
[1] 0.534184 0.534184
```

metric MDS Map

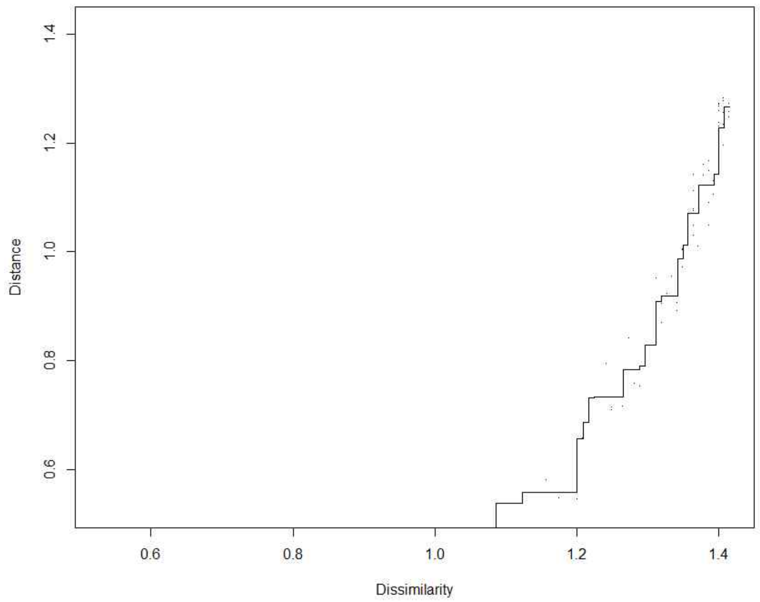


non-metric MDS Map



All MDS Maps show a color circle and similar patton.

non-metric MDS
Shepard Diagram



The Shepard Diagram is a scatterplot of distance and rank of distance. The graph is monotone increasing without decreasing. Therefore, There is no problem with data or algorithms.

Image Diagram

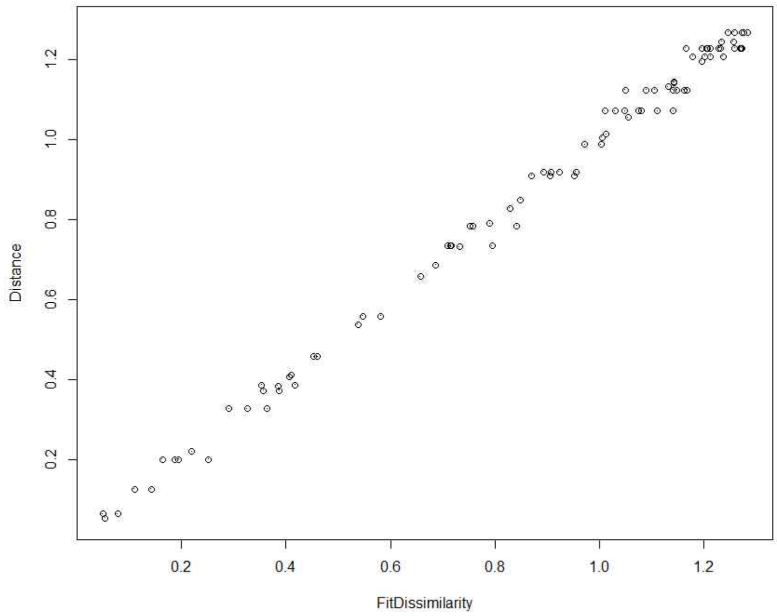


Image Diagram is a scatterplot of distance and rank of two-dimensional distance.
The graph is closing to the y=x. Therefore, The MDS Map is well fitted.

(4) Compare the Goodness-of-fits of MDS MAPs of (3).

❖ Kruskal(1964)'s Criterion for the number of dimensions

Stress(%)	0.0	2.5	5.0	10.0	20.0
Badness of fit	Perfect	Excellent	Good	Fair	Poor

In non-metric MDS, stress is 2.92494%(0.0292494). By Kruskal's criterion of the number of dimensions, Goodness-of-fits is excellent.
In metric MDS, Goodness-of-fit is 53.4184% (0.534184). It is not high value.
Therefore, non-metric MDS appropriate for this data more than metric MDS.

2. Consider the railway distance [Data 7.7.1] (railroad2.txt) between cities in Korea.

city	서울	부산	광주	대구	인천	강릉	청주	전주
서울	0.0							
부산	444.5	0.0						
광주	353.8	364.7	0.0					
대구	323.9	120.6	350.0	0.0				
인천	30.9	475.4	384.7	359.9	0.0			
강릉	348.5	505.0	523.0	388.8	379.4	0.0		
청주	141.2	326.1	235.4	211.4	172.1	288.6	0.0	
전주	279.9	391.0	124.6	276.4	310.8	449.4	136.3	0.0

(1) Perform metric MDS and non-metric MDS and interpret the axis of the MDS MAPs.

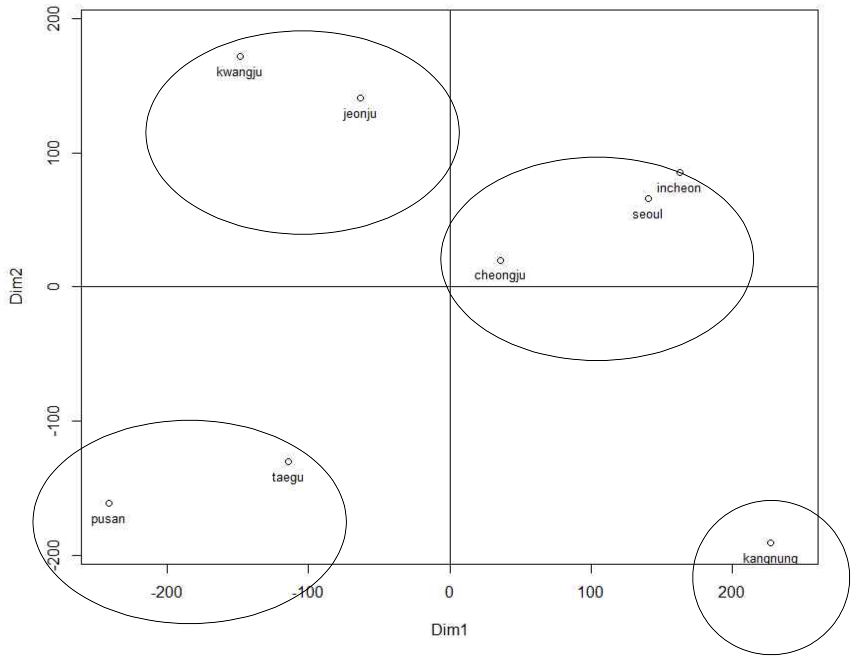
The above matrix is a dissimilarity matrix. Because, the diagonal elements of the dissimilarity matrix are zero.

```
# non-metric MDS
> con
$points
            [,1]       [,2]
[1,]  140.77360   65.61582
[2,] -241.32500 -161.23911
[3,] -148.23118  171.64021
[4,] -113.95571 -130.38363
[5,]  163.09964   84.94262
[6,]  226.77004 -190.90126
[7,]   36.04877   19.93897
[8,]  -63.18016  140.38637

$stress
[1] 6.004032
```

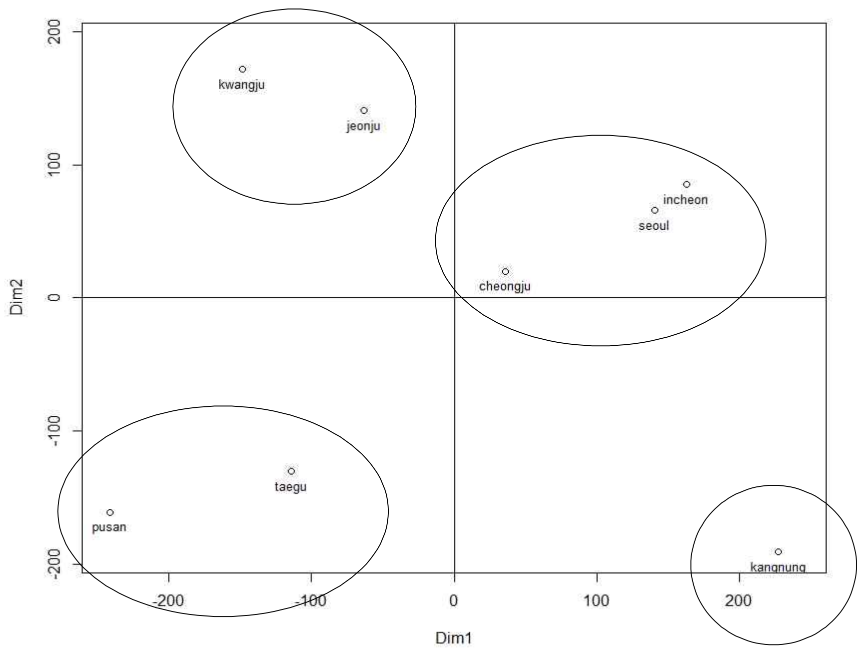
```
# metric MDS
> con
$points
            [,1]       [,2]
[1,]  140.75771   65.60058
[2,] -241.34975 -161.26080
[3,] -148.21178  171.64257
[4,] -113.96448 -130.40937
[5,]  163.09922   84.92990
[6,]  226.76135 -190.86162
[7,]   36.07025   19.95384
[8,]  -63.16252  140.40490

$eig
[1]  1.963297e+05  1.405292e+05
5.915327e+04
[4]  1.374540e+04  1.715287e+02
4.547474e-11
[7] -6.875141e+03 -1.219315e+04

$x
NULL

$ac
[1] 0

$GOF
[1] 0.7852237 0.8217492
```

MDS MAPs.

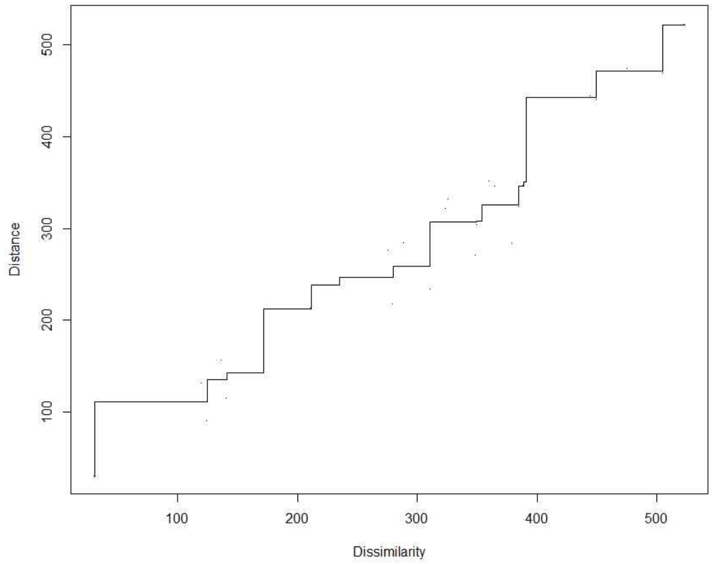
metric MDS



non-metric MDS



non-metric MDS
Shepard Diagram



The Shepard Diagram is a scatterplot of distance and rank of distance. The graph is monotone increasing without decreasing. Therefore, There is no problem with data or algorithms.

Image Diagram

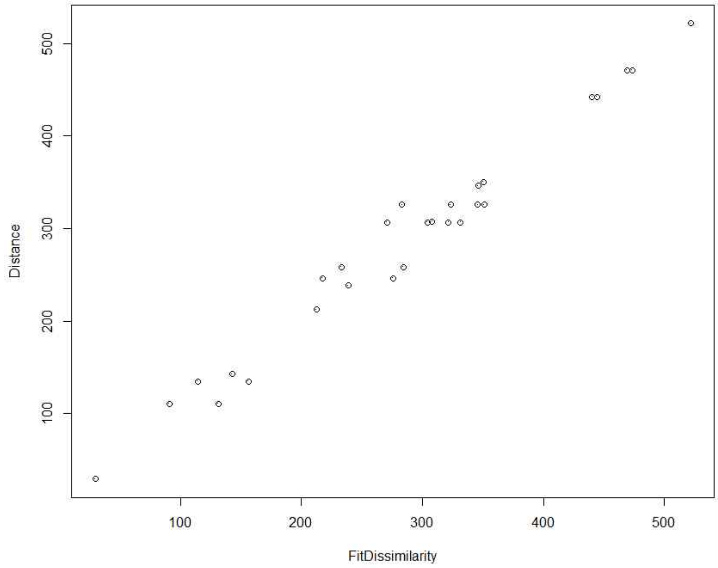


Image Diagram is a scatterplot of distance and rank of two-dimensional distance.
The graph is closing to the y=x. Therefore, The MDS Map is well fitted.
All MDS Maps show similar patton. The right side represents the north. The left side represents the sorth. The top represents the west. The bottom represents the east.

(2) Compare the Goodness-of-fits of two MDS MAPs.

❖ Kruskal(1964)'s Criterion for the number of dimensions

Stress(%)	0.0	2.5	5.0	10.0	20.0
Badness of fit	Perfect	Excellent	Good	Fair	Poor

In non-metric MDS, stress is 6.004032%(0.06004032). By Kruskal's criterion of the number of dimensions, Goodness-of-fits is good. In metric MDS, the first Goodness-of-fit obtained by replacing the negative eigenvalue with the absolute value is 78.52237% (0.7852237) and second Goodness-of-fit obtained by replacing the negative eigenvalue with the zero value is 82.17492% (0.8217492). That are high value. Therefore, both metric MDS and non-metric MDS appropriate for this data.

```
setwd("H:/학교/2020 2학기
정호재/다변량통계학2/실습/20201029/Rdata")
Data7.7.2<-read.table("color3.txt", header=T)
Data7.7.2
C<-as.matrix(Data7.7.2)
color=colnames(C)
n<-nrow(C)

# Standard Transformation : cij(similarity) to
dij(dissimilarity)
J<-matrix(1,n,n)
cii=diag(diag(C))%*%J
cij=C
cjj=J%*%diag(diag(C))
D<-sqrt(cii-2*cij+cjj)
D

# Non-Metric MDS
library(MASS)
con<-isoMDS(D, k=2)
con
x<-con$points[,1]
y<-con$points[,2]
lim1<-c(-max(abs(x)), max(abs(x)))
lim2<-c(-max(abs(y))-0.2, max(abs(y)))

plot(x,y, xlab="Dim1", ylab="Dim2", xlim=lim1,
ylim=lim2)
text(x,y,color, cex=0.8, pos=1)
abline(v=0,h=0)

# Shepard Diagram
color_sh <- Shepard(D[lower.tri(D)], con$points)
color_sh
plot(color_sh, pch = ".", xlab = "Dissimilarity", ylab
= "Distance",
    xlim = range(color_sh$x), ylim =
range(color_sh$x))
lines(color_sh$x, color_sh$yf, type = "S")

# image Diagram
ccolor_sh=cbind(color_sh$x, color_sh$y,
color_sh$yf)
ccolor_sh
plot(ccolor_sh[,2], ccolor_sh[,3], pch = ".", xlab =
"FitDissimilarity", ylab = "Distance",
    xlim = range(ccolor_sh[,2]), ylim =
range(ccolor_sh[,2]))
lines(ccolor_sh[,2],  ccolor_sh[,3], type = "p")

# Metric MDS
con<-cmdscale(D, k=2, eig=T)
con
x<-con$points[,1]
y<-con$points[,2]
lim1<-c(-max(abs(x)), max(abs(x)))
lim2<-c(-max(abs(y)), max(abs(y)))
plot(x,y, xlab="Dim1", ylab="Dim2", xlim=lim1,
ylim=lim2)
text(x,y,color, cex=0.8, pos=1)
abline(v=0,h=0)

####################################
Data7.7.1<-read.table("railroad2.txt", header=T)
Data7.7.1
D<-as.matrix(Data7.7.1)
city=colnames(D)
n<-nrow(D)

library(MASS)
con<-isoMDS(D, k=2)
con
x<-con$points[,1]
y<-con$points[,2]
lim1<-c(-max(abs(x)), max(abs(x)))
lim2<-c(-max(abs(y))-0.2, max(abs(y)))

plot(x,y, xlab="Dim1", ylab="Dim2", xlim=lim1,
ylim=lim2)
text(x,y,city, cex=0.8, pos=1)
abline(v=0,h=0)

# Shepard  Diagram
railroad_sh <- Shepard(D[lower.tri(D)], con$points)
railroad_sh
plot(railroad_sh, pch = ".", xlab = "Dissimilarity",
ylab = "Distance",
    xlim = range(railroad_sh$x), ylim =
range(railroad_sh$x))
lines(railroad_sh$x, railroad_sh$yf, type = "S")

# image  Diagram
```

```
crailroad_sh=cbind(railroad_sh$x, railroad_sh$y, railroad_sh$yf)
crailroad_sh
plot(crailroad_sh[,2], crailroad_sh[,3], pch = ".", xlab = "FitDissimilarity", ylab = "Distance",
      xlim = range(crailroad_sh[,2]), ylim = range(crailroad_sh[,2]))
lines(crailroad_sh[,2],  crailroad_sh[,3], type = "p")

# Metric MDS
con<-cmdscale(D, k=2, eig=T)
con
x<-con$points[,1]
y<-con$points[,2]
lim1<-c(-max(abs(x)), max(abs(x)))
lim2<-c(-max(abs(y)), max(abs(y)))
plot(x,y, xlab="Dim1", ylab="Dim2", xlim=lim1, ylim=lim2)
text(x,y,city, cex=0.8, pos=1)
abline(v=0,h=0)
```