Assignment 7

MATH 381 A - Winter 2022

Yixuan Liu

March 3, 2022

Introduction

In this assignment we will apply multidimensional scaling (MSD) to investigate a set of objects by R. The set we will investigate is Olympic track times, in tenths of a second, that were recorded over 25 Olympic Games from 1912 to 2020. In each year, the times for the races of 100 meters, 200 meters, 400 meters, 800 meters, 1500 meters, 5000 meters and 10000 meters in tenths of a second were recorded.

1

Store the data as a 25×7 matrix `X`. Since the times for each kind of races have huge differences, we need to nomarlize the data first. We use two ways to do this:

$$X_{ij} = \frac{X_{ij} - \frac{1}{25}\sum_{i=0}^{25} X_{ij}}{\text{standard deviation of } j^{th} \text{ column}}$$

$$X_{ij} = \frac{X_{ij} - \text{min of } j^{th} \text{ column}}{\text{max of } j^{th} \text{ column} - \text{min of } j^{th} \text{ column}}$$

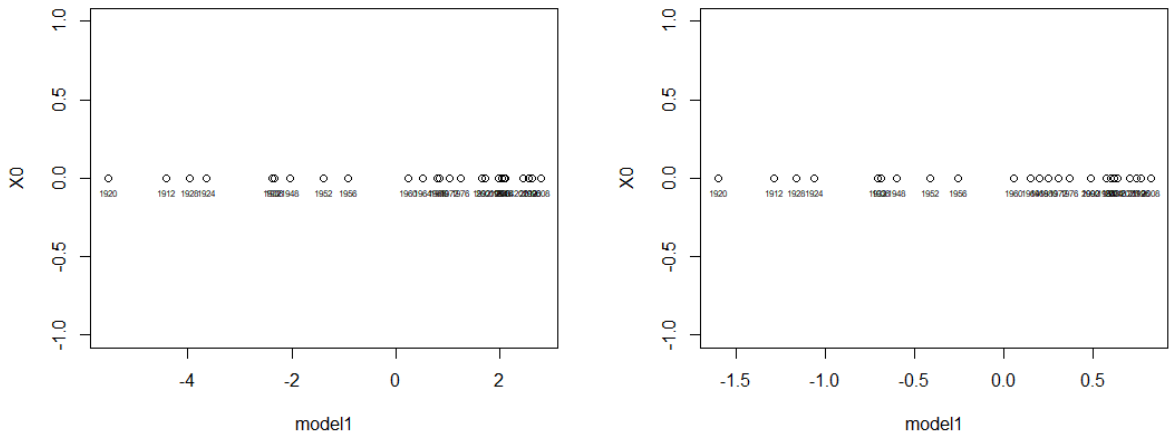
Now we can calculate the Minkowski distance matrix for the two normalized data by using Euclidean distance:

$$d_{ij} = (\sum_{m} |r_{i,m} - r_{j,m}|^{\lambda})^{\frac{1}{\lambda}}$$

where $\lambda = 2$ for Euclidean distance.

2

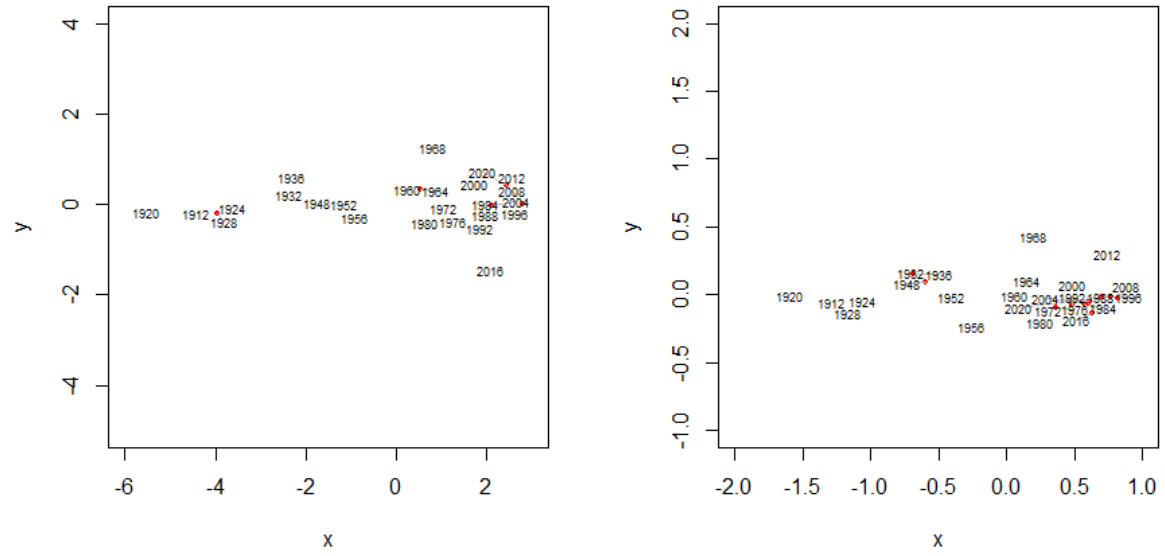
Now we can use the `cmdscale` function to get a 1-dimensional model for both distance matrices of two normalized data, and get the plots:



(a) standardized data (b) projected data

Figure 1: 1D model

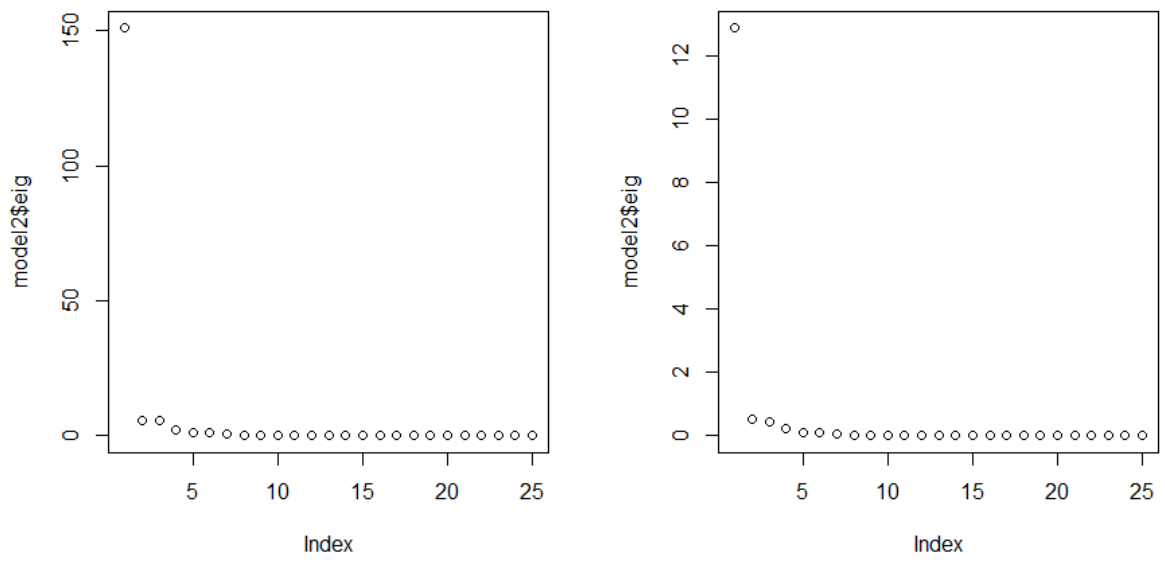
And 2-dimensional models:



(a) standardized data (b) projected data

Figure 2: 2D model

The 3D model is hard to show, but we can still calculate different features for 3D model. Since eigenvalues do not depend on the dimension of the model, we can plot them once:



(a) standardized data

(b) projected data

Figure 3: Eigenvalues

The two plot have approximately same pattern. We can see that the eigenvalues drops dramatically from the first one to the second one, and the second and the third eigenvalues have relatively significant value compare to the eigenvalues from the forth eigenvalue. We can assume that the 1D model would be sufficent, but a 2D or 3D model could do better.
To see how well a model is, we can look at `$GOF`. It will give us two values for each model, the first one is calculated with the absolute eigenvalues and the second one is calculated with the maximum of eigenvalues and 0. In our vector of eigenvalues, only the eigenvalues that are so small (with 10^{-15} to 10^{-17}) that can be treated as 0 have some negative values, so the two GOF is almost the same. Here is the table of GOF for the above models:

GOF			
	1D	2D	3D
standardized data	0.8983067	0.9335055	0.9670789
projected data	0.8999356	0.9368651	0.9668457

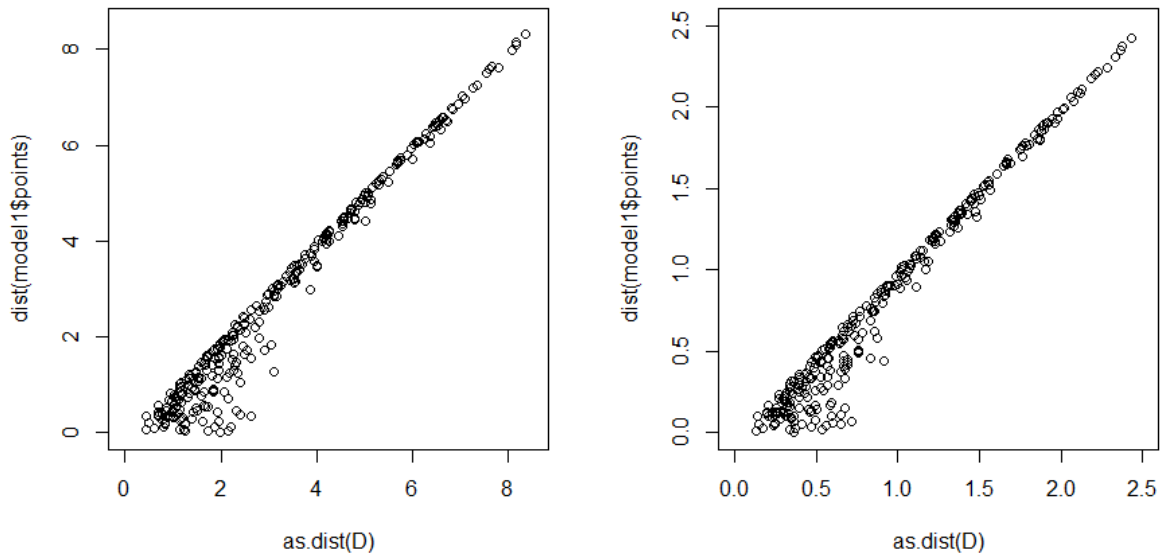
From the table we can see that the 1D model is already good enought and 2D and 3D models are doing better. And GOFs of standardized data and projected data have negligible difference.
We can also calculate the different between distance matrix for each model and input distance matrix. By gettiing the mean squared difference of the entries for each model, we have the following table:

Mean Squared Difference			
	1D	2D	3D
standardized data	0.3246855	0.09215854	0.02248068
projected data	0.02585289	0.01096329	0.001922921

and maximum absolute difference:

Maximum Absolute Difference			
	1D	2D	3D
standardized data	2.272291	1.659495	0.5846159
projected data	0.6484494	0.5780326	0.1771961

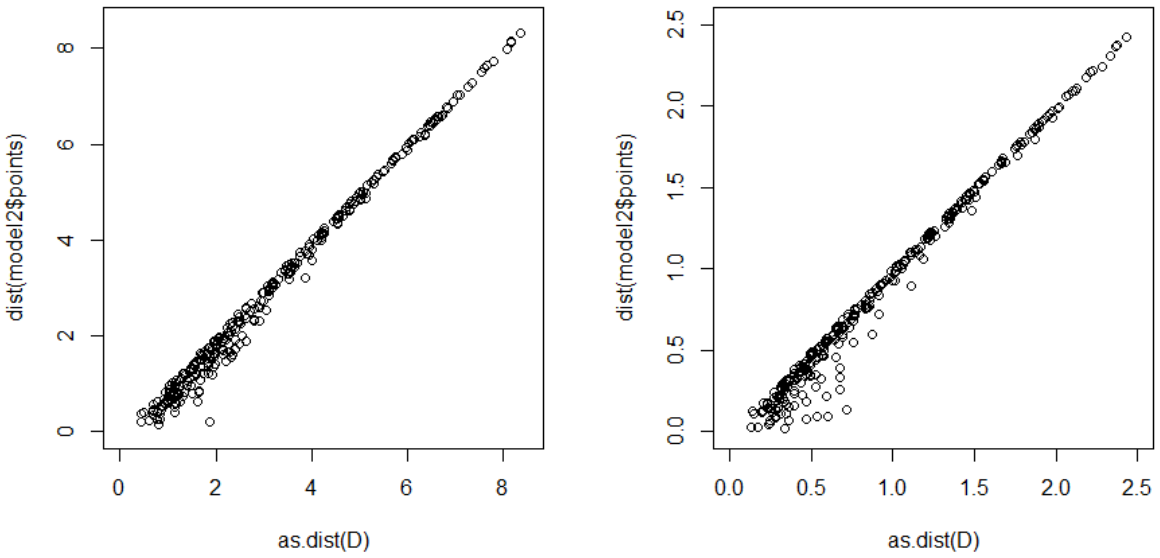
From the above two tables, we can see that the models with projected data have both smaller mean squared difference and maximum absolute difference. This might due to that the projected data has narrower range than the standardized data
Now we plot the distances in the model versus the input distances:



(a) standardized data

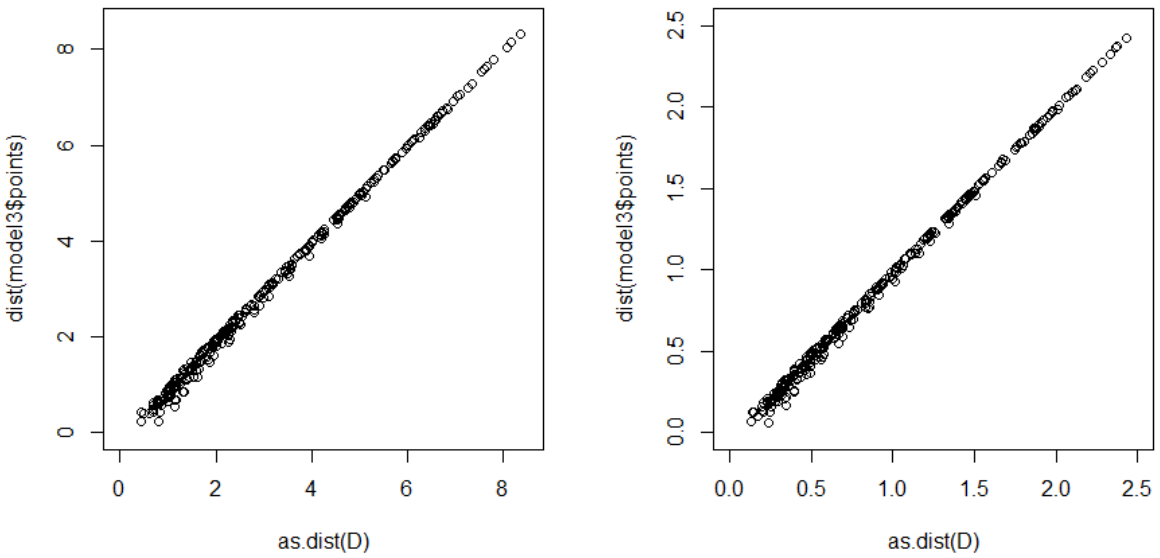
(b) projected data

Figure 4: 1D model



(a) standardized data (b) projected data

Figure 5: 2D model



(a) standardized data (b) projected data

Figure 6: 3D model

We can see that for both data, although the 1D model is good enough when we investigate

eigenvalue, but the plots show that there are a lot of points are under the $y = x$ line. However, for 3D models, all points are close to the line, so the 3D model is the best model among the 1D, 2D and 3D models.

3

In both 2D models for standardized data and projected data, the points do not spread out too much on the y-axis compared to the x-axis, so the 1D model is sufficient for both our data. By examining the original data for the Olympic races, we can see that in 1920, the recorded times for each race are the longest or almost longer than the same races in any other year. For the years in the right of the plot such as 1996, 2008 and 2012, the recorded times are relatively shorter than those in other years. So we may conclude that the x-axis represents how fast is the recorded time in general for all races in a year.

1 Appendix A - Data Set

```
1912,108,217,482,1119,2368,8766,18808
1920,108,220,496,1134,2418,8956,19058
1924,106,216,476,1124,2336,8712,18232
1928,108,218,478,1118,2332,8780,18188
1932,103,212,462,1098,2312,8700,18114
1936,103,207,465,1129,2278,8622,18154
1948,103,211,462,1092,2298,8576,17996
1952,104,207,459,1092,2252,8460,17570
1956,105,206,467,1077,2212,8196,17256
1960,102,205,449,1063,2156,8234,17122
1964,100,203,451,1051,2181,8288,17044
1968,99.5,198.3,438,1043,2149.1,8500,17674
1972,101.4,200,446.6,1058.6,2163,8064,16583.5
1976,100.6,202.3,442.6,1035,2191,8047.6,16603.6
1980,102.5,201.9,446,1054,2184,8010,16626.9
1984,99.9,198,442.7,1030,2125.3,7885.6,16675.4
1988,99.2,197.5,438.7,1035.5,2159.6,7917,16414.8
1992,99.6,200.1,435,1036.6,2201.2,7925.2,16667
1996,98.4,193.2,434.9,1025.8,2157.8,7897.6,16273.4
2000,98.7,200.9,438.4,1050.8,2120.7,8154.9,16382
2004,98.5,197.9,440,1044.5,2141.9,7943.9,16251
2008,96.9,193,437.5,1045.6,2131.1,7778.2,16211.7
2012,96.3,193.2,439.4,1009.1,2140.8,8216.6,16504.2
2016,98.1,197.8,430.3,1021.5,2300,7833,16251.7
2020,98,196.2,438.5,1050.6,2083.2,7781.5,16632.2
```