math381HW7

Siyue Zhu

March 2022

1 Introduction

In this paper, I'm going to calculate and analyze the distance between different countries using mortality rates from Leukemia. I'm going to create one dimensional, two dimensional and three dimensional models using MDS. The data I'm going to use is mortality rates from Leukemia per million children between the ages of 0 to 14. My dataset can be find using this link: https://people.sc.fsu.edu/ jburkardt/datasets/hartigan/file51.txt. The dataset has 18 rows and 12 columns, representing 18 different countries through 1956 to 1967, which are 12 years in total.

mortality rates from Leukemia for each country												
Country	1956	1957	1958	1959	1960	1961	1962	1963	1964	1965	1966	1976
Australia	34	33	43	44	38	38	39	34	36	35	37	34
Austria	27	35	40	45	39	34	37	39	35	39	37	29
Belgium	41	29	31	40	39	39	47	40	34	29	34	30
Canada	36	35	32	42	35	35	35	38	36	37	33	33
Denmark	33	44	46	45	35	48	43	49	56	39	39	38
Finland	43	41	34	32	37	28	28	42	43	38	32	41
France	44	41	41	42	40	36	40	36	36	34	34	31
West Germany	35	33	35	36	35	35	34	38	36	38	34	39
Hungary	28	26	31	31	29	27	27	34	36	35	29	35
Isreal	38	44	45	28	61	33	32	28	29	36	30	30
Japan	26	25	29	28	30	29	32	32	31	33	33	31
Netherlands	32	37	39	39	39	31	41	37	40	42	36	33
Northern Ireland	17	25	27	39	29	27	22	31	33	26	30	32
Norway	45	44	36	31	44	44	56	31	36	37	50	40
Portugal	20	28	21	34	30	30	30	29	35	37	40	36
Scotland	27	31	32	30	35	30	31	26	29	28	29	25
Sweden	33	45	46	45	44	37	44	31	39	35	37	38
Switzerland	42	46	44	47	38	35	45	42	33	31	36	31

The following matrix A is our original matrix without any normalization, and I'm going to include this original matrix for our comparison.

$$A = \begin{pmatrix}
34 & 33 & 43 & 44 & 38 & 38 & 39 & 34 & 36 & 35 & 37 & 34 \\
27 & 35 & 40 & 45 & 39 & 34 & 37 & 39 & 35 & 39 & 37 & 29 \\
41 & 29 & 31 & 40 & 39 & 39 & 47 & 40 & 34 & 29 & 34 & 30 \\
36 & 35 & 32 & 42 & 35 & 35 & 35 & 38 & 36 & 37 & 33 & 33 \\
33 & 44 & 46 & 45 & 35 & 48 & 43 & 49 & 56 & 39 & 39 & 38 \\
43 & 41 & 34 & 32 & 37 & 28 & 28 & 42 & 43 & 38 & 32 & 41 \\
44 & 41 & 41 & 42 & 40 & 36 & 40 & 36 & 36 & 34 & 34 & 31 \\
35 & 33 & 35 & 36 & 35 & 35 & 34 & 38 & 36 & 38 & 34 & 39 \\
28 & 26 & 31 & 31 & 29 & 27 & 27 & 34 & 36 & 35 & 29 & 35 \\
38 & 44 & 45 & 28 & 61 & 33 & 32 & 28 & 29 & 36 & 30 & 30 \\
26 & 25 & 29 & 28 & 30 & 29 & 32 & 32 & 31 & 33 & 33 & 31 \\
32 & 37 & 39 & 39 & 39 & 31 & 41 & 37 & 40 & 42 & 36 & 33 \\
17 & 25 & 27 & 39 & 29 & 27 & 22 & 31 & 33 & 26 & 30 & 32 \\
45 & 44 & 36 & 31 & 44 & 44 & 56 & 31 & 36 & 37 & 50 & 40 \\
20 & 28 & 21 & 34 & 30 & 30 & 30 & 29 & 35 & 37 & 40 & 36 \\
27 & 31 & 32 & 30 & 35 & 30 & 31 & 26 & 29 & 28 & 29 & 25 \\
33 & 45 & 46 & 45 & 44 & 37 & 44 & 31 & 39 & 35 & 37 & 38 \\
42 & 46 & 44 & 47 & 38 & 35 & 45 & 42 & 33 & 31 & 36 & 31
\end{pmatrix}$$

2 Calculating normalized matrices and distance matrices

Now, we are going to normalize our matrix A with two different methods. The first method is to find the mean and standard deviation of each column, and then subtract the mean and divide by the standard deviation in each column. This will transform our data into a new data set in which each column has mean zero and standard deviation 1. We can get our matrix B from the following equation:

$$B_{ij} = (A_{ij} - \mu_j)\sigma_j$$

A second way to normalize our matrix A, is to find the min and max value in each column, and then map each column entry x to (x-min)/(max-min). This will convert the data set into a new data set in which each column as a minimum of 0 and a maximum of 1. In this way, the dimensions will be (in a certain sense) comparable. We can get our matrix c from the following equation:

$$C_{ij} = (A_{ij} - min_j)/(max_j - min_j)$$

Also, we are going to calculate three distance matrices X, Y and Z from matrices A, B and C. Each entry in X_{ij}, Y_{ij} and Z_{ij} representing the distance between the country at row i and the country at column j from matrices A, B and C.

Matrix X can be calculated from the following equation:

$$X = \sqrt{\sum_{a=1}^{12}(A_{im} - A_{jm})^2}$$

Matrix Y can be calculated from the following equation:

$$Y = \sqrt{\sum_{a=1}^{12} (B_{im} - B_{jm})^2}$$

Matrix Z can be calculated from the following equation:

$$Z = \sqrt{\sum_{a=1}^{12} (C_{im} - C_{jm})^2}$$

3 GOF values

We see that there are two GOF values given by the cmdscale command in R, because there are different ways to handle negative eigenvalues. The first values is calculated by

$$(\sum_{a=1}^{12} \lambda_a)/(\sum_{a=1}^{12} |\lambda_a|)$$

, and the second value is calculated by

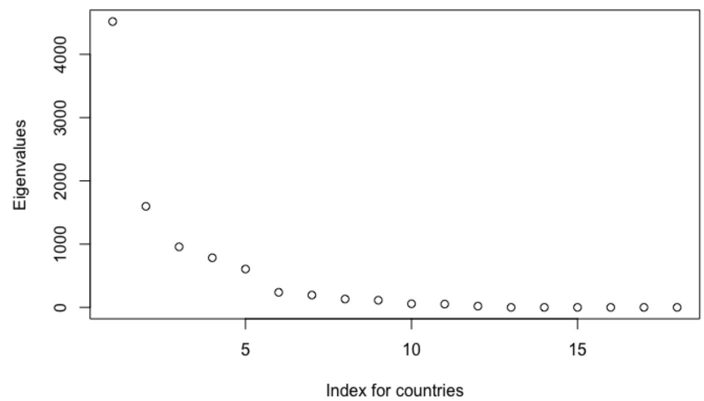
$$(\sum_{a=1}^{12} \lambda_a)/(\sum_{a=1}^{12} \max(0, \lambda_a))$$

4 Results

Now, I'm going to show all my results I get from R.

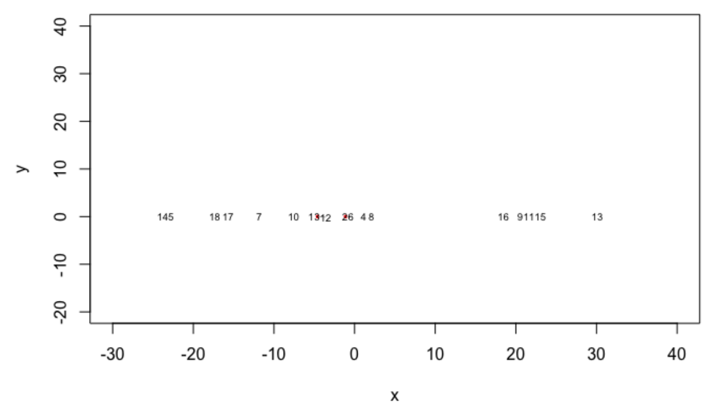
4.1 Model 1 (using the non-normalized matrix)

First, let's start with model 1. Model 1 is using matrix A, which is the non-normalized original matrix. Here is the plot of the eigenvalues versus the index of the countries.



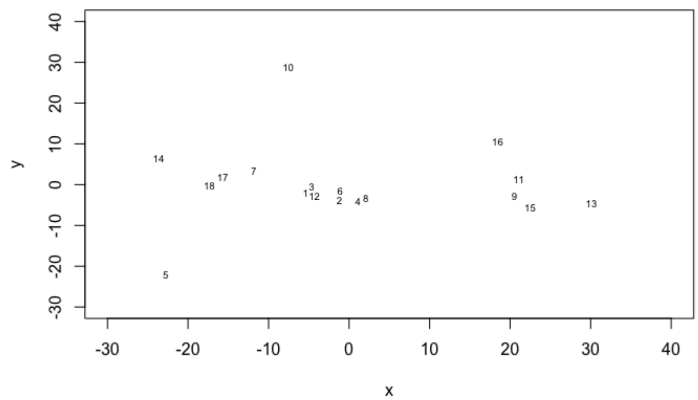
(a)

Here are the plot for one dimensional model and two dimensional model:



(b)

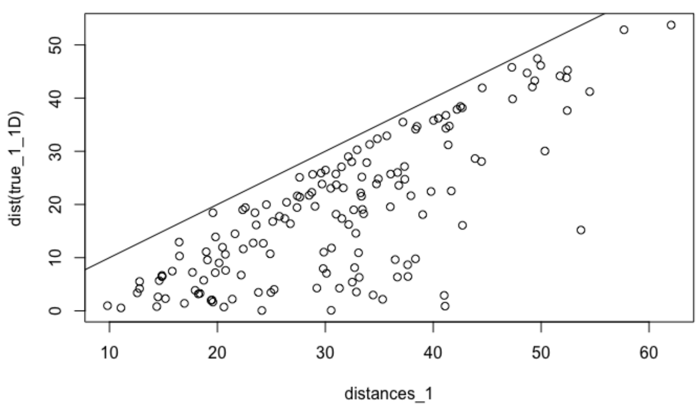
Figure 1: (a)one dimensional model



(a)

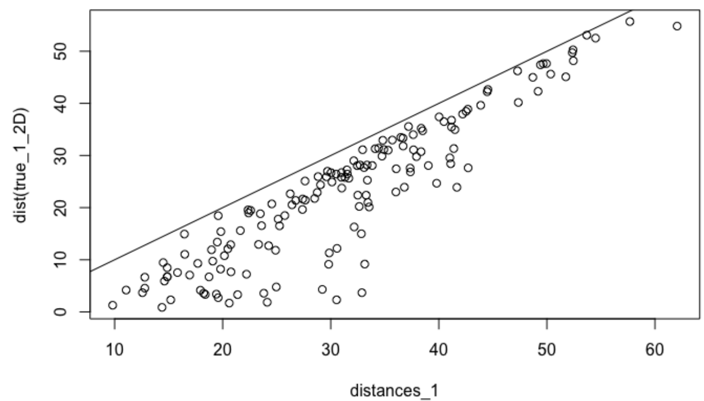
Figure 2: (b)two dimensional model

The following are plots of the distances in the model versus the input distances for 3 different dimensional.



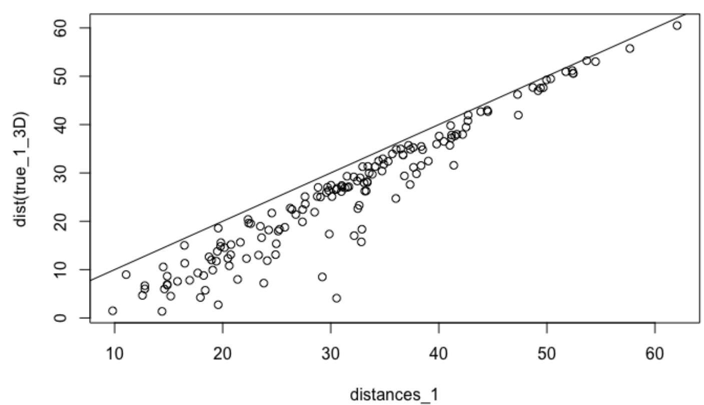
(a)

Figure 3: (c)distances plot for one dimensional model



(a)

Figure 4: (d)distances plot for two dimensional model



(a)

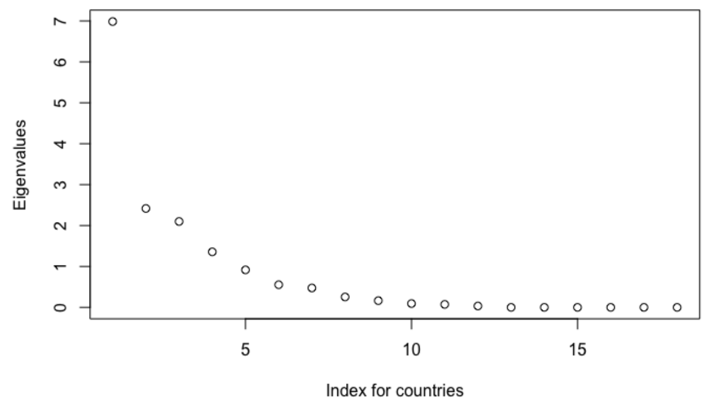
Figure 5: (e)distances plot for three dimensional model

Here is the table for the comparison between mean absolute difference, maximum absolute difference and GOF for dimensional 1, 2, and 3.

Dimension	1	2	3
Mean absolute difference	11.68936	7.611184	5.267824
Maximum absolute difference	40.20361	29.19582	26.44597
GOF	0.487, 0.487	0.66, 0.66	0.763, 0763

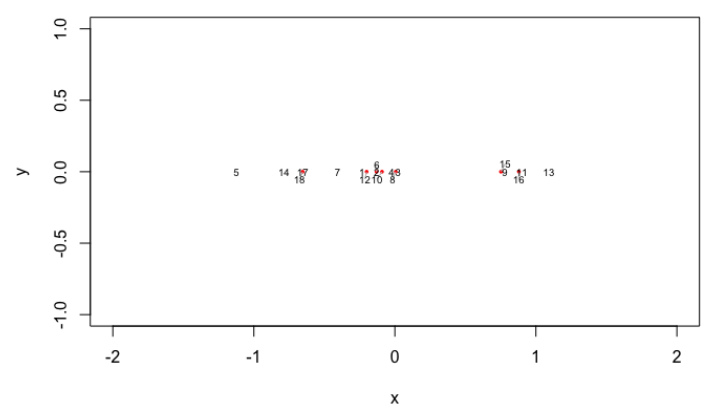
4.2 Model 2 (using the first normalized matrix)

Let's take a look at model 2. Model 2 is using matrix B, which is the a normalized using the first method. Here is the plot of the eigenvalues versus the index of the countries.



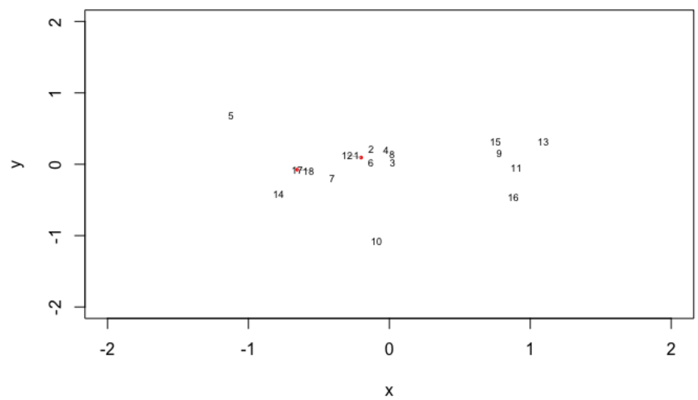
(a)

Here are the plot for one dimensional model and two dimensional model:



(b)

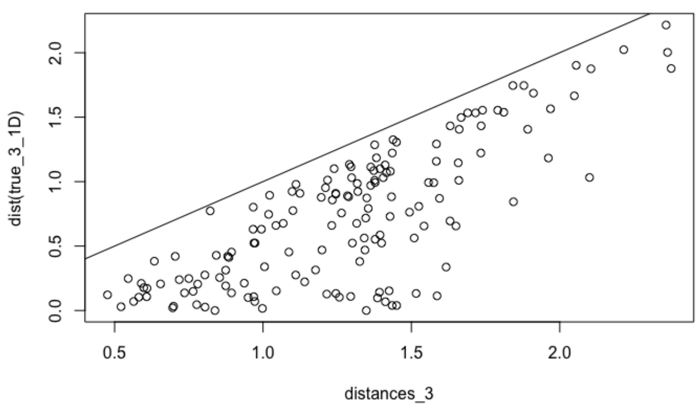
Figure 6: (f)one dimensional model for model 2



(a)

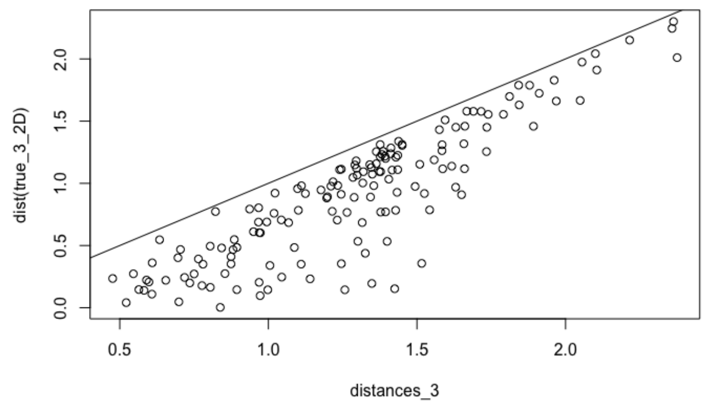
Figure 7: (g)two dimensional model for model 2

The following are plots of the distances in the model versus the input distances for 3 different dimensional.



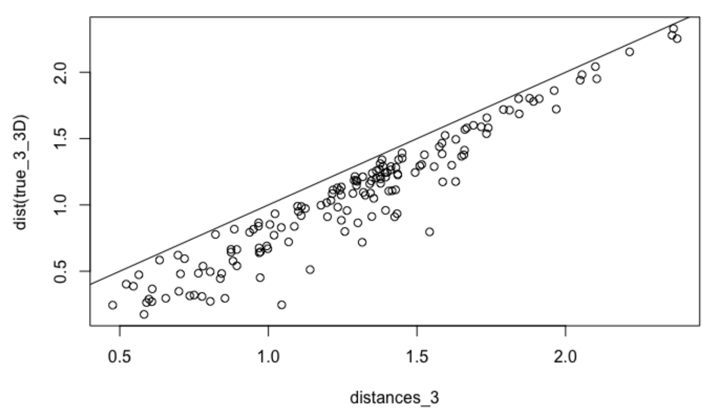
(a)

Figure 8: (h)distances plot for one dimensional model for model 2



(a)

Figure 9: (i)distances plot for two dimensional model for model 2



(a)

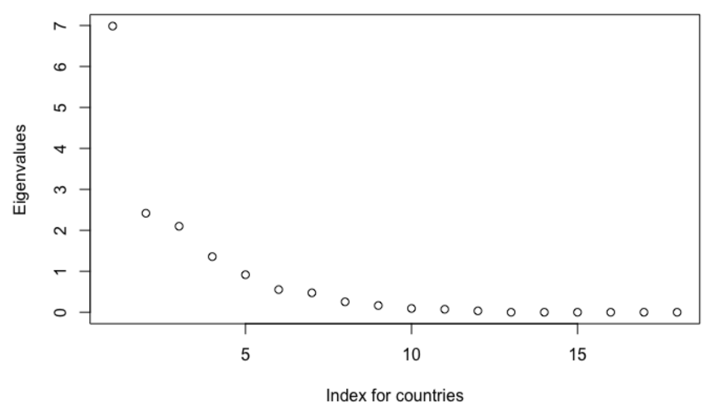
Figure 10: (j)distances plot for three dimensional model for model 2

Here is the table for the comparison between mean absolute difference, maximum absolute difference and GOF for dimensional 1, 2, and 3.

Dimension	1	2	3
Mean absolute difference	2.117832	1.405553	0.9949679
Maximum absolute difference	5.836701	4.229775	3.09991
GOF	0.434, 0.434	0.611, 0.611	0.732, 0732

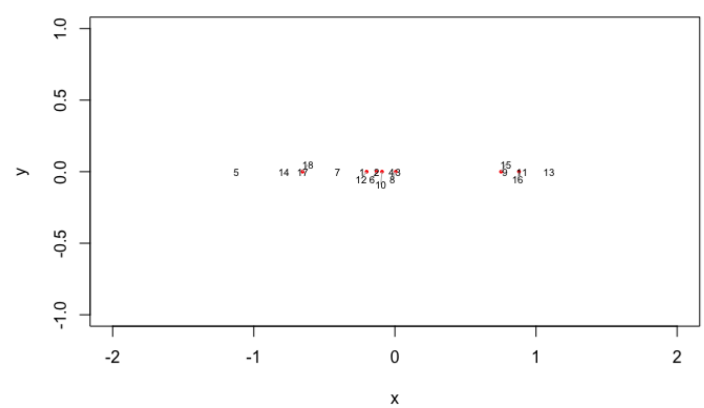
4.3 Model 3 (using the second normalized matrix)

Let's take a look at model 3. Model 3 is using matrix C, which is the a normalized using the second method. Here is the plot of the eigenvalues versus the index of the countries.



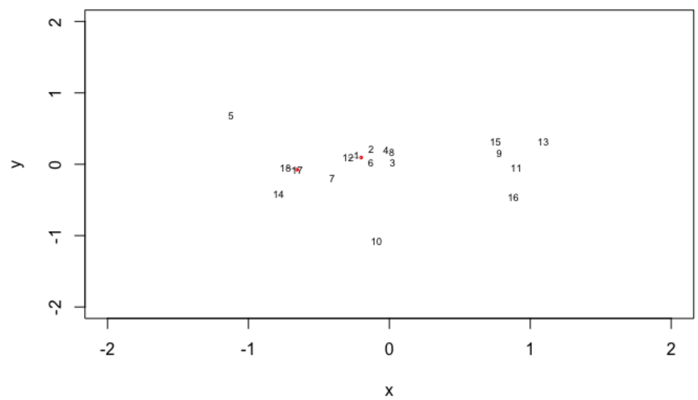
(a)

Here are the plot for one dimensional model and two dimensional model:



(b)

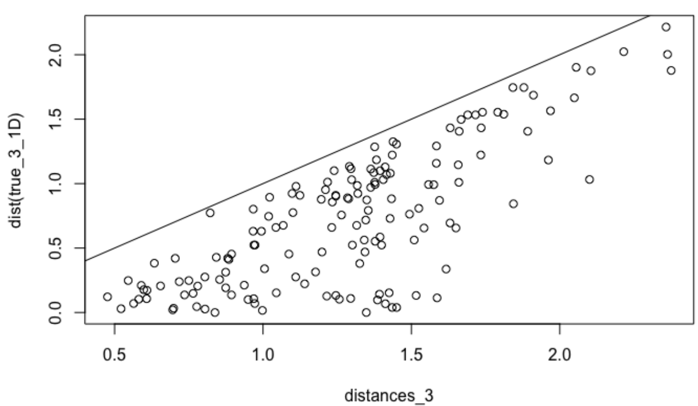
Figure 11: (f)one dimensional model for model 3



(a)

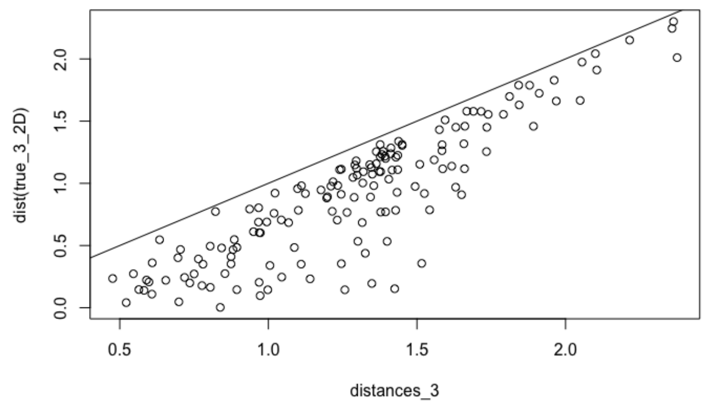
Figure 12: (k)two dimensional model for model 3

The following are plots of the distances in the model versus the input distances for 3 different dimensional.



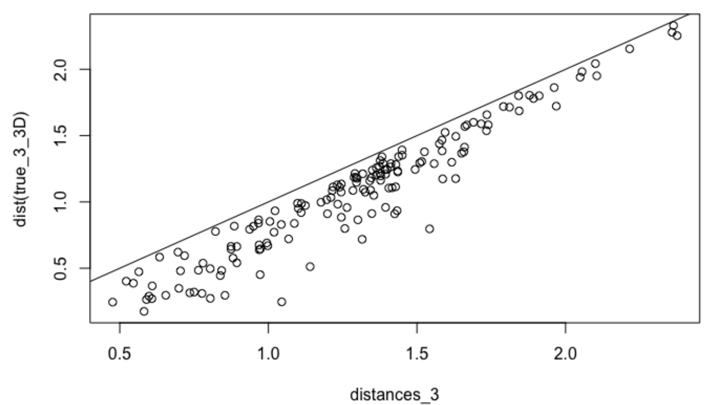
(a)

Figure 13: (l)distances plot for one dimensional model for model 3



(a)

Figure 14: (m)distances plot for two dimensional model for model 3



(a)

Figure 15: (n)distances plot for three dimensional model for model 3

Here is the table for the comparison between mean absolute difference, maximum absolute difference and GOF for dimensional 1, 2, and 3.

Dimension	1	2	3
Mean absolute difference	0.5197833	0.3523398	0.2084869
Maximum absolute difference	1.47299	1.272255	0.798645
GOF	0.453, 0.453	0.61, 0.61	0.746, 0746

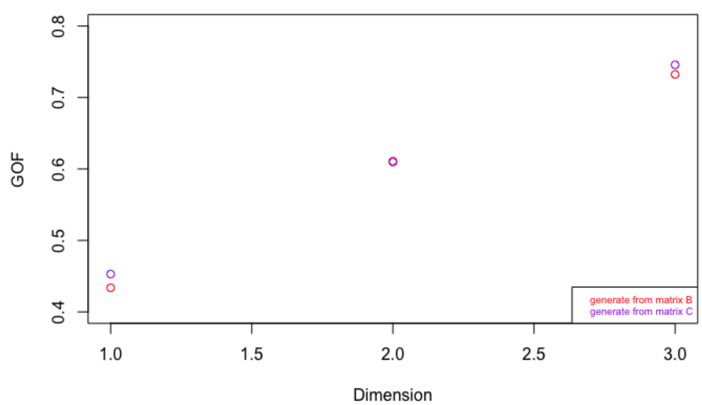
5 Discussion

I'm going to discuss methods based on normalized matrices, since it is meaningless to talk about non-normalized models.

5.1 Eigenvalues

We can see from the plot of eigenvalues for matrix B and C, the first eigenvalue is the largest in both graphs, which means our 1 dimensional space contains the most information from the original data set. For the plot of eigenvalues for matrix B, the first 12 indexes are greater than zero, and all other indexes are about equal to zero. This indicates that my model fits well for 12 dimensional space. Different from matrix B, matrix C has the first 6 indexes greater than zero, and all other indexes are about equal to zero. This indicates that my model fits well for 6 dimensional space.

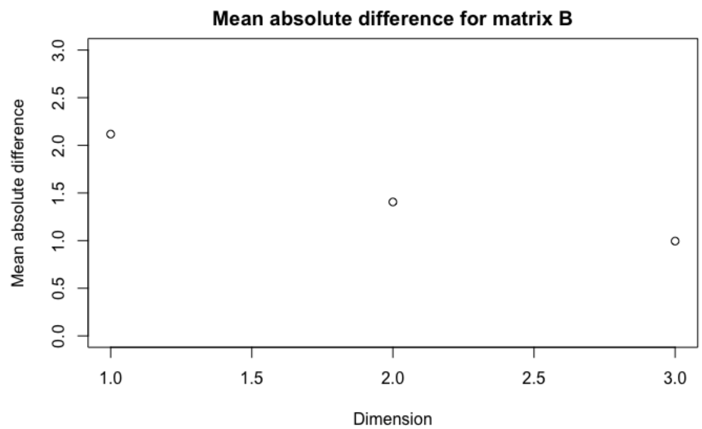
5.2 GOF



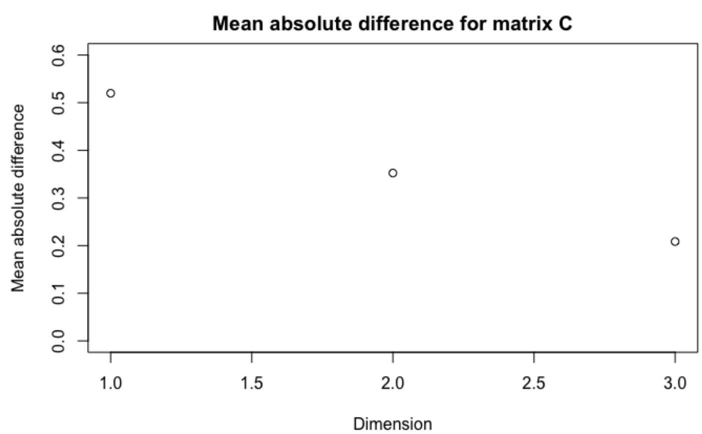
(a)

We can see that the GOF is getting larger as the dimension get larger, since larger dimensions include more information. Also, from the table above we can see that the two GOF for each model are the same even though there are two different ways dealing with the negative value of eigenvalues when to calculate the two GOF, this is because all our eigenvalues are positive.

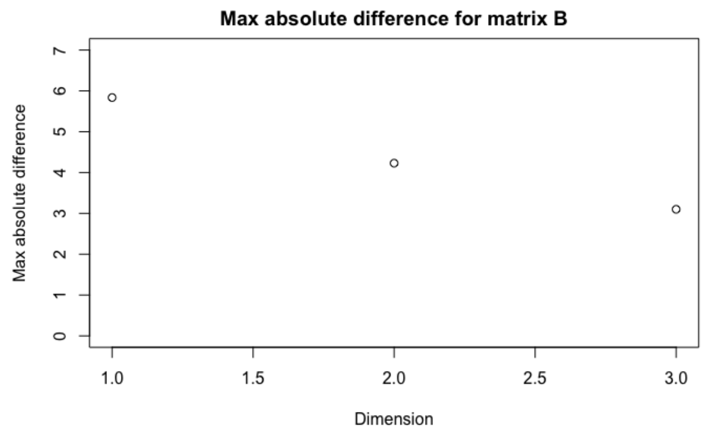
5.3 Mean and max absolute difference



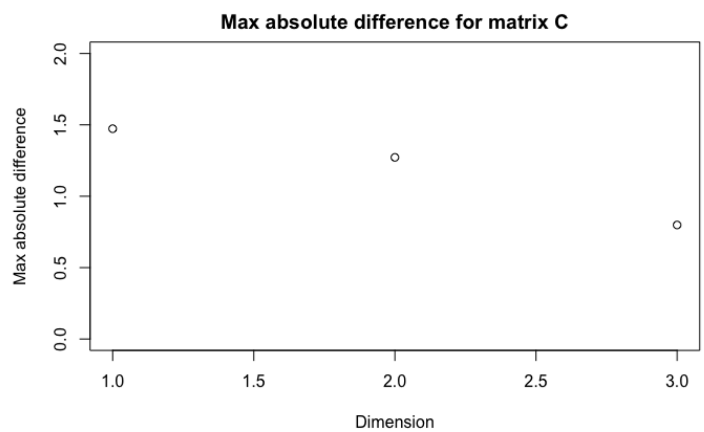
(b)



(c)



(d)



(e)

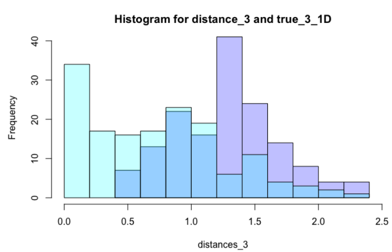
We can see from the plot above that both mean and max absolute difference is getting smaller with the dimension getting larger, which means that the value of our distance between each country is getting smaller as the dimension getting larger.

5.4 histograms of the original and model distances

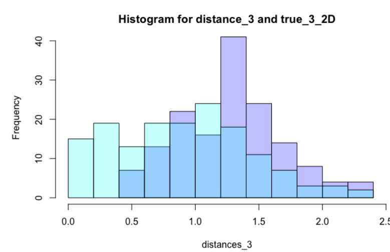
(f)

(g)

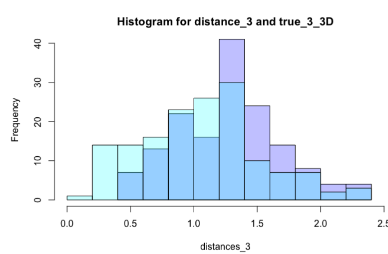
(h)



(i)



(j)



(k)

We can see from all the histograms above that with the dimension getting larger, the two histograms have more are overlapping. We can conclude that dimension 3 is the best model for both matrix B and matrix C, since the histograms for dimension 3 has the most area overlapping with the true distance.