Bioinformatics and Statistical Genetic

Population substructure

Leonardo Ortoleva & Egon Ferri

Contents

MDS solution as an initial configuration. How many dimensions are necessary to obtain a good representation with a stress below 5? Make a plot of the stress against the number of dimensions. 11

13 Compute the correlation matrix between the first two dimensions of a metric MDS and the two-dimensional solution of your best non-metric MDS. Make a scatterplot matrix of the 4 variables. Comment on your findings. 13

```
knitr::opts_chunk$set(echo = F)
```

1 Load data into the R environment.

Table 1: Data

FID	IID	PAT	MAT	SEX	PHENOTYPE	rs78200054_T	rs71235073_T
NA06984	NA06984	0	0	1	-9	1	1
NA06985	NA06985	0	0	2	-9	1	1
NA06986	NA06986	0	0	1	-9	1	1
NA06989	NA06989	0	0	2	-9	1	1
NA06994	NA06994	0	0	1	-9	1	1
NA07000	NA07000	0	0	2	-9	1	1
NA07037	NA07037	0	0	2	-9	1	1
NA07048	NA07048	0	0	1	-9	1	1
NA07051	NA07051	0	0	1	-9	1	1
NA07056	NA07056	0	0	2	-9	1	1

2 Compute the Manhattan distance matrix between the individuals using R function dist. Include a submatrix of dimension 5 by 5 with the distances between the first 5 individuals in your report.

Table 2: Submatrix

1	2	3	4	5
0.0000	255.2234	259.5400	270.5993	256.3201
255.2234	0.0000	258.1163	262.9239	262.3643
259.5400	258.1163	0.0000	261.5473	260.7623
270.5993	262.9239	261.5473	0.0000	274.8418
256.3201	262.3643	260.7623	274.8418	0.0000

3 The Manhattan distance (also known as the *taxicab metric*) is identical to the Minkowsky distance with parameter $\lambda = 1$. How does the Manhattan distance relate to the allele sharing distance, where the latter is calculated as two minus the number of shared alleles?

Let x_{ijk} be the number of shared alleles of individual i and j for variant k Allele sharing distance:

$$d_{ASk}(i, j, k) = 2 - x_{ijk}$$

Typically averaged over K genetic variants:

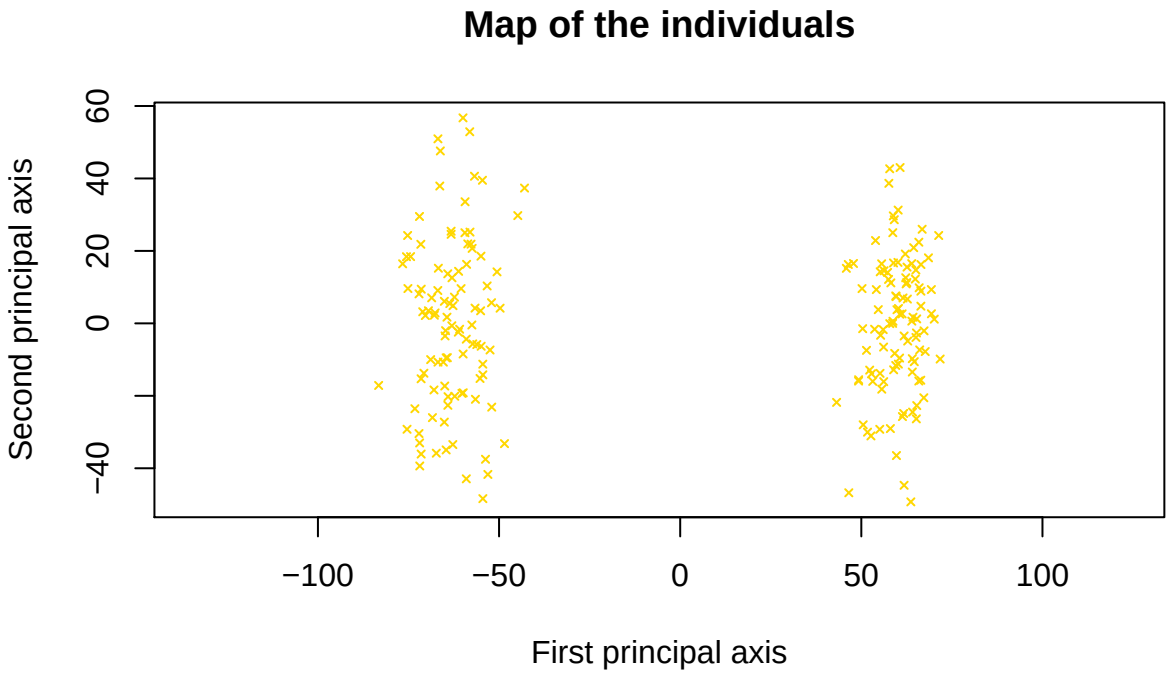
$$d_{AS}(i, j) = \frac{1}{K} \sum_{k=1}^{K} d_{ASk} = \frac{1}{K} \sum_{k=1}^{K} 2 - x_{ijk}$$

Manhattan distance:

$$d_M(i, j) = \|i - j\|_1 = \sum_{k=1}^{K} |i_k - j_k|$$

Since i_k and j_k can be only (0,1,2) (=AA, AB, BB), we see that the formulas are just identical. The only difference is that, if we apply the normalization of $\frac{1}{K}$, the *Allele sharing distance* is the $Manhattan\ distance$ normalized by the number of variants.

4 Apply metric multidimensional scaling using the Manhattan distance matrix to obtain a map of the individuals, and include your map in your report. Do you think the data come from one homogeneous human population? If not, how many subpopulations do you think the data might come from, and how many individuals pertain to each suppopulation?



We can see easily that there are 2 subpopulations.

Table 3: Sub populations

First subpopulation	Second subpopulation
99	104

5 Report the first 10 eigenvalues of the solution

Table 4: Eigenvalues

772590.2	90338.34	81999.06	78287.09	77424.73	73191.08	72052.16	70292.23	68918.76	68059.62

6 Does a perfect representation of this $n \times n$ distance matrix exist, in n or fewer dimensions? Why so or not?

```
## [1] "The goodness of fit is : 1"
```

An exact representation of the matrix will existe in $n-1$ dimensions. In fact, we see that with $k = n-1$ we have a GOF of 1, and the eigenvalues are all positives.

We see that an eigenvalue < 0 is found, but is computationally 0.

```
## [1] -1.041037e-11
```

7 What is the goodness-of-fit of a two-dimensional approximation to your distance matrix? Explain which criterium you have used.

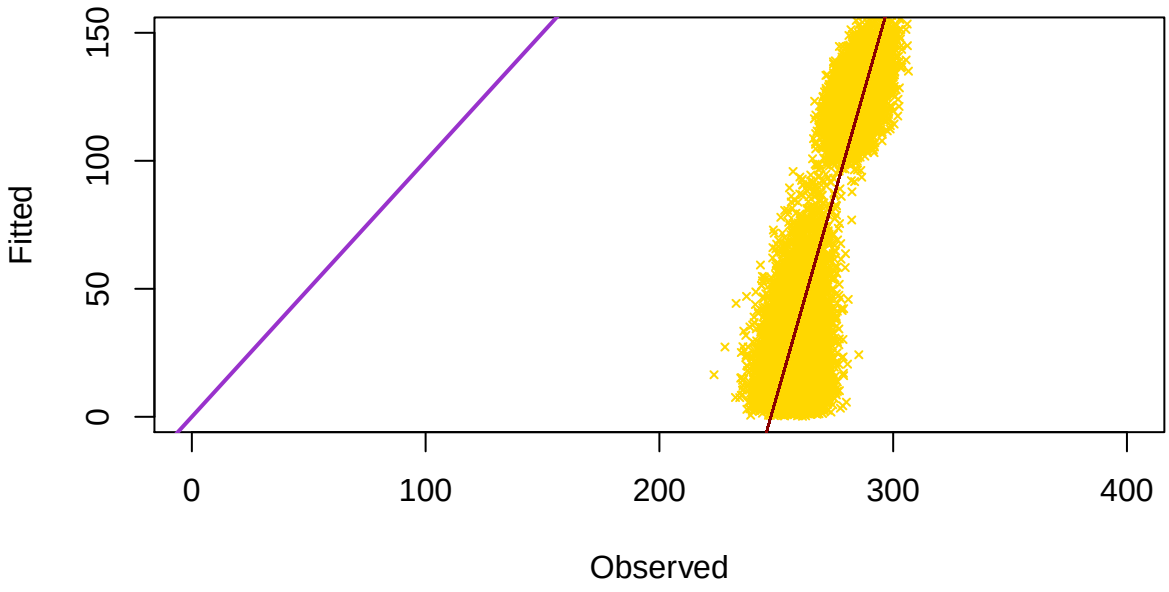
Using the R function "cmdscale" with k=2, we obtain a new MDS with maximum 2 dimension of the space.

```
## [1] "The new goodness of fit is 0.1154"
```

```
## [1] "The dimensions of the points of our MDS are "
```

```
## [1] 203   2
```

8 Make a plot of the estimated distances (according to your map of individuals) versus the observed distances. What do you observe? Regress estimated distances on observed distances and report the coefficient of determination of the regression.



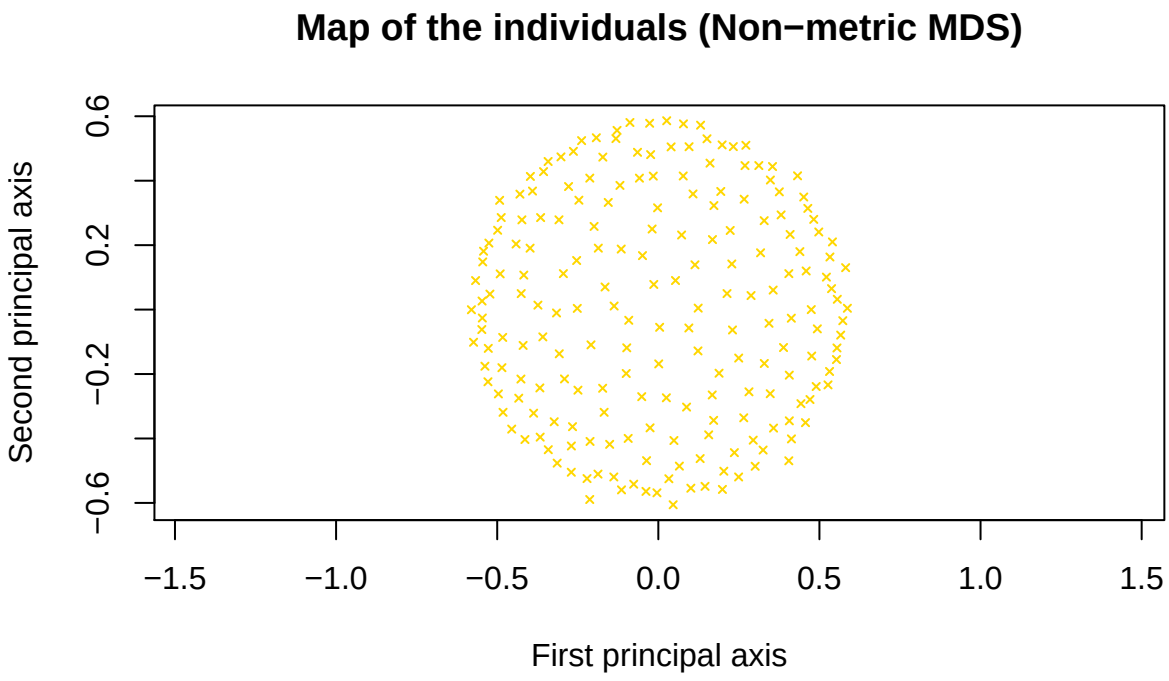
```
## [1] "The coefficient of correlation is: 0.9096"
```

```
## [1] "The R^2 (coefficient of determination of the regression) is: 0.8274"
```

We can observe that, even if we are clearly underestimating, a pattern between estimated and observed distances can be surely seen . This is confirmed also by the coefficient of correlation we obtained.

9 We now try non-metric multidimensional scaling using the isoMDs instruction. We use a random initial configuration. Make a plot of the two-dimensional solution. Do the results support that the data come from one homogeneous population? Try some additional runs of isoMDS with different initial configurations, or eventually using the classical metric solution as the initial solution. What do you observe?

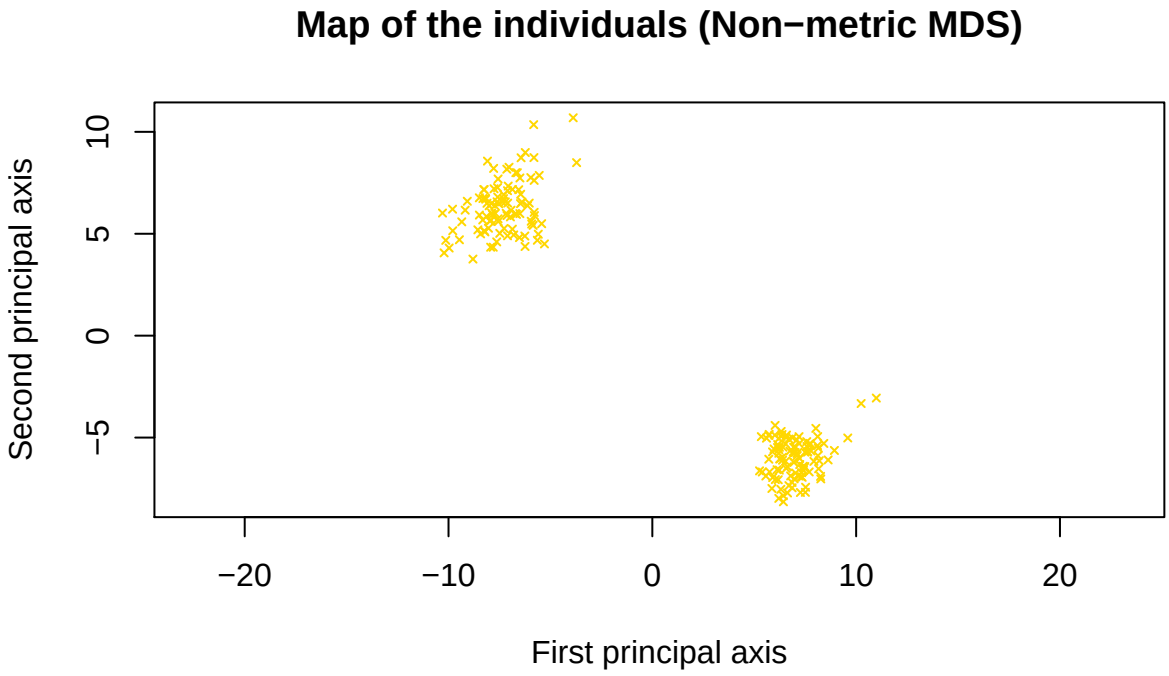
```
## initial  value 43.011318
## iter   5 value 41.680587
## iter   5 value 41.650252
## iter   5 value 41.650248
## final  value 41.650248
## converged
```



In this case, starting from a random inital configuration, the results of non-metric MDS seems to support that the data come from one homogeneous population.

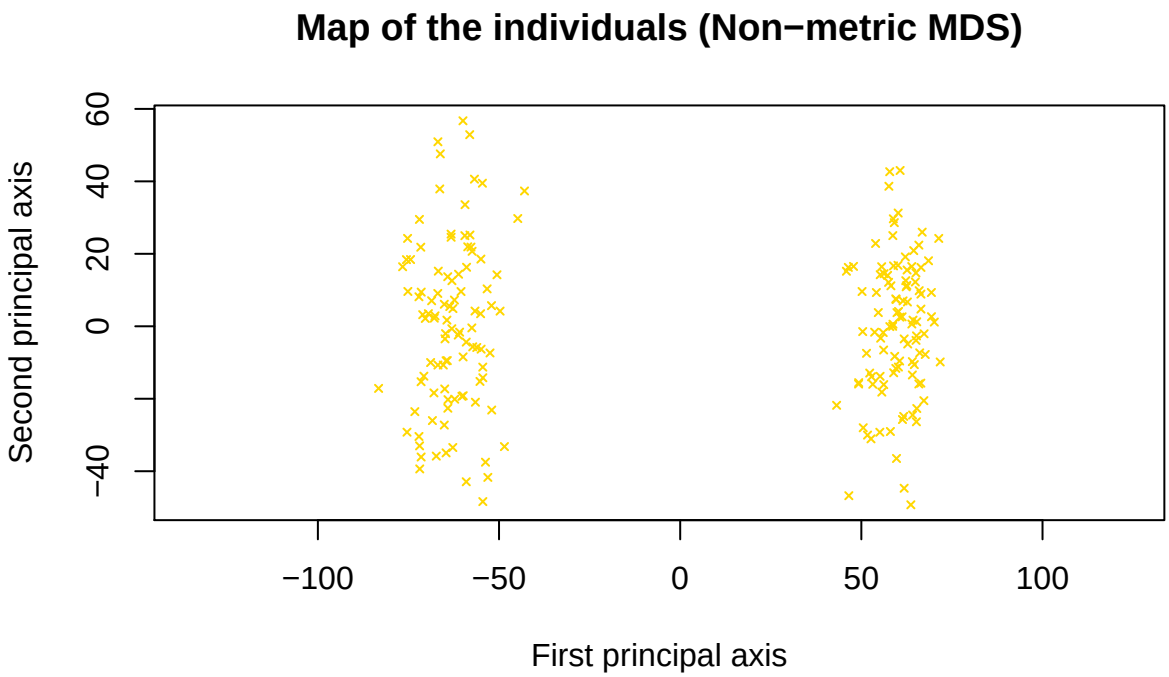
```
## initial  value 42.902690
## iter   5 value 41.405361
```

```
## iter  10 value 40.283359
## iter  15 value 38.880757
## iter  20 value 25.942854
## iter  25 value 17.332526
## iter  30 value 14.586613
## iter  35 value 13.201555
## iter  40 value 12.659129
## iter  45 value 12.503921
## final  value 12.480472
## converged
```



With different starting values, like in the metric approach, we again can support the hypotesis of two subpopulations.

```
## initial  value 16.693558
## final  value 16.692044
## converged
```

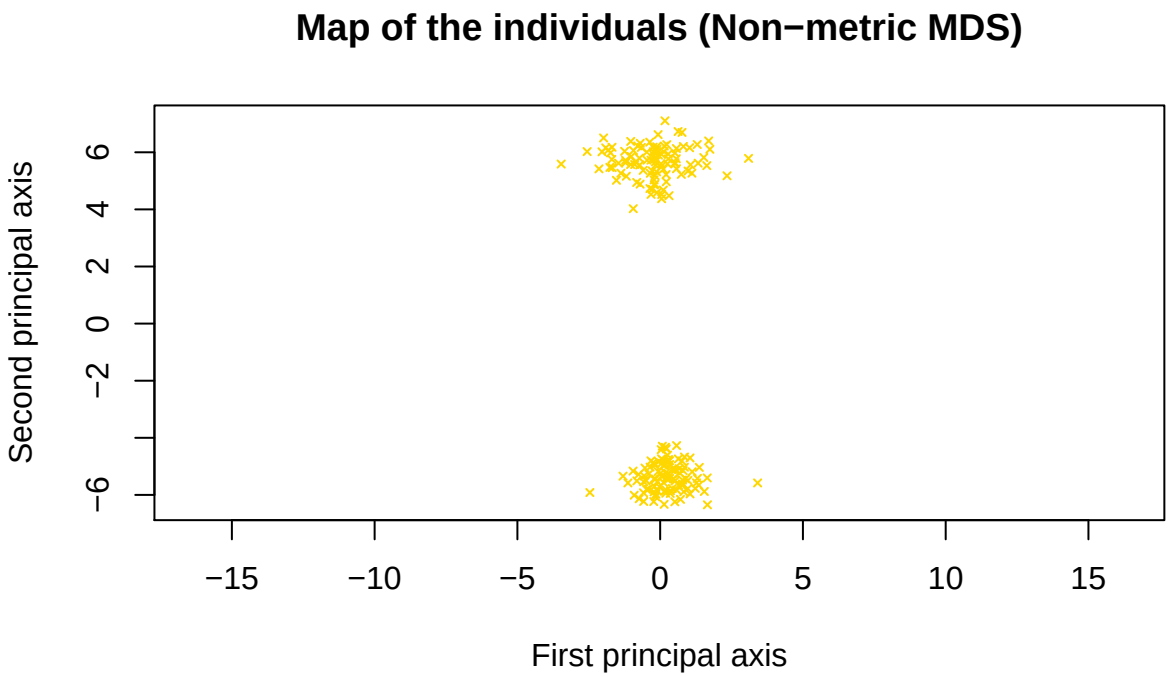


Using the classical metric (previously calculated) solution as the initial configuration, we observe that the result is very similar to the metric MDS and different from the first one with non metric.

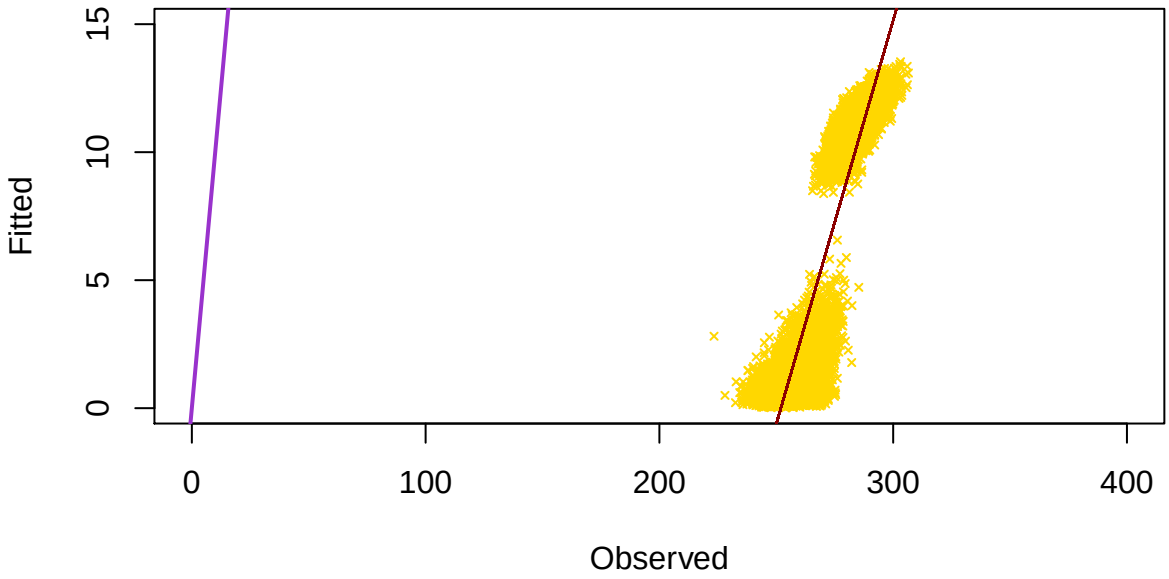
We can also observe that the stress value is smaller than the first case, but greater than the second, and that it converges in only two steps.

10 Set the seed of the random number generator to 123. Then run isoMDS a hundred times, each time using a different random initial configuration using the instructions above. Save the final stress-value and the coordinates of each run. Report the stress of the best run, and plot the corresponding map.

```
## [1] "The stress of the best run is: 12.0386"
```


Map of the individuals (Non−metric MDS)
Second principal axis
6
4
2
0
−2
−6
−15
−10
−5
0
5
10
15
First principal axis

11 Make again a plot of the estimated distances (according to your map of individuals of the best run) versus the observed distances, now for the two-dimensional solution of non-metric MDS. Regress estimated distances on observed distances and report the coefficient of determination of the regression.



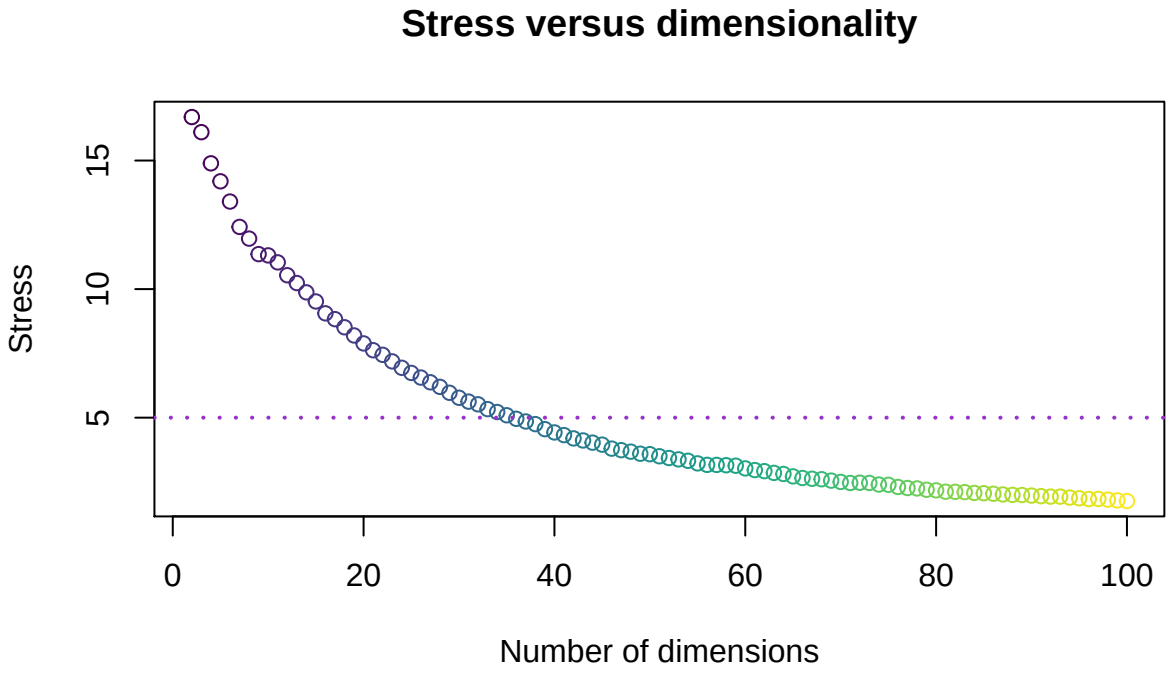

```
## [1] "The coefficient of correlation is: 0.9325"
```

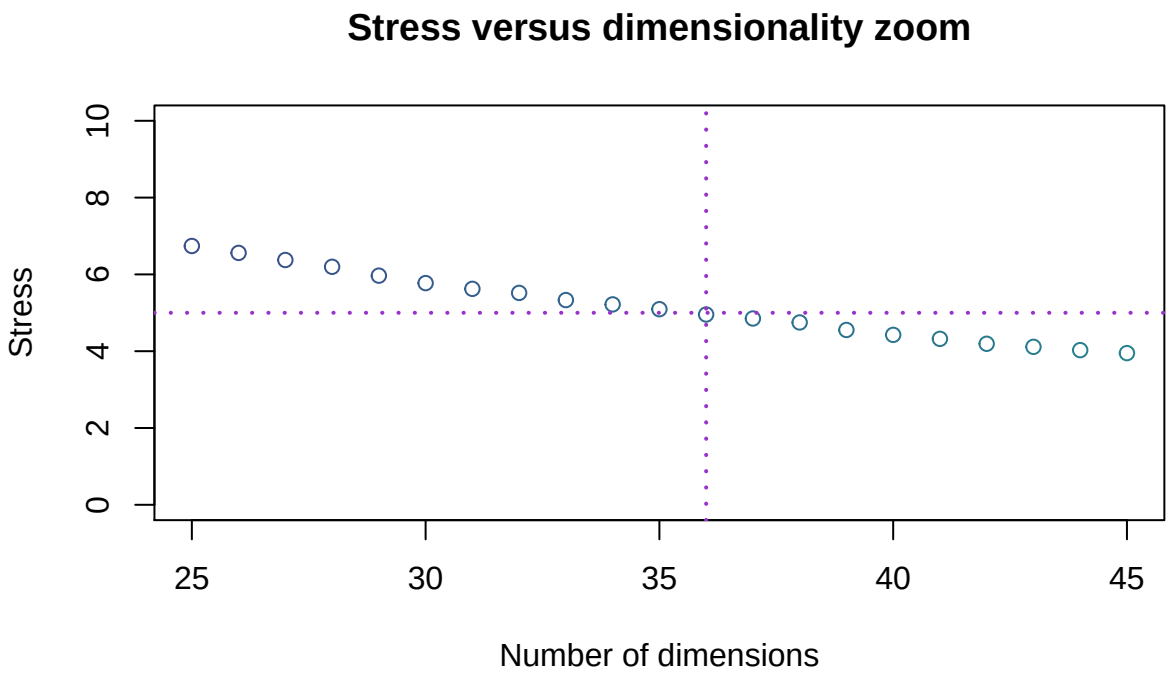
```
## [1] "The R^2 (coefficient of determination of the regression) is: 0.8696"
```

Again, we can observe that a pattern between estimated and observed distances can be surely seen . This is confirmed also by the coefficient of correlation we obtained.

Using the non-metric procedure we can observe greater underestimation.

12 **Compute the stress for a $1, 2, 3, 4, ..., n$-dimensional solution, always using the classical MDS solution as an initial configuration. How many dimensions are necessary to obtain a good representation with a stress below 5? Make a plot of the stress against the number of dimensions.**





We plotted the stress against the number of dimensions only for the first 100, due to the exploding computational cost of $isoMDS$ with respect to the number of dimensions.

We can see that, a good representation with a stress below 5 is reached starting from a $k = 36$ where the stress is:

```
## [1] 4.95528
```

13 Compute the correlation matrix between the first two dimensions of a metric MDS and the two-dimensional solution of your best non-metric MDS. Make a scatterplot matrix of the 4 variables. Comment on your findings.

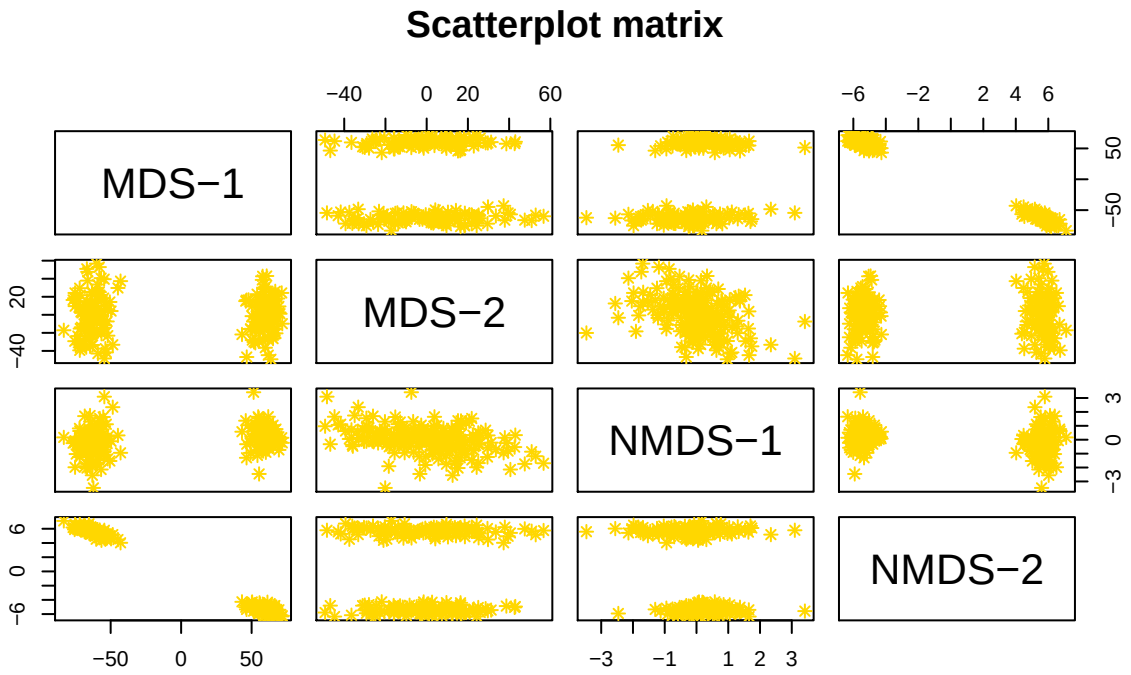


Table 5: Correlation matrix

	MDS-1	MDS-2	NMDS-1	NMDS-2
MDS-1	1.00	0.00	0.26	-1.00
MDS-2	0.00	1.00	-0.38	0.01
NMDS-1	0.26	-0.38	1.00	-0.25
NMDS-2	-1.00	0.01	-0.25	1.00

We can observe that MDS-1 is perfectly negativly correlated with the NMDS-2. MDS-2 is also negativly correlated with the NMDS-1, but more weakly.