Bioinformatics and Statistical Genetics (taught by Jan Graffelman), HW5, Population Substructure

Sten-Oliver Salumaa, Denys Kovalenko

11 January 2017

Task1

The file SNPChr20.rda contains genotype information of 310 individuals of unknown background. The genotype information concerns 50.000 SNPs on chromosome 20. Load this data into the R environment. The data file contains a matrix Y containing the allele counts (0,1 or 2) for 50.000 SNPs for one of the alleles of each SNP

Task 2

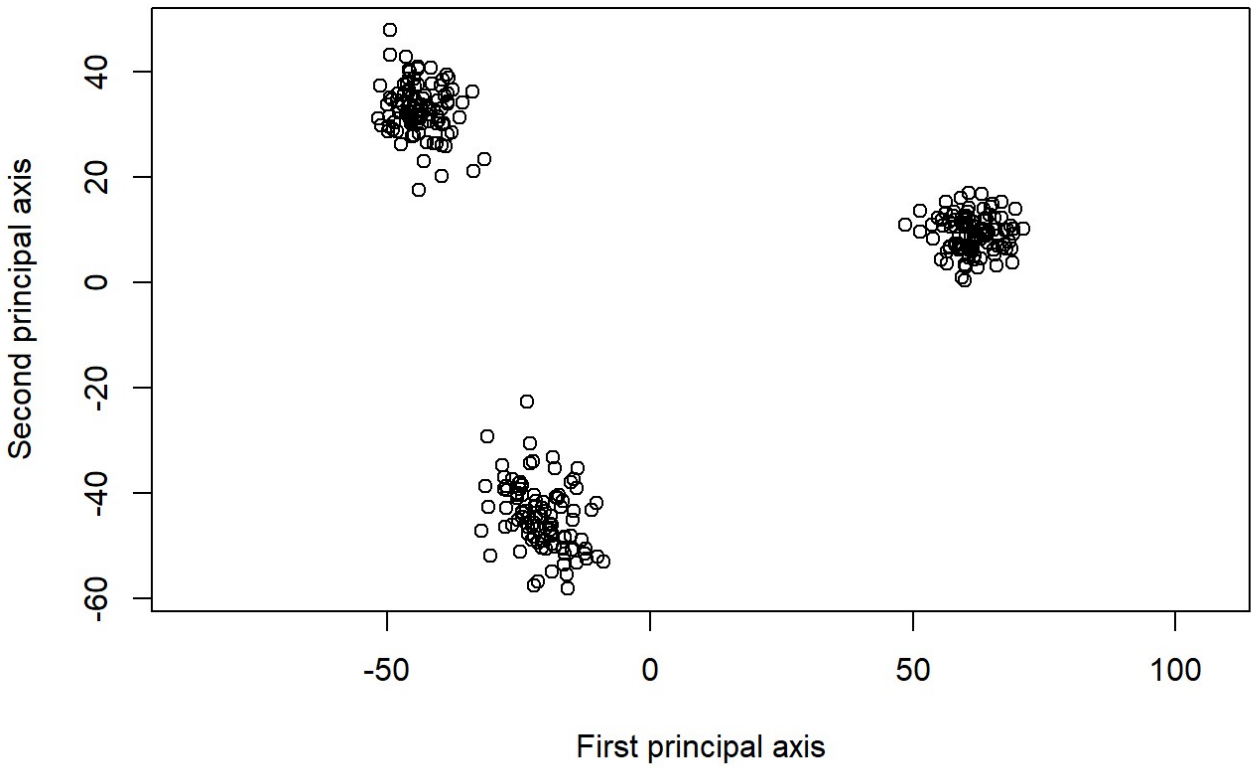
Compute the Manhattan distance matrix between the 310 individuals (this may take a few minutes) Include a submatrix of dimension 5 by 5 with the distances between the first 5 individuals in your report

```
## [1] "First 5 rows and columns of manhattan distance calculations: "
```

```
##          1        2        3        4        5
## 1     0.00 21309.95 21817.27 23580.92 21658.35
## 2 21309.95     0.00 22046.01 23013.26 24039.87
## 3 21817.27 22046.01     0.00 23125.79 22148.54
## 4 23580.92 23013.26 23125.79     0.00 22468.54
## 5 21658.35 24039.87 22148.54 22468.54     0.00
```

Task 3

Use metric multidimensional scaling to obtain a map of the individuals, and include your map in your report. Do you think the data come from one homogeneous population?



This map clearly shows that there are 3 different population groups involved. This data did not come from a homogenous population.

Task 4

Report the eigenvalues of the solution

```
##   [1] 6.524331e+05 3.256228e+05 5.508015e+04 5.047210e+04 4.717706e+04
##   [6] 4.503322e+04 4.438521e+04 4.328556e+04 4.221278e+04 3.998915e+04
##  [11] 3.902996e+04 3.819061e+04 3.804863e+04 3.614293e+04 3.570468e+04
##  [16] 3.470124e+04 3.434065e+04 3.409082e+04 3.372721e+04 3.328329e+04
##  [21] 3.255422e+04 3.201295e+04 3.157350e+04 3.139962e+04 3.096795e+04
##  [26] 3.014435e+04 2.939971e+04 2.932939e+04 2.884059e+04 2.867022e+04
##  [31] 2.811682e+04 2.785178e+04 2.757721e+04 2.748621e+04 2.642479e+04
##  [36] 2.612584e+04 2.605793e+04 2.552455e+04 2.548869e+04 2.546581e+04
##  [41] 2.528546e+04 2.506110e+04 2.482176e+04 2.460843e+04 2.431108e+04
##  [46] 2.375799e+04 2.349604e+04 2.332180e+04 2.315661e+04 2.294385e+04
##  [51] 2.286457e+04 2.272770e+04 2.260502e+04 2.220800e+04 2.218336e+04
##  [56] 2.202067e+04 2.180242e+04 2.154286e+04 2.126854e+04 2.111021e+04
##  [61] 2.087280e+04 2.070304e+04 2.056596e+04 2.041437e+04 2.030141e+04
##  [66] 2.007698e+04 1.991275e+04 1.985443e+04 1.967515e+04 1.939239e+04
##  [71] 1.928890e+04 1.925872e+04 1.915481e+04 1.896660e+04 1.877343e+04
##  [76] 1.858343e+04 1.840947e+04 1.833190e+04 1.813520e+04 1.801637e+04
##  [81] 1.800065e+04 1.791092e+04 1.782183e+04 1.768937e+04 1.759356e+04
##  [86] 1.740482e+04 1.732892e+04 1.716827e+04 1.709265e+04 1.688996e+04
##  [91] 1.681012e+04 1.672969e+04 1.660805e+04 1.652847e+04 1.637972e+04
##  [96] 1.632735e+04 1.614731e+04 1.606775e+04 1.595217e+04 1.593876e+04
## [101] 1.587884e+04 1.570392e+04 1.566171e+04 1.559656e+04 1.535959e+04
## [106] 1.528466e+04 1.518583e+04 1.513546e+04 1.508110e+04 1.482615e+04
## [111] 1.475799e+04 1.466435e+04 1.457458e+04 1.452418e+04 1.445685e+04
## [116] 1.435810e+04 1.426481e+04 1.412325e+04 1.397366e+04 1.392949e+04
## [121] 1.390017e+04 1.382192e+04 1.369941e+04 1.369294e+04 1.352703e+04
## [126] 1.345406e+04 1.340497e+04 1.332826e+04 1.325664e+04 1.318402e+04
## [131] 1.313288e+04 1.296303e+04 1.287716e+04 1.282917e+04 1.278038e+04
## [136] 1.266425e+04 1.257729e+04 1.255026e+04 1.248216e+04 1.243381e+04
## [141] 1.231310e+04 1.226461e+04 1.214640e+04 1.211881e+04 1.205942e+04
## [146] 1.199521e+04 1.188695e+04 1.185828e+04 1.184953e+04 1.173104e+04
## [151] 1.164178e+04 1.160066e+04 1.146251e+04 1.143574e+04 1.133263e+04
## [156] 1.125026e+04 1.120487e+04 1.118672e+04 1.107409e+04 1.104034e+04
## [161] 1.101705e+04 1.092330e+04 1.081364e+04 1.079584e+04 1.075655e+04
## [166] 1.062713e+04 1.055110e+04 1.054487e+04 1.053474e+04 1.040035e+04
## [171] 1.039007e+04 1.036951e+04 1.033071e+04 1.026000e+04 1.024265e+04
## [176] 1.012499e+04 1.009025e+04 9.949817e+03 9.929392e+03 9.902202e+03
## [181] 9.851000e+03 9.753018e+03 9.744371e+03 9.631667e+03 9.612502e+03
## [186] 9.541772e+03 9.488489e+03 9.441596e+03 9.376694e+03 9.304079e+03
## [191] 9.243250e+03 9.207796e+03 9.163975e+03 9.120757e+03 9.080327e+03
## [196] 9.004981e+03 8.970866e+03 8.902072e+03 8.863539e+03 8.831060e+03
## [201] 8.780892e+03 8.686072e+03 8.647069e+03 8.609512e+03 8.558079e+03
## [206] 8.509885e+03 8.450458e+03 8.409983e+03 8.329061e+03 8.280091e+03
## [211] 8.253636e+03 8.224925e+03 8.150115e+03 8.131877e+03 8.071189e+03
## [216] 8.044426e+03 7.949554e+03 7.904572e+03 7.823776e+03 7.788971e+03
## [221] 7.746268e+03 7.711607e+03 7.686578e+03 7.647894e+03 7.602426e+03
## [226] 7.572630e+03 7.517144e+03 7.465507e+03 7.400401e+03 7.374045e+03
## [231] 7.356652e+03 7.243419e+03 7.223269e+03 7.169423e+03 7.153400e+03
## [236] 7.041951e+03 7.004298e+03 6.975544e+03 6.964631e+03 6.906782e+03
## [241] 6.870536e+03 6.833465e+03 6.767837e+03 6.704481e+03 6.672993e+03
## [246] 6.616796e+03 6.564603e+03 6.551354e+03 6.508328e+03 6.479023e+03
## [251] 6.415621e+03 6.377467e+03 6.313915e+03 6.272123e+03 6.206887e+03
## [256] 6.159553e+03 6.134428e+03 6.089739e+03 6.060420e+03 6.012341e+03
## [261] 5.914628e+03 5.893287e+03 5.855619e+03 5.808768e+03 5.732301e+03
## [266] 5.702690e+03 5.681233e+03 5.632219e+03 5.599155e+03 5.574440e+03
```

```
## [271] 5.516696e+03 5.494541e+03 5.448738e+03 5.368852e+03 5.311237e+03
## [276] 5.276036e+03 5.217112e+03 5.174368e+03 5.140099e+03 5.092034e+03
## [281] 5.019027e+03 5.008921e+03 4.937990e+03 4.914669e+03 4.888803e+03
## [286] 4.866518e+03 4.810177e+03 4.786995e+03 4.769432e+03 4.730525e+03
## [291] 4.677002e+03 4.596888e+03 4.585019e+03 4.551388e+03 4.519305e+03
## [296] 4.481720e+03 4.406442e+03 4.356352e+03 4.259636e+03 4.209979e+03
## [301] 4.150176e+03 4.066897e+03 4.025879e+03 3.973774e+03 3.929820e+03
## [306] 3.890330e+03 3.773802e+03 3.733280e+03 3.708671e+03 1.160627e-10
```

Task 5

Is the distance matrix you have used an Euclidean distance matrix?

For the values in Task 2 we used Manhattan distance and for the calculations to plot the multidimensional scaling results we used euclidian distance. Also, the function ‘is.euclid’ from library ‘ade4’ confirms our answer.

```
library(ade4)
D2 <- dist_Y_euc
is.euclid(D2)
```

```
## [1] TRUE
```

Task 6

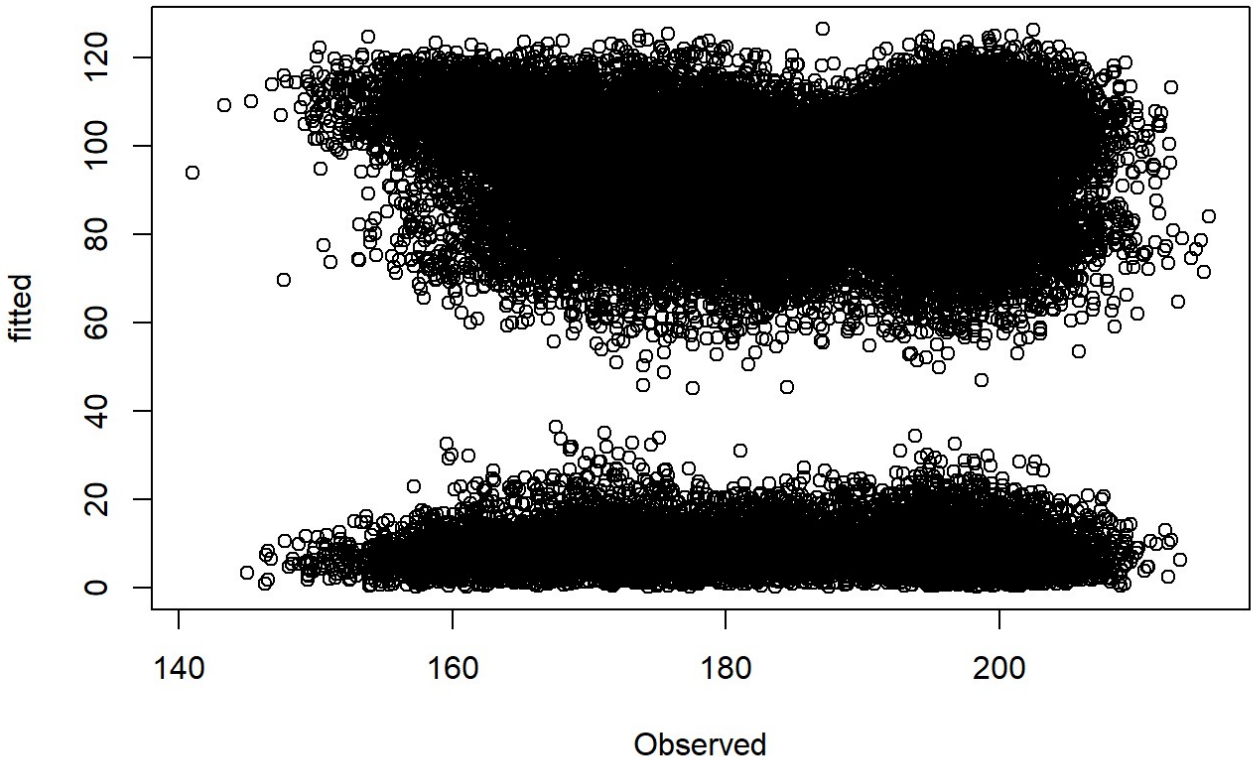
What is the goodness-of-fit of a two-dimensional approximation to your distance matrix?

Here we are applying the same ‘cmdscale’ function to the previously calculated eucledian distances but this time with ‘k’ value as 2. Reporting the GOF:

```
## [1] 0.1835 0.1835
```

Task 7

Make a plot of the estimated distances (according to your map of individuals) versus the observed distances.



Regress estimated distances on observed distances and report the coefficient of determination of the regression

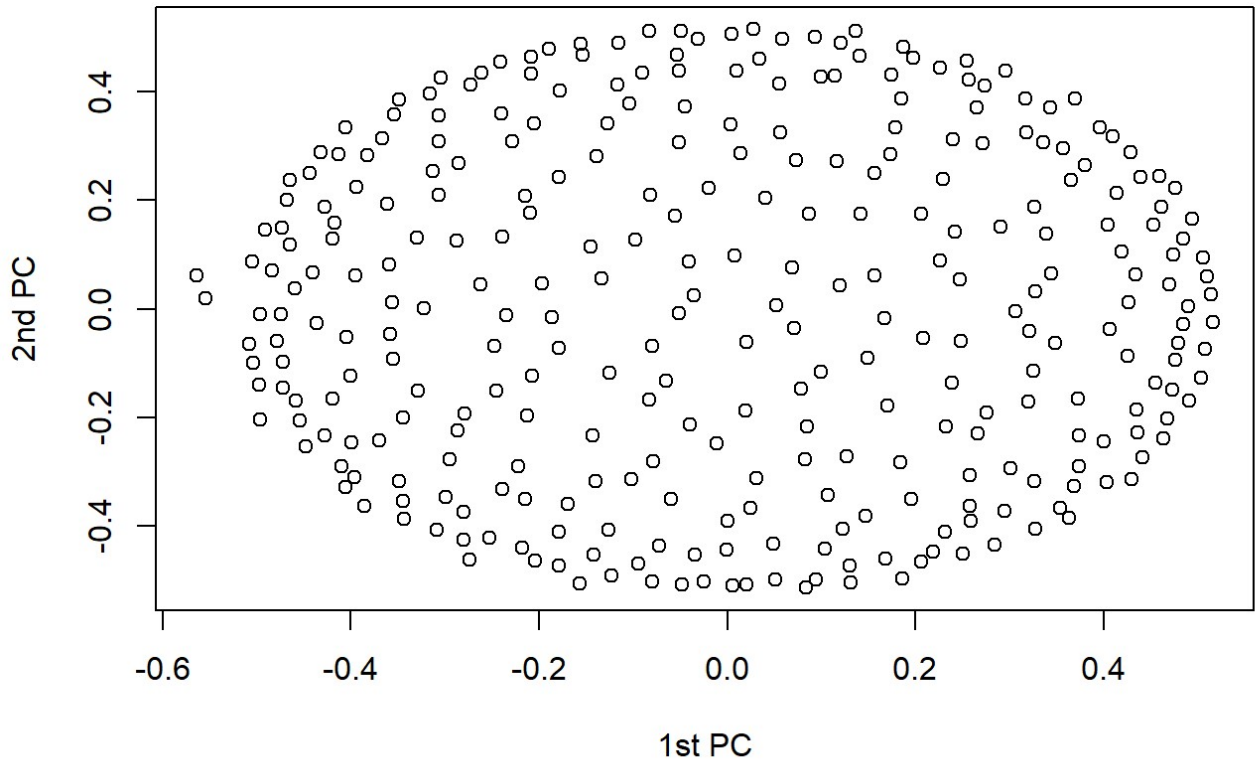
```
## [1] "R-squared value: "
```

```
## [1] 0.0001151277
```

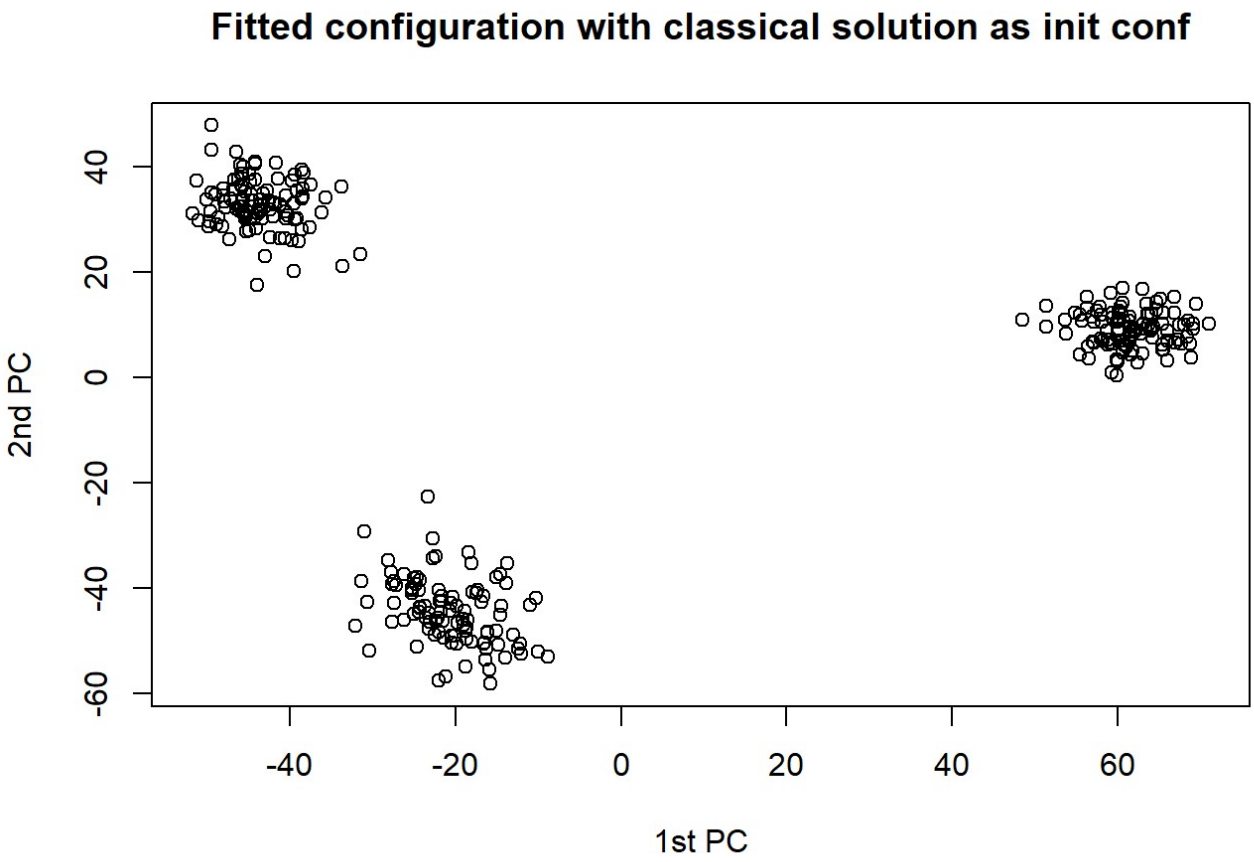
Task 8

Try now non-metric multidimensional scaling with your distance matrix. Use both a random initial configuration as well as the classical metric solution as an initial solution. Make a plot of the two-dimensional solution. Do the results support that the data come from one homogeneous population?

```
## initial  value 42.651222
## final  value 41.824552
## converged
```



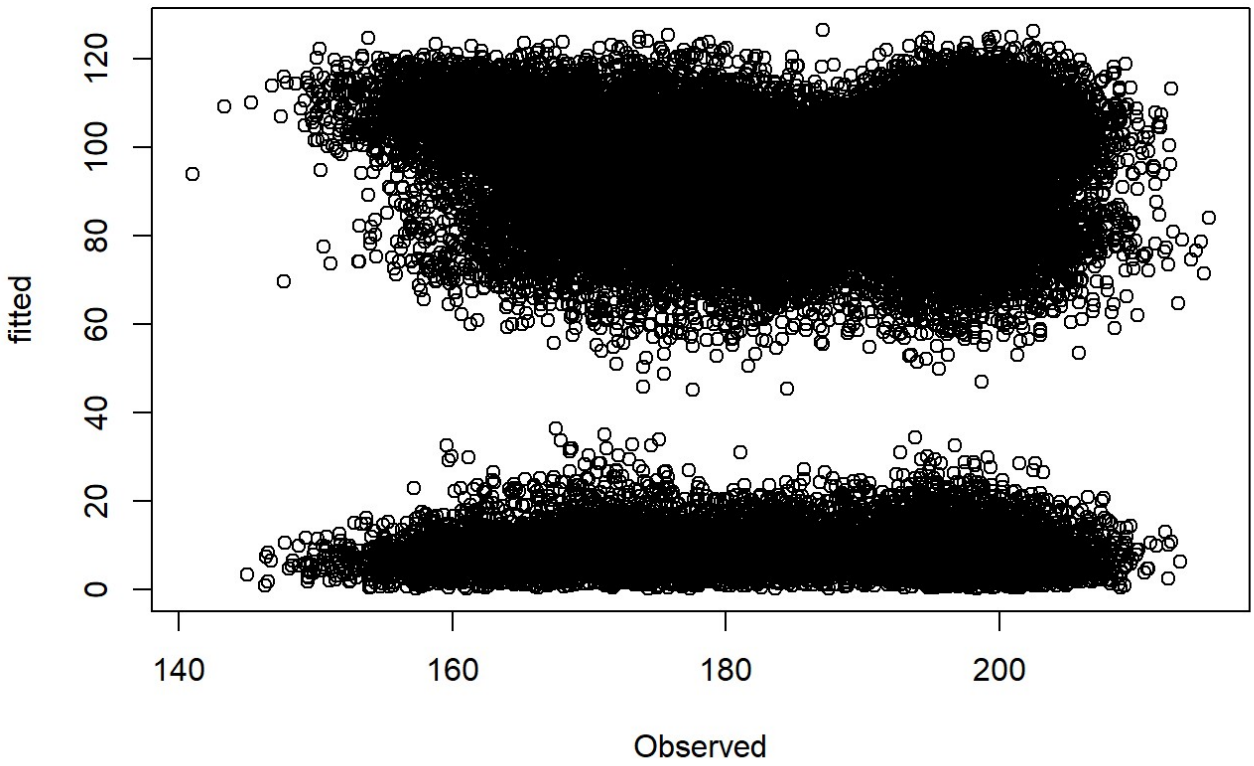

```
## initial  value 19.959595
## final  value 19.957419
## converged
```



No, we got to see 3 clusters with NMDS for both classical and and random initial solutions, which supports that data comes not from one homogeneous population.

Task 9.

(1p) Make again a plot of the estimated distances (according to your map of individuals) 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. Is the fit better or worse than with metric MDS?

```
## [1] "R-squared value: "
```

```
## [1] 0.0001150912
```

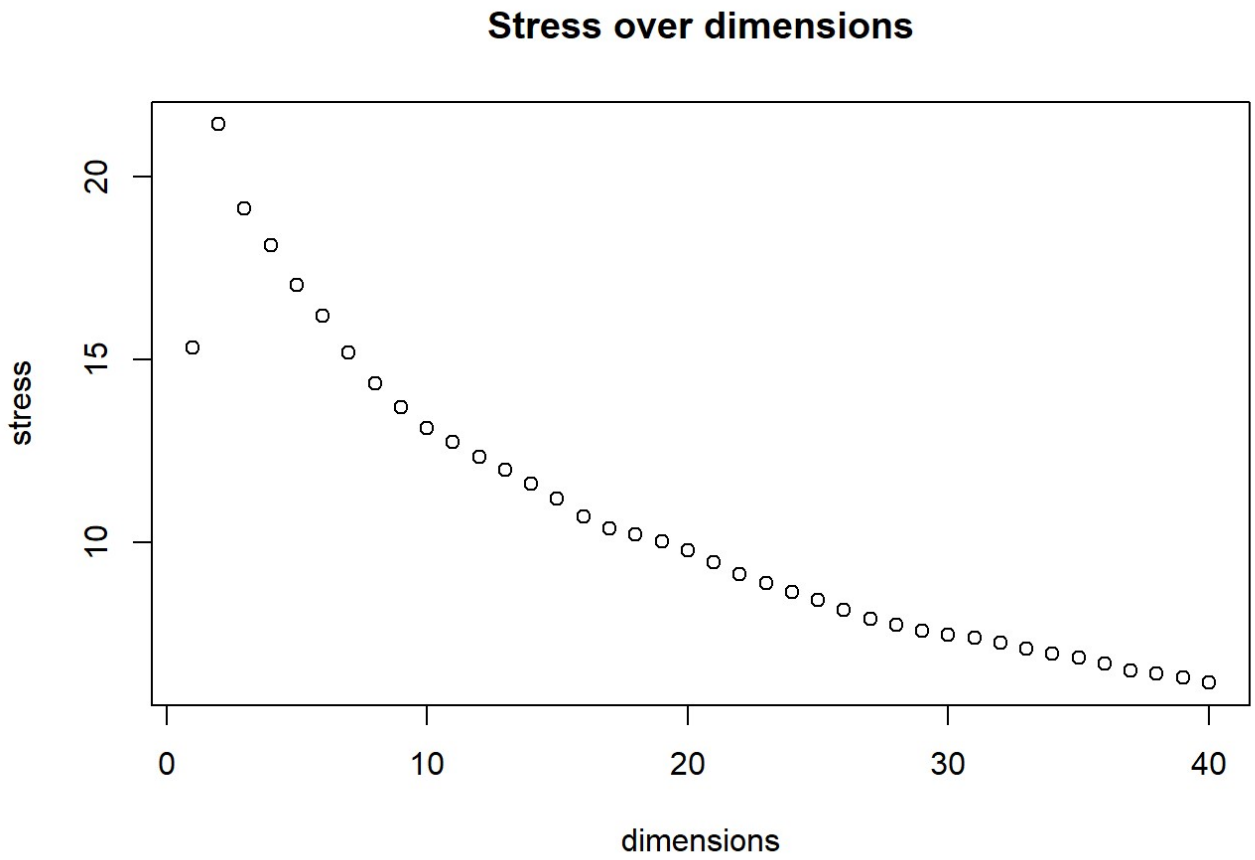
The fit is as good as with metric MDS.

Task 10(1p) 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? Make a plot of the stress against the number of dimensions

```
## initial  value 15.326162
## final  value 15.326162
## converged
## initial  value 21.442371
## final  value 21.442371
## converged
## initial  value 19.146292
## final  value 19.146292
## converged
## initial  value 18.127964
## final  value 18.127964
## converged
## initial  value 17.041411
## final  value 17.041411
## converged
## initial  value 16.212276
## final  value 16.212276
## converged
## initial  value 15.199191
## final  value 15.199191
## converged
## initial  value 14.337953
## final  value 14.337953
## converged
## initial  value 13.695651
## final  value 13.695651
## converged
## initial  value 13.116964
## final  value 13.116964
## converged
## initial  value 12.754196
## final  value 12.754196
## converged
## initial  value 12.339105
## final  value 12.339105
## converged
## initial  value 11.977386
## final  value 11.977386
## converged
## initial  value 11.588622
## final  value 11.588622
## converged
## initial  value 11.181171
## final  value 11.181171
## converged
## initial  value 10.694685
## final  value 10.694685
## converged
## initial  value 10.387073
## final  value 10.387073
## converged
## initial  value 10.220250
## final  value 10.220250
## converged
```

```
## initial  value 10.020136
## final  value 10.020136
## converged
## initial  value 9.783582
## final  value 9.783582
## converged
## initial  value 9.456209
## final  value 9.456209
## converged
## initial  value 9.125281
## final  value 9.125281
## converged
## initial  value 8.878197
## final  value 8.878197
## converged
## initial  value 8.634316
## final  value 8.634316
## converged
## initial  value 8.402523
## final  value 8.402523
## converged
## initial  value 8.151222
## final  value 8.151222
## converged
## initial  value 7.908000
## final  value 7.907999
## converged
## initial  value 7.745245
## final  value 7.745245
## converged
## initial  value 7.583432
## final  value 7.583432
## converged
## initial  value 7.475113
## final  value 7.475113
## converged
## initial  value 7.377914
## final  value 7.377914
## converged
## initial  value 7.253147
## final  value 7.253147
## converged
## initial  value 7.091661
## final  value 7.091661
## converged
## initial  value 6.947376
## final  value 6.947376
## converged
## initial  value 6.836533
## final  value 6.836533
## converged
## initial  value 6.684968
## final  value 6.684968
## converged
## initial  value 6.493151
```

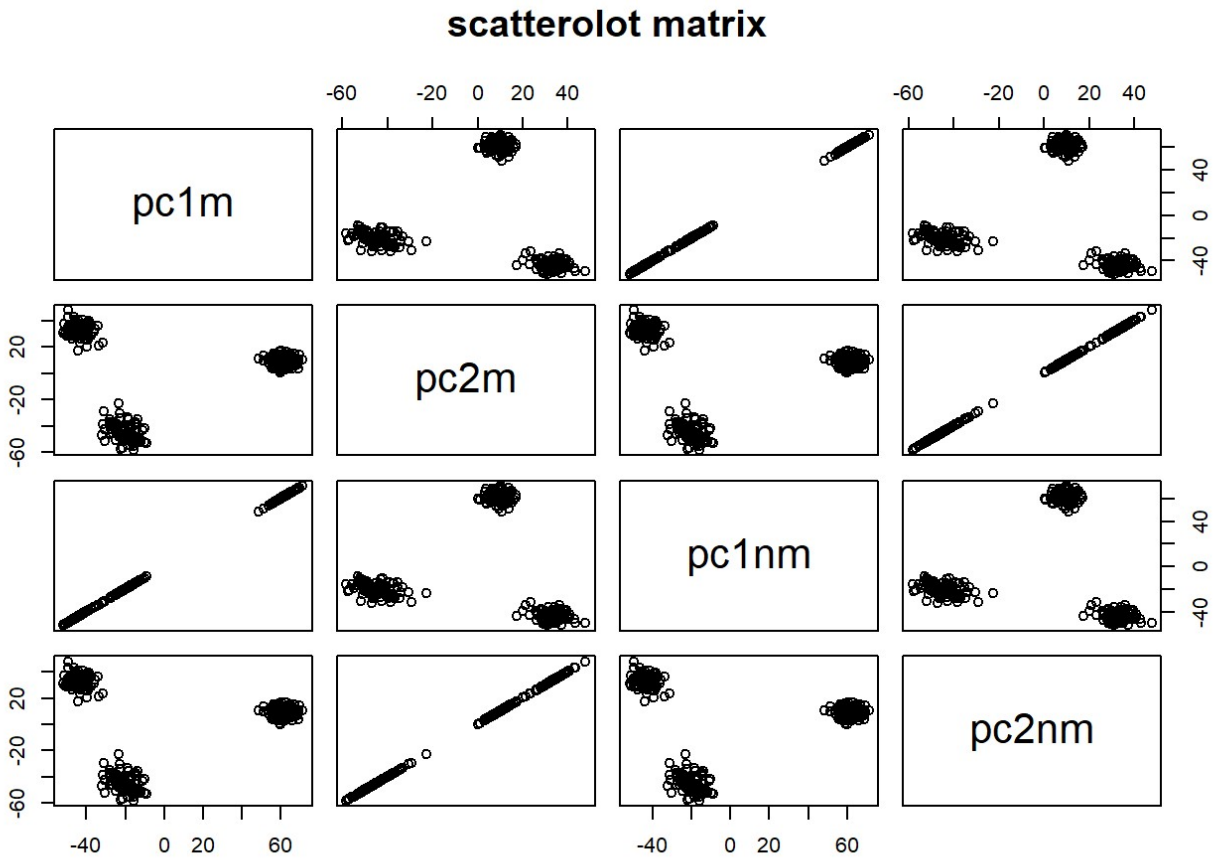
```
## final  value 6.493151
## converged
## initial  value 6.401223
## final  value 6.401222
## converged
## initial  value 6.287904
## final  value 6.287904
## converged
## initial  value 6.150093
## final  value 6.150093
## converged
```



After 15, stress decreases more slowly. Altough on observable dimensions rate of decrease stays the same value of stress itself is pretty low at 30 already.

Task 11.

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



Metric and non-metric are clearly very strongly correlated. We can even see that in this case it could be even said that they’re the same.Both show 3 homogeneous clusters so dataset is heterogenous.