CODICEPERSONA_PROBLEMA

Marco Scarpelli

DATA

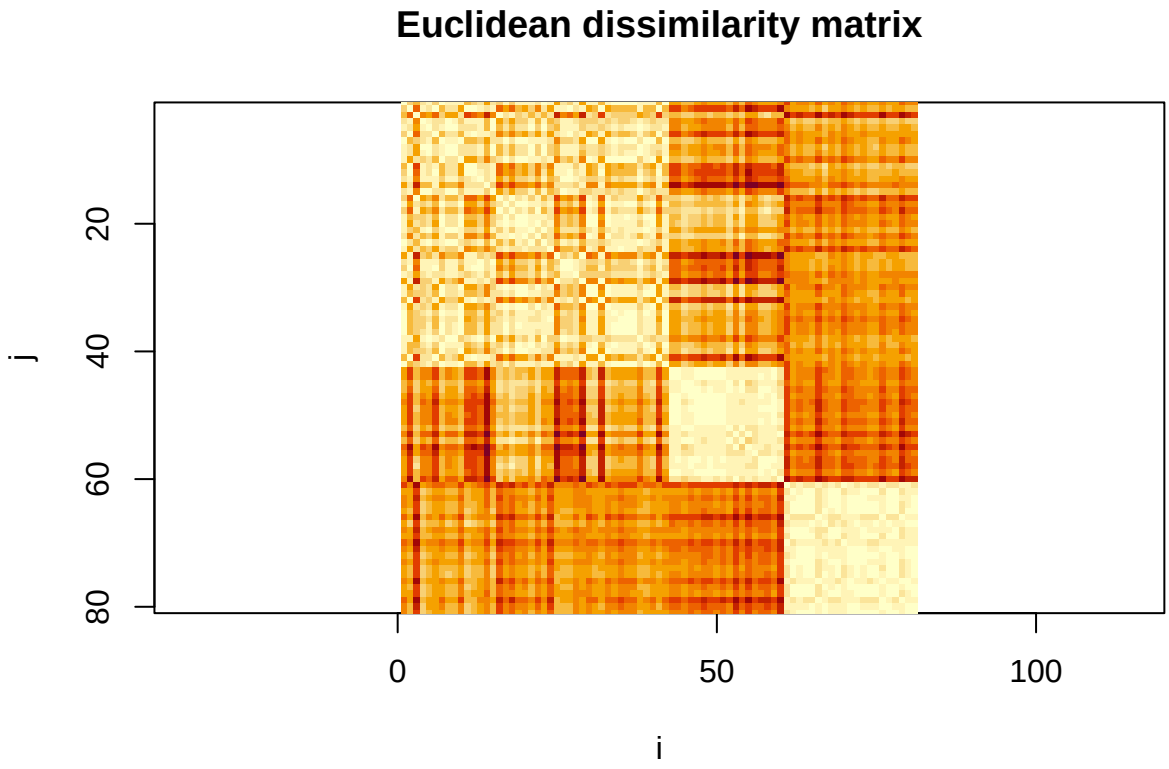
Print dataframe

```
##   major_axis eccentricity
## 1      0.589        0.924
## 2      0.741        0.901
## 3      0.427        0.930
## 4      0.673        0.861
## 5      0.643        0.897
## 6      0.728        0.930
```

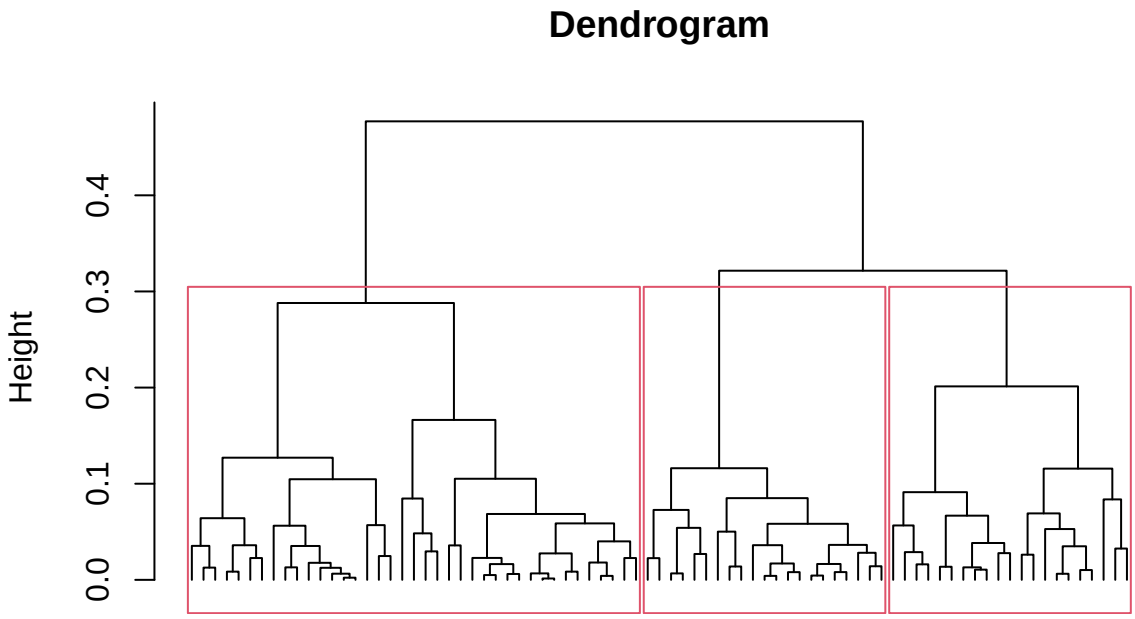
```
## [1] 81  2
```

Point a

We compute the dissimilarity matrix and plot it.



Let us run the algorithm with complete linkage and plot the dendrogram, drawing a box around $k = 3$:

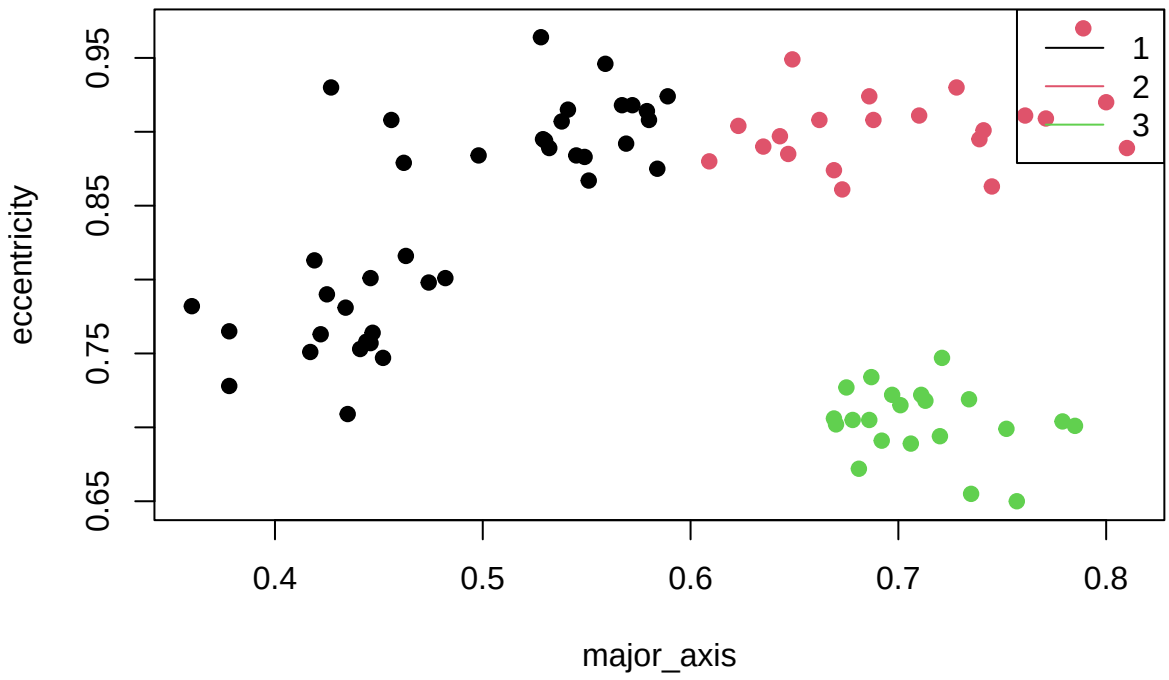


We cut into 3 clusters and report the results. First, the amount of elements per cluster:

```
## clusters
##  1  2  3
## 39 21 21
```

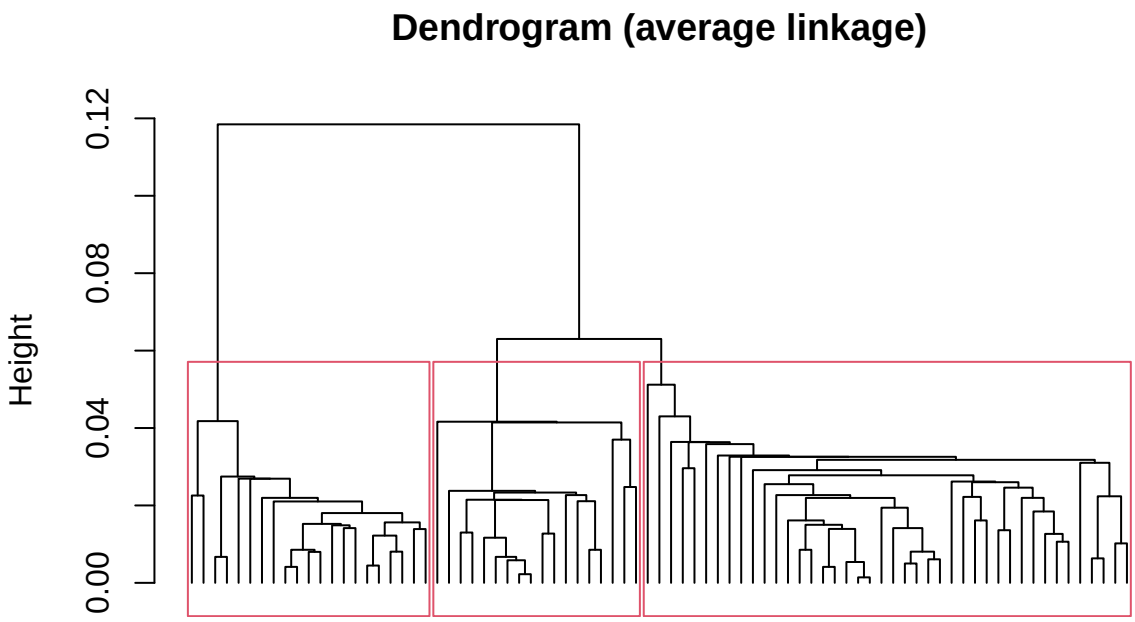
Then, the cluster means:

```
##        major_axis eccentricity
## mean_1  0.4884103    0.8428462
## mean_2  0.7037143    0.9037619
## mean_3  0.7118571    0.7036667
```

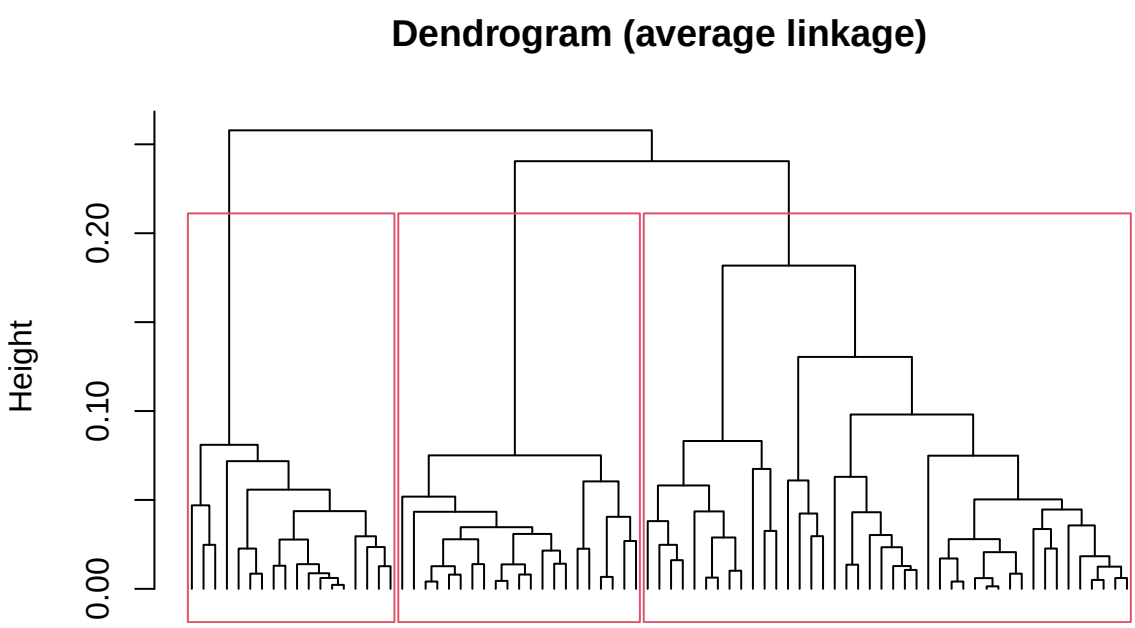


Point b

We can see from the plot that some points in the top-right cluster were incorrectly identified as being part of the top-right one. I will try the single-linkage method, still with Euclidean distance.



however, the dendrogram does not look too good, in that we are cutting at a point of high "instability". Let us try average linkage:

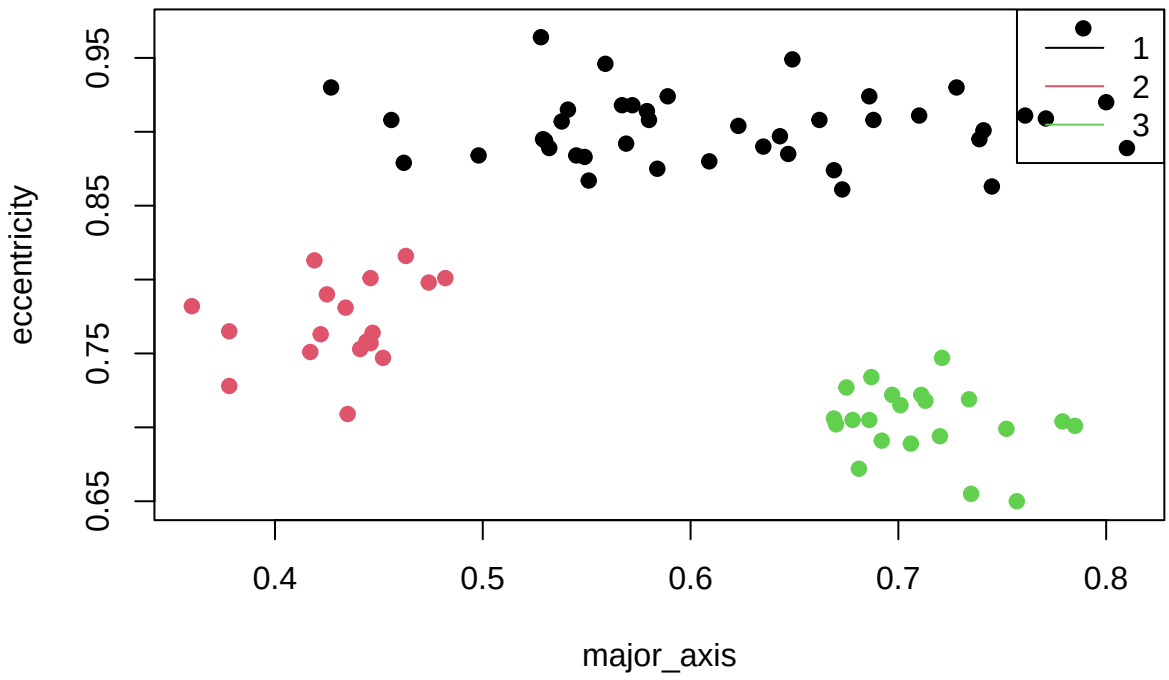


This dendrogram is better, so we will carry on with average linkage. Next a report on elements in each cluster:

```
## clusters
##  1  2  3
## 42 18 21
```

and cluster means:

```
##        major_axis eccentricity
## mean_1  0.6205476    0.9041190
## mean_2  0.4312778    0.7709444
## mean_3  0.7118571    0.7036667
```



Point C

We test each cluster separately, so that is equivalent to performing tests on the mean of an univariate Gaussian; let us first verify Gaussianity of major_axis within each cluster:

```
##
##  Shapiro-Wilk normality test
##
## data:  df[i1, 1]
## W = 0.96596, p-value = 0.2401
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df[i2, 1]
## W = 0.92703, p-value = 0.1721
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df[i3, 1]
## W = 0.92705, p-value = 0.1201
```

all three p-values are above 10%, so we can say that the data is Gaussian. Furthermore, we assume the data to be independent.

Now, we perform the Bonferroni test; even the multivariate one will still only look at the diagonal of the matrix, so we can do that and only extrapolate the component we want, or perform an univariate test. This

is because we artificially remove any interaction to be able to make more discoveries.

$k = 3$ since we test 3 means and 3 variables

Mean:

```
##       inf    center       sup
## 0.5780689 0.6205476 0.6630263
```

```
##       inf    center       sup
## 0.4083724 0.4312778 0.4541831
```

```
##       inf    center       sup
## 0.6898297 0.7118571 0.7338846
```

Variance:

```
##         inf      center         sup
## 0.005872386 0.009860400 0.019196333
```

```
##          inf       center          sup
## 0.0004965239 0.0010605654 0.0032610417
```

```
##          inf       center          sup
## 0.0005855364 0.0011892286 0.0032866447
```