Introduction to Statistical Learning 2020/2021 Coursework

Juliette Limozin, CID: 01343907

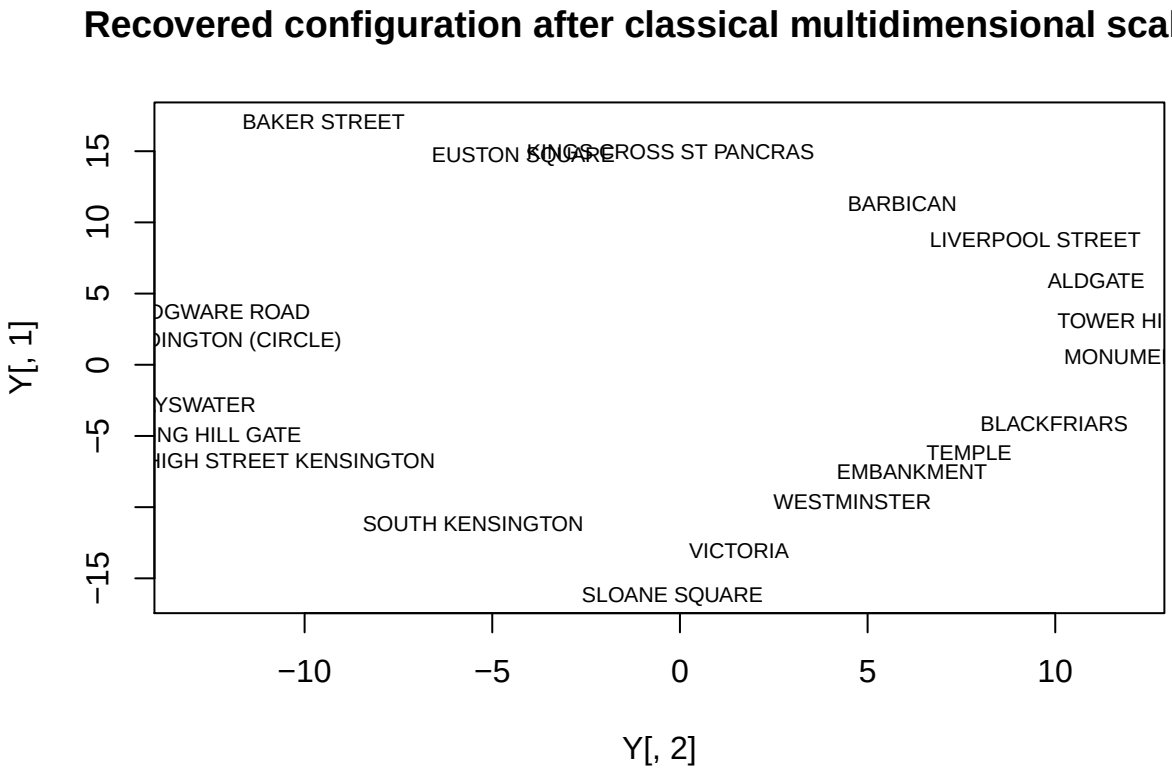
Due 25/02/2021, 5 pm

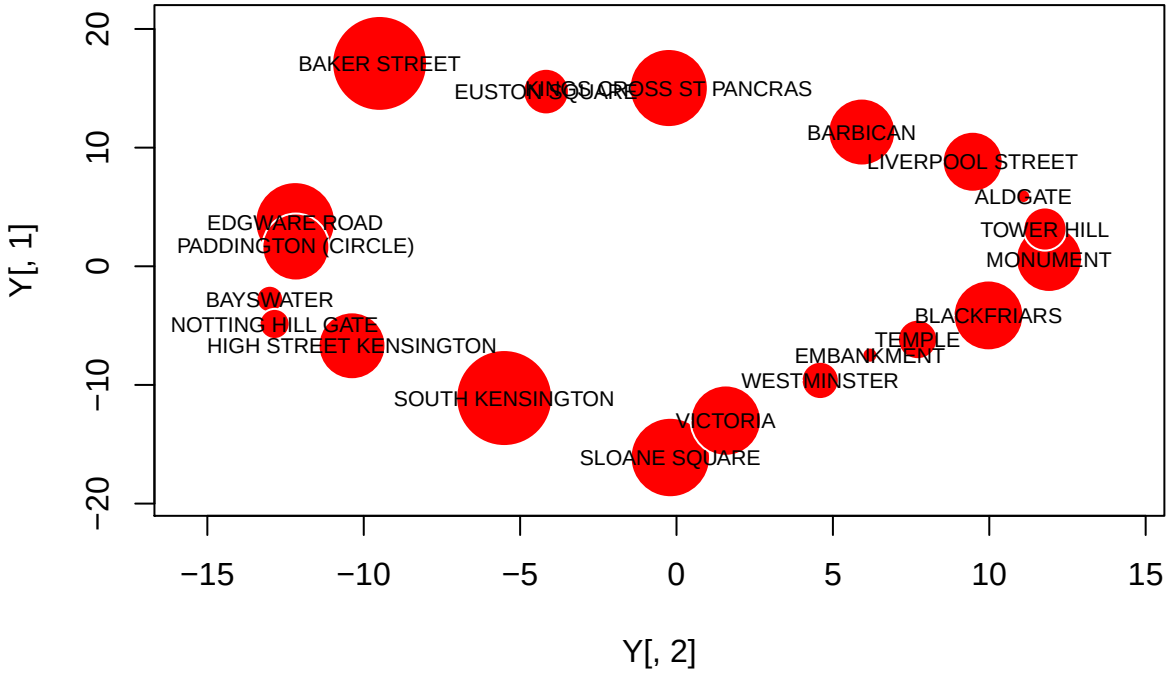
I have decided to perform analysis on the distance between Circle Line stations because it is my least favourite tube line, and I always felt like it was making quite a detour to reach different stations. I therefore took my distance measure in this coursework as the travel time between stations, accounting for any time spent when a train is stopping at a station, to prove my point of the inefficiency of this tube line.

My data was gathered from Google Maps: I calculated the travel time (in minutes), using the Circle line only, between the stations, setting an approximate departure time of 12:00 pm on Monday March 1st for consistency in the measurements. I made sure to go both clockwise and counter clockwise on the Circle line when taking the travel time between two stations, and took the rotation that took the least time. Note, whenever I was travelling clockwise and passing through Edgware Road, I had to change plaforms, which Google Maps accounted for when calculating travel times. My gathered data is a symmetrix matrix with zeros on the diagonal.

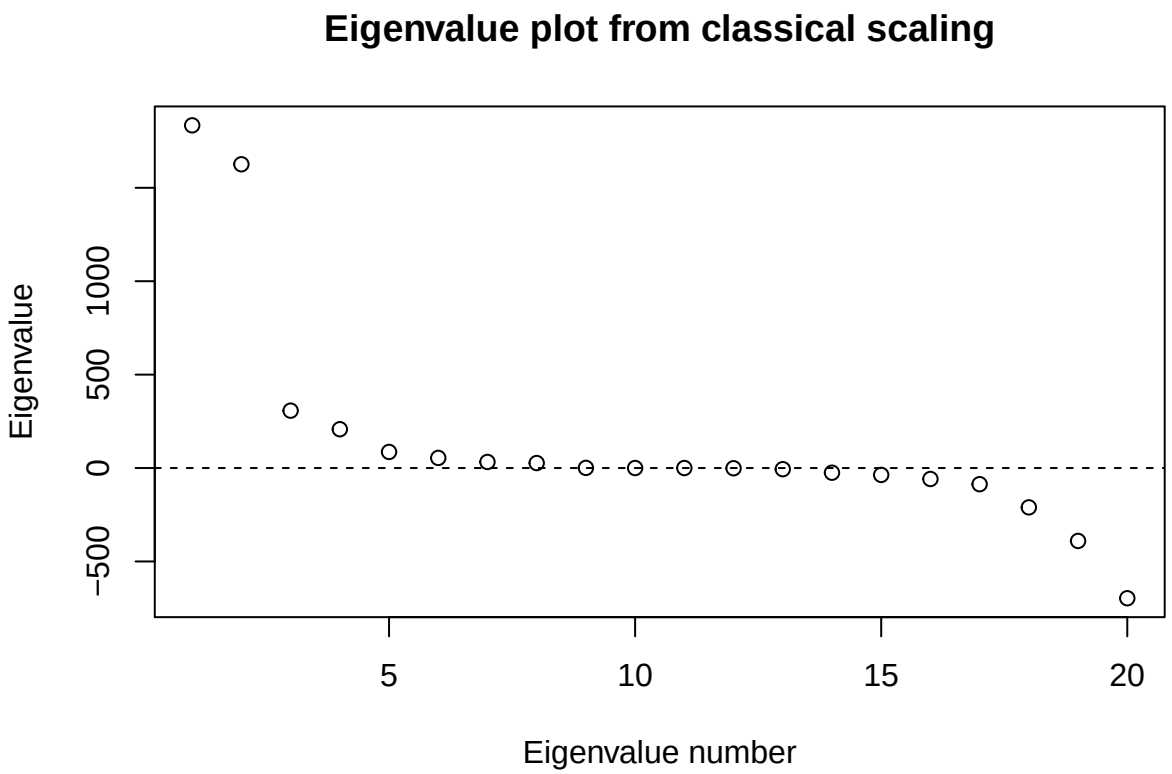
The symmetry assumption is indeed a bit weird given that the Circle line has an inner and outer rail ring, therefore making travling on the outer ring longer than on the inner ring. So the travel times are not symmetric in practice, but I considered the difference trivial enough to consider the travel times to be the same both ways.

After applying classical multidimensional scaling on the distance matrix, I get the following recovered positions of the stations from the first two dimensions, with an additional plot that includes the third dimension.



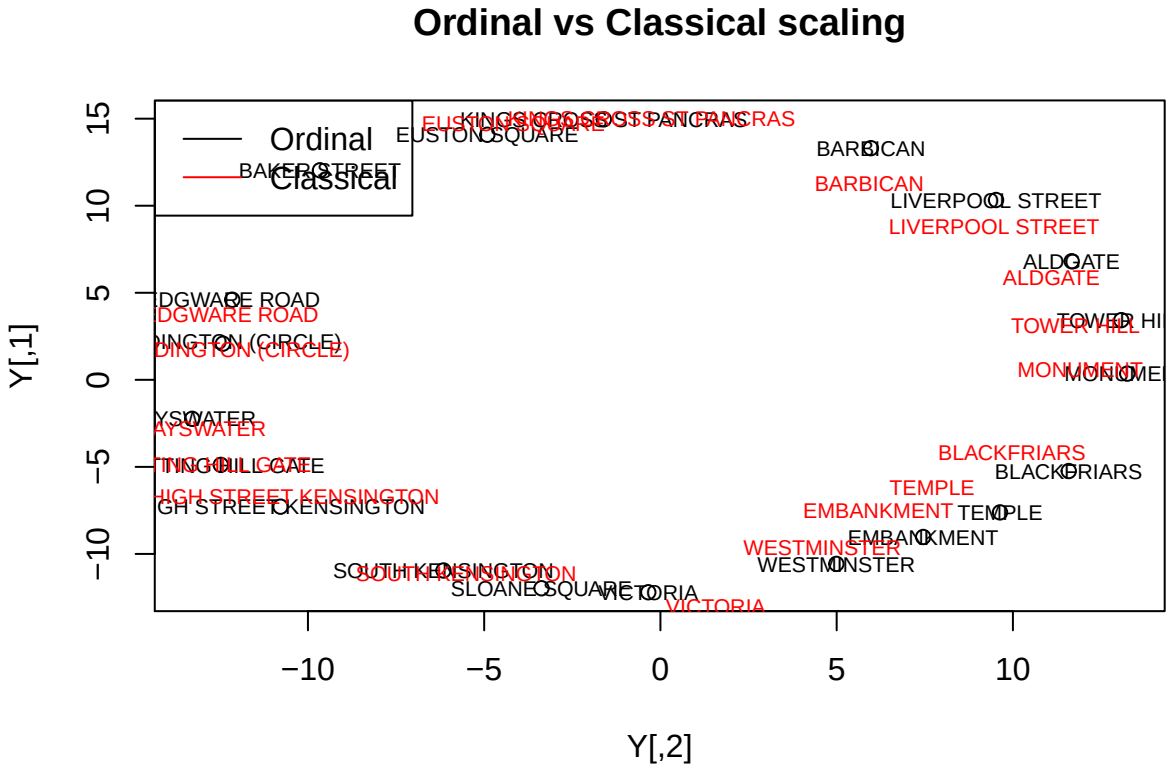


At first the recovered positions seem relatively accurate. The eigenvalue plot below shows that it is reasonable to take the first two dimensions for the recovered configuration as their eigenvalues are significantly bigger, with some interest in the third dimensions.

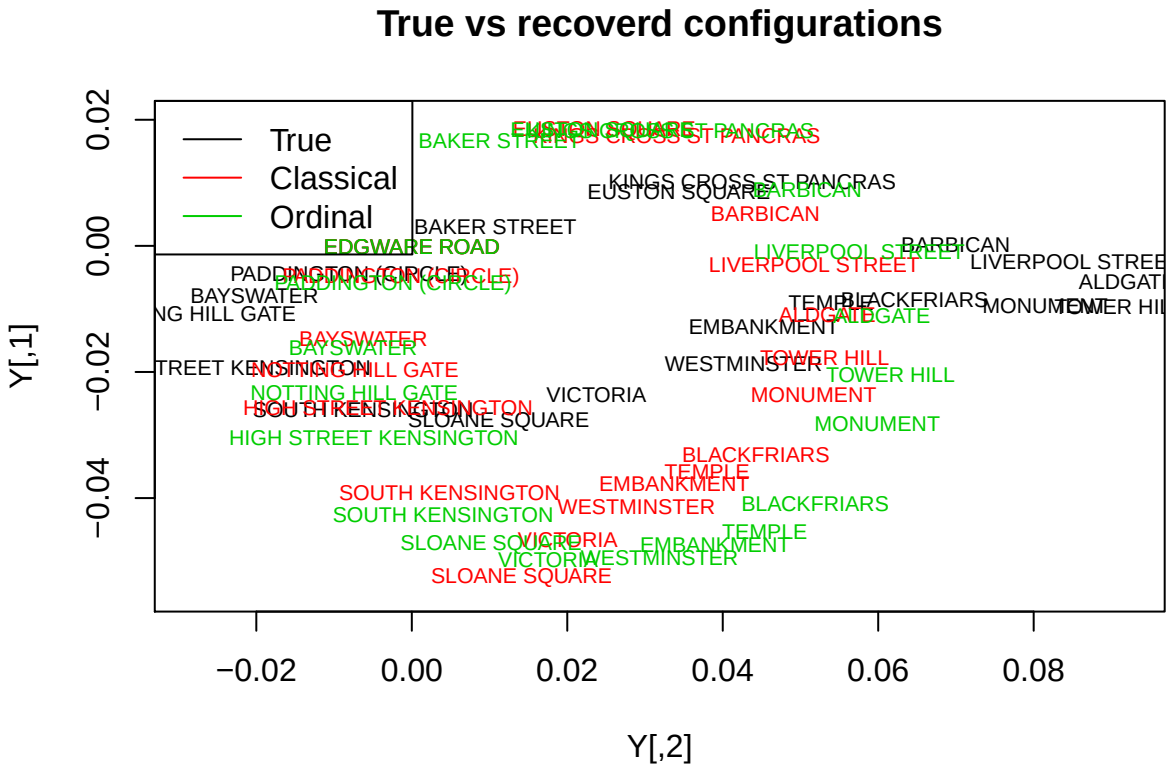


I also recovered the configuration using ordinal scaling. The plot comparing the configuration from the two scaling methods is shown below. Because we see a reduction in the stress it is expected to have an ordinal scaling solution that is different from the classical solution.

```
## initial  value 9.544264
## iter   5 value 2.596164
## final  value 2.567369
## converged
```



To compare the recovered positions with the real positions, which I gathered in a data set containing the latitude and longitude of each station, I used the Procrustes method. I then centered the configurations at Edgware Road station as I found that the translation vector I got from Procrutes was often too long. The plot below shows a comparison of the configurations.



As we can see, the recovered configurations are often quite far fom the real positions, and even more with

the ordinal scaling for stations in the South. For both scaling methods, we notice that the configurations seem to have been squashed vertically, making the stations on the West and East sides of the Circle Line closer than they really are in terms of travel time, and the opposite for stations in the North and South.

My takeaway from this is that if I want to travel between some Western and Eastern stations, the Circle Line is efficient in terms of travel time; however if I want to travel betwen Northern and Southern stations of the line, I'm much better off using a different tube line. This makes sense in reality, as for example it takes 8 minutes to travel between Victoroia and King's Cross using the Victoria line, versus 26 minutes using the Circle line.