# Statistical Learning Coursework

*Matthew Cowley, CID 1059624*

February 18, 2020

## Choice of data and motivations

The data set I have decided to investigate is the amount of meat eaten per capita for different types of meat, for a random 53 countries. The types of meat include mutton/lamb, beef, pigmeat, poultry and other. I then extended the data set to include fruit, eggs, veg and fish. The units for all the data are kg per capita for a given year and. The data was obtained from ourworldindata.org and can be found at https://ourworldindata.org/meat-production.

The business motivations for analysing this data set, is for assessing where marketing food supplements would be effective. It would also be interesting to see how the configuration produced links with a countries GDP and religion, along with other social-economic attributes.

## Analysis of results

### Figure 1

As we can see from Figure 1, a distance-based analyse appears appropriate with many eigenvalues being very small (approximately 10 are being non zero, found using a log plot). Figure 1 suggest that it is appropriate to view the configuration in 2D, maybe 3, shown by the big drop after the first 2 eigenvalues. You could argue that the third eigenvalue is large enough to be of interest.

### Figure 2

The configuration you obtain, when using classical scaling and euclidean distance is shown in figure 2. It is not very clear with lots of overlaps, but a clear group emerges of wealthy western nations e.g UK and USA, and generally poorer countries e.g Cambodia and Indonesia. Co ordinate 1 appears to roughly represent GDP per capita with a few anomaly's e.g Greece and Italy are very high up, maybe as they pride themselves on there cuisine and not there business.

**Figure 3**

However when we use the Canberra metric as in figure 3, we appear to get a clearer configuration. With a clear group representing the Indo-Asia area and highly Muslim group around Egypt. It does group country's with generally good diets together. However the eigenvalues are not as good with it suggesting 2 dimensions would not be appropriate, as result I will be using the Euclidean metric instead.

**Figure 4**

There is not much difference, but considering the ordinal scaling starts from classical scale, it is interesting that it improves the solution and that the stress converges after 3 iterations. When using the Euclidean metric the solution is the same. However the end result is not really effected, just more clustered.

**Figure 5**

The result does not change much which is good, as the ordinal scaling was initialised with a configuration created from random sample of normal distributions, which implies the configuration is fairly stable. The Ordinal scaling does seem slightly more dense, so classical scaling is preferred in this example.

**Figure 6 and 7**

It is very interesting that as time has gone on, it appears that data configurations has spread out. This could represent how the divide between rich and poor is increasing. It could also represent how peoples diets are changing e.g people being vegetarians or people really loving meat. But the best explanation is perhaps that amount of food has increased global, apart from very poor country's where population has increased.

**Figure 8**

It is reassuring to see clustering when a basic K-means has been implemented. Green being rich country's, black being poorer. However there is not a clear divide between cyan and blue, with red being upper end of dimension 2.

## Conclusions

In general figure 2 is the most useful for drawing conclusions. It depends where you want to market a given product, for example I would recommend if you want to sell an up market product you would sell to places like USA and UK and country's in the black cluster of figure 8. If you wanted a product which would improve poorer country's, I would take country's near Bangladesh. It also is a good indicator of where aid is needed or food in general is needed, e.g a cheap sustainable product would do hopefully do well for country's in this region.

These extremes appear to be increasing as time goes on seen from figures 6 and 7. So I would advise any food producing company to take this into account, e.g if you are global company diversify you menu to be more appropriate for a given country.

So overall you can use these configuration to influence where and how to market a product.

# Figures

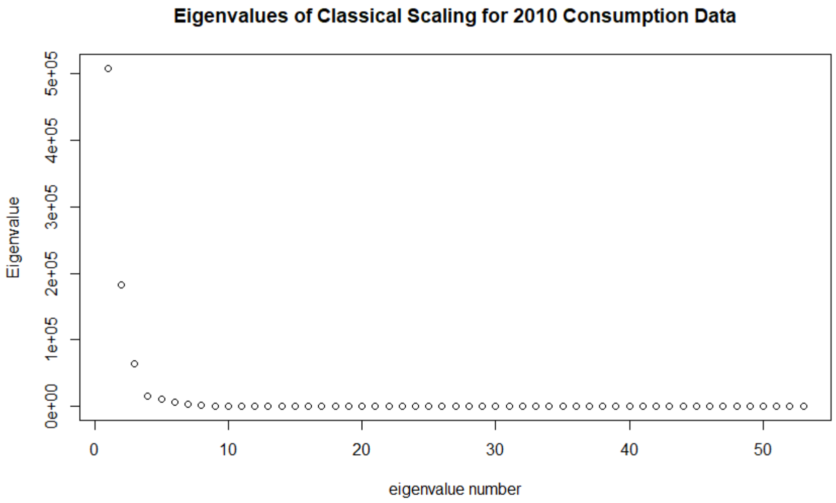

Figure 1: A graph to show the size of eigenvalues, for euclidean distance and classical scaling of the data

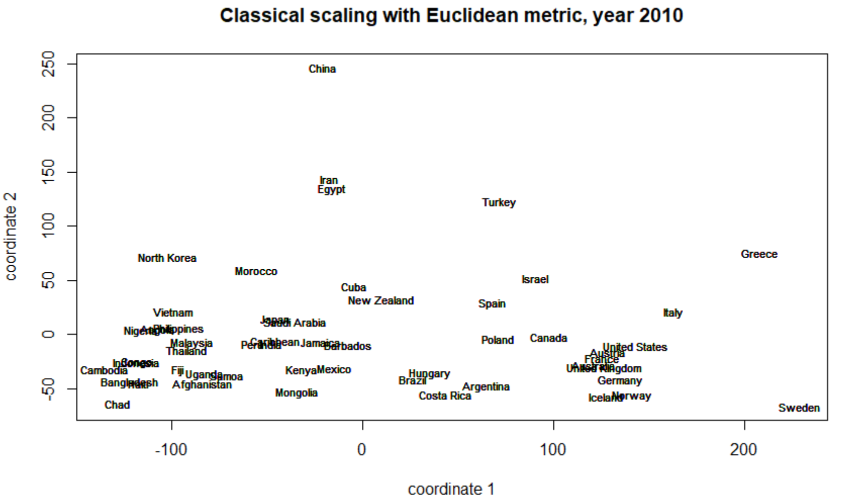

Figure 2: Configuration using classical scaling

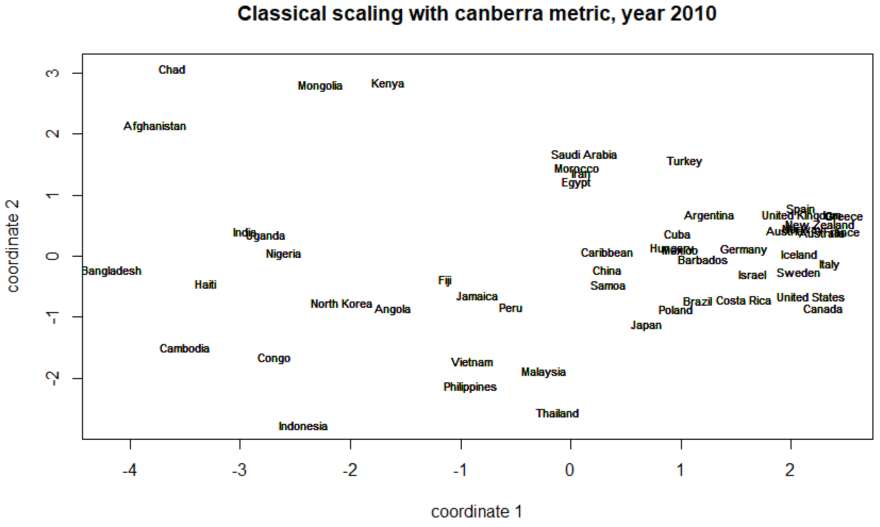

Figure 3: How changing metric effect results

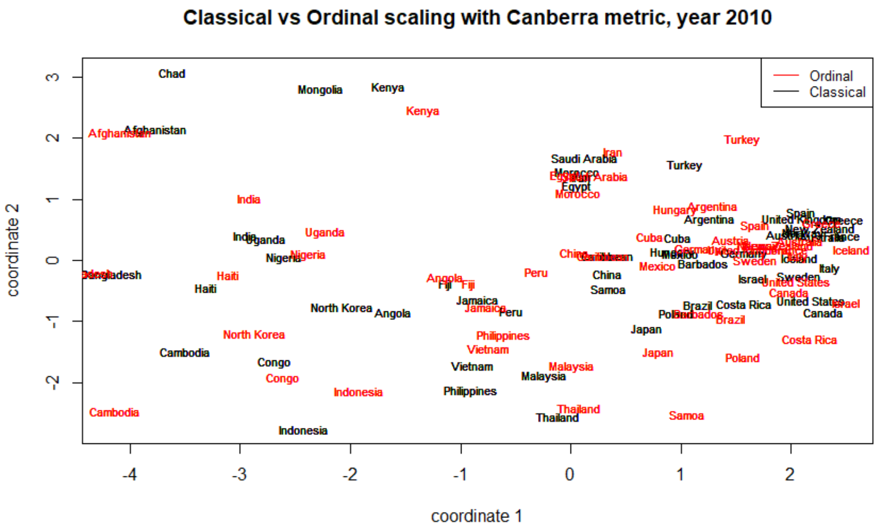

Figure 4: How Ordinal and Classical scaling compare with Canberra metric

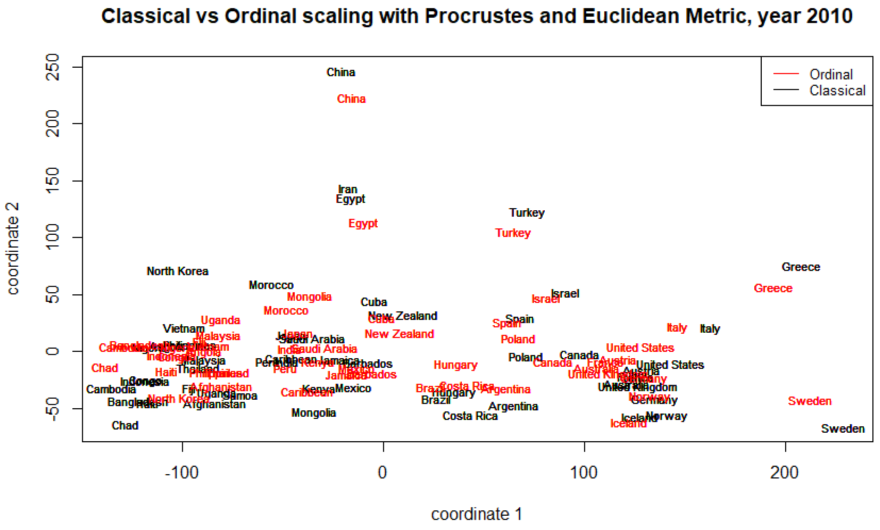

Figure 5: How Ordinal and Classical scaling compare with Euclidean metric, using Procrustes to help compare

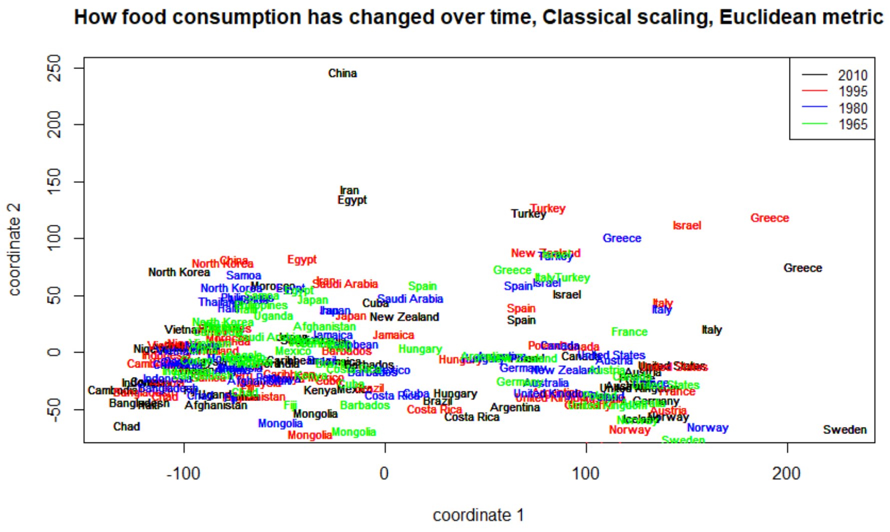

Figure 6: How food consumption has changed over time using Classical scaling and Euclidean metric, using Procrustes scaling to compare results

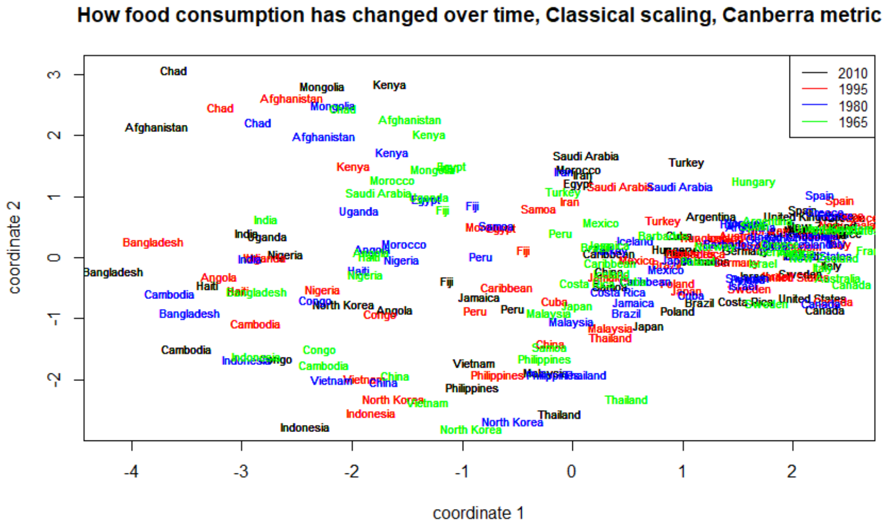

Figure 7: How food consumption has changed over time using Classical scaling and Canberra metric, using Procrustes scaling to compare results

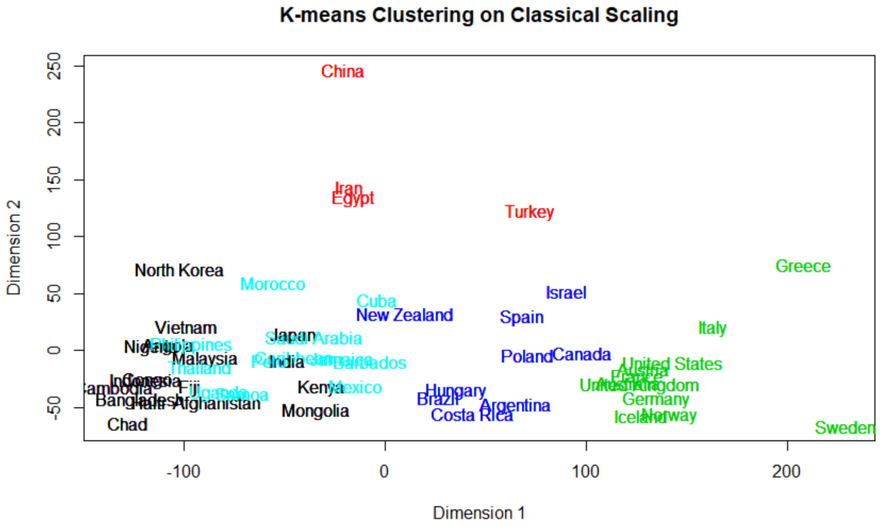

Figure 8: K means of Classical scaling with Euclidean metric and 5 clusters