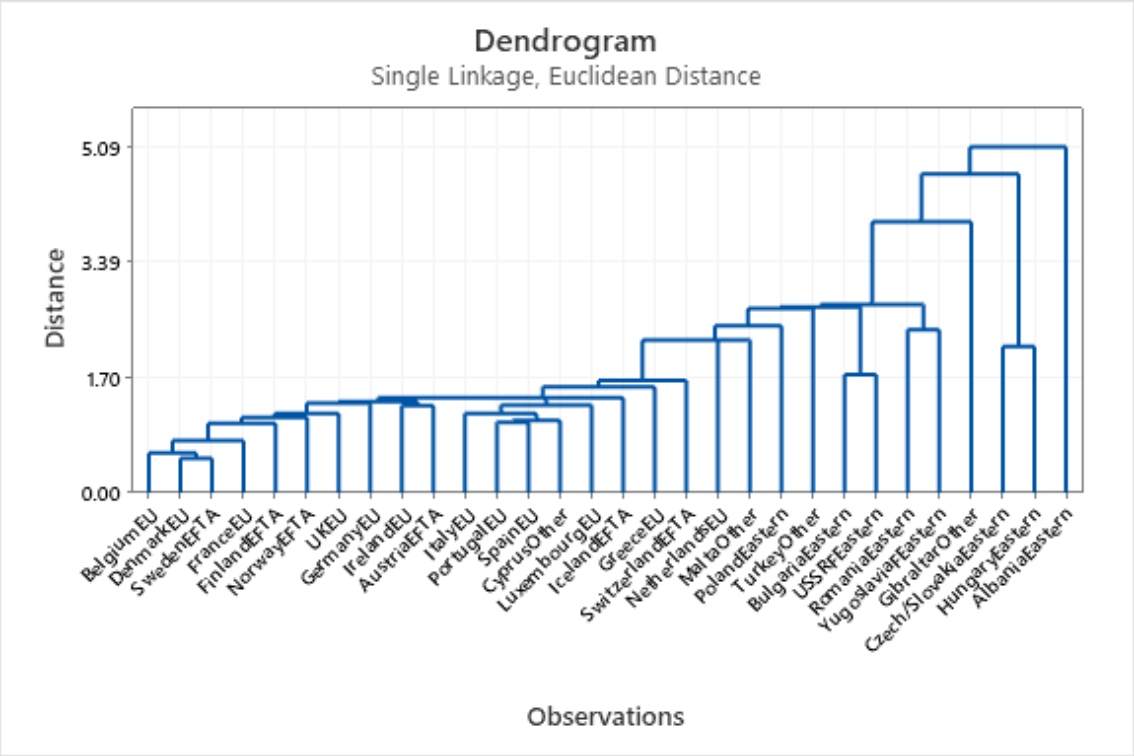
Additional Multivariate Exercise 5

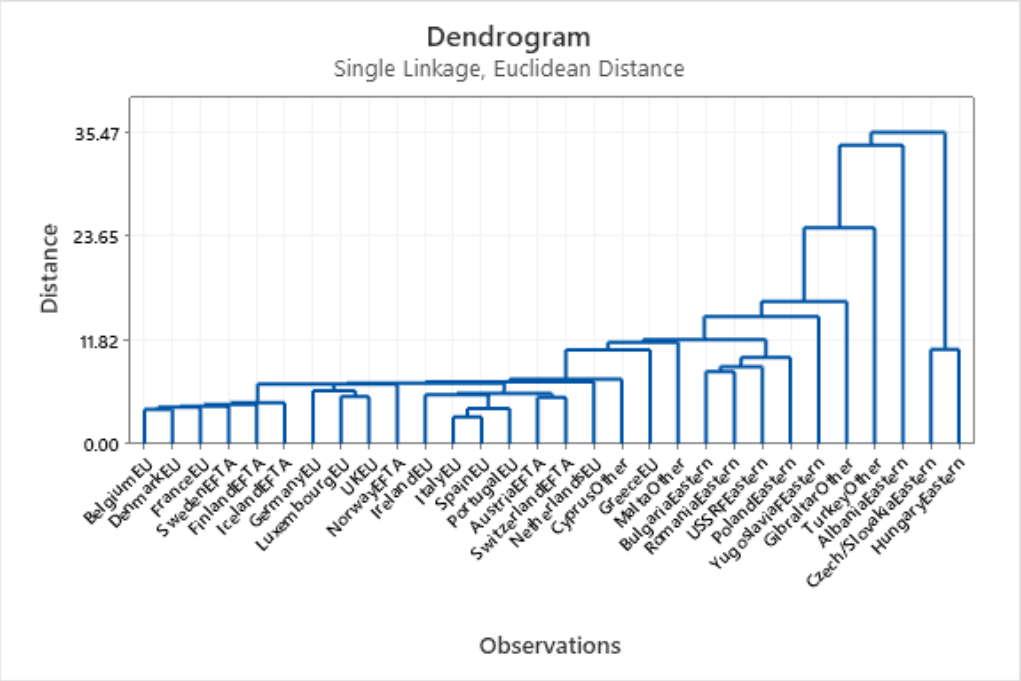
Data: The European employment data were described in the Manly textbook and also previously studied in Exercise 2.

Hierarchical cluster analysis (single linkage): The output from this analysis, based on standardized variables and Euclidean distances, performed in Minitab is shown below. The dendrogram is indeed identical to the one shown in Figure 9.3 of Manly (3rd/4th edition), even if that figure has a different distance scale (as explained in the text).



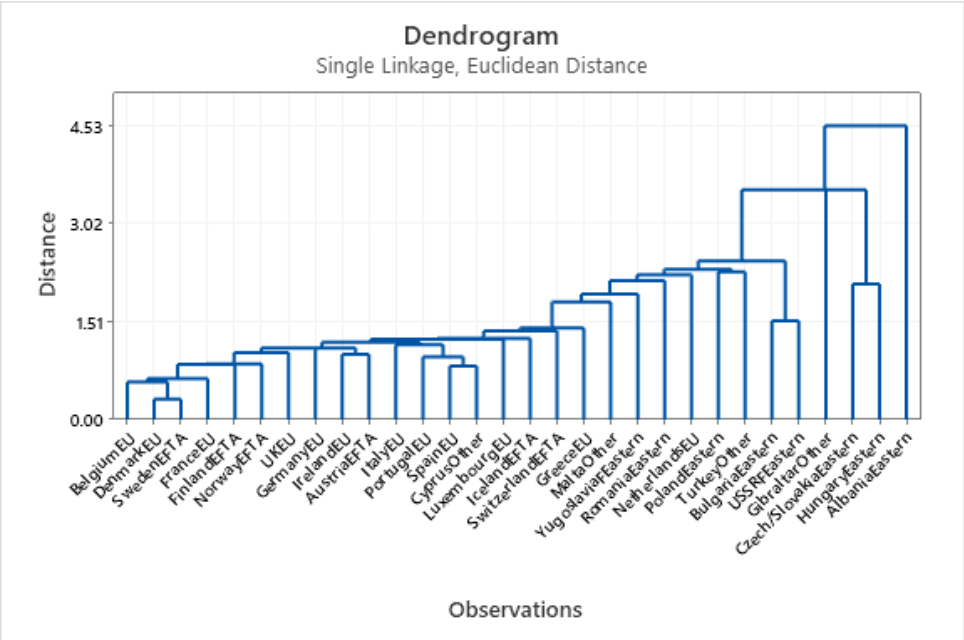
The text describes how the dendrogram suggests to define four clusters from the top, namely Albania, Czech/Slovakia and Hungary, Gibraltar, and the remaining countries in one large cluster. This does not match the political grouping well, and also some of the distances at the bottom of the dendrogram look a bit puzzling (for example, Denmark and Sweden being closest to Belgium and France). Switching to another linkage does not seem to affect the patterns much (as judged from results with average and complete linkages, not shown).

Although all variables are on a percentage scale, the spread across countries is very variable, ranging from standard deviations of 0.6 and 1.2 for PS and TC to the highest standard deviation of 12.3 (for AGR). Because of the wide ranges of spread, the usual approach would be to standardize the variables, as we have done. In this instance, it could perhaps be argued that the variables with only little spread should in fact not be as influential to the patterns obtained as those with larger spread. Little difference means little difference!, and it would seem quite reasonable that for example the PS variable contributes only little to distinguishing the countries. In order to explore the impact of this, we obtain a dendrogram for the raw data as well (on the next page).



The resulting dendrogram is (perhaps) surprisingly similar. The only major difference at the top of the dendrogram is the formation of a separate cluster for Turkey. Also Gibraltar has been moved closer to the large remainder cluster of countries.

Another potentially crucial factor in the formation of the distances and hence the dendrogram could be the presence of errors. We noted in Exercise 2 that the zero values for MAN for Albania, Hungary and Czech/Slovakia, and the zero value for SPS for Albania looked quite suspect. Because these countries accounted for several of the dominant clusters, the dendrogram will obviously change a lot by omitting them! Instead we try an analysis without the two variables in question. As is seen below, the results are surprisingly similar to the original dendrogram. Apparently these extreme values were not overly influential for our results.

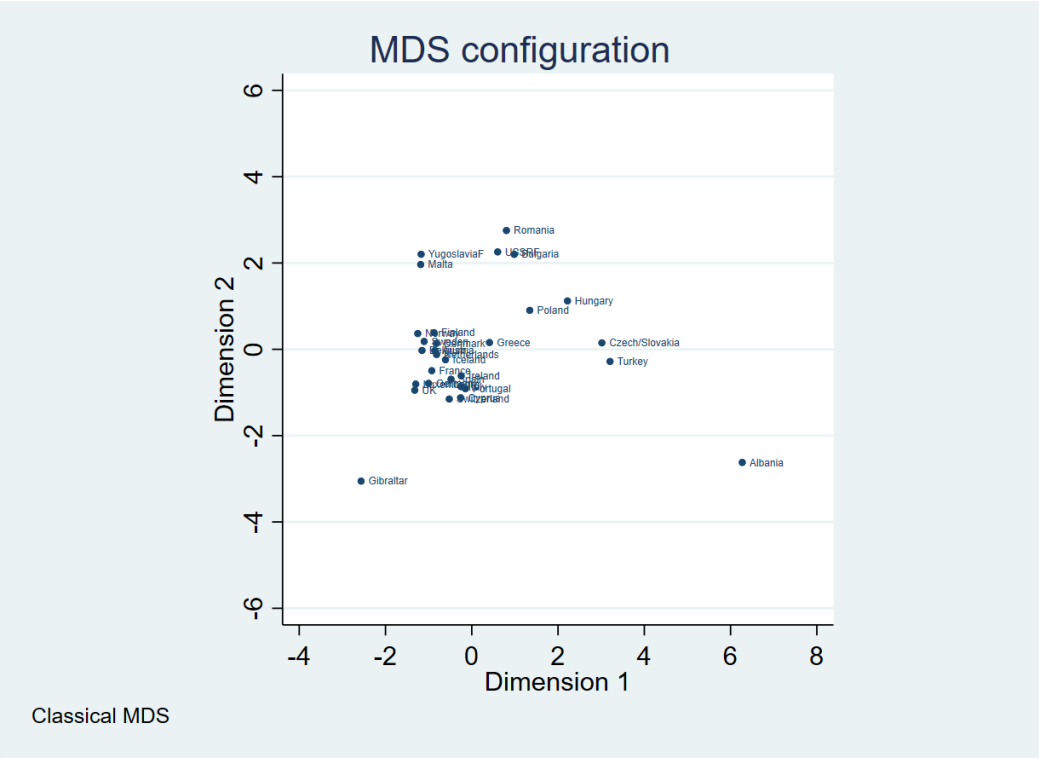


K-means clustering: The Manly text reports (partial) results from K-means clustering with $K = 2, \ldots, 6$, based on Euclidean distance and standardized variables. The table below summarizes the results obtained from a large number of random starting configurations with the default algorithm implemented in R.

Cluster allocation		K-means clustering with K clusters				
Country	Group	2	3	4	5	6
Belgium	EU	1	1	1	1	1
Denmark		1	1	1	1	1
France		1	1	1	1	1
Germany		1	1	1	4	4
Greece		1	1	1	4	4
Ireland		1	1	1	4	4
Italy		1	1	1	4	4
Luxembourg		1	1	1	4	4
Netherlands		1	1	1	1	1
Portugal		1	1	1	4	4
Spain		1	1	1	4	4
UK		1	1	1	4	4
Austria	EFTA	1	1	1	4	4
Finland		1	1	1	1	1
Iceland		1	1	1	4	4
Norway		1	1	1	1	1
Sweden		1	1	1	1	1
Switzerland		1	1	1	4	4
Albania	East	2	2	2	2	2
Bulgaria		1	3	4	5	5
Czech/Slovakia		2	2	3	3	3
Hungary		2	2	3	3	3
Poland		1	3	4	5	5
Romania		1	3	4	5	5
USSRF		1	3	4	5	5
YugoslaviaF		1	3	4	5	5
Cyprus	Other	1	1	1	4	4
Gibraltar		1	1	1	4	6
Malta		1	3	4	1	1
Turkey		2	2	2	2	2
Prop. variance explained		26.6%	42.1%	56.5%	65.6%	72.4%

These results are essentially the same as given in the Manly text (the percentages of variance explained vary a bit, but this could be due to round-off errors in the standardization). The agreement with the political grouping is perhaps not so good, but the deviations make a lot of sense. The Western countries are split in two groups that are broadly Scandinavian versus Southern, the cluster formed by Albania and Turkey can perhaps be understood as based on demographic likeness, and finally the closeness of Czech/Slovakia and Hungary was noted earlier, including the presence of suspect zero values. One would perhaps have expected Albania to be split out into a separate cluster, but that did not happen!

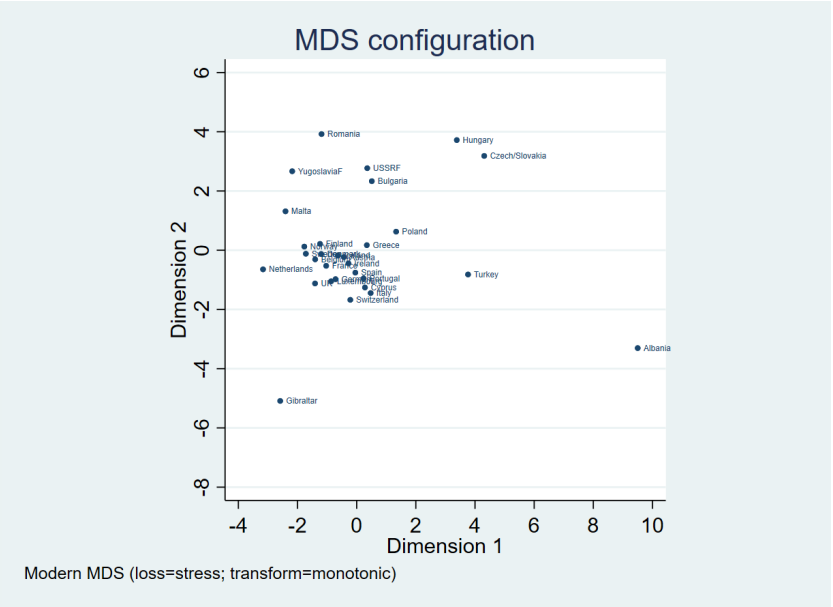
Multidimensional scaling (for countries): Figure 6.2 of the Manly text (3rd/4th editions) plots the first two principal components for the data against each other; classical multidimensional scaling gives the same result (below).



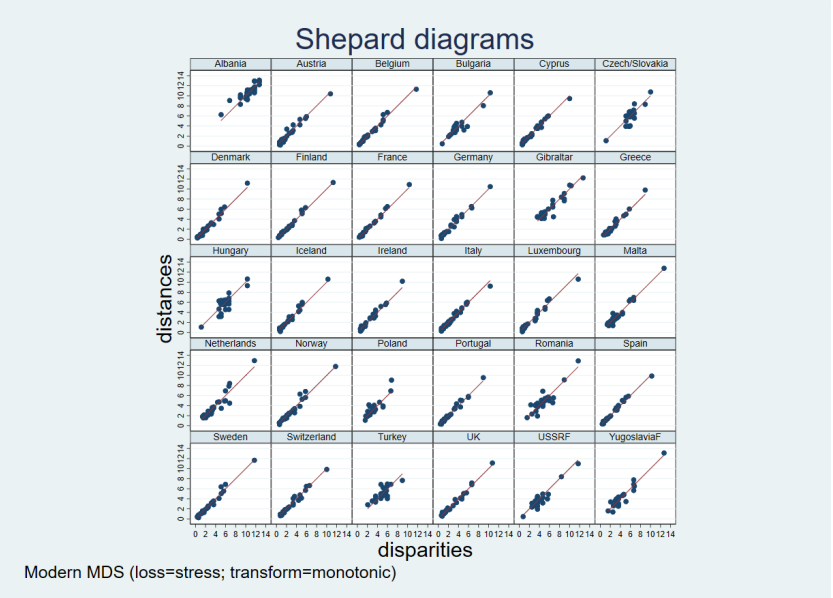
The plots shows many features that agree with the previous analyses: the separations of Albania and Gibraltar, as well as the closeness of most Western European countries, but maybe somewhat not surprisingly not the closeness of Hungary and Czech/Slovakia. It would seem there is room for improvement, either by including higher orders of the classical solution or by the iterative procedures of modern MDS.

The Manly exercise asks specifically about the number of dimensions needed. Only vague guidance is offered in the text on selecting the number of dimensions from the fit (the "Stress 1" criterion), but it is indicated that a range between 0.05 and 0.1 could be desirable. The values obtained for 2, 3 and 4 dimensions were 0.177, 0.103 and 0.065, respectively. Furthermore, power and monotonic ($\sim$ non-metric MDS) transformations gave values of 0.142 and 0.124 for the 2-dimensional solution. From these results one would prefer either a 3-dimensional solution (possibly non-metric) or the non-metric 2-dimensional solution. The configuration plot for the latter is shown on the next page.

The configuration seems to match our previous results better. We still have Albania and Gibraltar separated, but Czech/Slovakia and Hungary have been drawn much closer, as have USSRF and Bulgaria (that also appeared close in the dendrograms).



We also illustrate the Shepard diagram offered by the Stata implementation. It can be produced overall or for each observation (here, country). It plots the observed distances against the disparities obtained from the fit of the scaling equation (Equation (1) of Lecture 9). Every pair of observations contributes a distance, thus $n(n-1)/2$ points in the overall plot or $(n-1)$ points in the separate plots. If the MDS configuration matches well, the deviations from the line $y=x$ should be small.



The plots show the agreement with the overlaid line ($y=x$) is variable across countries, with many of the Western European countries very close to the line, but several of the Eastern European countries showing points scattered about the line. The largest distance, markedly away from all others for many countries, is to Albania; in the distance matrix, its distances are large to almost all other countries. Perhaps this could be taken to suggest to try clustering without Albania, on the basis that this country had several suspect values and could due to its large distances be quite influential. However, we leave further exploration for self study.