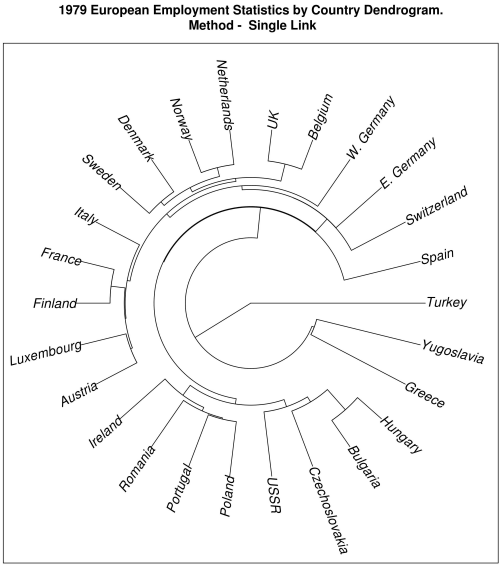
REPORT

Problem 1

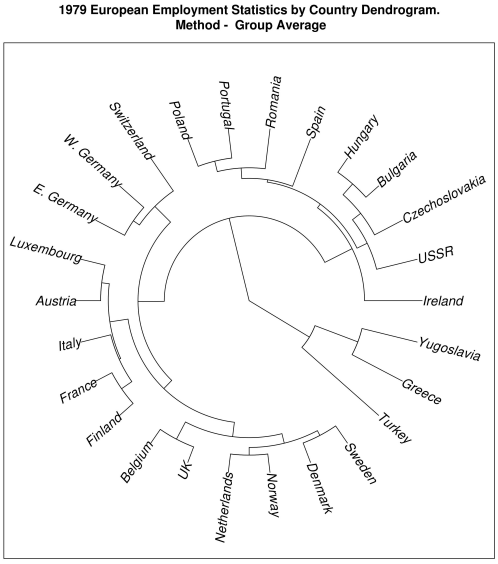
Part 1

1. *Use an agglomerative clusterer to cluster this data. Produce a dendrogram of this data for each of single link, complete link, and group average clustering. You should label the countries on the axis. What structure in the data does each method expose? it's fine to look for code, rather than writing your own. Hint: I made plots I liked a lot using R's hclust clustering function, and then turning the result into a phylogenetic tree and using a fan plot, a trick I found on the web; try plot(as.phylo(hclustresult), type='fan'). You should see dendrograms that "make sense" (at least if you remember some European history), and have interesting differences.*

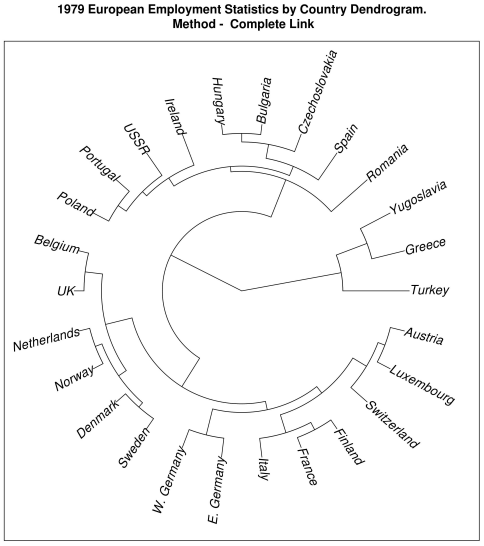
The code for this problem is in a separate file and runs without errors on my machine. Part 1 of my code implements agglomerative clustering using hclust and plots the three cases: single link, complete link, and group average. The results are presented in the figures below.



a)



b)



c)

Figure 1. Agglomerative Clustering Results

To answer the question: “What structure in the data does each method expose?”, I can say that, in general, the clustering results represent grouping of European countries based on some of their economic indicators (employment in different sectors of economy in this case). One can draw a conclusion about the similarity of these indicators in the economies of the clustered countries.

For example, clustering results for the complete link and group average methods (Fig. 1 b) and c)) yield the same final result of three clusters that are somewhat aligned with a description on the original dataset website (lib.stat.cmu.edu): one cluster for western European countries, one cluster for eastern European countries which, for some reason (similarity of the employment structure), also includes Ireland, Spain, and Portugal, and one cluster for Yugoslavia, Greece, and Turkey which is also due probably to the similarity of employment patterns/industries in these 3 countries and coincides with their geographic adjacency. As explained in the description of the dataset, this may have to do with industrial vs. agricultural basis of the economy and the development of the service sector. Although the final clustering result is the same, one can clearly see that grouping of the countries inside each cluster is different for the complete link and group average methods. This has to do with the differences between the two methods when they are applied to the data at hand.

The single link method (Fig. 1 c)) failed to note these similarities and yielded a different structure: one cluster for Turkey (which can mean that it is different, after all, in some individual indicators), one cluster for Yugoslavia and Greece (which can mean that their economies could have been more similar than compared to Turkey, at least based on some individual indicators), and one cluster for the rest of the countries combining both western and eastern European countries which means that some individual indicators (used by single link) were quite close for all of these countries. Therefore, single link is not the best method for this type of analysis.

Part 2

2. Using k-means, cluster this dataset. What is a good choice of k for this data and why?

The next figure shows k-means clustering results for 6 different values of k from 2 to 7. I used a distance matrix and multidimensional scaling (which is, in a way, similar to using the two most important principal components) to build these graphs; see the section commented as *Part 2. Kmeans clustering* in my code.

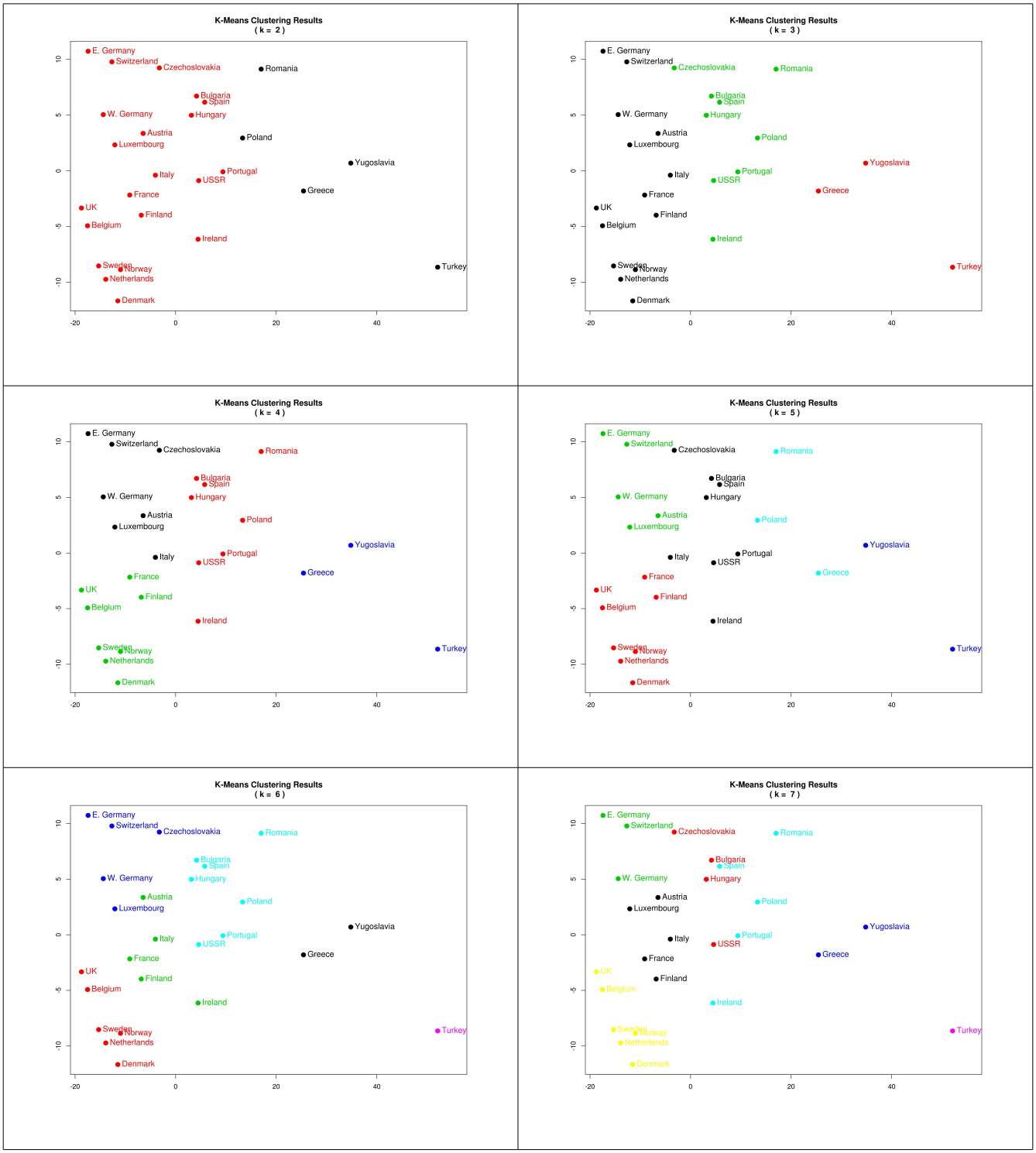


Figure 2. K-Means Clustering Results

The number of starts *nstart* in the *kmeans* function does not have any effect: 1 or 100 yields the same result. Formally, I selected 10. Scaling of the initial data (in *rawData*) did not improve the clustering results, but made the information shown on plots more crowded; therefore, I did not use it in the final version.

For the maximum number of k, I chose 7 because it is more representative than a smaller value, and a greater value would not have revealed any significant differences. Below is my general comparison of k-means clustering results for each k:

- k=2 is not very informative compared to other values of k;
- k=3 provides the results that coincide exactly with the hclust results for the complete link and group average methods; in all the three cases, Yugoslavia, Greece, and Turkey belong to the same cluster, and Ireland, Spain, and Portugal are in one cluster with the Eastern Block countries;
- k=4 – the western European countries are split into two groups, and one of them receives Czechoslovakia which says that its economy could be close to that of western countries;
- k=5 - same as k=4, but Turkey becomes a separate cluster which means that its economy is different, after all, even from that of Greece and Yugoslavia;
- k=6 – a split among the Eastern Block countries into two clusters;
- k=7 – now the western European countries are split into 3 different clusters, and Czechoslovakia goes back to one of the 2 eastern block clusters. One can also notice how some western European countries are clustered based on their geographical location which can mean that economies can be similar based on geographic adjacency. It looks very much like any further splitting of these clusters would be probably interesting only to a scientist who is after very fine differences between the economies of these countries, but loses any common logics from a neutral observer's point of view.

As a better justification of a good choice of k, the next figure shows a plot demonstrating the so called "elbow criterion" to select k. Using an example from the course textbook (Fig. 6.8) and sample code mentioned in line 7 of my code, I plotted a comparison of the value of the cost function for each of several different values of k on Fig. 3 below. It uses the within-cluster sum of squares and the value of k. As you can see, there is a sharp drop in the cost function from 1 to 2, and another noticeable drop from 2 to 3. After that, the drops in the function become much smaller, and the curve becomes much flatter. Looking at the graph, a good choice of k can be definitely 3; one may also consider 4 as it may reveal some more details for analysis because the drop between 3 and 4 is still a bit larger than the subsequent drops.

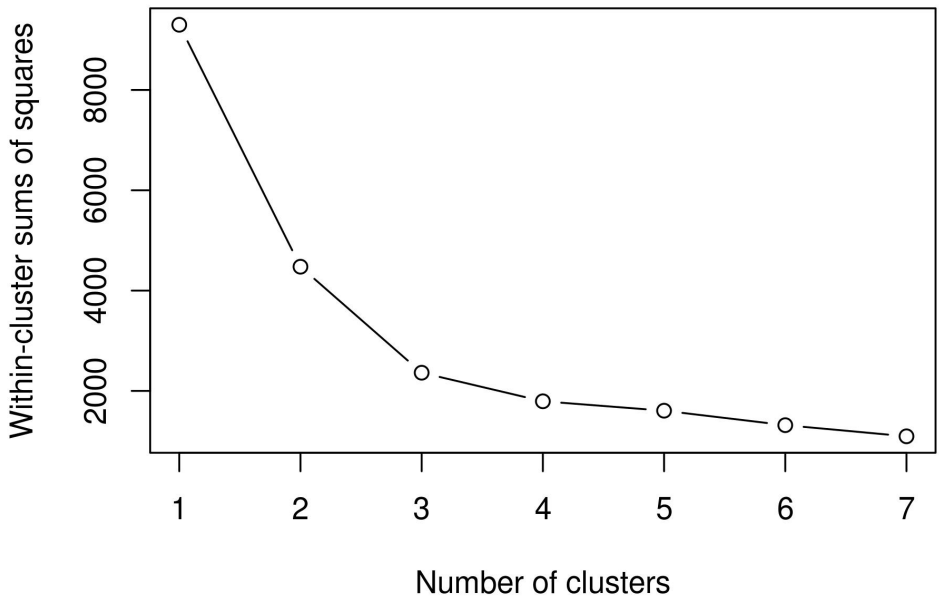


Figure 3. Cost Function to Choose K

Problem 2

6.2. Obtain the activities of daily life dataset from the UC Irvine machine learning website.
(a) Build a classifier that classifies sequences into one of the 14 activities provided. To make features, you should vector quantize, then use a histogram of cluster centers (as described in the subsection; this gives a pretty explicit set of steps to follow). You will find it helpful to use hierarchical k-means to vector quantize. You may use whatever multi-class classifier you wish, though I'd start with R's decision forest, because it's easy to use and effective. You should report (a) the total error rate and (b) the class confusion matrix of your classifier.

(b) Now see if you can improve your classifier by (a) modifying the number of cluster centers in your hierarchical k-means and (b) modifying the size of the fixed length samples that you use.

The assignment is implemented in accordance with the instructions. The codes runs without errors on my machine and is commented, so that one could follow the logic of the program.
The initial run was made with the segment length of 96 and k = 450; the resulting error was: OOB estimate of 37.05%, test set error rate: 38.86%. After this, I ran the program with different segment lengths and values of k, and copied the best results in the attached file clf_results.txt (including the initial run, see the first entry). I did not record all the results, as they were not better than the initial ones. The best error rate achieved is at segment length 48 and k = 400 and is:
OOB estimate of error rate: 27.26%
Test set error rate: 34.86%

This is the corresponding confusion matrix:

OOB estimate of error rate: 27.26%