Homework 11 – Intro. to Computational Statistics

Using a high-dimensional dataset of your choice, perform a factor analysis and clustering and interpret the results. You may use, for instance, the datasets inside the `psych` package, such as `bfi` (25 personality items thought to boil down to a few core personality types) or `iqitems` (14 scores that are thought to boil down to a few core mental skills), or anything else you can find. (Load the data using, for instance, `data(bfi)` after loading the psych package; you may need to clean it a bit first with `na.omit()` to remove the observations with na items, or else impute those missing items.)

For the factor analysis, you may use any of the methods covered in the lesson – they should all produce similar results, though princomp and prcomp might be simplest. You don't have to interpret everything, say, `fa()` outputs, which is a lot of stuff – easier to use `str()` to examine the output of your function and find the quantities you want.

After running your factor analysis or PCA, be sure to discuss and interpret your output:

1. Examine the factor eigenvalues or variances (or the sdev or standard deviations as reported by prcomp or princomp, which you then need to square to get the variances). Plot these in a scree plot and use the "elbow" test to guess how many factors one should retain. What proportion of the total variance does your subset of variables explain?
2. Examine the loadings of the factors on the variables (sometimes called the "rotation" in the function output) – ie, the projection of the factors on the variables – focusing on just the first one or two factors. Sort the variables by their loadings, and try to interpret what the first one or two factors "mean." This may require looking more carefully into the dataset to understand exactly what each of the variables were measuring. You can find more about the data in the `psych` package using `?psych` or visiting http://personality-project.org/ .

Next perform a cluster analysis of the same data.

3. First use k-means and examine the centers of the first two or three clusters. How are they similar to and different from the factor loadings of the first couple factors?
4. Next use hierarchical clustering. Print the dendrogram, and use that to guide your choice of the number of clusters. Use `cutree` to generate a list of which clusters each observation belongs to. Aggregate the data by cluster and then examine those centers (the aggregate means) as you did in (3). Can you interpret all of them meaningfully using the methods from (3) to look at the centers?
5. From the factor and cluster analysis, what can you say more generally about what you have learned about your data?