Assn 9: CLUSTER ANALYSIS I (PARTITIONING METHODS)

Using the data on members of the Classification Society posted on Moodle (files "csna.mlt" (data) and "csna.doc"(documentation)):

1. Using R or SPSS, use kmeans to partition the data set into k=1 to 9 clusters, using only the 14 binary variables dealing with research interests. You can use any method for choosing initial cluster seeds that you wish (or use the default in the package), but NOTE and state what option you are using. For each solution, save the (total) within-cluster sum of squares, and plot that against the number of clusters. Do you notice any anomalies? If yes, try to explain them.

2. Using the plot you created in step 1, choose what seems to be the optimal number of clusters, k. Re-run that solution, and save the cluster membership of each case in a vector or variable.

3. To help you interpret the k-cluster solution, run the following analyses:

A. INTERNAL VALIDITY: calculate means (and, optionally, the standard deviations) of the 14 interest variables, BY cluster. This basically shows the cluster centroids (with SDs) on the variables used to define the clustering.

B. EXTERNAL VALIDITY: Do a crosstabulation (2-way table) of the obtained cluster memberships against the levels of the "academic specialty or discipline" variable ("spec"). Are your obtained clusters interpretable in terms of particular disciplines?

4. Using your results from part 3, interpret each cluster.

=================

R code to read in data:

```
csna_int<-read.csv("C:/Users/corter/Desktop/mdscstuff/CSNA_MLT.csv",header=TRUE,sep=",",
fill=TRUE)
head(csna_int)

#put numeric interest variables into matrix x
x<-csna_int[, 4:17]
```