Homework 5

STAT 448 - Advanced Data Analysis

Due: Thursday, April 12 at 5:00 pm

The data sets are provided in the **HW5Data.sas** file in the Homework 5 folder on the course website.

The data is based on the Romano-British Pottery data set **pottery.dat** described on page 338 of *A Handbook of Statistical Analysis Using SAS, 3rd Edition* by Der and Everitt and in the attached pdf file. Use the 9 **oxide** levels as the explanatory variables, and keep the kiln site label around for labeling plots (like we did for **state** in the **crime** example and the **posn** variable in the **decathlon** example in class).

1. Use the **pottery** data and answer the following questions.
   (a) Perform a principal components analysis on the oxide levels, and determine how many components you would keep to retain at least 80% of the total variation from the original variables. Also comment on how many components would be chosen based on the average eigenvalue and scree plot methods.

   (b) For the components you would keep based on the 80% criterion in part a, explain what features these components pick out of the data (they should tell us something about types of oxide levels being picked out by the components)

   (c) Create score plots for the components kept and label with the kiln site numbers. Comment on what these tell us about the types of oxides from the 5 sites, and comment on similarities and differences of oxide contents for the various kiln sites.

2. With correlation-based PCA, the scale of the original variables is removed. If we use covariance instead, original variables with greater variance will contribute more to the total variation and have a larger impact on the analysis. We can add an option to the **proc** statement to use the covariance instead.

   Repeat the steps of Exercise 1 using the covariance instead of the correlation, and comment on similarities of and differences between the correlation-based and covariance-based results.

3. Consider grouping observations into clusters using cluster analysis.

   (a) Use average linkage on the original oxide levels. What do the dendrogram, pseudo F, pseudo t^2, and CCC statistics suggest about the number of clusters? How many clusters should you choose? Comment on oxide differences between clusters.

   (b) Comment on how well the clusters do (or do not) match up with the original kilns, and comment on any similarities with what you observed in the PCA results in Exercises 1 and 2.

4. Repeat Exercise 3 using standardized oxide levels, and comment on whether the clustering in Exercise 3 or Exercise 4 does a better job of matching the original kilns and why. (Note: with standardized variables, the clustering is based on relative changes in the clustering variables rather than absolute changes in the clustering variables).