STAT 8250 — Applied Multivariate Methods
Lab 8 – Due: Wednesday, Dec. 6

A famous data set that has been used many times to illustrate multivariate techniques is R.A. Fisher's (1936) iris data. The data consist of measurements of petal length (PL), petal width (PW), stamen length (SL), and stamen width (SW) on each of 150 iris plants. The data set contains 50 plants of each of three varieties denoted by IS, IC, and IV. Here, we will use nonhierarchical clustering methods to analyze these data pretending that the varieties are unknown. At the end, we will compare our results to the true variety identifications for the 150 plants.

Copy the file iris1.sas from the public directory to your home directory, run it, and look at the program and its output in an editor. I'll describe what is done in each step of the program and you should follow along and answer the questions that follow.

First, the data are read into a SAS data set called IRIS. The variable VARIETY identifying the variety (S,C, or V) of each plant is defined here.

Next, the data are standardized with PROC STANDARD so that each of the four variables (SL, SW, PL, PW) has mean 0 and standard deviation 1.

Next, a principal components analysis is performed on the standardized iris data with PROC PRINCOMP. The output from this analysis appears on p.1 of iris1.lst.

1. What percentage of the total variance of the standardized variables corresponding to SL, SW, PL, and PW is accounted for by the first two principal components?

Next, PROC PLOT is used on data set SCRS (which contains principal component scores) to produce a scatterplot of the first two principal component scores for each of the 150 irises. This plot appears on p.2.

2. Pretending for the moment that you don't know the true number of varieties represented among these data, how many clusters are suggested by the plot on p.2? Think about how you would form the clusters based on this plot.

Next, PROC FASTCLUS is used to implement the $K-$means nonhierarchical clustering algorithm using $K = 2$. PROC FASTCLUS automatically selects the seeds according to an algorithm designed to minimize the number of passes through the data that are necessary to obtain final clusters. That is, seeds are selected in a way designed to minimize the number of times that step 2 of the clustering algorithm (p.291 in the notes) is repeated. The algorithm for selecting initial seeds is controlled by the REPLACE=

option. REPLACE=FULL implements the algorithm in its entirety which should lead to well-separated seeds. The MAXITER= option controls the number of times that cluster centroids are recomputed (the number of times that step 2, p.291, is repeated). It is important to note that by default, SAS sets MAXITER=1, so that step 2 is only performed once! This makes sense only if you have a very large data set and you are trying to limit computation time.

The results of PROC FASTCLUS are on pp.3–4. The initial seeds that were used are given, the iteration history is given — cluster centroids were replaced (step 2) four times, and some information on final clusters is given — 97 irises assigned to cluster 1, 53 to cluster 2; distance between final cluster centroids; means for each of the four variables by final cluster.

Next, PROC PLOT is used to produce a principal component scores scatterplot featuring the cluster identification according to PROC FASTCLUST with $K = 2$. This plot appears on p.5.

3. How does the plot on p.5 compare with how you would group items into two clusters?

For the moment we will ignore the statistics printed on p. 6 and p.10 and p.14. These statistics, called Beale's intracluster residual sum of squares, can be used to form a pseudo F type statistic (not the same as the pseudo F statistic given by PROC CLUSTER or the pseudo F statistic given in the output of PROC FASTCLUS) that can be used to compare two cluster structures. We will talk about Beale's pseudo F statistic in an upcoming lecture.

The second call to PROC FASTCLUS produces the output on pp.7–8. This time we assume $K = 3$ clusters. A scatterplot of the first 2 PC scores identifying the 3 clusters obtained here appears on p.9.

The third call to PROC FASTCLUS produces the output on pp.11–12. This time we assume $K = 3$ clusters again, but we also make a couple of changes to the algorithm. The DRIFT option replaces step 2 with step 2* (see notes). That is, instead of recomputing cluster centroids after all items have been assigned, we recompute cluster centroids after each item has been assigned. Secondly, the REPLACE=RANDOM tells SAS to select the 3 initial seeds at random rather than according to the leader algorithm. The RANDOM= supplies a seed to the random number generator (in this case 132). A scatterplot of the first 2 PC scores identifying the 3 clusters obtained here appears on p.13.

4. Compare the plots on p.9 and p.13. Have the changes in the clustering algorithm affected the results?

Finally, the last call to PROC PLOT produces the scatterplot on p.14 which identifies the true varieties of the 150 irises.

5. How well do the results of the cluster analyses with $K = 3$ reproduce the true cluster structure?