Chapter 17 Notes

Previous Analyses

With regression models (e.g. linear regression, generalized linear models, etc.), we had explanatory variables that we wanted to have little correlation between each other, but high correlation with a response. Then we could use the explanatory variables in some additive way to predict or model the response.

With principal components analysis, we had highly correlated variables and we mapped those variables to new uncorrelated variables called principal components. The principal components could be used to pick out features in the original data. Usually we could get a pretty good approximation of the information in the original data by using far fewer of the principal components (so we can work with fewer dimensions) and when we are lucky we might be able to roughly interpret what the first few principal components mean.

Cluster Analysis

With cluster analysis, we will start with multivariate data and group the observations based on distance. Conceptually, observations that are close together are more similar and we want to group them together. Observations that are farther apart are less similar and we may want to put them in a separate **cluster** or group.

In **hierarchical** cluster analysis, which we will be looking at, we start with every individual observation being in its own cluster and then start to merge observations and clusters together. This will give a hierarchy of proximities of the data points.

To do this, we need to define:

- a **distance** function which will tell us about distance between points
- a **linkage** function which will tell us about distance between clusters

Our distance function will be Euclidean distance, though it is possible to use other distances in practice. For linkage, there are many possibilities, and we can see them under **The Cluster Procedure>>Details>>Clustering Methods** in the help. Three of the easiest linkages to understand are:

- **single** linkage- measures cluster distance as the distance between the closest points in two different clusters
- **complete** linkage- measures cluster distance as the distance between the two most distant points in the two clusters
- **average** linkage- measures cluster distance as the average of all the distances between points in one cluster and points in the other cluster

We will also want to visualize the clustering. We can do this with a tree plot called a **dendrogram**. The dendrogram will show us how the points were merged and how far apart individual points and clusters are. The hierarchy will range from all points in their own cluster (where the clustering algorithm starts) to all points in a single cluster (where the clustering ends) with the merging of points and clusters in between. The length of the branches in the tree show how far apart (based on the linkage) observations or clusters are when they get merged.

We will usually want to determine how many clusters to keep to have well-separated groups. We may be able to guess a reasonable number from the dendrogram, or we may want to look at other diagnostics.

The Cluster Procedure

The procedure of interest in SAS is **proc cluster**. It will allow us to do hierarchical clustering with various linkages. A couple items of note are that we will need to set the **method** option to set a linkage for the clustering, and the **copy** statement can be used to include variables in the output data set that were not specified in a **var** or **id** statement.

The Tree Procedure

The **tree** procedure can be useful for creating dendrograms, but it can also be useful for just creating output data sets with cluster information in them.

An Initial Example

As an initial example we'll use Fisher's iris data set contained in **sashelp.iris**. We will use a complete linkage, which compares the most distant points in clusters when determining distance between clusters. We will focus on the cluster history and dendrogram for now.

- How many clusters might we guess based on the dendrogram?
- We know there were 3 species, so we might obtain 3 clusters and see how well we can reconstruct the original species groups.
- What happens if we use single linkage (looking at the distance between the closest points in two clusters) or average linkage (looking at the average of the distances between all possible pairs of points in two clusters)?
- How many clusters would you guess based on the dendrograms for single and average linkages?
- How well do the clusterings match the original groupings (species)?

Some Diagnostics for Number of Clusters

We may want to determine the number of clusters that gives us well separated groups. A few possibly diagnostics are:

- cubic clustering criterion (the **ccc** option)
- pseudo t^2 and F statistics (the **pseudo** option)

We may want to look at plots (by way of the **plots** option to the **proc cluster** statement) of these values to see what number of clusters might be best. Note there are some limitations mentioned for these measures (in particular the docs mention that they may not be appropriate for the single linkage). For ccc and the pseudo F statistic, we want to look for peaks in their values with respect to cluster number. For the pseudo t^2 statistic, we want to look at dips in the value.

The cubic clustering criterion is described in Sarle, W. S. (1983), Cubic Clustering Criterion, Technical Report A-108, SAS Institute Inc.. In a hypothesis testing framework, the null is that the data are from a random multivariate uniform population. The alternative is that the data are from a mixture of multivariate normal populations with the same variance and equal sample probability. Positive ccc values give an indication that the clustering is better than uniformly clustering the data into the same number of clusters, and larger values of ccc are better.

The pseudo F and t^2 statistics carry with them assumptions of random sampled data from multivariate normal distributions. These are not going to be very good assumptions for data in general, and will be especially bad when using linkages like single which chops off tails of distributions. These statistics can still give a rough idea of better or worse clusters and are appropriate for **average** and a few other linkages. (See SAS docs for the **pseudo** option for more details.)

Examples

Iris data

- What would we see from plots of the ccc and pseudo F and t^2 statistics using the average linkage?
- How many clusters would we choose?

US Air Quality Data

Using the **usair** data set from the text, we will do the following:

- identify and remove extreme cities from the data set
- perform complete linkage cluster analysis using variables which could be predictors of SO_2 level
- do means analysis by cluster
- pick the two most important principal components out of the original explanatory variables and see where the clustered values fall in that principal component space
- visualize SO_2 level by cluster
- perform analysis of variance on SO_2 level as a function of cluster and interpret the results