STAT 530/J530
October 27th, 2005

Instructor: Brian Habing
Department of Statistics
LeConte 203
Telephone: 803-777-3578
E-mail: habing@stat.sc.edu

Agglomerative Hierarchical Clustering

(0) Start with each item as a separate cluster.

(1) Identify the pair of clusters with the smallest distance between them and merge them. This reduces the number of clusters by 1.

(2) Repeat step 1 until all of the items are in a single cluster.

Linkage Methods

Complete Linkage, Farthest Neighbor, Compact

- Tends to produce convex clusters of similar diameter… sometimes against the natural structure of the data.
- Highly sensitive to outliers.
- Ties can greatly change future linkings.

Linkage Methods

Single Linkage, Nearest Neighbor, Connected

Tends to produce long, stringy, and non-convex clusters.

Because many simulation studies use convex clusters it often does not perform well in them.

Linkage Methods

Average (Mean of the Distances)

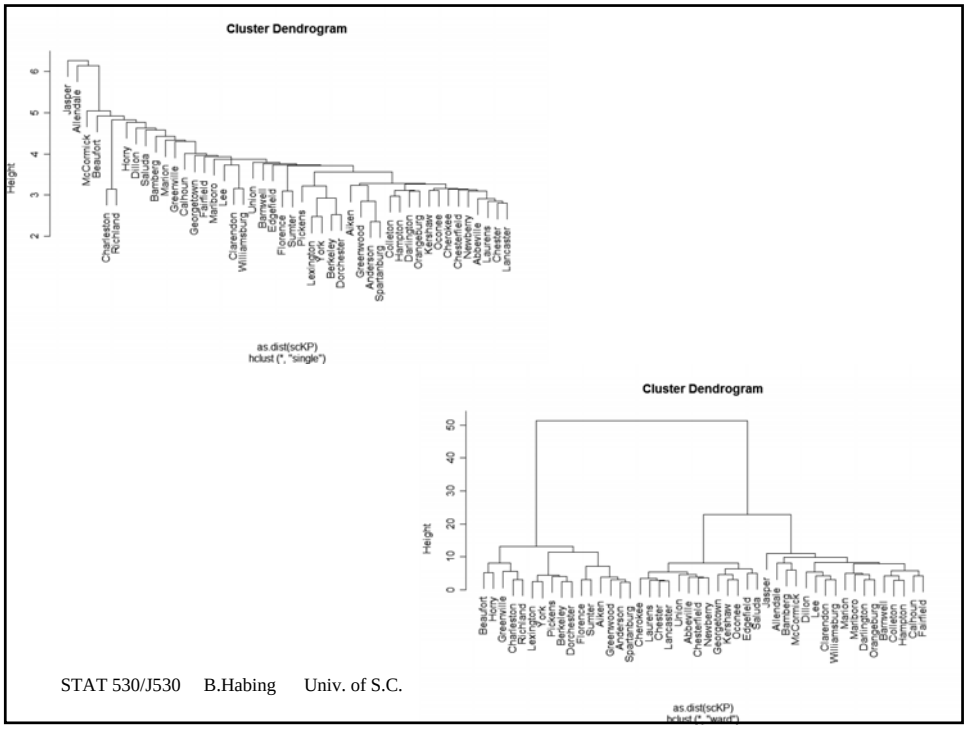
Centroid (Distance of the Means)

- Compromise between Single and Complete
- Average approximates a "least squares" criterion.
- Centroid is more robust to outliers, but in general performs somewhat worse than average.

Linkage Methods

Ward's Method (Minimum Variance Method) is related to the sum of squares in ANOVA

- Tends to produce equal sized convex clusters.



Verdict?

Plusses

- Computationally Easy
- Produces a sequence of Clusters

Minuses

- Which linkage to use?
- How many clusters?

K-Means Clustering

K-Means clustering is a
Partitioning Method

The goal is to find the set of exactly K clusters that is optimal

Verdict?

Plusses

- Computationally Easy
- Produces a sequence of Clusters

Minuses

- Which linkage to use?
- How many clusters?

The Steps

0) Find an initial partition of the individuals into the required number of groups (say by using an agglomerative method and “cutting” the tree).

The Steps

1) Calculate the change in the clustering criterion produced by moving each individual from its own cluster to another.
2) Make the change that leads to the greatest improvement in the value of the clustering criterion.
3) Repeat 1 and 2 until there is no improvement.

Details

The standard methods use the within group sum of squares as the criterion. This is similar to Ward's Linkage in the hierarchical methods.

This is the "maximum-likelihood" clustering in the case where the clusters are multivariate normal with the same covariance.

Warning!

The solution you get depends on the initial clustering you give it, so you need to try several different values!

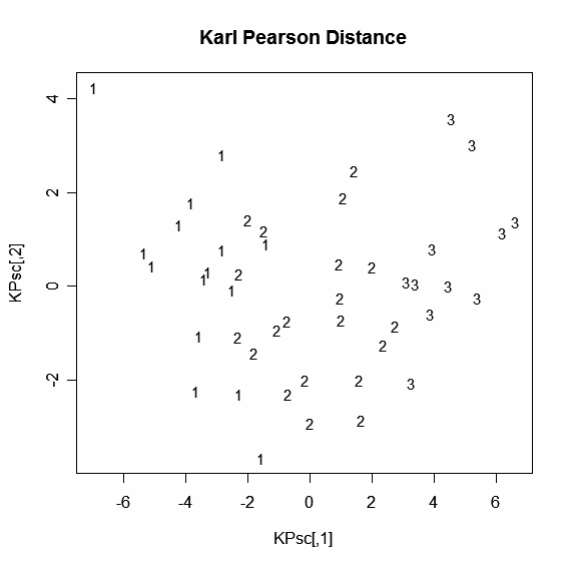
Take 1

```
   Abbeville        Aiken    Allendale     Anderson      Bamberg     Barnwell
           1            1            2            1            2            2
    Beaufort     Berkeley      Calhoun   Charleston     Cherokee      Chester
           3            3            2            3            1            1
Chesterfield    Clarendon     Colleton   Darlington       Dillon   Dorchester
           1            2            1            1            2            3
   Edgefield    Fairfield     Florence   Georgetown   Greenville    Greenwood
           1            2            1            1            3            1
     Hampton        Horry       Jasper     Kershaw    Lancaster      Laurens
           2            3            2            1            1            1
         Lee    Lexington    McCormick       Marion     Marlboro     Newberry
           2            3            2            2            2            1
      Oconee   Orangeburg      Pickens     Richland       Saluda  Spartanburg
           1            1            3            3            2            3
      Sumter        Union Williamsburg         York
           1            1            2            3

Within cluster sum of squares by cluster:
[1] 1371.1241 1691.9302  949.8228
```


Take 2

```
   Abbeville        Aiken    Allendale     Anderson      Bamberg     Barnwell
           2            2            3            2            3            3
    Beaufort     Berkeley      Calhoun   Charleston     Cherokee      Chester
           1            1            3            1            2            2
Chesterfield    Clarendon     Colleton   Darlington       Dillon   Dorchester
           3            3            3            2            3            1
   Edgefield    Fairfield     Florence   Georgetown   Greenville    Greenwood
           2            3            2            2            1            2
     Hampton        Horry       Jasper     Kershaw    Lancaster      Laurens
           3            1            3            2            2            2
         Lee    Lexington    McCormick       Marion     Marlboro     Newberry
           3            1            3            3            3            2
      Oconee   Orangeburg      Pickens     Richland       Saluda  Spartanburg
           2            3            2            1            3            2
      Sumter        Union Williamsburg         York
           2            3            3            1

Within cluster sum of squares by cluster:
[1]  687.6149 1121.9567 2181.6446
```


Displaying the Results



Validating Your Results

Cluster Validation

1) Randomly divide the data set in two.
1) Use the chosen method and # of clusters on each half, find the centroids of each cluster.
2) Assign each point from the other half of the data to the cluster with the nearest centroid.
3) Compare the two sets of results.