## STAT 530/J530
## November 6th, 2008

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



## K-Means Clustering

K-Means clustering is a <u>Partitioning Method</u>

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



## 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
```



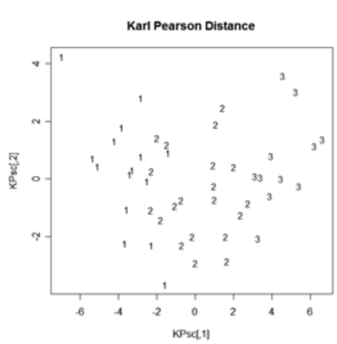
# Take 2

```
  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.

STAT 530/J530 B.Habing Univ. of S.C. 10

## Discriminating Between Groups

In cluster analysis groups of observations are formed based on the values of the variables.

In other cases we know the group membership in advance.

STAT 530/J530 B.Habing Univ. of S.C. 11

## Research Questions

- Are the groups significantly different based on the values of the variables?

Multivariate Analysis of Variance (MANOVA) tests the null hypothesis that the mean vectors of different groups are equal against the alternate hypothesis that they differ.

STAT 530/J530 B.Habing Univ. of S.C. 12

## Crab Data

http://www.stat.sc.edu/~habing/courses/data/crabs.txt

The data set contains four groups of fifteen crabs each: O = orange male, o = orange female, B= blue male, b = blue female.

Each crab has eight measurements. The first four are: FL = frontal lobe size (mm), RW = rear width (mm), CL = carapace length (mm), and CW = carapace width (mm).



## Research Questions

- What combination of the variables does the best job of distinguishing between the groups of observations?

Fisher's Linear Discriminant Analysis searches for the linear combination of variables that "does the best job" of distinguishing between the group



## An Alternate Approach

An alternate approach is to use multiple logistic regression, with the different group memberships as the response and the variables as the predictors.

## Hotelling's T²

Hotelling's $T^2$ is the special case of MANOVA where there are only two groups (populations).

If $\mu_1$ and $\mu_2$ are the (qx1) mean vectors then the procedure tests:

$$H_0: \mu_1 = \mu_2$$



## Assumptions

1. The variables in each population are multivariate normal.
2. The two populations have the same covariance matrix.
3. The observations are independent.



## Hotelling's T²

The formula is:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}\left(\overline{\mathrm{x}}_1 - \overline{\mathrm{x}}_2\right)^T S^{-1}\left(\overline{\mathrm{x}}_1 - \overline{\mathrm{x}}_2\right)$$

Where $S^{-1}$ is the pooled covariance estimate.

With proper scaling can be compared to an F distribution.

## More Than Two Groups?

In general Multivariate Analysis of Variance (MANOVA) is designed to test: $H_0$: $\mu_1 = \mu_2 = \cdots = \mu_k$
where the $\mu_i$ are the (qx1) mean vectors.

The assumptions are the same as before: independence, multivariate normality, and equal variances.



## The Basic Idea

The basic idea is similar to one-way ANOVA. A between-group sum of squares (H) and within-group sum of squares (E) are calculated, and the ratio is taken. If the ratio is large then the null hypothesis is rejected.

The difficulty is that H and E are matrices!



## Four Statistics

Wilk's Lamda: $\Lambda = \frac{|\mathrm{E}|}{|\mathrm{H}+\mathrm{E}|}$

Roy's Greatest Root:
Largest Eigen value of $E^{-1}H$

Lawley-Hotelling Trace:
Trace of $E^{-1}H$

Pillai Trace:
Trace of $H(H+E)^{-1}$

## Crab Example

```
DATA crabs;
INPUT specsex $ FL RW CL CW;
CARDS;
B 8.1 6.7 16.1 19
<INSERT REST OF DATA HERE>
;
PROC GLM DATA=crabs;
CLASS specsex;
MODEL FL RW CL CW = specsex;
MANOVA H=specsex;
```



## Crab Example

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall specsex Effect

H = Type III SSCP Matrix for specsex

E = Error SSCP Matrix

S=3 M=0 N=95.5

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---|---|---|---|---|---|
| Wilks' Lambda | 0.03224835 | 113.34 | 12 | 510.92 | <.0001 |
| Pillai's Trace | 1.70891528 | 64.53 | 12 | 585 | <.0001 |
| Hotelling-Lawley Trace | 9.02228354 | 144.47 | 12 | 333.46 | <.0001 |
| Roy's Greatest Root | 5.81592752 | 283.53 | 4 | 195 | <.0001 |



## Which One???

All result in the same p-value when $k$=2.

Steven, J.P. (2002). *Applied Multivariate Statistics for the Social Sciences*. Mahwah, NJ: Lawrence Erlbaum.

has an extensive overview…

## Which One???

Wilk's, Lawley-Hotelling, and Pillai are all fairly robust if the covariances are not equal assuming the sample sizes are fairly equal (largest/smallest < 1.5).

Roy's is most powerful if the differences can be measured using only a single combination of variables.

Pillai's is slightly more powerful than the others if the differences are found on several orthogonal combinations of variables.



## About the Assumptions - Independence

The independence assumption is critical. Dependence between observations can result in the $\alpha$ level being several times larger than it should be.

In some cases group averages or hierarchical models can be used to remove the effects of individuals being measured in groups.



## About the Assumptions - Normality

Violation of Normality seems to have only a small effect on the $\alpha$-level.

The tails of the distribution do seem to have a sometimes substantial effect on power.

The skewness seems to have much less effect although there seem to be fewer studies.

## About the Assumptions - Covariances

The $\alpha$-level is maintained fairly well if the sample sizes are equal except for extremely different covariance matrices.

Very unequal sample sizes can amplify even slight differences in the covariances and greatly affect the $\alpha$-level.

In general large variance in a small group makes the test liberal, large variability in the larger group makes it conservative.



## In General

- Univariate and Multivariate tests can disagree
- Don't include too many variables
- Always form the most specific hypotheses you can