# Statistics 407 Lab 8

**Due date: Wednesday, October 10**, in class.

**Purpose:** This lab is about diagnosing the results of cluster analysis. There are two different data sets used. For one we will run to different algorithms, and compare the results to determine which is better. The second data set is count data from ecology so we we will look at effect of using different types of distance metrics on the result.

The first dataset (`crimes2008.csv`) contains FBI crime rate statistics. These are the indices for 9 different types of crimes reported by the states of the USA, for 2008: violent, property, murder, rape, robbery, assault, burglary, ltheft (larceny theft), vtheft (vehicle theft). The values have been population adjusted so that the numers are per million people.

The second data set (`newpbi.csv`) contains insect counts collected on 30 Iowa prairies. The insects have been very broadly grouped into Orders: Coleoptera-beetles, Hemiptera-true bugs, Hymenoptera-bees, wasps, ants, ..., Neuroptera-net-winged insects, Orthoptera-grasshoppers, locusts, crickets, ..., Phasmatodea-stick insects. Counts of these insects was taken periodically over the summer at the 30 sites. The goal is to group the sites into similarly composed insect populations.

1. Cluster states based on crime statistics

   (a) Make a scatterplot matrix of the crime indices, with and without Washingto DC. Write a paragraph describing the relationships between the statistics, and about any observations about cluster patterns in the data.

   (b) Cluster the states using hierarchical clustering, with Euclidean distance and wards linkage. Plot the dendrogram. How many clusters would be suggested by the dendrogram?

   (c) Use k-means clustering with $k$ set several different values, say 2-8. Calculate the ratio of between SS to total SS for each value of k. Tabulate this. What is between SS? total SS? What happens to this value as $k$ ranges from 2 to 8? Why is this?

   (d) Use the `fpc` package in R, and the function `cluster.stats` to produce the statistic `wb.ratio` to examine the within group distances to the between group distances for each cluster solution. How many clusters would be chosen by this approach? (The `wb.ratio` statistic reports the ratio between two quantities comparing within to between distances. The average of the distances between points that are in the same cluster, ie within. And the distances between points that are not in the same cluster, ie between. The smaller the value of this the better the result describes clustering as explaining the variation in the data.)

   (e) Decide on an appropriate number of clusters, and report the results. Tabulate the cluster means, standard deviation, and number of points in each cluster. Plot the cluster means using a parallel coordinate plot. List the states in each cluster. Write a paragraph describing the characteristics of each cluster, eg cluster 3 is characterized by low larceny and vehicle theft.

2. Cluster prairies based on types of insect counts

   (a) Compute four different distance metrics on the insect data: Euclidean, Canberra, Bray-Curtis, and Cao. The definitions of these distance metrics for two $p$-dimensional points $\mathbf{x}_j, \mathbf{x}_k, j, k = 1, ..., n; j \neq k$ are:

$$d_{Canberra}(\mathbf{x}_j, \mathbf{x}_k) = \frac{1}{\#non-zero\ counts} \sum_{i=1}^{p} \frac{x_{ij} - x_{ik}}{x_{ij} + x_{ik}}$$

$$d_{Bray}(\mathbf{x}_j, \mathbf{x}_k) = \sum_{i=1}^{p} \frac{|x_{ij} - x_{ik}|}{x_{ij} + x_{ik}}$$

$$d_{Cao}(\mathbf{x}_j, \mathbf{x}_k) \quad = \quad \frac{1}{\#types\ of\ insects} \sum_{i=1}^{p} (\log \frac{x_{ij} + x_{ik}}{2} - \frac{x_{ij} * \log x_{ik} + x_{ik} * \log x_{ij}}{x_{ij} + x_{ik}})$$

Plot the distances against each other. How similar are the results?

EXTRA CREDIT: Why does Euclidean distance have a clump of points that have large values relaitive to all of the other distances?

(b) Run hierarchical clustering with Wards linkage on the Canberra and Cao distance matrices. Tabulate the cluster results. Write a few sentences comparing the results, and compute the percentage of cases that the two methods agree belong together.

(c) Which result is better according to the wb.ratio?

(d) Plot the best result in principal components.
EXTRA CREDIT: Which type of insects contribute the most to PC1? PC2?

**Useful code:**

```
# Read in the crimes data, and check it
> library(ggplot2)
> crime<-read.csv(file.choose())
> head(crime)
> crime <- rescaler(crime[,-c(1,2,4)])
> rownames(crime) <- crime[,1]
# Compute the euclidean distance matrix, and do hierarchical clustering
> crime.dist <- dist(crime[,-1])
> crime.hc <- hclust(crime.dist, method="ward")
> plot(crime.hc, hang=-1)
# Do kmeans clustering
> crime.km3 <- kmeans(crime[,-1], 3)
> crime.km3
# Examine results
> crime.km3$betweenss/crime.km3$totss
> library(fpc)
> cluster.stats(crime.dist, clustering=crime.km3$cluster)$wb.ratio
# Summarize clustering results
> library(GGally)
> library(plyr)
> options(digits=2)
> crime.km3$centers  # OR
> ddply(crime[,-1], .(crime.km3$cluster), colMeans)
> ddply(crime[,-1], .(crime.km3$cluster), function(x) sapply(x, sd))
> crime.km.centers <- ddply(crime[,-1], .(crime.km3$cluster), colMeans)
> crime.km.centers
> colnames(crime.km.centers)[1] <- "cl"
> crime.km.centers$cl <- factor(crime.km.centers$cl)
> ggparcoord(crime.km.centers, columns=2:10, groupColumn=1)
> crimeresults <- crime
> crimeresults$kcl3 <- crime.km3$cluster
# Read in insects data, and check
> pbi <- read.csv(file.choose()) # newpbi-orders.csv
> rownames(pbi) <- pbi[,1]
> head(pbi)
> library(vegan)
> ggpairs(pbi.cast, columns=3:8)
# Calculate different distance matrices
> pbi.euc.dist <- vegdist(pbi[,3:8], method="euclidean")
> pbi.canb.dist <- vegdist(pbi[,3:8], method="canberra")
> pbi.bray.dist <- vegdist(pbi[,3:8], method="bray")
> pbi.cao.dist <- vegdist(pbi[,3:8], method="cao")
> pairs(cbind(pbi.euc.dist, pbi.canb.dist, pbi.bray.dist, pbi.cao.dist))
# Do hierarchical clustering, and choose number of clusters
> pbi.canb.clust <- hclust(pbi.canb.dist, method="ward")
> plot(pbi.canb.clust)
> pbi.cao.clust <- hclust(pbi.cao.dist, method="ward")
> plot(pbi.cao.clust)
> pbi$cl.canb <- cutree(pbi.canb.clust, 3)
> pbi$cl.cao <- cutree(pbi.cao.clust, 3)
# Compute confusion matrix, and other comparison statistics
> table(pbi$cl.canb, pbi$cl.cao)
> cluster.stats(pbi.canb.dist, pbi$cl.canb)$wb.ratio
> cluster.stats(pbi.cao.dist, pbi$cl.cao)$wb.ratio
# Summarize using PCA
> pbi.pca <- prcomp(pbi[,3:8], scale=T, retx=T)
> summary(pbi.pca)
> plot(pbi.pca, type="l")
> pbi$PC1 <- pbi.pca$x[,1]
> pbi$PC2 <- pbi.pca$x[,2]
> qplot(PC1, PC2, data=pbi, colour=as.factor(cl.cao))
```