K-Means Cluster Analysis

There are many approaches for cluster analysis. Of the many clustering approaches there are three that are often used: partitioning, agglomerative, and hierarchical. Partitioning is what will be described in this example and it generally incorporates the well known k-means clustering. The k-means clustering approach requires a very specific type of data and pre-defined requirements by the researcher. First, the researcher must set the number of clusters in advance. There is no strict method to determine the "correct" number of clusters. This example describes one approach to determine an appropriate number of clusters by using a scree plot to see where the "cliff" reaches a bottom plateau. This is similiar to the scree test in factor analysis. Additionally, a multivariate analysis of variance can be used to test good seperation between groups. Second, k-means is only appropriate for continuous (non-categorical) data.

```
require(graphics)
ss = function(x) sum(scale(x, scale = FALSE)^2)

# a 3-dimensional example
x = rbind(matrix(rnorm(99, mean = 0, sd = 0.3), ncol = 3),
           matrix(rnorm(99, mean = 1, sd = 0.3), ncol = 3),
           matrix(rnorm(99, mean = 2, sd = 1), ncol = 3));
colnames(x) = c("x", "y","z");

wss = (nrow(x)-1)*sum(apply(x,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(x,
                                     centers=i)$withinss)
plot(1:15, wss, type="b", xlab="Number of Clusters",
     ylab="Within groups sum of squares")

(cl = kmeans(x, 3))

x = data.frame(x, cl$cluster)

fit = manova(cbind(x$x,x$y,x$z)~factor(x$cl.cluster));

fitted.x = fitted(cl);  head(fitted.x);
resid.x = x - fitted(cl)

cbind(cl[c("betweenss", "tot.withinss", "totss")], # the same two columns
      c(ss(fitted.x), ss(resid.x), ss(x)))

summary.manova(fit, test="Wilks");

plot(x[,1:3], col = cl$cluster)
```