Nonhierarchical clusters

Group items into K clusters.

Assigns items to nearest centroid (usually based on Euclidean distance with data that may be standardized).

Centroids are recomputed and items reassigned if necessary until no more reassignments are needed.

FASTCLUS

Produces separate clusters based on distances from quantitative variables.

Uses Euclidean distances, so the cluster centers are based on least-squares estimation.

Every observation belongs to one and only one cluster

Clusters are not hierarchical

Often called a k-means model, since the centroids are the means of the observations in each cluster.

Each iteration reduces the least-squares criterion until convergence is achieved.

Designed to find good clusters (but not necessarily the best possible clusters) with only two or three passes over the data set.

Sensitive to outliers and can be an effective procedure for detecting outliers because outliers often appear as clusters with only one member.

Minimization criteria (L = or Least =)

LEAST=2, minimize the root mean square difference between the data and the corresponding cluster means.

LEAST=1, minimize the mean absolute difference between the data and the corresponding cluster medians.

LEAST=MAX, minimize the maximum absolute difference between the data and the corresponding cluster midranges

ACECLUS - Approximate Covariance Estimation for CLUStering

The usual calculation for variance / covariance uses deviations from the mean. However, variance / covariance can also be calculated from pairwise differences.

Obtains approximate estimates of the pooled within-cluster covariance matrix

Clusters are assumed to be multivariate normal with equal covariance matrices.

May need a linear transformation to yield a spherical within-cluster covariance matrix.

Cluster membership and cluster number does not have to be known.
A useful method of determining cluster number for other cluster approaches.

THRESHOLD =, the threshold for including pairs of observations in the estimation of the within-cluster covariance matrix. A pair is included if the Euclidean distance between them is less than or equal to *t* times the root mean square distance computed over all pairs of observations.

PROPORTION=, the proportion (0 to 1) or percentage (>1) of pairs to be included in the estimation of the within-cluster covariance matrix.

MODECLUS , non-parametric density estimation

Clusters observations based on nonparametric density estimates. The data can be numeric coordinates or distances.

Can perform approximate significance tests for the number of clusters and can hierarchically join nonsignificant clusters.

Output data sets contains density estimates and cluster membership, various cluster statistics including approximate p-values, and a summary of the number of clusters generated by various algorithms, smoothing parameters, and significance levels.

The methods in MODECLUS are not inherently hierarchical.

Uses spherical kernels of fixed or variable radius. The size of the sphere is determined by the smoothing parameters specified. For fixed-radius kernels, specify the radius as a Euclidean distance with either the DR= or R= option.

METHOD=n, MET=n, M=n, specifies what clustering method to use, indicated by numbers from 0 to 6. For most purposes, METHOD=1 is recommended.

You must specify the METHOD= option to obtain a cluster analysis.

You can specify a list of values for the METHOD= option. Each value in the list is combined with each combination of smoothing and cascading parameters to produce a separate cluster analysis.

CLUSTERING METHODS (from SAS)

METHOD=0, Begin with each observation in a separate cluster. For each observation and each of its neighbors, join the cluster to which the observation belongs with the cluster to which the neighbor belongs.

METHOD=1, Begin with each observation in a separate cluster, find the nearest neighbor with a greater estimated density. If such a neighbor exists, join the cluster to which the observation belongs with the cluster to which the specified neighbor belongs.

Then consider each observation with density estimates equal to that of one or more neighbors but not less than the estimate at any neighbor. Join the cluster containing the observation with (1) each cluster containing a neighbor of the observation such that the maximum density estimate in the cluster equals the density estimate at the observation and (2) the cluster containing the nearest neighbor of the observation such that the maximum density estimate in the cluster exceeds the density estimate at the observation.

See SAS for other methods.