QUICK CLUSTER

When the desired number of clusters is known, QUICK CLUSTER groups cases efficiently into clusters.

Notation

The following notation is used throughout this chapter unless otherwise stated:

NC	Number of clusters requested
$\mathbf{M}_i$	Mean of *i*th cluster
$\mathbf{x}_k$	Vector of *k*th observation
$d\left(\mathbf{x}_i, \mathbf{x}_j\right)$	Euclidean distance between vectors $\mathbf{x}_i$ and $\mathbf{x}_j$
d_{mn}	$\min_{i,j} d\left(\mathbf{M}_i, \mathbf{M}_j\right)$
ε	Convergence criteria

Algorithm

The first iteration involves three steps.

Step 1: Select Initial Cluster Centers

(a) If $\min_i d\left(\mathbf{x}_k, \mathbf{M}_i\right) > d_{mn}$ and $d\left(\mathbf{x}_k, \mathbf{M}_m\right) > d\left(\mathbf{x}_k, \mathbf{M}_n\right)$, then $\mathbf{x}_k$ replaces $\mathbf{M}_n$. If $\min_i d\left(\mathbf{x}_k, \mathbf{M}_i\right) > d_{mn}$ and $d\left(\mathbf{x}_k, \mathbf{M}_m\right) < d\left(\mathbf{x}_k, \mathbf{M}_n\right)$, then $\mathbf{x}_k$ replaces $\mathbf{M}_m$; that is, if the distance between $\mathbf{x}_k$ and its closest cluster mean is greater than the distance between the two closest means ($\mathbf{M}_m$ and $\mathbf{M}_n$), then $\mathbf{x}_k$ replaces either $\mathbf{M}_m$ or $\mathbf{M}_n$, whichever is closer to $\mathbf{x}_k$.

(b) If $\mathbf{x}_k$ does not replace a cluster mean in (a), a second test is made:

Let $\mathbf{M}_q$ be the closest cluster mean to $\mathbf{x}_k$.

Let $\mathbf{M}_p$ be the second closest cluster mean to $\mathbf{x}_k$.

If $d\left(\mathbf{x}_k, \mathbf{M}_p\right) > \min_i d\left(\mathbf{M}_q, \mathbf{M}_i\right)$, then $\mathbf{M}_q = \mathbf{x}_k$;

That is, if $\mathbf{x}_k$ is further from the second closest cluster's center than the closest cluster's center is from any other cluster's center, replace the closest cluster's center with $\mathbf{x}_k$.

At the end of one pass through the data, the initial means of all *NC* clusters are set.

Note that if NOINITIAL is specified, the first *NC* cases with no missing values are the initial cluster means.

Step 2: Update Initial Cluster Centers

Starting with the first case, each case in turn is assigned to the nearest cluster, and that cluster mean is updated. Note that the initial cluster center is included in this mean. The updated cluster means are the classification cluster centers.

Note that if NOUPDATE is specified, this step is skipped.

Step 3: Assign Cases to the Nearest Cluster

The third pass through the data assigns each case to the nearest cluster, where distance from a cluster is the Euclidean distance between that case and the (updated) classification centers. Final cluster means are then calculated as the average values of clustering variables for cases assigned to each cluster. Final cluster means do not contain classification centers.

When the number of iterations is greater than one, the final cluster means in step 3 are set to the classification cluster means in the end of step 2, and QUICK CLUSTER repeats step 3 again. The algorithm stops when either the maximum number of iterations is reached or the maximum change of cluster centers in two successive iterations is smaller than ε times the minimum distance among the initial cluster centers.

Reference

Hartigan, J. A. 1975. *Clustering algorithms*. New York: John Wiley & Sons, Inc.