NAG Toolbox

nag_mv_cluster_hier (g03ec)

1 Purpose

nag_mv_cluster_hier (g03ec) performs hierarchical cluster analysis.

2 Syntax

```
[d, ilc, iuc, cd, iord, dord, ifail] = nag_mv_cluster_hier(method, n, d)
[d, ilc, iuc, cd, iord, dord, ifail] = g03ec(method, n, d)
```

3 Description

Given a distance or dissimilarity matrix for n objects (see nag_mv_distance_mat (g03ea)), cluster analysis aims to group the n objects into a number of more or less homogeneous groups or clusters. With agglomerative clustering methods, a hierarchical tree is produced by starting with n clusters, each with a single object and then at each of $n-1$ stages, merging two clusters to form a larger cluster, until all objects are in a single cluster. This process may be represented by a dendrogram (see nag_mv_cluster_hier_dendrogram (g03eh)).

At each stage, the clusters that are nearest are merged, methods differ as to how the distances between the new cluster and other clusters are computed. For three clusters i, j and k let n_i, n_j and n_k be the number of objects in each cluster and let d_{ij}, d_{ik} and d_{jk} be the distances between the clusters. Let clusters j and k be merged to give cluster jk, then the distance from cluster i to cluster jk, $d_{i.jk}$ can be computed in the following ways.

1. Single link or nearest neighbour : $d_{i.jk} = \min\left(d_{ij}, d_{ik}\right)$.
2. Complete link or furthest neighbour : $d_{i.jk} = \max\left(d_{ij}, d_{ik}\right)$.
3. Group average : $d_{i.jk} = \dfrac{n_j}{n_j+n_k}d_{ij} + \dfrac{n_k}{n_j+n_k}d_{ik}$.
4. Centroid : $d_{i.jk} = \dfrac{n_j}{n_j+n_k}d_{ij} + \dfrac{n_k}{n_j+n_k}d_{ik} - \dfrac{n_j n_k}{\left(n_j+n_k\right)^2}d_{jk}$.
5. Median : $d_{i.jk} = \frac{1}{2}d_{ij} + \frac{1}{2}d_{ik} - \frac{1}{4}d_{jk}$.
6. Minimum variance : $d_{i.jk} = \left\{\left(n_i+n_j\right)d_{ij} + \left(n_i+n_k\right)d_{ik} - n_i d_{jk}\right\}/\left(n_i+n_j+n_k\right)$.

For further details see Everitt (1974) or Krzanowski (1990).

If the clusters are numbered $1, 2, \ldots, n$ then, for convenience, if clusters j and k, $j < k$, merge then the new cluster will be referred to as cluster j. Information on the clustering history is given by the values of j, k and d_{jk} for each of the $n-1$ clustering steps. In order to produce a dendrogram, the ordering of the objects such that the clusters that merge are adjacent is required. This ordering is computed so that the first element is 1. The associated distances with this ordering are also computed.

4 References

Everitt B S (1974) *Cluster Analysis* Heinemann

Krzanowski W J (1990) *Principles of Multivariate Analysis* Oxford University Press

5 Parameters

5.1 Compulsory Input Parameters

1: **method** – INTEGER

Indicates which clustering method is used.

method $= 1$
Single link.

method $= 2$
Complete link.

method $= 3$
Group average.

method $= 4$
Centroid.

method $= 5$
Median.

method $= 6$
Minimum variance.

Constraint: **method** $= 1$, 2, 3, 4, 5 or 6.

2: **n** – INTEGER

n, the number of objects.

Constraint: $\mathbf{n} \geq 2$.

3: $\mathbf{d}(\mathbf{n} \times (\mathbf{n}-\mathbf{1})/\mathbf{2})$ – REAL (KIND=nag_wp) array

The strictly lower triangle of the distance matrix. D must be stored packed by rows, i.e., $\mathbf{d}((i-1)(i-2)/2+j)$, $i > j$ must contain d_{ij}.

Constraint: $\mathbf{d}(i) \geq 0.0$, for $i = 1, 2, \ldots, n(n-1)/2$.

5.2 Optional Input Parameters

None.

5.3 Output Parameters

1: $\mathbf{d}(\mathbf{n} \times (\mathbf{n}-\mathbf{1})/\mathbf{2})$ – REAL (KIND=nag_wp) array

Is overwritten.

2: $\mathbf{ilc}(\mathbf{n}-\mathbf{1})$ – INTEGER array

$\mathbf{ilc}(l)$ contains the number, j, of the cluster merged with cluster k (see **iuc**), $j < k$, at step l, for $l = 1, 2, \ldots, n-1$.

3: $\mathbf{iuc}(\mathbf{n}-\mathbf{1})$ – INTEGER array

$\mathbf{iuc}(l)$ contains the number, k, of the cluster merged with cluster j, $j < k$, at step l, for $l = 1, 2, \ldots, n-1$.

4: $\mathbf{cd}(\mathbf{n}-\mathbf{1})$ – REAL (KIND=nag_wp) array

$\mathbf{cd}(l)$ contains the distance d_{jk}, between clusters j and k, $j < k$, merged at step l, for $l = 1, 2, \ldots, n-1$.

5: **iord**(**n**) – INTEGER array

The objects in dendrogram order.

6: **dord**(**n**) – REAL (KIND=nag_wp) array

The clustering distances corresponding to the order in **iord**. **dord**(l) contains the distance at which cluster **iord**(l) and **iord**$(l+1)$ merge, for $l = 1, 2, \ldots, n-1$. **dord**(n) contains the maximum distance.

7: **ifail** – INTEGER

ifail $= 0$ unless the function detects an error (see Section 5).

6 Error Indicators and Warnings

Errors or warnings detected by the function:

ifail $= 1$

On entry, **method** $\neq$ 1, 2, 3, 4, 5 or 6,
or **n** < 2.

ifail $= 2$

On entry, **d**$(i) < 0.0$ for some $i = 1, 2, \ldots, n(n-1)/2$.

ifail $= 3$

A true dendrogram cannot be formed because the distances at which clusters have merged are not increasing for all steps, i.e., **cd**$(l) <$ **cd**$(l-1)$ for some $l = 2, 3, \ldots, n-1$. This can occur for the median and centroid methods.

ifail $= -99$

An unexpected error has been triggered by this routine. Please contact NAG.

ifail $= -399$

Your licence key may have expired or may not have been installed correctly.

ifail $= -999$

Dynamic memory allocation failed.

7 Accuracy

For **method** ≥ 3 slight rounding errors may occur in the calculations of the updated distances. These would not normally significantly affect the results, however there may be an effect if distances are (almost) equal.

If at a stage, two distances d_{ij} and d_{kl}, $(i < k)$ or $(i = k)$, and $j < l$, are equal then clusters k and l will be merged rather than clusters i and j. For single link clustering this choice will only affect the order of the objects in the dendrogram. However, for other methods the choice of kl rather than ij may affect the shape of the dendrogram. If either of the distances d_{ij} and d_{kl} is affected by rounding errors then their equality, and hence the dendrogram, may be affected.

8 Further Comments

The dendrogram may be formed using nag_mv_cluster_hier_dendrogram (g03eh). Groupings based on the clusters formed at a given distance can be computed using nag_mv_cluster_hier_indicator (g03ej).

9 Example

Data consisting of three variables on five objects are read in. Euclidean squared distances based on two variables are computed using nag_mv_distance_mat (g03ea), the objects are clustered using nag_mv_cluster_hier (g03ec) and the dendrogram computed using nag_mv_cluster_hier_dendrogram (g03eh). The dendrogram is then printed.

9.1 Program Text

```
     function g03ec_example

fprintf('g03ec example results\n\n');

x = [1, 5, 2;
     2, 1, 1;
     3, 4, 3;
     4, 1, 2;
     5, 5, 0];
[n,m]  = size(x);

isx    = ones(m,1,nag_int_name);
isx(1) = nag_int(0);
s      = ones(m,1);
ld     = (n*(n-1))/2;
d      = zeros(ld,1);

% Compute the distance matrix
update = 'I';
dist = 'S';
scal = 'U';
[s, d, ifail] = g03ea( ...
        update, dist, scal, x, isx, s, d);

% Clustering method
method = nag_int(5);
% Perform clustering
n      = nag_int(n);
[d, ilc, iuc, cd, iord, dord, ifail] = ...
  g03ec(method, n, d);

row = {'A'; 'B'; 'C'; 'D'; 'E'};
fprintf('  Distance   Clusters Joined\n\n');
for i = 1:n-1
  fprintf('%10.3f     %s %s\n', cd(i), row{ilc(i)}, row{iuc(i)})
end
```

9.2 Program Results

```
     g03ec example results

  Distance   Clusters Joined

     1.000     B D
     2.000     A C
     6.500     A E
    14.125     A B
```