NAG Toolbox

nag_mv_cluster_hier_indicator (g03ej)

1 Purpose

nag_mv_cluster_hier_indicator (g03ej) computes a cluster indicator variable from the results of nag_mv_cluster_hier (g03ec).

2 Syntax

```
[k, dlevel, ic, ifail] = nag_mv_cluster_hier_indicator(cd, iord, dord, k,
dlevel, 'n', n)
```

```
[k, dlevel, ic, ifail] = g03ej(cd, iord, dord, k, dlevel, 'n', n)
```

3 Description

Given a distance or dissimilarity matrix for n objects, cluster analysis aims to group the n objects into a number of more or less homogeneous groups or clusters. With agglomerative clustering methods (see nag_mv_cluster_hier (g03ec)), 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. nag_mv_cluster_hier_indicator (g03ej) takes the information from the tree and produces the clusters that exist at a given distance. This is equivalent to taking the dendrogram (see nag_mv_cluster_hier_dendrogram (g03eh)) and drawing a line across at a given distance to produce clusters.

As an alternative to giving the distance at which clusters are required, you can specify the number of clusters required and nag_mv_cluster_hier_indicator (g03ej) will compute the corresponding distance. However, it may not be possible to compute the number of clusters required due to ties in the distance matrix.

If there are k clusters then the indicator variable will assign a value between 1 and k to each object to indicate to which cluster it belongs. Object 1 always belongs to cluster 1.

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

The clustering distances in increasing order as returned by nag_mv_cluster_hier (g03ec).

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

2: $\mathbf{iord}(\mathbf{n})$ – INTEGER array

The objects in dendrogram order as returned by nag_mv_cluster_hier (g03ec).

3: $\mathbf{dord}(\mathbf{n})$ – REAL (KIND=nag_wp) array

The clustering distances corresponding to the order in **iord**.

4: **k** – INTEGER

Indicates if a specified number of clusters is required.

If $\mathbf{k} > 0$ then nag_mv_cluster_hier_indicator (g03ej) will attempt to find **k** clusters.

If $\mathbf{k} \leq 0$ then nag_mv_cluster_hier_indicator (g03ej) will find the clusters based on the distance given in **dlevel**.

Constraint: $\mathbf{k} \leq \mathbf{n}$.

5: **dlevel** – REAL (KIND=nag_wp)

If $\mathbf{k} \leq 0$, **dlevel** must contain the distance at which clusters are produced. Otherwise **dlevel** need not be set.

Constraint: if $\mathbf{dlevel} > 0.0$, $\mathbf{k} \leq 0$.

5.2 Optional Input Parameters

1: **n** – INTEGER

Default: the dimension of the arrays **iord**, **dord**. (An error is raised if these dimensions are not equal.)

n, the number of objects.

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

5.3 Output Parameters

1: **k** – INTEGER

The number of clusters produced, k.

2: **dlevel** – REAL (KIND=nag_wp)

If $\mathbf{k} > 0$ on entry, **dlevel** contains the distance at which the required number of clusters are found. Otherwise **dlevel** remains unchanged.

3: $\mathbf{ic}(\mathbf{n})$ – INTEGER array

$\mathbf{ic}(i)$ indicates to which of k clusters the ith object belongs, for $i = 1, 2, \ldots, n$.

4: **ifail** – INTEGER

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

6 Error Indicators and Warnings

Errors or warnings detected by the function:

$\mathbf{ifail} = 1$

On entry, $\mathbf{k} > \mathbf{n}$,
or $\mathbf{k} \leq 0$ and $\mathbf{dlevel} \leq 0.0$.
or $\mathbf{n} < 2$.

$\mathbf{ifail} = 2$

On entry, **cd** is not in increasing order,
or **dord** is incompatible with **cd**.

ifail $= 3$

On entry, $\mathbf{k} = 1$,
or $\mathbf{k} = \mathbf{n}$,
or $\mathbf{dlevel} \geq \mathbf{cd}(\mathbf{n} - 1)$,
or $\mathbf{dlevel} < \mathbf{cd}(1)$.

Note: on exit with this value of **ifail** the trivial clustering solution is returned.

ifail $= 4$ (*warning*)

The precise number of clusters requested is not possible because of tied clustering distances. The actual number of clusters, less than the number requested, is returned in **k**.

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

The accuracy will depend upon the accuracy of the distances in **cd** and **dord** (see nag_mv_cluster_hier (g03ec)).

8 Further Comments

A fixed number of clusters can be found using the non-hierarchical method used in nag_mv_cluster_kmeans (g03ef).

9 Example

Data consisting of three variables on five objects are input. Euclidean squared distances are computed using nag_mv_distance_mat (g03ea) and median clustering performed using nag_mv_cluster_hier (g03ec). A dendrogram is produced by nag_mv_cluster_hier_dendrogram (g03eh) and printed. nag_mv_cluster_hier_indicator (g03ej) finds two clusters and the results are printed.

9.1 Program Text

```
     function g03ej_example

fprintf('g03ej 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

% k clusters
k = nag_int(2);
dlevel = 0;

% Compute cluster indicators
[k, dlevel, ic, ifail] = g03ej( ...
 cd, iord, dord, k, dlevel);

% Display the indicators
fprintf('\n Allocation to %2d clusters\n', k);
fprintf(' Clusters found at distance %6.3f\n\n', dlevel);
fprintf(' Object  Cluster\n\n');
for i=1:n
  fprintf('%6s     %2d\n',row{i}, ic(i));
end
```

9.2 Program Results

```
     g03ej example results

  Distance   Clusters Joined

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

 Allocation to  2 clusters
 Clusters found at distance  6.500

 Object  Cluster

     A      1
     B      2
     C      1
     D      2
     E      1
```