NAG Toolbox for MATLAB

g03ej

1 Purpose

g03ej computes a cluster indicator variable from the results of g03ec.

2 Syntax

```
[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 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. 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 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 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: **cd**$(\mathbf{n}-\mathbf{1})$ – **double array**

The clustering distances in increasing order as returned by g03ec.

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

2: **iord**$(\mathbf{n})$ – **int32 array**

The objects in dendrogram order as returned by g03ec.

3: **dord**$(\mathbf{n})$ – **double array**

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

4: **k** – **int32 scalar**

Indicates if a specified number of clusters is required.

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

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

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

5: **dlevel – double scalar**

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 – int32 scalar**

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 Input Parameters Omitted from the MATLAB Interface

None.

5.4 Output Parameters

1: **k – int32 scalar**

The number of clusters produced, k.

2: **dlevel – double scalar**

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

3: **ic(n) – int32 array**

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

4: **ifail – int32 scalar**

0 unless the function detects an error (see Section 6).

6 Error Indicators and Warnings

Errors or warnings detected by the function:

ifail $= 1$

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

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

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**.

7 Accuracy

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

8 Further Comments

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

9 Example

```
cd = [1;
     2;
     6.5;
     14.125];
iord = [int32(1);
     int32(3);
     int32(5);
     int32(2);
     int32(4)];
dord = [2;
     6.5;
     14.125;
     1;
     14.125];
k = int32(2);
dlevel = 0;
[kOut, dlevelOut, ic, ifail] = g03ej(cd, iord, dord, k, dlevel)
```

```
kOut =
           2
dlevelOut =
    6.5000
ic =
           1
           2
           1
           2
           1
ifail =
           0
```