

NAG Toolbox for MATLAB

g03ec

1 Purpose

g03ec performs hierarchical cluster analysis.

2 Syntax

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

3 Description

Given a distance or dissimilarity matrix for n objects (see 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 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 : $d_{i,jk} = \min(d_{ij}, d_{ik})$.
2. Complete link or furthest neighbour : $d_{i,jk} = \max(d_{ij}, d_{ik})$.
3. Group average : $d_{i,jk} = \frac{n_j}{n_j + n_k}d_{ij} + \frac{n_k}{n_j + n_k}d_{ik}$.
4. Centroid : $d_{i,jk} = \frac{n_j}{n_j + n_k}d_{ij} + \frac{n_k}{n_j + n_k}d_{ik} - \frac{n_j n_k}{(n_j + n_k)^2}d_{jk}$.
5. Median : $d_{i,jk} = \frac{1}{2}d_{ij} + \frac{1}{2}d_{ik} - \frac{1}{4}d_{jk}$.
6. Minimum variance : $d_{i,jk} = \{(n_i + n_j)d_{ij} + (n_i + n_k)d_{ik} - n_i d_{jk}\} / (n_i + n_j + n_k)$.

For further details see Everitt 1974 or Krzanowski 1990.

If the clusters are numbered $1, 2, \dots, n$ 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** – int32 scalar

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** – **int32** scalar

n , the number of objects.

Constraint: $n \geq 2$.

3: **d**($n \times (n - 1)/2$) – **double** array

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

Constraint: $d(i) \geq 0.0$, for $i = 1, 2, \dots, n(n - 1)/2$.

5.2 Optional Input Parameters

None.

5.3 Input Parameters Omitted from the MATLAB Interface

iwk

5.4 Output Parameters

1: **d**($n \times (n - 1)/2$) – **double** array

Is destroyed.

2: **ilc**($n - 1$) – **int32** array

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

3: **iuc**($n - 1$) – **int32** array

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

4: **cd**($n - 1$) – **double** array

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

5: **iord(n) – int32 array**

The objects in dendrogram order.

6: **dord(n) – double 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 $l = 1, 2, \dots, n - 1$. **dord(n)** contains the maximum distance.

7: **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, **method** \neq 1, 2, 3, 4, 5 or 6,
or **n** < 2.

ifail = 2

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

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 $l = 2, 3, \dots, n - 1$. This can occur for the median and centroid methods.

7 Accuracy

For **method** \geq 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 g03eh. Groupings based on the clusters formed at a given distance can be computed using g03ej.

9 Example

```
method = int32(5);
n = int32(5);
d = [17;
     2;
     13;
     16;
     1;
     10;
```

```
4;  
17;  
10;  
20];  
[dOut, ilc, iuc, cd, iord, dord, ifail] = g03ec(method, n, d)  
  
dOut =  
14.1250  
2.0000  
11.2500  
16.0000  
1.0000  
10.0000  
6.5000  
18.2500  
10.0000  
20.0000  
ilc =  
2  
1  
1  
1  
iuc =  
4  
3  
5  
2  
cd =  
1.0000  
2.0000  
6.5000  
14.1250  
iord =  
1  
3  
5  
2  
4  
dord =  
2.0000  
6.5000  
14.1250  
1.0000  
14.1250  
ifail =  
0
```