NAG Library Routine Document

G03EJF

Note: before using this routine, please read the Users' Note for your implementation to check the interpretation of ***bold italicised*** terms and other implementation-dependent details.

1 Purpose

G03EJF computes a cluster indicator variable from the results of G03ECF.

2 Specification

```
SUBROUTINE G03EJF (N, CD, IORD, DORD, K, DLEVEL, IC, IFAIL)

INTEGER            N, IORD(N), K, IC(N), IFAIL
REAL (KIND=nag_wp) CD(N-1), DORD(N), DLEVEL
```

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 G03ECF), 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. G03EJF takes the information from the tree and produces the clusters that exist at a given distance. This is equivalent to taking the dendrogram (see G03EHF) 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 G03EJF 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

1: N – INTEGER *Input*

On entry: n, the number of objects.

Constraint: N ≥ 2.

2: CD(N $-$ 1) – REAL (KIND=nag_wp) array *Input*

On entry: the clustering distances in increasing order as returned by G03ECF.

Constraint: CD$(i+1) \geq$ CD(i), for $i = 1, 2, \ldots,$ N-2.

3: IORD(N) – INTEGER array *Input*

On entry: the objects in dendrogram order as returned by G03ECF.

4: DORD(N) – REAL (KIND=nag_wp) array *Input*

On entry: the clustering distances corresponding to the order in IORD.

5: K – INTEGER *Input/Output*

On entry: indicates if a specified number of clusters is required.

If K > 0 then G03EJF will attempt to find K clusters.

If K ≤ 0 then G03EJF will find the clusters based on the distance given in DLEVEL.

Constraint: K $\leq$ N.

On exit: the number of clusters produced, k.

6: DLEVEL – REAL (KIND=nag_wp) *Input/Output*

On entry: if K ≤ 0, DLEVEL must contain the distance at which clusters are produced. Otherwise DLEVEL need not be set.

Constraint: if DLEVEL > 0.0, K ≤ 0.

On exit: if K > 0 on entry, DLEVEL contains the distance at which the required number of clusters are found. Otherwise DLEVEL remains unchanged.

7: IC(N) – INTEGER array *Output*

On exit: IC(i) indicates to which of k clusters the ith object belongs, for $i = 1, 2, \ldots, n$.

8: IFAIL – INTEGER *Input/Output*

On entry: IFAIL must be set to 0, -1 or 1. If you are unfamiliar with this parameter you should refer to Section 3.3 in the Essential Introduction for details.

For environments where it might be inappropriate to halt program execution when an error is detected, the value -1 or 1 is recommended. If the output of error messages is undesirable, then the value 1 is recommended. Otherwise, if you are not familiar with this parameter, the recommended value is 0. **When the value -1 or 1 is used it is essential to test the value of IFAIL on exit.**

On exit: IFAIL $= 0$ unless the routine detects an error or a warning has been flagged (see Section 6).

6 Error Indicators and Warnings

If on entry IFAIL $= 0$ or -1, explanatory error messages are output on the current error message unit (as defined by X04AAF).

Errors or warnings detected by the routine:

IFAIL $= 1$

On entry, K $>$ N,
or K ≤ 0 and DLEVEL ≤ 0.0.
or N < 2.

IFAIL $= 2$

On entry, CD is not in increasing order,
or DORD is incompatible with CD.

IFAIL $= 3$

On entry, K $= 1$,
or K $=$ N,

or DLEVEL $\geq$ CD(N $-$ 1),
or DLEVEL $<$ 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 G03ECF).

8 Further Comments

A fixed number of clusters can be found using the non-hierarchical method used in G03EFF.

9 Example

Data consisting of three variables on five objects are input. Euclidean squared distances are computed using G03EAF and median clustering performed using G03ECF. A dendrogram is produced by G03EHF and printed. G03EJF finds two clusters and the results are printed.

9.1 Program Text

```
    Program g03ejfe

!     G03EJF Example Program Text

!     Mark 24 Release. NAG Copyright 2012.

!     .. Use Statements ..
      Use nag_library, Only: g03eaf, g03ecf, g03ejf, nag_wp
!     .. Implicit None Statement ..
      Implicit None
!     .. Parameters ..
      Integer, Parameter               :: nin = 5, nout = 6, rnlen = 3
!     .. Local Scalars ..
      Real (Kind=nag_wp)               :: dlevel
      Integer                          :: i, ifail, k, ld, ldx, liwk, m,       &
                                          method, n, n1
      Character (1)                    :: dist, scal, update
!     .. Local Arrays ..
      Real (Kind=nag_wp), Allocatable  :: cd(:), d(:), dord(:), s(:), x(:,:)
      Integer, Allocatable             :: ic(:), ilc(:), iord(:), isx(:),      &
                                          iuc(:), iwk(:)
      Character (rnlen), Allocatable   :: row_name(:)
!     .. Executable Statements ..
      Write (nout,*) 'G03EJF Example Program Results'
      Write (nout,*)

!     Skip heading in data file
      Read (nin,*)

!     Read in the problem size
      Read (nin,*) n, m

!     Read in information on the type of distance matrix to use
      Read (nin,*) update, dist, scal

      ldx = n
      ld = n*(n-1)/2
      n1 = n - 1
      liwk = 2*n
      Allocate (x(ldx,m),isx(m),s(m),d(ld),ilc(n1),iuc(n1),cd(n1),iord(n), &
```

```
        dord(n),iwk(liwk),ic(n),row_name(n))

!     Read in the data used to construct distance matrix
      Read (nin,*)(x(i,1:m),i=1,n)

!     Read in variable inclusion flags
      Read (nin,*) isx(1:m)

!     Read in scaling
      If (scal=='G' .Or. scal=='g') Then
        Read (nin,*) s(1:m)
      End If

!     Compute the distance matrix
      ifail = 0
      Call g03eaf(update,dist,scal,n,m,x,ldx,isx,s,d,ifail)

!     Read in information on the clustering method to use
      Read (nin,*) method

!     Read in first RNLEN characters of row names. Used to make example
!     output easier to read
      Read (nin,*) row_name(1:n)

!     Perform clustering
      ifail = 0
      Call g03ecf(method,n,d,ilc,iuc,cd,iord,dord,iwk,ifail)

!     Display full clustering information
      Write (nout,*) '  Distance   Clusters Joined'
      Write (nout,*)
      Do i = 1, n - 1
        Write (nout,99999) cd(i), row_name(ilc(i)), row_name(iuc(i))
      End Do
      Write (nout,*)

!     Read in number of clusters required (K) and
!     distance (DLEVEL). If K > 0 then DLEVEL is
!     ignored (i.e. attempt to find K clusters,
!     irrespective of distance), else all clusters at
!     level DLEVEL are used
      Read (nin,*) k, dlevel

!     Compute cluster indicator
      ifail = 0
      Call g03ejf(n,cd,iord,dord,k,dlevel,ic,ifail)

!     Display the indicators
      Write (nout,99998) ' Allocation to ', k, ' clusters'
      Write (nout,99996) ' Clusters found at distance ', dlevel
      Write (nout,*)
      Write (nout,*) ' Object  Cluster'
      Write (nout,*)
      Write (nout,99997)(row_name(i),ic(i),i=1,n)

99999 Format (1X,F10.3,5X,2A)
99998 Format (1X,A,I0,A)
99997 Format (6X,A,5X,I2)
99996 Format (1X,A,F0.3)
    End Program g03ejfe
```

9.2 Program Data

```
G03EJF Example Program Data
5 3                  : N,M (G03EAF)
'I' 'S' 'U'          : UPDATE,DIST,SCAL (G03EAF)
 1  5.0 2.0
 2  1.0 1.0
 3  4.0 3.0
 4  1.0 2.0
```

```
 5  5.0 0.0          : End of X (G03EAF)
 0   1   1           : ISX
5                    : METHOD (G03ECF)
'A' 'B' 'C' 'D' 'E'  : Row names (NAME)
2 0.0                : K, DLEVEL
```

9.3 Program Results

```
 G03EJF Example Program 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
```