NAG Fortran 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              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, the user 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: the number of objects, n.

Constraint: $\mathrm{N} \geq 2$.

2: CD(N−1) – ***real*** array *Input*

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

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

3: IORD(N) – INTEGER array *Input*

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

4: DORD(N) – ***real*** 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*** *Input/Output*

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

Constraint: if K ≤ 0 then DLEVEL > 0.0.

On exit: if K > 0 on entry, then 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. Users who are unfamiliar with this parameter should refer to Chapter P01 for details.

On exit: IFAIL $= 0$ unless the routine detects an error (see Section 6).

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, for users 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.**

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

Note: the listing of the example program presented below uses ***bold italicised*** terms to denote precision-dependent details. Please read the Users' Note for your implementation to check the interpretation of these terms. As explained in the Essential Introduction to this manual, the results produced may not be identical for all implementations.

```
*     G03EJF Example Program Text
*     Mark 18 Revised.  NAG Copyright 1997.
*     .. Parameters ..
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
      INTEGER          NMAX, MMAX, LENC
      PARAMETER        (NMAX=10,MMAX=10,LENC=20)
*     .. Local Scalars ..
      real             DLEVEL, DMIN, DSTEP, YDIST
      INTEGER          I, IFAIL, J, K, LDX, M, METHOD, N, NSYM
      CHARACTER        DIST, SCALE, UPDATE
*     .. Local Arrays ..
      real             CD(NMAX-1), D(NMAX*(NMAX-1)/2), DORD(NMAX),
     +                 S(MMAX), X(NMAX,MMAX)
      INTEGER          IC(NMAX), ILC(NMAX-1), IORD(NMAX), ISX(MMAX),
     +                 IUC(NMAX-1), IWK(2*NMAX)
      CHARACTER*60     C(LENC)
      CHARACTER*3      NAME(NMAX)
*     .. External Subroutines ..
      EXTERNAL         G03EAF, G03ECF, G03EHF, G03EJF
*     .. Intrinsic Functions ..
      INTRINSIC        real, MOD
*     .. Executable Statements ..
      WRITE (NOUT,*) 'G03EJF Example Program Results'
*     Skip heading in data file
      READ (NIN,*)
      READ (NIN,*) N, M
      IF (N.LE.NMAX .AND. M.LE.MMAX) THEN
         READ (NIN,*) METHOD
         READ (NIN,*) UPDATE, DIST, SCALE
         DO 20 J = 1, N
            READ (NIN,*) (X(J,I),I=1,M), NAME(J)
   20    CONTINUE
         READ (NIN,*) (ISX(I),I=1,M)
         READ (NIN,*) (S(I),I=1,M)
         READ (NIN,*) K, DLEVEL
*
*     Compute the distance matrix
*
         IFAIL = 0
         LDX = NMAX
*
```

```
         CALL G03EAF(UPDATE,DIST,SCALE,N,M,X,LDX,ISX,S,D,IFAIL)
*
*     Perform clustering
*
         IFAIL = 0
*
         CALL G03ECF(METHOD,N,D,ILC,IUC,CD,IORD,DORD,IWK,IFAIL)
*
         WRITE (NOUT,*)
         WRITE (NOUT,*) '  Distance   Clusters Joined'
         WRITE (NOUT,*)
         DO 40 I = 1, N - 1
            WRITE (NOUT,99999) CD(I), NAME(ILC(I)), NAME(IUC(I))
   40    CONTINUE
*
*     Produce dendrogram
*
         IFAIL = 0
         NSYM = LENC
         DMIN = 0.0e0
         DSTEP = (CD(N-1))/real(NSYM)
*
         CALL G03EHF('S',N,DORD,DMIN,DSTEP,NSYM,C,LENC,IFAIL)
*
         WRITE (NOUT,*)
         WRITE (NOUT,*) 'Dendrogram'
         WRITE (NOUT,*)
         YDIST = CD(N-1)
         DO 60 I = 1, NSYM
            IF (MOD(I,3).EQ.1) THEN
               WRITE (NOUT,99999) YDIST, C(I)
            ELSE
               WRITE (NOUT,99998) C(I)
            END IF
            YDIST = YDIST - DSTEP
   60    CONTINUE
         WRITE (NOUT,*)
         WRITE (NOUT,99998) (NAME(IORD(I)),I=1,N)
         IFAIL = 0
*
         CALL G03EJF(N,CD,IORD,DORD,K,DLEVEL,IC,IFAIL)
*
         WRITE (NOUT,*)
         WRITE (NOUT,99997) ' Allocation to ', K, ' clusters'
         WRITE (NOUT,*)
         WRITE (NOUT,*) ' Object  Cluster'
         WRITE (NOUT,*)
         DO 80 I = 1, N
            WRITE (NOUT,99996) NAME(I), IC(I)
   80    CONTINUE
      END IF
      STOP
*
99999 FORMAT (F10.3,5X,2A)
99998 FORMAT (15X,20A)
99997 FORMAT (A,I2,A)
99996 FORMAT (5X,A,5X,I2)
      END
```

9.2 Program Data

```
G03EJF Example Program Data
5 3
5
'I' 'S' 'U'
 1  5.0 2.0 'A  '
 2  1.0 1.0 'B  '
 3  4.0 3.0 'C  '
 4  1.0 2.0 'D  '
 5  5.0 0.0 'E  '
 0   1   1
1.0 1.0 1.0
2 0.0
```

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

 Dendrogram

    14.125           -------
                     I     I
                     I     I
    12.006           I     I
                     I     I
                     I     I
     9.887           I     I
                     I     I
                     I     I
     7.769           I     I
                  ---*     I
                  I  I     I
     5.650        I  I     I
                  I  I     I
                  I  I     I
     3.531        I  I     I
                  I  I     I
                ---*  I     I
     1.412      I  I  I  ---*
                I  I  I  I  I

                A  C  E  B  D

 Allocation to  2 clusters

  Object  Cluster

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