

NAG Library Routine Document

G03EFF

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

G03EFF performs K -means cluster analysis.

2 Specification

```

SUBROUTINE G03EFF(WEIGHT, N, M, X, LDX, ISX, NVAR, K, CMEANS, LDC, WT,
1          INC, NIC, CSS, CSW, MAXIT, IWK, WK, IFAIL)

  INTEGER          N, M, LDX, ISX(M), NVAR, K, LDC, INC(N), NIC(K),
1          MAXIT, IWK(N+3*K), IFAIL
  double precision X(LDX,M), CMEANS(LDC,NVAR), WT(*), CSS(K), CSW(K),
1          WK(N+2*K)
  CHARACTER*1      WEIGHT

```

3 Description

Given n objects with p variables measured on each object, x_{ij} for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, p$, G03EFF allocates each object to one of K groups or clusters to minimize the within-cluster sum of squares:

$$\sum_{k=1}^K \sum_{i \in S_k} \sum_{j=1}^p (x_{ij} - \bar{x}_{kj})^2,$$

where S_k is the set of objects in the k th cluster and $\bar{x}_{kj}$ is the mean for the variable j over cluster k . This is often known as K -means clustering.

In addition to the data matrix, a K by p matrix giving the initial cluster centres for the K clusters is required. The objects are then initially allocated to the cluster with the nearest cluster mean. Given the initial allocation, the procedure is to iteratively search for the K -partition with locally optimal within-cluster sum of squares by moving points from one cluster to another.

Optionally, weights for each object, w_i , can be used so that the clustering is based on within-cluster weighted sums of squares:

$$\sum_{k=1}^K \sum_{i \in S_k} \sum_{j=1}^p w_i (x_{ij} - \tilde{x}_{kj})^2,$$

where $\tilde{x}_{kj}$ is the weighted mean for variable j over cluster k .

The routine is based on the algorithm of Hartigan and Wong (1979).

4 References

Everitt B S (1974) *Cluster Analysis* Heinemann

Hartigan J A and Wong M A (1979) Algorithm AS 136: A K -means clustering algorithm *Appl. Statist.* **28** 100–108

Kendall M G and Stuart A (1976) *The Advanced Theory of Statistics (Volume 3)* (3rd Edition) Griffin

Krzanowski W J (1990) *Principles of Multivariate Analysis* Oxford University Press

5 Parameters

- 1: WEIGHT – CHARACTER*1 *Input*
On entry: indicates if weights are to be used.
 WEIGHT = 'U'
 No weights are used.
 WEIGHT = 'W'
 Weights are used and must be supplied in WT.
Constraint: WEIGHT = 'U' or 'W'.

- 2: N – INTEGER *Input*
On entry: n , the number of objects.
Constraint: $N > 1$.

- 3: M – INTEGER *Input*
On entry: the total number of variables in array X.
Constraint: $M \geq \text{NVAR}$.

- 4: X(LDX,M) – **double precision** array *Input*
On entry: $X(i, j)$ must contain the value of the j th variable for the i th object, for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, M$.

- 5: LDX – INTEGER *Input*
On entry: the first dimension of the array X as declared in the (sub)program from which G03EFF is called.
Constraint: $\text{LDX} \geq N$.

- 6: ISX(M) – INTEGER array *Input*
On entry: $\text{ISX}(j)$ indicates whether or not the j th variable is to be included in the analysis. If $\text{ISX}(j) > 0$, the variable contained in the j th column of X is included, for $j = 1, 2, \dots, M$.
Constraint: $\text{ISX}(j) > 0$ for NVAR values of j .

- 7: NVAR – INTEGER *Input*
On entry: p , the number of variables included in the sums of squares calculations.
Constraint: $1 \leq \text{NVAR} \leq M$.

- 8: K – INTEGER *Input*
On entry: K , the number of clusters.
Constraint: $K \geq 2$.

- 9: CMEANS(LDC,NVAR) – **double precision** array *Input/Output*
On entry: $\text{CMEANS}(i, j)$ must contain the value of the j th variable for the i th initial cluster centre, for $i = 1, 2, \dots, K$ and $j = 1, 2, \dots, p$.
On exit: $\text{CMEANS}(i, j)$ contains the value of the j th variable for the i th computed cluster centre, for $i = 1, 2, \dots, K$ and $j = 1, 2, \dots, p$.

- 10: LDC – INTEGER *Input*
On entry: the first dimension of the array CMEANS as declared in the (sub)program from which G03EFF is called.
Constraint: $LDC \geq K$.
- 11: WT(*) – **double precision** array *Input*
Note: the dimension of the array WT must be at least N if WEIGHT = 'W', and at least 1 otherwise.
On entry: if WEIGHT = 'W', the first n elements of WT must contain the weights to be used.
 If $WT(i) = 0.0$, the i th observation is not included in the analysis. The effective number of observation is the sum of the weights.
 If WEIGHT = 'U', WT is not referenced and the effective number of observations is n .
Constraint: if WEIGHT = 'W', $WT(i) \geq 0.0$ and $WT(i) > 0.0$ for at least two values of i , for $i = 1, 2, \dots, n$.
- 12: INC(N) – INTEGER array *Output*
On exit: INC(i) contains the cluster to which the i th object has been allocated, for $i = 1, 2, \dots, n$.
- 13: NIC(K) – INTEGER array *Output*
On exit: NIC(i) contains the number of objects in the i th cluster, for $i = 1, 2, \dots, K$.
- 14: CSS(K) – **double precision** array *Output*
On exit: CSS(i) contains the within-cluster (weighted) sum of squares of the i th cluster, for $i = 1, 2, \dots, K$.
- 15: CSW(K) – **double precision** array *Output*
On exit: CSW(i) contains the within-cluster sum of weights of the i th cluster, for $i = 1, 2, \dots, K$.
 If WEIGHT = 'U', the sum of weights is the number of objects in the cluster.
- 16: MAXIT – INTEGER *Input*
On entry: the maximum number of iterations allowed in the analysis.
Constraint: MAXIT > 0.
Suggested value: MAXIT = 10.
- 17: IWK(N + 3 × K) – INTEGER array *Workspace*
- 18: WK(N + 2 × K) – **double precision** array *Workspace*
- 19: 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.
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, 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.**

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, $WEIGHT \neq 'W'$ or $'U'$,
 or $N < 2$,
 or $NVAR < 1$,
 or $M < NVAR$,
 or $K < 2$,
 or $LDX < N$,
 or $LDC < K$,
 or $MAXIT \leq 0$.

$IFAIL = 2$

On entry, $WEIGHT = 'W'$ and a value of $WT(i) < 0.0$ for some i ,
 or $WEIGHT = 'W'$ and $WT(i) = 0.0$ for all or all but one values of i .

$IFAIL = 3$

On entry, the number of positive values in ISX does not equal $NVAR$.

$IFAIL = 4$

On entry, at least one cluster is empty after the initial assignment. Try a different set of initial cluster centres in $CMEANS$ and also consider decreasing the value of K . The empty clusters may be found by examining the values in NIC .

$IFAIL = 5$

Convergence has not been achieved within the maximum number of iterations given by $MAXIT$. Try increasing $MAXIT$ and, if possible, use the returned values in $CMEANS$ as the initial cluster centres.

7 Accuracy

G03EFF produces clusters that are locally optimal; the within-cluster sum of squares may not be decreased by transferring a point from one cluster to another, but different partitions may have the same or smaller within-cluster sum of squares.

8 Further Comments

The time per iteration is approximately proportional to npK .

9 Example

The data consists of observations of five variables on twenty soils (see Hartigan and Wong (1979)). The data is read in, the K -means clustering performed and the results printed.

9.1 Program Text

```

*      G03EFF Example Program Text
*      Mark 20 Revised. NAG Copyright 2001.
*      .. Parameters ..
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
      INTEGER          NMAX, MMAX, KMAX
      PARAMETER        (NMAX=20,MMAX=5,KMAX=3)
*      .. Local Scalars ..
      INTEGER          I, IFAIL, J, K, LDC, LDX, M, MAXIT, N, NVAR
      CHARACTER        WEIGHT
*      .. Local Arrays ..
      DOUBLE PRECISION CMEANS(KMAX,MMAX), CSS(KMAX), CSW(KMAX),
+      WK(NMAX+2*KMAX), WT(NMAX), X(NMAX,MMAX)
      INTEGER          INC(NMAX), ISX(MMAX), IWK(NMAX+3*KMAX), NIC(KMAX)
*      .. External Subroutines ..
      EXTERNAL         G03EFF
*      .. Executable Statements ..
      WRITE (NOUT,*) 'G03EFF Example Program Results'
*      Skip heading in the data file
      READ (NIN,*)
      READ (NIN,*) WEIGHT, N, M, NVAR, K, MAXIT
      IF (N.LE.NMAX .AND. M.LE.MMAX) THEN
        IF (WEIGHT.EQ.'W' .OR. WEIGHT.EQ.'w') THEN
          DO 20 I = 1, N
            READ (NIN,*) (X(I,J),J=1,M), WT(I)
20          CONTINUE
        ELSE
          DO 40 I = 1, N
            READ (NIN,*) (X(I,J),J=1,M)
40          CONTINUE
        END IF
        DO 60 I = 1, K
          READ (NIN,*) (CMEANS(I,J),J=1,NVAR)
60          CONTINUE
          READ (NIN,*) (ISX(J),J=1,M)
          LDX = NMAX
          LDC = KMAX
          IFAIL = 1
*
          CALL G03EFF(WEIGHT,N,M,X,LDX,ISX,NVAR,K,CMEANS,LDC,WT,INC,NIC,
+          CSS,CSW,MAXIT,IWK,WK,IFAIL)
*
          IF (IFAIL.EQ.0) THEN
            WRITE (NOUT,*)
            WRITE (NOUT,*) ' The cluster each point belongs to'
            WRITE (NOUT,99999) (INC(I),I=1,N)
            WRITE (NOUT,*)
            WRITE (NOUT,*) ' The number of points in each cluster'
            WRITE (NOUT,99999) (NIC(I),I=1,K)
            WRITE (NOUT,*)
            WRITE (NOUT,*)
            +      ' The within-cluster sum of weights of each cluster'
            WRITE (NOUT,99998) (CSW(I),I=1,K)
            WRITE (NOUT,*)
            WRITE (NOUT,*)
            +      ' The within-cluster sum of squares of each cluster'
            WRITE (NOUT,99997) (CSS(I),I=1,K)
            WRITE (NOUT,*)
            WRITE (NOUT,*) ' The final cluster centres'
            WRITE (NOUT,*)
            +      '          1          2          3          4          5'
            DO 80 I = 1, K
              WRITE (NOUT,99996) I, (CMEANS(I,J),J=1,NVAR)
80            CONTINUE
          ELSE
            WRITE (NOUT,*)
            WRITE (NOUT,99995) ' ** G03EFF returned with IFAIL = ',
            +      IFAIL
            END IF

```

```

      END IF
*
99999 FORMAT (1X,10I6)
99998 FORMAT (1X,5F9.2)
99997 FORMAT (1X,5F13.4)
99996 FORMAT (1X,I5,5X,5F8.4)
99995 FORMAT (1X,A,I5)
      END

```

9.2 Program Data

G03EFF Example Program Data

```

'u' 20 5 5 3 10      : WEIGHT N M NVAR K MAXIT
77.3 13.0  9.7 1.5 6.4
82.5 10.0  7.5 1.5 6.5
66.9 20.6 12.5 2.3 7.0
47.2 33.8 19.0 2.8 5.8
65.3 20.5 14.2 1.9 6.9
83.3 10.0  6.7 2.2 7.0
81.6 12.7  5.7 2.9 6.7
47.8 36.5 15.7 2.3 7.2
48.6 37.1 14.3 2.1 7.2
61.6 25.5 12.9 1.9 7.3
58.6 26.5 14.9 2.4 6.7
69.3 22.3  8.4 4.0 7.0
61.8 30.8  7.4 2.7 6.4
67.7 25.3  7.0 4.8 7.3
57.2 31.2 11.6 2.4 6.5
67.2 22.7 10.1 3.3 6.2
59.2 31.2  9.6 2.4 6.0
80.2 13.2  6.6 2.0 5.8
82.2 11.1  6.7 2.2 7.2
69.7 20.7  9.6 3.1 5.9      : End of X (and WTs if used)
82.5 10.0  7.5 1.5 6.5
47.8 36.5 15.7 2.3 7.2
67.2 22.7 10.1 3.3 6.2      : End of CMEANS
1 1 1 1 1      : ISX

```

9.3 Program Results

G03EFF Example Program Results

The cluster each point belongs to

1	1	3	2	3	1	1	2	2	3
3	3	3	3	3	3	3	1	1	3

The number of points in each cluster

6	3	11
---	---	----

The within-cluster sum of weights of each cluster

6.00	3.00	11.00
------	------	-------

The within-cluster sum of squares of each cluster

46.5717	20.3800	468.8964
---------	---------	----------

The final cluster centres

	1	2	3	4	5
1	81.1833	11.6667	7.1500	2.0500	6.6000
2	47.8667	35.8000	16.3333	2.4000	6.7333
3	64.0455	25.2091	10.7455	2.8364	6.6545