NAG Fortran Library Routine Document

G03FCF

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

G03FCF performs non-metric (ordinal) multidimensional scaling.

2 Specification

```
 SUBROUTINE G03FCF (TYP, N, NDIM, D, X, LDX, STRESS, DFIT, ITER, IOPT,
1                   WK, IWK, IFAIL)
 INTEGER            N, NDIM, LDX, ITER, IOPT, IWK(N*(N-1)/2+N*NDIM+5),
1                   IFAIL
 double precision   D(N*(N-1)/2), X(LDX,NDIM), STRESS, DFIT(2*N*(N-1)),
1                   WK(15*N*NDIM)
 CHARACTER*1        TYP
```

3 Description

For a set of n objects, a distance or dissimilarity matrix D can be calculated such that d_{ij} is a measure of how 'far apart' the objects i and j are. If p variables x_k have been recorded for each observation this measure may be based on Euclidean distance, $d_{ij} = \sum_{k=1}^{p} \left(x_{ki} - x_{kj}\right)^2$, or some other calculation such as the number of variables for which $x_{kj} \neq x_{ki}$. Alternatively, the distances may be the result of a subjective assessment. For a given distance matrix, multidimensional scaling produces a configuration of n points in a chosen number of dimensions, m, such that the distance between the points in some way best matches the distance matrix. For some distance measures, such as Euclidean distance, the size of distance is meaningful, for other measures of distance all that can be said is that one distance is greater or smaller than another. For the former metric scaling can be used, see G03FAF, for the latter, a non-metric scaling is more appropriate.

For non-metric multidimensional scaling, the criterion used to measure the closeness of the fitted distance matrix to the observed distance matrix is known as STRESS. STRESS is given by,

$$\sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{i-1}\left(\hat{d}_{ij} - \tilde{d}_{ij}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{i-1}\hat{d}_{ij}^{\,2}}}$$

where $\hat{d}_{ij}^{\,2}$ is the Euclidean squared distance between points i and j and $\tilde{d}_{ij}$ is the fitted distance obtained when $\hat{d}_{ij}$ is monotonically regressed on d_{ij}, that is $\tilde{d}_{ij}$ is monotonic relative to d_{ij} and is obtained from $\hat{d}_{ij}$ with the smallest number of changes. So STRESS is a measure of by how much the set of points preserve the order of the distances in the original distance matrix. Non-metric multidimensional scaling seeks to find the set of points that minimize the STRESS.

An alternate measure is squared STRESS, $sstress$,

$$\sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{i-1}\left(\hat{d}_{ij}^{\,2}-\tilde{d}_{ij}^{\,2}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{i-1}\hat{d}_{ij}^{\,4}}}$$

in which the distances in STRESS are replaced by squared distances.

In order to perform a non-metric scaling, an initial configuration of points is required. This can be obtained from principal co-ordinate analysis, see G03FAF. Given an initial configuration, G03FCF uses the optimization routine E04DGF/E04DGA to find the configuration of points that minimizes STRESS or $sstress$. The routine E04DGF/E04DGA uses a conjugate gradient algorithm. G03FCF will find an optimum that may only be a local optimum, to be more sure of finding a global optimum several different initial configurations should be used; these can be obtained by randomly perturbing the original initial configuration using routines from Chapter G05.

4 References

Chatfield C and Collins A J (1980) *Introduction to Multivariate Analysis* Chapman and Hall

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

5 Parameters

1: TYP – CHARACTER*1 *Input*

On entry: indicates whether STRESS or $sstress$ is to be used as the criterion.

If TYP = 'T' STRESS is used.

If TYP = 'S' $sstress$ is used.

Constraint: TYP = 'S' or 'T'.

2: N – INTEGER *Input*

On entry: n, the number of objects in the distance matrix.

Constraint: N > NDIM.

3: NDIM – INTEGER *Input*

On entry: m, the number of dimensions used to represent the data.

Constraint: NDIM $\geq$ 1.

4: D(N × (N – 1)/2) – ***double precision*** array *Input*

On entry: the lower triangle of the distance matrix D stored packed by rows. That is $\mathrm{D}((i-1)\times(i-2)/2+j)$ must contain d_{ij}, for $i=2,3,\ldots,n; j=1,2,\ldots,i-1$. If d_{ij} is missing then set $d_{ij}<0$; for further comments on missing values see Section 8.

5: X(LDX,NDIM) – ***double precision*** array *Input/Output*

On entry: the ith row must contain an initial estimate of the co-ordinates for the ith point, for $i=1,2,\ldots,n$. One method of computing these is to use G03FAF.

On exit: the ith row contains m co-ordinates for the ith point, for $i=1,2,\ldots,n$.

6: LDX – INTEGER *Input*

On entry: the first dimension of the array X as declared in the (sub)program from which G03FCF is called.

Constraint: LDX $\geq$ N.

7: STRESS – ***double precision*** *Output*

On exit: the value of STRESS or $sstress$ at the final iteration.

8: DFIT($2 \times$ N $\times$ (N $-$ 1)) – ***double precision*** array *Output*

On exit: auxiliary outputs.

If TYP = 'T', the first $n(n-1)/2$ elements contain the distances, $\hat{d}_{ij}$, for the points returned in X, the second set of $n(n-1)/2$ contains the distances $\hat{d}_{ij}$ ordered by the input distances, d_{ij}, the third set of $n(n-1)/2$ elements contains the monotonic distances, $\tilde{d}_{ij}$, ordered by the input distances, d_{ij} and the final set of $n(n-1)/2$ elements contains fitted monotonic distances, $\tilde{d}_{ij}$, for the points in X. The $\tilde{d}_{ij}$ corresponding to distances which are input as missing are set to zero.

If TYP = 'S', the results are as above except that the squared distances are returned.

Each distance matrix is stored in lower triangular packed form in the same way as the input matrix D.

9: ITER – INTEGER *Input*

On entry: the maximum number of iterations in the optimization process.

ITER = 0

A default value of 50 is used.

ITER < 0

A default value of $\max(50, 5nm)$ (the default for E04DGF/E04DGA) is used.

10: IOPT – INTEGER *Input*

On entry: selects the options, other than the number of iterations, that control the optimization.

IOPT = 0

Default values are selected as described in Section 8. In particular if an accuracy requirement of $\epsilon = 0.00001$ is selected, see Section 7.

IOPT > 0

The default values are used except that the accuracy is given by 10^{-i} where $i =$ IOPT.

IOPT < 0

The option setting mechanism of E04DGF/E04DGA can be used to set all options except **Iteration Limit**; this option is only recommended if you are an experienced user of NAG optimization routines. For further details see E04DGF/E04DGA.

11: WK($15 \times$ N $\times$ NDIM) – ***double precision*** array *Workspace*

12: IWK(N $\times$ (N $-$ 1)/2 + N $\times$ NDIM + 5) – INTEGER array *Workspace*

13: 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 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, 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, NDIM < 1,
or N $\leq$ NDIM,
or TYP $\neq$ 'T' or 'S',
or LDX $<$ N.

IFAIL $= 2$

On entry, all elements of D ≤ 0.0.

IFAIL $= 3$

The optimization has failed to converge in ITER function iterations. Try either increasing the number of iterations using ITER or increasing the value of ϵ, given by IOPT, used to determine convergence. Alternatively try a different starting configuration.

IFAIL $= 4$

The conditions for an acceptable solution have not been met but a lower point could not be found. Try using a larger value of ϵ, given by IOPT.

IFAIL $= 5$

The optimization cannot begin from the initial configuration. Try a different set of points.

IFAIL $= 6$

The optimization has failed. This error is only likely if IOPT < 0. It corresponds to IFAIL $= 4$, 7 and 9 in E04DGF/E04DGA.

7 Accuracy

After a successful optimization the relative accuracy of STRESS should be approximately ϵ, as specified by IOPT.

8 Further Comments

The optimization routine E04DGF/E04DGA used by G03FCF has a number of options to control the process. The options for the maximum number of iterations (**Iteration Limit**) and accuracy (**Optimality Tolerance**) can be controlled by ITER and IOPT respectively. The printing option (**Print Level**) is set to -1 to give no printing. The other option set is to stop the checking of derivatives (**Verify** $=$ No) for efficiency. All other options are left at their default values. If however IOPT < 0 is used, only the maximum number of iterations is set. All other options can be controlled by the option setting mechanism of E04DGF/E04DGA with the defaults as given by that routine.

Missing values in the input distance matrix can be specified by a negative value and providing there are not more than about two thirds of the values missing the algorithm may still work. However the routine G03FAF does not allow for missing values so an alternative method of obtaining an initial set of co-ordinates is required. It may be possible to estimate the missing values with some form of average and then use G03FAF to give an initial set of co-ordinates.

9 Example

The data, given by Krzanowski (1990), are dissimilarities between water vole populations in Europe. Initial estimates are provided by the first two principal co-ordinates computed by G03FAF. The two dimension solution is computed using G03FCF and then plotted using G01AGF.

9.1 Program Text

```
*     G03FCF Example Program Text
*     Mark 17 Release. NAG Copyright 1995.
*     .. Parameters ..
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
      INTEGER          NMAX, MMAX, NNMAX
      PARAMETER        (NMAX=14,MMAX=2,NNMAX=NMAX*(NMAX-1)/2)
*     .. Local Scalars ..
      DOUBLE PRECISION STRESS
      INTEGER          I, IFAIL, IOPT, ITER, J, LDX, N, NDIM, NN
      CHARACTER        TYPE
*     .. Local Arrays ..
      DOUBLE PRECISION D(NNMAX), DFIT(4*NNMAX), WK(NNMAX+15*NMAX*MMAX),
     +                 X(NMAX,NMAX)
      INTEGER          IWK(NNMAX+NMAX*NMAX+5)
*     .. External Subroutines ..
      EXTERNAL         G01AGF, G03FAF, G03FCF
*     .. Executable Statements ..
      WRITE (NOUT,*) 'G03FCF Example Program Results'
*     Skip heading in data file
      READ (NIN,*)
      READ (NIN,*) N, NDIM, TYPE
      IF (N.LE.NMAX) THEN
         NN = N*(N-1)/2
         READ (NIN,*) (D(I),I=1,NN)
         LDX = NMAX
         IFAIL = 0
         CALL G03FAF('L',N,D,NDIM,X,LDX,WK,WK(N+1),IWK,IFAIL)
         ITER = 0
         IOPT = 0
         IFAIL = 0
*
         CALL G03FCF(TYPE,N,NDIM,D,X,LDX,STRESS,DFIT,ITER,IOPT,WK,IWK,
     +               IFAIL)
*
         WRITE (NOUT,*)
         WRITE (NOUT,99999) STRESS
         WRITE (NOUT,*)
         WRITE (NOUT,*) ' Co-ordinates'
         WRITE (NOUT,*)
         DO 20 I = 1, N
            WRITE (NOUT,99998) (X(I,J),J=1,NDIM)
   20    CONTINUE
         WRITE (NOUT,*)
         WRITE (NOUT,*) ' Plot of first two dimensions'
         WRITE (NOUT,*)
         IFAIL = 0
         CALL G01AGF(X(1,1),X(1,2),N,IWK,50,18,IFAIL)
      END IF
      STOP
*
99999 FORMAT (10X,'STRESS = ',E13.4)
99998 FORMAT (8F10.4)
      END
```

9.2 Program Data

```
G03FCF Example Program Data

14 2 'T'

0.099
0.033 0.022
0.183 0.114 0.042
0.148 0.224 0.059 0.068
0.198 0.039 0.053 0.085 0.051
0.462 0.266 0.322 0.435 0.268 0.025
0.628 0.442 0.444 0.406 0.240 0.129 0.014
0.113 0.070 0.046 0.047 0.034 0.002 0.106 0.129
0.173 0.119 0.162 0.331 0.177 0.039 0.089 0.237 0.071
0.434 0.419 0.339 0.505 0.469 0.390 0.315 0.349 0.151 0.430
0.762 0.633 0.781 0.700 0.758 0.625 0.469 0.618 0.440 0.538 0.607
0.530 0.389 0.482 0.579 0.597 0.498 0.374 0.562 0.247 0.383 0.387 0.084
0.586 0.435 0.550 0.530 0.552 0.509 0.369 0.471 0.234 0.346 0.456 0.090 0.038
```

9.3 Program Results

```
 G03FCF Example Program Results

          STRESS =    0.1256E+00

  Co-ordinates

    0.2060    0.2438
    0.1063    0.1418
    0.2224    0.0817
    0.3032    0.0355
    0.2645   -0.0698
    0.1554   -0.0435
   -0.0070   -0.1612
    0.0749   -0.3275
    0.0488    0.0289
    0.0124   -0.0267
   -0.1649   -0.2500
   -0.5073    0.1267
   -0.3093    0.1590
   -0.3498    0.0700

  Plot of first two dimensions

            +....+....+....+....+....+....+....+....+....+....+.
            .                             .                    .
            .                             .                    .
            .                             .         1          .
       0.200+                             +                    +
            .              1              .    1               .
            .    1                        .                    .
            .            1                .          1         .
            .                             . 1            1     .
       0.000+....+....+....+....+....+....+....+....+....+....++
            .                             .1      1            .
            .                             .            1       .
            .                             .                    .
            .                             1                    .
      -0.200+                             +                    +
            .                     1       .                    .
            .                             .                    .
            .                             .   1                .
            .                             .                    .
      -0.400+....+....+....+....+....+....+....+....+....+....+.
           -.6000    -.4000    -.2000    0.0000    0.2000    0.4000
                -.5000    -.3000    -.1000    0.1000    0.3000
```