NAG Fortran Library Routine Document

G03DBF

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

G03DBF computes Mahalanobis squared distances for group or pooled variance-covariance matrices. It is intended for use after G03DAF.

2 Specification

```
 SUBROUTINE G03DBF(EQUAL, MODE, NVAR, NG, GMEAN, LDG, GC, NOBS, M, ISX,
1                  X, LDX, D, LDD, WK, IFAIL)
 INTEGER           NVAR, NG, LDG, NOBS, M, ISX(*), LDX, LDD, IFAIL
 real              GMEAN(LDG,NVAR), GC((NG+1)*NVAR*(NVAR+1)/2), X(LDX,*),
1                  D(LDD,NG), WK(2*NVAR)
 CHARACTER*1       EQUAL, MODE
```

3 Description

Consider p variables observed on n_g populations or groups. Let $\bar{x}_j$ be the sample mean and S_j the within-group variance-covariance matrix for the jth group and let x_k be the kth sample point in a data set. A measure of the distance of the point from the jth population or group is given by the Mahalanobis distance, ${D_{kj}}^2$:

$$ {D_{kj}}^2 = (x_k - \bar{x}_j)^{\mathrm{T}} S_j^{-1} (x_k - \bar{x}_j). $$

If the pooled estimated of the variance-covariance matrix S is used rather than the within-group variance-covariance matrices, then the distance is:

$$ {D_{kj}}^2 = (x_k - \bar{x}_j)^{\mathrm{T}} S^{-1} (x_k - \bar{x}_j). $$

Instead of using the variance-covariance matrices S and S_j, G03DBF uses the upper triangular matrices R and R_j supplied by G03DAF such that $S = R^{\mathrm{T}} R$ and $S_j = R_j^{\mathrm{T}} R_j$. ${D_{kj}}^2$ can then be calculated as $z^{\mathrm{T}} z$ where $R_j z = (x_k - \bar{x}_j)$ or $R z = (x_k - \bar{x}_j)$ as appropriate.

A particular case is when the distance between the group or population means is to be estimated. The Mahalanobis distance between the ith and jth groups is:

$$ {D_{ij}}^2 = (\bar{x}_i - \bar{x}_j)^{\mathrm{T}} S_j^{-1} (\bar{x}_i - \bar{x}_j) $$

or

$$ {D_{ij}}^2 = (\bar{x}_i - \bar{x}_j)^{\mathrm{T}} S^{-1} (\bar{x}_i - \bar{x}_j). $$

Note: ${D_{jj}}^2 = 0$ and that in the case when the pooled variance-covariance matrix is used ${D_{ij}}^2 = {D_{ji}}^2$ so in this case only the lower triangular values of ${D_{ij}}^2$, $i > j$, are computed.

4 References

Aitchison J and Dunsmore I R (1975) *Statistical Prediction Analysis* Cambridge

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: EQUAL – CHARACTER*1 *Input*

On entry: indicates whether or not the within-group variance-covariance matrices are assumed to be equal and the pooled variance-covariance matrix used.

If EQUAL = 'E' the within-group variance-covariance matrices are assumed equal and the matrix R stored in the first $p(p+1)/2$ elements of GC is used.

If EQUAL = 'U' the within-group variance-covariance matrices are assumed to be unequal and the matrices R_j, for $j = 1, 2, \ldots, n_g$, stored in the remainder of GC are used.

Constraint: EQUAL = 'E' or 'U'.

2: MODE – CHARACTER*1 *Input*

On entry: indicates whether distances from sample points are to be calculated or distances between the group means.

If MODE = 'S' the distances between the sample points given in X and the group means are calculated.

If MODE = 'M' the distances between the group means will be calculated.

Constraint: MODE = 'M' or 'S'.

3: NVAR – INTEGER *Input*

On entry: the number of variables, p, in the variance-covariance matrices as specified to G03DAF.

Constraint: NVAR $\geq$ 1.

4: NG – INTEGER *Input*

On entry: the number of groups, n_g.

Constraint: NG $\geq$ 2.

5: GMEAN(LDG,NVAR) – ***real*** array *Input*

On entry: the jth row of GMEAN contains the means of the p selected variables for the jth group, for $j = 1, 2, \ldots, n_g$. These are returned by G03DAF.

6: LDG – INTEGER *Input*

On entry: the first dimension of the array GMEAN as declared in the (sub)program from which G03DBF is called.

Constraint: LDG $\geq$ NG.

7: GC((NG+1)∗NVAR∗(NVAR+1)/2) – ***real*** array *Input*

On entry: the first $p(p+1)/2$ elements of GC should contain the upper triangular matrix R and the next n_g blocks of $p(p+1)/2$ elements should contain the upper triangular matrices R_j. All matrices must be stored packed by column. These matrices are returned by G03DAF. If EQUAL = 'E' only the first $p(p+1)/2$ elements are referenced, if EQUAL = 'U' only the elements $p(p+1)/2+1$ to $(n_g+1)p(p+1)/2$ are referenced.

Constraints:

if EQUAL = 'E' the diagonal elements of $R \neq 0.0$,
if EQUAL = 'U' the diagonal elements of the $R_j \neq 0.0$, for $j = 1, 2, \ldots,$NG.

8: NOBS – INTEGER *Input*

On entry: if MODE = ’S’ the number of sample points in X for which distances are to be calculated. If MODE = ’M’, NOBS is not referenced.

Constraint: if MODE = ’S’, NOBS $\geq$ 1.

9: M – INTEGER *Input*

On entry: if MODE = ’S’ the number of variables in the data array X. If MODE = ’M’, then M is not referenced.

Constraint: if MODE = ’S’, M $\geq$ NVAR.

10: ISX($*$) – INTEGER array *Input*

Note: the dimension of the array ISX must be at least $\max(1, \mathrm{M})$.

On entry: if MODE = ’S’, ISX(l) indicates if the lth variable in X is to be included in the distance calculations. If ISX(l) > 0 the lth variable is included, for $l = 1, 2, \ldots, \mathrm{M}$; otherwise the lth variable is not referenced.

If MODE = ’M’, then ISX is not referenced.

Constraint: if MODE = ’S’, ISX(l) > 0 for NVAR values of l.

11: X(LDX,*) – ***real*** array *Input*

Note: the second dimension of the array X must be at least $\max(1, \mathrm{M})$.

On entry: if MODE = ’S’ the kth row of X must contain x_k. That is X(k, l) must contain the kth sample value for the lth variable for $k = 1, 2, \ldots, \mathrm{NOBS}$; $l = 1, 2, \ldots, \mathrm{M}$. Otherwise X is not referenced.

12: LDX – INTEGER *Input*

On entry: the first dimension of the array X as declared in the (sub)program from which G03DBF is called.

Constraint: if MODE = ’S’, LDX $\geq$ NOBS.

13: D(LDD,NG) – ***real*** array *Output*

On exit: the squared distances.

If MODE = ’S’, D(k, j) contains the squared distance of the kth sample point from the jth group mean, ${D_{kj}}^2$, for $k = 1, 2, \ldots, \mathrm{NOBS}$; $j = 1, 2, \ldots, n_g$.

If MODE = ’M’ and EQUAL = ’U’, D(i, j) contains the squared distance between the ith mean and the jth mean, ${D_{ij}}^2$, for $i = 1, 2, \ldots, n_g$; $j = 1, 2, \ldots, i - 1, i + 1, \ldots, n_g$. The elements D($i, i$) are not referenced for $i = 1, 2, \ldots, n_g$.

If MODE = ’M’ and EQUAL = ’E’, D(i, j) contains the squared distance between the ith mean and the jth mean, ${D_{ij}}^2$, for $i = 1, 2, \ldots, n_g$; $j = 1, 2, \ldots, i - 1$. Since $D_{ij} = D_{ji}$ the elements D(i, j) are not referenced, for $i = 1, 2, \ldots, n_g$; $j = i, i + 1, \ldots, n_g$.

14: LDD – INTEGER *Input*

On entry: the first dimension of the array D as declared in the (sub)program from which G03DBF is called.

Constraint: if MODE = ’S’, LDD $\geq$ NOBS; if MODE = ’M’, LDD $\geq$ NG.

15: WK(2*NVAR) – ***real*** array *Workspace*

16: 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, NVAR < 1,
or NG < 2,
or LDG $<$ NG,
or MODE = 'S' and NOBS < 1,
or MODE = 'S' and M $<$ NVAR,
or MODE = 'S' and LDX $<$ NOBS,
or MODE = 'S' and LDD $<$ NOBS,
or MODE = 'M' and LDD $<$ NG,
or EQUAL $\neq$ 'E' or 'U',
or MODE $\neq$ 'M' or 'S'.

IFAIL $= 2$

On entry, MODE = 'S' and the number of variables indicated by ISX is not equal to NVAR,
or EQUAL = 'E' and a diagonal element of R is zero,
or EQUAL = 'U' and a diagonal element of R_j for some j is zero.

7 Accuracy

The accuracy will depend upon the accuracy of the input R or R_j matrices.

8 Further Comments

If the distances are to be used for discrimination, see also G03DCF.

9 Example

The data, taken from Aitchison and Dunsmore (1975), is concerned with the diagnosis of three 'types' of Cushing's syndrome. The variables are the logarithms of the urinary excretion rates (mg/24hr) of two steroid metabolites. Observations for a total of 21 patients are input and the group means and R matrices are computed by G03DAF. A further six observations of unknown type are input, and the distances from the group means of the 21 patients of known type are computed under the assumption that the within-group variance-covariance matrices are not equal. These results are printed and indicate that the first four are close to one of the groups while observations 5 and 6 are some distance from any group.

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.

```
*     G03DBF Example Program Text
*     Mark 15 Release. NAG Copyright 1991.
*     .. Parameters ..
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
      INTEGER          NMAX, MMAX, GPMAX
      PARAMETER        (NMAX=21,MMAX=2,GPMAX=3)
*     .. Local Scalars ..
      real             DF, SIG, STAT
      INTEGER          I, IFAIL, J, M, N, NG, NOBS, NVAR
      CHARACTER        EQUAL, WEIGHT
*     .. Local Arrays ..
      real             D(NMAX,GPMAX), DET(GPMAX),
     +                 GC((GPMAX+1)*MMAX*(MMAX+1)/2), GMEAN(GPMAX,MMAX),
     +                 WK(NMAX*(MMAX+1)), WT(NMAX), X(NMAX,MMAX)
      INTEGER          ING(NMAX), ISX(MMAX), IWK(GPMAX), NIG(GPMAX)
*     .. External Subroutines ..
      EXTERNAL         G03DAF, G03DBF
*     .. Executable Statements ..
      WRITE (NOUT,*) 'G03DBF Example Program Results'
*     Skip headings in data file
      READ (NIN,*)
      READ (NIN,*) N, M, NVAR, NG, WEIGHT
      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), ING(I), WT(I)
   20       CONTINUE
         ELSE
            DO 40 I = 1, N
               READ (NIN,*) (X(I,J),J=1,M), ING(I)
   40       CONTINUE
         END IF
         READ (NIN,*) (ISX(J),J=1,M)
         IFAIL = 0
*
         CALL G03DAF(WEIGHT,N,M,X,NMAX,ISX,NVAR,ING,NG,WT,NIG,GMEAN,
     +               GPMAX,DET,GC,STAT,DF,SIG,WK,IWK,IFAIL)
*
         READ (NIN,*) NOBS, EQUAL
         IF (NOBS.LE.NMAX) THEN
            DO 60 I = 1, NOBS
               READ (NIN,*) (X(I,J),J=1,M)
   60       CONTINUE
            IFAIL = 0
*
            CALL G03DBF(EQUAL,'Sample points',NVAR,NG,GMEAN,GPMAX,GC,
     +                  NOBS,M,ISX,X,NMAX,D,NMAX,WK,IFAIL)
*
            WRITE (NOUT,*)
            WRITE (NOUT,*) '  Obs          Distances'
            WRITE (NOUT,*)
            DO 80 I = 1, NOBS
               WRITE (NOUT,99999) I, (D(I,J),J=1,NG)
   80       CONTINUE
         END IF
      END IF
      STOP
*
99999 FORMAT (1X,I3,3F10.3)
      END
```

9.2 Program Data

```
G03DBF Example Program Data
  21 2 2 3 'U'
  1.1314    2.4596    1
  1.0986    0.2624    1
  0.6419   -2.3026    1
  1.3350   -3.2189    1
  1.4110    0.0953    1
  0.6419   -0.9163    1
  2.1163    0.0000    2
  1.3350   -1.6094    2
  1.3610   -0.5108    2
  2.0541    0.1823    2
  2.2083   -0.5108    2
  2.7344    1.2809    2
  2.0412    0.4700    2
  1.8718   -0.9163    2
  1.7405   -0.9163    2
  2.6101    0.4700    2
  2.3224    1.8563    3
  2.2192    2.0669    3
  2.2618    1.1314    3
  3.9853    0.9163    3
  2.7600    2.0281    3
   1         1
   6 'U'
  1.6292   -0.9163
  2.5572    1.6094
  2.5649   -0.2231
  0.9555   -2.3026
  3.4012   -2.3026
  3.0204   -0.2231
```

9.3 Program Results

```
 G03DBF Example Program Results

   Obs          Distances

   1     3.339     0.752    50.928
   2    20.777     5.656     0.060
   3    21.363     4.841    19.498
   4     0.718     6.280   124.732
   5    55.000    88.860    71.785
   6    36.170    15.785    15.749
```