NAG Library Routine Document

G03DAF

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

G03DAF computes a test statistic for the equality of within-group covariance matrices and also computes matrices for use in discriminant analysis.

2 Specification

```
SUBROUTINE G03DAF (WEIGHT, N, M, X, LDX, ISX, NVAR, ING, NG, WT, NIG, GMN,      &
                   LDGMN, DET, GC, STAT, DF, SIG, WK, IWK, IFAIL)

INTEGER            N, M, LDX, ISX(M), NVAR, ING(N), NG, NIG(NG), LDGMN,         &
                   IWK(NG), IFAIL
REAL (KIND=nag_wp) X(LDX,M), WT(*), GMN(LDGMN,NVAR), DET(NG),                   &
                   GC((NG+1)*NVAR*(NVAR+1)/2), STAT, DF, SIG,                   &
                   WK(N*(NVAR+1))
CHARACTER(1)       WEIGHT
```

3 Description

Let a sample of n observations on p variables come from n_g groups with n_j observations in the jth group and $\sum n_j = n$. If the data is assumed to follow a multivariate Normal distribution with the variance-covariance matrix of the jth group Σ_j, then to test for equality of the variance-covariance matrices between groups, that is, $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_{n_g} = \Sigma$, the following likelihood-ratio test statistic, G, can be used;

$$G = C\left\{ \left(n - n_g\right) \log|S| - \sum_{j=1}^{n_g} \left(n_j - 1\right) \log\left|S_j\right| \right\},$$

where

$$C = 1 - \frac{2p^2 + 3p - 1}{6(p+1)\left(n_g - 1\right)} \left(\sum_{j=1}^{n_g} \frac{1}{\left(n_j - 1\right)} - \frac{1}{\left(n - n_g\right)} \right),$$

and S_j are the within-group variance-covariance matrices and S is the pooled variance-covariance matrix given by

$$S = \frac{\sum_{j=1}^{n_g} \left(n_j - 1\right) S_j}{\left(n - n_g\right)}.$$

For large n, G is approximately distributed as a χ^2 variable with $\frac{1}{2}p(p+1)\left(n_g - 1\right)$ degrees of freedom, see Morrison (1967) for further comments. If weights are used, then S and S_j are the weighted pooled and within-group variance-covariance matrices and n is the effective number of observations, that is, the sum of the weights.

Instead of calculating the within-group variance-covariance matrices and then computing their determinants in order to calculate the test statistic, G03DAF uses a QR decomposition. The group means are subtracted from the data and then for each group, a QR decomposition is computed to give an upper triangular matrix R_j^*. This matrix can be scaled to give a matrix R_j such that $S_j = R_j^{\mathrm{T}} R_j$. The pooled R matrix is then computed from the R_j matrices. The values of $|S|$ and the $\left|S_j\right|$ can then be calculated from the diagonal elements of R and the R_j.

This approach means that the Mahalanobis squared distances for a vector observation x can be computed as $z^{\mathrm{T}}z$, where $R_j z = (x - \bar{x}_j)$, $\bar{x}_j$ being the vector of means of the jth group. These distances can be calculated by G03DBF. The distances are used in discriminant analysis and G03DCF uses the results of G03DAF to perform several different types of discriminant analysis. The differences between the discriminant methods are, in part, due to whether or not the within-group variance-covariance matrices are equal.

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

Morrison D F (1967) *Multivariate Statistical Methods* McGraw–Hill

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 to be used and must be supplied in WT.

Constraint: WEIGHT = 'U' or 'W'.

2: N – INTEGER *Input*

On entry: n, the number of observations.

Constraint: N $\geq$ 1.

3: M – INTEGER *Input*

On entry: the number of variables in the data array X.

Constraint: M $\geq$ NVAR.

4: X(LDX,M) – REAL (KIND=nag_wp) array *Input*

On entry: X(k,l) must contain the kth observation for the lth variable, for $k = 1, 2, \ldots, n$ and $l = 1, 2, \ldots,$ M.

5: LDX – INTEGER *Input*

On entry: the first dimension of the array X as declared in the (sub)program from which G03DAF is called.

Constraint: LDX $\geq$ N.

6: ISX(M) – INTEGER array *Input*

On entry: ISX(l) indicates whether or not the lth variable in X is to be included in the variance-covariance matrices.

If ISX$(l) > 0$ the lth variable is included, for $l = 1, 2, \ldots,$ M; otherwise it is not referenced.

Constraint: ISX$(l) > 0$ for NVAR values of l.

7: NVAR – INTEGER *Input*

On entry: p, the number of variables in the variance-covariance matrices.

Constraint: NVAR ≥ 1.

8: ING(N) – INTEGER array *Input*

On entry: ING(k) indicates to which group the kth observation belongs, for $k = 1, 2, \ldots, n$.

Constraint: $1 \leq$ ING$(k) \leq$ NG, for $k = 1, 2, \ldots, n$

The values of ING must be such that each group has at least NVAR members.

9: NG – INTEGER *Input*

On entry: the number of groups, n_g.

Constraint: NG ≥ 2.

10: WT$(*)$ – REAL (KIND=nag_wp) 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 in the analysis and the effective number of observations for a group is the sum of the weights of the observations in that group. If WT$(k) = 0.0$ the kth observation is excluded from the calculations.

If WEIGHT = 'U', WT is not referenced and the effective number of observations for a group is the number of observations in that group.

Constraint: if WEIGHT = 'W', WT$(k) \geq 0.0$, for $k = 1, 2, \ldots, n$.

11: NIG(NG) – INTEGER array *Output*

On exit: NIG(j) contains the number of observations in the jth group, for $j = 1, 2, \ldots, n_g$.

12: GMN(LDGMN,NVAR) – REAL (KIND=nag_wp) array *Output*

On exit: the jth row of GMN contains the means of the p selected variables for the jth group, for $j = 1, 2, \ldots, n_g$.

13: LDGMN – INTEGER *Input*

On entry: the first dimension of the array GMN as declared in the (sub)program from which G03DAF is called.

Constraint: LDGMN $\geq$ NG.

14: DET(NG) – REAL (KIND=nag_wp) array *Output*

On exit: the logarithm of the determinants of the within-group variance-covariance matrices.

15: GC((NG + 1) × NVAR × (NVAR + 1)/2) – REAL (KIND=nag_wp) array *Output*

On exit: the first $p(p+1)/2$ elements of GC contain R and the remaining n_g blocks of $p(p+1)/2$ elements contain the R_j matrices. All are stored in packed form by columns.

16: STAT – REAL (KIND=nag_wp) *Output*

On exit: the likelihood-ratio test statistic, G.

17: DF – REAL (KIND=nag_wp) *Output*

On exit: the degrees of freedom for the distribution of G.

18: SIG – REAL (KIND=nag_wp) *Output*

On exit: the significance level for G.

19: WK(N × (NVAR + 1)) – REAL (KIND=nag_wp) array *Workspace*

20: IWK(NG) – INTEGER array *Workspace*

21: 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, NVAR < 1,
or N < 1,
or NG < 2,
or M $<$ NVAR,
or LDX $<$ N,
or LDGMN $<$ NG,
or WEIGHT $\neq$ 'U' or 'W'.

IFAIL $= 2$

On entry, WEIGHT = 'W' and a value of WT < 0.0.

IFAIL $= 3$

On entry, there are not exactly NVAR elements of ISX > 0,
or a value of ING is not in the range 1 to NG,
or the effective number of observations for a group is less than 1,
or a group has less than NVAR members.

IFAIL $= 4$

R or one of the R_j is not of full rank.

7 Accuracy

The accuracy is dependent on the accuracy of the computation of the QR decomposition. See F08AEF (DGEQRF) for further details.

8 Further Comments

The time taken will be approximately proportional to np^2.

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 statistics computed by G03DAF. The printed results show that there is evidence that the within-group variance-covariance matrices are not equal.

9.1 Program Text

```
    Program g03dafe

!     G03DAF Example Program Text

!     Mark 24 Release. NAG Copyright 2012.

!     .. Use Statements ..
      Use nag_library, Only: g03daf, nag_wp, x04caf
!     .. Implicit None Statement ..
      Implicit None
!     .. Parameters ..
      Integer, Parameter               :: nin = 5, nout = 6
!     .. Local Scalars ..
      Real (Kind=nag_wp)               :: df, sig, stat
      Integer                          :: i, ifail, ldgmn, ldx, lgc, lwk, lwt, &
                                          m, n, ng, nvar
      Character (1)                    :: weight
!     .. Local Arrays ..
      Real (Kind=nag_wp), Allocatable  :: det(:), gc(:), gmn(:,:), wk(:),      &
                                          wt(:), x(:,:)
      Integer, Allocatable             :: ing(:), isx(:), iwk(:), nig(:)
!     .. Intrinsic Procedures ..
      Intrinsic                        :: count
!     .. Executable Statements ..
      Write (nout,*) 'G03DAF Example Program Results'
      Write (nout,*)
      Flush (nout)

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

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

      If (weight=='W' .Or. weight=='w') Then
        lwt = n
      Else
        lwt = 0
      End If
      ldx = n
      Allocate (x(ldx,m),ing(n),wt(lwt),isx(m))

!     Read in data
      If (lwt>0) Then
        Read (nin,*)(x(i,1:m),ing(i),wt(i),i=1,n)
      Else
        Read (nin,*)(x(i,1:m),ing(i),i=1,n)
      End If

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

!     Calculate NVAR
      nvar = count(isx(1:m)==1)

      ldgmn = ng
      lgc = (ng+1)*nvar*(nvar+1)/2
      lwk = n*(nvar+1)
      Allocate (nig(ng),gmn(ldgmn,nvar),det(ng),gc(lgc),wk(lwk),iwk(ng))
```

```
!     Compute test statistic
      ifail = 0
      Call g03daf(weight,n,m,x,ldx,isx,nvar,ing,ng,wt,nig,gmn,ldgmn,det,gc, &
        stat,df,sig,wk,iwk,ifail)

!     Display results
      ifail = 0
      Call x04caf('General',' ',ng,nvar,gmn,ldgmn,'Group means',ifail)
      Write (nout,*)
      Write (nout,*) ' LOG of determinants'
      Write (nout,*)
      Write (nout,99999) det(1:ng)
      Write (nout,*)
      Write (nout,99998) ' STAT = ', stat
      Write (nout,99998) '   DF = ', df
      Write (nout,99998) '  SIG = ', sig

99999 Format (1X,3F10.4)
99998 Format (1X,A,F7.4)
    End Program g03dafe
```

9.2 Program Data

```
G03DAF Example Program Data
  21 2 3 'U'           : N,M,NG,WEIGHT
  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 : End of X,ING
  1     1              : ISX
```

9.3 Program Results

```
 G03DAF Example Program Results

 Group means
             1          2
 1      1.0433    -0.6034
 2      2.0073    -0.2060
 3      2.7097     1.5998

  LOG of determinants

    -0.8273   -3.0460   -2.2877

  STAT = 19.2410
    DF =  6.0000
   SIG =  0.0038
```