NAG Fortran Library Routine Document

G03DCF

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

G03DCF allocates observations to groups according to selected rules. It is intended for use after G03DAF.

2 Specification

```
      SUBROUTINE G03DCF(TYPE, EQUAL, PRIORS, NVAR, NG, NIG, GMEAN, LDG, GC,
     1                  DET, NOBS, M, ISX, X, LDX, PRIOR, P, LDP, IAG, ATIQ,
     2                  ATI, WK, IFAIL)
      INTEGER           NVAR, NG, NIG(NG), LDG, NOBS, M, ISX(M), LDX, LDP,
     1                  IAG(NOBS), IFAIL
      real              GMEAN(LDG,NVAR), GC((NG+1)*NVAR*(NVAR+1)/2), DET(NG),
     1                  X(LDX,M), PRIOR(NG), P(LDP,NG), ATI(LDP,*), WK(2*NVAR)
      LOGICAL           ATIQ
      CHARACTER*1       TYPE, EQUAL, PRIORS
```

3 Description

Discriminant analysis is concerned with the allocation of observations to groups using information from other observations whose group membership is known, X_t; these are called the training set. 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; these are calculated from a training set of n observations with n_j observations in the jth group, and let x_k be the kth observation from the set of observations to be allocated to the n_g groups. The observation can be allocated to a group according to a selected rule. The allocation rule or discriminant function will be based on the distance of the observation from an estimate of the location of the groups, usually the group means. A measure of the distance of the observation from the jth group mean is given by the Mahalanobis distance, ${D_{kj}}^2$:

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

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

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

Instead of using the variance-covariance matrices S and S_j, G03DCF 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.

In addition to the distances a set of prior probabilities of group membership, π_j, for $j = 1, 2, \ldots, n_g$, may be used, with $\sum \pi_j = 1$. The prior probabilities reflect the user's view as to the likelihood of the observations coming from the different groups. Two common cases for prior probabilities are $\pi_1 = \pi_2 = \cdots = \pi_{n_g}$, that is equal prior probabilities, and $\pi_j = n_j/n$, for $j = 1, 2, \ldots, n_g$, that is prior probabilities proportional to the number of observations in the groups in the training set.

G03DCF uses one of four allocation rules. In all four rules the p variables are assumed to follow a multivariate Normal distribution with mean μ_j and variance-covariance matrix Σ_j if the observation comes from the jth group. The different rules depend on whether or not the within-group variance-covariance matrices are assumed equal, i.e., $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_{n_g}$, and whether a predictive or estimative approach is used. If $p(x_k|\mu_j, \Sigma_j)$ is the probability of observing the observation x_k from group j, then the posterior probability of belonging to group j is:

$$p(j|x_k, \mu_j, \Sigma_j) \propto p(x_k|\mu_j, \Sigma_j)\pi_j. \tag{3}$$

In the estimative approach the parameters μ_j and Σ_j in (3) are replaced by their estimates calculated from X_t. In the predictive approach a non-informative prior distribution is used for the parameters and a posterior distribution for the parameters, $p(\mu_j, \Sigma_j|X_t)$, is found. A predictive distribution is then obtained by integrating $p(j|x_k, \mu_j, \Sigma_j)p(\mu_j, \Sigma_j|X)$ over the parameter space. This predictive distribution then replaces $p(x_k|\mu_j, \Sigma_j)$ in (3). See Aitchison and Dunsmore (1975), Aitchison *et al.* (1977) and Moran and Murphy (1979) for further details.

The observation is allocated to the group with the highest posterior probability. Denoting the posterior probabilities, $p(j|x_k, \mu_j, \Sigma_j)$, by q_j, the four allocation rules are:

(i) Estimative with equal variance-covariance matrices – Linear Discrimination

$$\log q_j \propto -\tfrac{1}{2}{D_{kj}}^2 + \log \pi_j$$

(ii) Estimative with unequal variance-covariance matrices – Quadratic Discrimination

$$\log q_j \propto -\tfrac{1}{2}{D_{kj}}^2 + \log \pi_j - \tfrac{1}{2}\log|S_j|$$

(iii) Predictive with equal variance-covariance matrices

$$q_j^{-1} \propto ((n_j+1)/n_j)^{p/2}\{1 + [n_j/((n-n_g)(n_j+1))]{D_{kj}}^2\}^{(n+1-n_g)/2}$$

(iv) Predictive with unequal variance-covariance matrices

$$q_j^{-1} \propto C\{((n_j^2-1)/n_j)|S_j|\}^{p/2}\ \{1 + (n_j/(n_j^2-1)){D_{kj}}^2\}^{n_j/2},$$

where

$$C = \frac{\Gamma(\frac{1}{2}(n_j-p))}{\Gamma(\frac{1}{2}n_j)}.$$

In the above the appropriate value of ${D_{kj}}^2$ from (1) or (2) is used. The values of the q_j are standardized so that,

$$\sum_{j=1}^{n_g} q_j = 1.$$

Moran and Murphy (1979) show the similarity between the predictive methods and methods based upon likelihood ratio tests.

In addition to allocating the observation to a group G03DCF computes an atypicality index, $I_j(x_k)$. This represents the probability of obtaining an observation more typical of group j than the observed x_k, see Aitchison and Dunsmore (1975) and Aitchison *et al.* (1977). The atypicality index is computed for unequal within-group variance-covariance matrices as:

$$I_j(x_k) = P(B \leq z : \tfrac{1}{2}p, \tfrac{1}{2}(n_j-p))$$

where $P(B \leq \beta : a, b)$ is the lower tail probability from a beta distribution and

$$z = {D_{kj}}^2/({D_{kj}}^2 + (n_j^2-1)/n_j),$$

and for equal within-group variance-covariance matrices as:

$$I_j(x_k) = P(B \leq z : \tfrac{1}{2}p, \tfrac{1}{2}(n-n_g-p+1)),$$

with

$$z = {D_{kj}}^2/({D_{kj}}^2 + (n-n_g)(n_j+1)/n_j).$$

If $I_j(x_k)$ is close to 1 for all groups it indicates that the observation may come from a grouping not represented in the training set. Moran and Murphy (1979) provide a frequentist interpretation of $I_j(x_k)$.

4 References

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

Aitchison J, Habbema J D F and Kay J W (1977) A critical comparison of two methods of statistical discrimination *Appl. Statist.* **26** 15–25

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

Moran M A and Murphy B J (1979) A closer look at two alternative methods of statistical discrimination *Appl. Statist.* **28** 223–232

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

5 Parameters

1: TYPE – CHARACTER*1 *Input*

On entry: whether the estimative or predictive approach is used.

If TYPE = 'E' the estimative approach is used.

If TYPE = 'P' the predictive approach is used.

Constraint: TYPE = 'E' or 'P'.

2: 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_i, for $i = 1, 2, \ldots, n_g$, stored in the remainder of GC are used.

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

3: PRIORS – CHARACTER*1 *Input*

On entry: indicates the form of the prior probabilities to be used.

If PRIORS = 'E', equal prior probabilities are used.

If PRIORS = 'P', prior probabilities proportional to the group sizes in the training set, n_j, are used.

If PRIORS = 'I', the prior probabilities are input in PRIOR.

Constraint: PRIORS = 'E', 'I' or 'P'.

4: NVAR – INTEGER *Input*

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

Constraint: NVAR ≥ 1.

5: NG – INTEGER *Input*

On entry: the number of groups, n_g.

Constraint: NG ≥ 2.

6: NIG(NG) – INTEGER array *Input*

On entry: the number of observations in each group in the training set, n_j.

Constraints:

if EQUAL = 'E', NIG(j) > 0, for $j = 1, 2, \ldots, n_g$ and $\sum_{j=1}^{n_g} \mathrm{NIG}(j) > \mathrm{NG} + \mathrm{NVAR}$,
if EQUAL = 'U', NIG(j)> NVAR, for $j = 1, 2, \ldots, n_g$.

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

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

8: LDG – INTEGER *Input*

On entry: the first dimension of the array GMEAN as declared in the (sub)program from which G03DCF is called.

Constraint: LDG $\geq$ NG.

9: 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 must be $\neq$ 0.0,
if EQUAL = 'U' the diagonal elements of the R_j must be $\neq$ 0.0, for $j = 1, 2, \ldots, n_g$.

10: DET(NG) – ***real*** array *Input*

On entry: if EQUAL = 'U' the logarithms of the determinants of the within-group variance-covariance matrices as returned by G03DAF. Otherwise DET is not referenced.

11: NOBS – INTEGER *Input*

On entry: the number of observations in X which are to be allocated.

Constraint: NOBS ≥ 1.

12: M – INTEGER *Input*

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

Constraint: M $\geq$ NVAR.

13: ISX(M) – INTEGER array *Input*

On entry: 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 $l = 1, 2, \ldots,$ M; otherwise the lth variable is not referenced.

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

14: X(LDX,M) – ***real*** array *Input*

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

15: LDX – INTEGER *Input*

On entry: the first dimension of the array X as declared in the (sub)program from which G03DCF is called.

Constraint: LDX $\geq$ NOBS.

16: PRIOR(NG) – ***real*** array *Input/Output*

On entry: if PRIORS = 'I' the prior probabilities for the n_g groups.

Constraint: if PRIORS = 'I', then PRIOR$(j) > 0.0$ for $j = 1, 2, \ldots, n_g$ and

$$\left|1 - \sum_{j=1}^{n_g} \text{PRIOR}(j)\right| \leq 10 \times \textbf{\textit{machine precision}}.$$

On exit: if PRIORS = 'P' the computed prior probabilities in proportion to group sizes for the n_g groups. If PRIORS = 'I' the input prior probabilities will be unchanged, and if PRIORS = 'E', PRIOR is not set.

17: P(LDP,NG) – ***real*** array *Output*

On exit: P(k, j) contains the posterior probability p_{kj} for allocating the kth observation to the jth group, for $k = 1, 2, \ldots,$ NOBS; $j = 1, 2, \ldots, n_g$.

18: LDP – INTEGER *Input*

On entry: the first dimension of the array P as declared in the (sub)program from which G03DCF is called.

Constraint: LDP $\geq$ NOBS.

19: IAG(NOBS) – INTEGER array *Output*

On exit: the groups to which the observations have been allocated.

20: ATIQ – LOGICAL *Input*

On entry: ATIQ must be .TRUE. if atypicality indices are required. If ATIQ is .FALSE. the array ATI is not set.

21: ATI(LDP,*) – ***real*** array *Output*

Note: the second dimension of the array ATI must be at least NG, if ATIQ is .TRUE., and 1 otherwise.

On exit: if AITQ is .TRUE., ATI(k, j) will contain the atypicality index for the kth observation with respect to the jth group, for $k = 1, 2, \ldots,$ NOBS; $j = 1, 2, \ldots, n_g$. If ATIQ is .FALSE., ATI is not set.

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

23: 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 NOBS < 1,
or M $<$ NVAR,
or LDG $<$ NG,
or LDX $<$ NOBS,
or LDP $<$ NOBS,
or TYPE $\neq$ 'E' or 'P',
or EQUAL $\neq$ 'E' or 'U',
or PRIORS $\neq$ 'E', 'I' or 'P'.

IFAIL $= 2$

On entry, the number of variables indicated by ISX is not equal to NVAR,
or EQUAL = 'E' and NIG$(j) \leq 0$, for some j,
or EQUAL = 'E' and $\sum_{j=1}^{n_g} \mathrm{NIG}(j) \leq \mathrm{NG} + \mathrm{NVAR}$,
or EQUAL = 'U' and NIG$(j) \leq$ NVAR for some j.

IFAIL $= 3$

On entry, PRIORS = 'I' and PRIOR$(j) \leq 0.0$ for some j,
or PRIORS = 'I' and $\sum_{j=1}^{n_g} \mathrm{PRIOR}(j)$ is not within $10\times$ ***machine precision*** of 1.

IFAIL $= 4$

On entry, 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 of the returned posterior probabilities will depend on the accuracy of the input R or R_j matrices. The atypicality index should be accurate to four significant places.

8 Further Comments

The distances ${D_{kj}}^2$ can be computed using G03DBF if other forms of discrimination are required.

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 allocations made using the predictive approach and under the assumption that the within-group covariance matrices are not equal. The posterior probabilities of group membership, q_j, and the atypicality index are printed along with the allocated group. The atypicality index shows that observations 5 and 6 do not seem to be typical of the three types present in the initial 21 observations.

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.

```
*     G03DCF 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, TYPE, WEIGHT
*     .. Local Arrays ..
      real             ATI(NMAX,GPMAX), DET(GPMAX),
     +                 GC((GPMAX+1)*MMAX*(MMAX+1)/2), GMEAN(GPMAX,MMAX),
     +                 P(NMAX,GPMAX), PRIOR(GPMAX), WK(NMAX*(MMAX+1)),
     +                 WT(NMAX), X(NMAX,MMAX)
      INTEGER          IAG(NMAX), ING(NMAX), ISX(MMAX), IWK(GPMAX),
     +                 NIG(GPMAX)
*     .. External Subroutines ..
      EXTERNAL         G03DAF, G03DCF
*     .. Executable Statements ..
      WRITE (NOUT,*) 'G03DCF 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, TYPE
         IF (NOBS.LE.NMAX) THEN
            DO 60 I = 1, NOBS
               READ (NIN,*) (X(I,J),J=1,M)
   60       CONTINUE
            IFAIL = 0
*
            CALL G03DCF(TYPE,EQUAL,'Equal priors',NVAR,NG,NIG,GMEAN,
     +                  GPMAX,GC,DET,NOBS,M,ISX,X,NMAX,PRIOR,P,NMAX,IAG,
     +                  .TRUE.,ATI,WK,IFAIL)
*
            WRITE (NOUT,*)
            WRITE (NOUT,*) '   Obs       Posterior        Allocated',
     +        '     Atypicality'
            WRITE (NOUT,*)
     +        '               probabilities    to group      index'
            WRITE (NOUT,*)
            DO 80 I = 1, NOBS
               WRITE (NOUT,99999) I, (P(I,J),J=1,NG), IAG(I),
     +           (ATI(I,J),J=1,NG)
   80       CONTINUE
         END IF
      END IF
      STOP
```

```
*
99999 FORMAT (1X,2(I6,5X,3F6.3))
      END
```

9.2 Program Data

```
G03DCF 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' 'P'
  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

```
 G03DCF Example Program Results

    Obs       Posterior         Allocated     Atypicality
              probabilities     to group      index

      1      0.094 0.905 0.002      2      0.596 0.254 0.975
      2      0.005 0.168 0.827      3      0.952 0.836 0.018
      3      0.019 0.920 0.062      2      0.954 0.797 0.912
      4      0.697 0.303 0.000      1      0.207 0.860 0.993
      5      0.317 0.013 0.670      3      0.991 1.000 0.984
      6      0.032 0.366 0.601      3      0.981 0.978 0.887
```