# NAG Fortran Library Routine Document

# G03ACF

**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

G03ACF performs a canonical variate (canonical discrimination) analysis.

## 2 Specification

```
 SUBROUTINE G03ACF(WEIGHT, N, M, X, LDX, ISX, NX, ING, NG, WT, NIG, CVM,
1                  LDCVM, E, LDE, NCV, CVX, LDCVX, TOL, IRANKX, WK, IWK,
2                  IFAIL)
 INTEGER           N, M, LDX, ISX(M), NX, ING(N), NG, NIG(NG), LDCVM,
1                  LDE, NCV, LDCVX, IRANKX, IWK, IFAIL
 real              X(LDX,M), WT(*), CVM(LDCVM,NX), E(LDE,6),
1                  CVX(LDCVX,NG-1), TOL, WK(IWK)
 CHARACTER*1       WEIGHT
```

## 3 Description

Let a sample of $n$ observations on $n_x$ variables in a data matrix come from $n_g$ groups with $n_1, n_2, \ldots, n_{n_g}$ observations in each group, $\sum n_i = n$. Canonical variate analysis finds the linear combination of the $n_x$ variables that maximizes the ratio of between-group to within-group variation. The variables formed, the canonical variates, can be used to discriminate between groups.

The canonical variates can be calculated from the eigenvectors of the within-group sums of squares and cross-products matrix. However, G03ACF calculates the canonical variates by means of a singular value decomposition (SVD) of a matrix $V$. Let the data matrix with variable (column) means subtracted be $X$ and let its rank be $k$; then the $k$ by $(n_g - 1)$ matrix $V$ is given by:

> $V = Q_X^{\mathrm{T}} Q_g$, where $Q_g$ is an $n$ by $(n_g - 1)$ orthogonal matrix that defines the groups and $Q_X$ is the first $k$ rows of the orthogonal matrix $Q$ either from the $QR$ decomposition of $X$:
>
> $$X = QR$$
>
> if $X$ is of full column rank, i.e., $k = n_x$, else from the SVD of $X$:
>
> $$X = QDP^{\mathrm{T}}.$$

Let the SVD of $V$ be:

$$V = U_x \Delta U_g^{\mathrm{T}}$$

then the non-zero elements of the diagonal matrix $\Delta$, $\delta_i$, for $i = 1, 2, \ldots, l$, are the $l$ canonical correlations associated with the $l$ canonical variates, where $l = \min(k, n_g)$.

The eigenvalues, $\lambda_i^2$, of the within-group sums of squares matrix are given by:

$$\lambda_i^2 = \frac{\delta_i^2}{1 - \delta_i^2}$$

and the value of $\pi_i = \lambda_i^2 / \sum \lambda_i^2$ gives the proportion of variation explained by the $i$th canonical variate. The values of the $\pi_i$'s give an indication as to how many canonical variates are needed to adequately describe the data, i.e., the dimensionality of the problem.

To test for a significant dimensionality greater than $i$ the $\chi^2$ statistic:

$$(n - 1 - n_g - \tfrac{1}{2}(k - n_g)) \sum_{j=i+1}^{l} \log(1 + \lambda_j^2)$$

can be used. This is asymptotically distributed as a $\chi^2$ distribution with $(k - i)(n_g - 1 - i)$ degrees of freedom. If the test for $i = h$ is not significant, then the remaining tests for $i > h$ should be ignored.

The loadings for the canonical variates are calculated from the matrix $U_x$. This matrix is scaled so that the canonical variates have unit within group variance.

In addition to the canonical variates loadings the means for each canonical variate are calculated for each group.

Weights can be used with the analysis, in which case the weighted means are subtracted from each column and then each row is scaled by an amount $\sqrt{w_i}$, where $w_i$ is the weight for the $i$th observation (row).

## 4 References

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

Gnanadesikan R (1977) *Methods for Statistical Data Analysis of Multivariate Observations* Wiley

Hammarling S (1985) The singular value decomposition in multivariate statistics *SIGNUM Newsl.* **20 (3)** 2–25

Kendall M G and Stuart A (1969) *The Advanced Theory of Statistics (Volume 1)* (3rd Edition) Griffin

## 5 Parameters

1: WEIGHT – CHARACTER*1 *Input*

*On entry*: indicates if weights are to be used.

If WEIGHT = 'U', no weights are used.

If WEIGHT = 'W' or 'V', weights are used and must be supplied in WT.

*Constraint*: WEIGHT = 'U', 'W' or 'V'.

In the case of WEIGHT = 'W', the weights are treated as frequencies and the effective number of observations is the sum of the weights. If WEIGHT = 'V', the weights are treated as being inversely proportional to the variance of the observations and the effective number of observations is the number of observations with non-zero weights.

2: N – INTEGER *Input*

*On entry*: the number of observations, $n$.

*Constraint*: N $\geq$ NX + NG.

3: M – INTEGER *Input*

*On entry*: the total number of variables, $m$.

*Constraint*: M $\geq$ NX.

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

*On entry*: X$(i,j)$ must contain the $i$th observation for the $j$th variable, for $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, m$.

5: LDX – INTEGER *Input*

*On entry*: the first dimension of the array X as declared in the (sub)program from which G03ACF is called.

*Constraint*: LDX $\geq$ N.

6: ISX(M) – INTEGER array *Input*

*On entry*: ISX($j$) indicates whether or not the $j$th variable is to be included in the analysis.

If ISX($j$) $> 0$, then the variables contained in the $j$th column of X is included in the canonical variate analysis, for $j = 1, 2, \ldots, m$.

*Constraint*: ISX($j$) $> 0$ for NX values of $j$.

7: NX – INTEGER *Input*

*On entry*: the number of variables in the analysis, $n_x$.

*Constraint*: NX $\geq 1$.

8: ING(N) – INTEGER array *Input*

*On entry*: ING($i$) indicates which group the $i$th observation is in, for $i = 1, 2, \ldots, n$. The effective number of groups is the number of groups with non-zero membership.

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

9: NG – INTEGER *Input*

*On entry*: the number of groups, $n_g$.

*Constraint*: NG $\geq 2$.

10: WT(*) – ***real*** array *Input*

**Note:** the dimension of the array WT must be at least N if WEIGHT = 'W' or 'V' and 1 otherwise.

*On entry*: if WEIGHT = 'W' or 'V', then the first $n$ elements of WT must contain the weights to be used in the analysis.

If WT($i$) = 0.0, then the $i$th observation is not included in the analysis.

If WEIGHT = 'U', then WT is not referenced.

*Constraints*:

WT($i$) $\geq$ 0.0, for $i = 1, 2, \ldots, n$,

$$\sum_1^n \mathrm{WT}(i) \geq \mathrm{NX} + \text{ effective number of groups.}$$

11: NIG(NG) – INTEGER array *Output*

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

12: CVM(LDCVM,NX) – ***real*** array *Output*

*On exit*: CVM($i, j$) contains the mean of the $j$th canonical variate for the $i$th group, for $i = 1, 2, \ldots, n_g$; $j = 1, 2, \ldots, l$; the remaining columns, if any, are used as workspace.

13: LDCVM – INTEGER *Input*

*On entry*: the dimension of the array CVM as declared in the (sub)program from which G03ACF is called.

*Constraint*: LDCVM $\geq$ NG.

14: E(LDE,6) – ***real*** array *Output*

*On exit*: the statistics of the canonical variate analysis.

E$(i,1)$, the canonical correlations, $\delta_i$, for $i = 1, 2, \ldots, l$.

E$(i,2)$, the eigenvalues of the within-group sum of squares matrix, $\lambda_i^2$, for $i = 1, 2, \ldots, l$.

E$(i,3)$, the proportion of variation explained by the $i$th canonical variate, for $i = 1, 2, \ldots, l$.

E$(i,4)$, the $\chi^2$ statistic for the $i$th canonical variate, for $i = 1, 2, \ldots, l$.

E$(i,5)$, the degrees of freedom for $\chi^2$ statistic for the $i$th canonical variate, for $i = 1, 2, \ldots, l$.

E$(i,6)$, the significance level for the $\chi^2$ statistic for the $i$th canonical variate, for $i = 1, 2, \ldots, l$.

15: LDE – INTEGER *Input*

*On entry*: the first dimension of the array E as declared in the (sub)program from which G03ACF is called.

*Constraint*: LDE $\geq$ min(NX, NG $-$ 1).

16: NCV – INTEGER *Output*

*On exit*: the number of canonical variates, $l$. This will be the minimum of $n_g - 1$ and the rank of X.

17: CVX(LDCVX,NG-1) – ***real*** array *Output*

*On exit*: the canonical variate loadings. CVX$(i,j)$ contains the loading coefficient for the $i$th variable on the $j$th canonical variate, for $i = 1, 2, \ldots, n_x$; $j = 1, 2, \ldots, l$; the remaining columns, if any, are used as workspace.

18: LDCVX – INTEGER *Input*

*On entry*: the first dimension of the array CVX as declared in the (sub)program from which G03ACF is called.

*Constraint*: LDCVX $\geq$ NX.

19: TOL – ***real*** *Input*

*On entry*: the value of TOL is used to decide if the variables are of full rank and, if not, what is the rank of the variables. The smaller the value of TOL the stricter the criterion for selecting the singular value decomposition. If a non-negative value of TOL less than ***machine precision*** is entered, then the square root of ***machine precision*** is used instead.

*Constraint*: TOL $\geq$ 0.0.

20: IRANKX – INTEGER *Output*

*On exit*: the rank of the dependent variables.

If the variables are of full rank then IRANKX = NX.

If the variables are not of full rank then IRANKX is an estimate of the rank of the dependent variables. IRANKX is calculated as the number of singular values greater than TOL$\times$ (largest singular value).

21: WK(IWK) – ***real*** array *Workspace*
22: IWK – INTEGER *Input*

*On entry*: the dimension of the array WK as declared in the (sub)program from which G03ACF is called.

*Constraints*:

if NX $\geq$ NG $-$ 1, then IWK $\geq$ N $\times$ NX + max(5 $\times$ (NX $-$ 1) + (NX + 1) $\times$ NX, N),
if NX $<$ NG $-$ 1, then IWK $\geq$ N $\times$ NX + max(5 $\times$ (NX $-$ 1) + (NG $-$ 1) $\times$ NX, N).

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, NX $< 1$,
or NG $< 2$,
or M $<$ NX,
or N $<$ NX $+$ NG,
or LDX $<$ N,
or LDCVX $<$ NX,
or LDCVM $<$ NG,
or LDE $< \min(\mathrm{NX}, \mathrm{NG} - 1)$,
or NX $\geq$ NG $- 1$ and IWK $< \mathrm{N} \times \mathrm{NX} + \max(5 \times (\mathrm{NX} - 1) + (\mathrm{NX} + 1) \times \mathrm{NX}, \mathrm{N})$,
or NX $<$ NG $- 1$ and IWK $< \mathrm{N} \times \mathrm{NX} + \max(5 \times (\mathrm{NX} - 1) + (\mathrm{NG} - 1) \times \mathrm{NX}, \mathrm{N})$,
or WEIGHT $\neq$ 'U', 'W' or 'V',
or TOL $< 0.0$.

IFAIL $= 2$

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

IFAIL $= 3$

On entry, a value of ING $< 1$,
or a value of ING $>$ NG.

IFAIL $= 4$

On entry, the number of variables to be included in the analysis as indicated by ISX is not equal to NX.

IFAIL $= 5$

A singular value decomposition has failed to converge. This is an unlikely error exit.

IFAIL $= 6$

A canonical correlation is equal to 1. This will happen if the variables provide an exact indication as to which group every observation is allocated.

IFAIL $= 7$

On entry, less than two groups have non-zero membership, i.e., the effective number of groups is less than 2,
or the effective number of groups plus the number of variables, NX, is greater than the effective number of observations.

IFAIL = 8

The rank of the variables is 0. This will happen if all the variables are constants.

## 7 Accuracy

As the computation involves the use of orthogonal matrices and a singular value decomposition rather than the traditional computing of a sum of squares matrix and the use of an eigenvalue decomposition, G03ACF should be less affected by ill-conditioned problems.

## 8 Further Comments

None.

## 9 Example

A sample of nine observations, each consisting of three variables plus group indicator, is read in. There are three groups. An unweighted canonical variate analysis is performed and the results printed.

### 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.

```
*     G03ACF Example Program Text
*     Mark 18 Revised.  NAG Copyright 1997.
*     .. Parameters ..
      INTEGER          NMAX, MMAX, IWKMAX
      PARAMETER        (NMAX=9,MMAX=3,IWKMAX=50)
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
*     .. Local Scalars ..
      real             TOL
      INTEGER          I, IFAIL, IRX, J, M, N, NCV, NG, NX
      CHARACTER        WEIGHT
*     .. Local Arrays ..
      real             CVM(MMAX,MMAX), CVX(MMAX,MMAX), E(MMAX,6),
     +                 WK(IWKMAX), WT(NMAX), X(NMAX,MMAX)
      INTEGER          ING(NMAX), ISX(2*MMAX), NIG(MMAX)
*     .. External Subroutines ..
      EXTERNAL         G03ACF
*     .. Executable Statements ..
      WRITE (NOUT,*) 'G03ACF Example Program Results'
*     Skip heading in data file
      READ (NIN,*)
      READ (NIN,*) N, M, NX, NG, WEIGHT
      IF (N.LE.NMAX .AND. M.LE.MMAX) THEN
         IF (WEIGHT.EQ.'W' .OR. WEIGHT.EQ.'w' .OR. WEIGHT.EQ.'V' .OR.
     +       WEIGHT.EQ.'v') THEN
            DO 20 I = 1, N
               READ (NIN,*) (X(I,J),J=1,M), WT(I), ING(I)
   20       CONTINUE
         ELSE
            DO 40 I = 1, N
               READ (NIN,*) (X(I,J),J=1,M), ING(I)
   40       CONTINUE
         END IF
         READ (5,*) (ISX(J),J=1,M)
         TOL = 0.000001e0
         IFAIL = 0
*
         CALL G03ACF(WEIGHT,N,M,X,NMAX,ISX,NX,ING,NG,WT,NIG,CVM,MMAX,E,
     +               MMAX,NCV,CVX,MMAX,TOL,IRX,WK,IWKMAX,IFAIL)
*
         WRITE (NOUT,*)
         WRITE (NOUT,99999) 'Rank of X = ', IRX
```

```
         WRITE (NOUT,*)
         WRITE (NOUT,*)
     +   'Canonical    Eigenvalues Percentage     CHISQ      DF     SIG'
         WRITE (NOUT,*) 'Correlations              Variation'
         DO 60 I = 1, NCV
            WRITE (NOUT,99998) (E(I,J),J=1,6)
   60    CONTINUE
         WRITE (NOUT,*)
         WRITE (NOUT,*) 'Canonical Coefficients for X'
         DO 80 I = 1, NX
            WRITE (NOUT,99997) (CVX(I,J),J=1,NCV)
   80    CONTINUE
         WRITE (NOUT,*)
         WRITE (NOUT,*) 'Canonical variate means'
         DO 100 I = 1, NG
            WRITE (NOUT,99997) (CVM(I,J),J=1,NCV)
  100    CONTINUE
      END IF
      STOP
*
99999 FORMAT (1X,A,I2)
99998 FORMAT (1X,2F12.4,F11.4,F10.4,F8.1,F8.4)
99997 FORMAT (1X,5F9.4)
      END
```

## 9.2 Program Data

```
G03ACF Example Program Data
 9 3 3 3 'U'
 13.3 10.6 21.2 1
 13.6 10.2 21.0 2
 14.2 10.7 21.1 3
 13.4  9.4 21.0 1
 13.2  9.6 20.1 2
 13.9 10.4 19.8 3
 12.9 10.0 20.5 1
 12.2  9.9 20.7 2
 13.9 11.0 19.1 3
  1    1    1
```

## 9.3 Program Results

```
 G03ACF Example Program Results

 Rank of X =  3

 Canonical    Eigenvalues Percentage     CHISQ      DF     SIG
 Correlations             Variation
       0.8826      3.5238     0.9795    7.9032     6.0  0.2453
       0.2623      0.0739     0.0205    0.3564     2.0  0.8368

 Canonical Coefficients for X
   -1.7070   0.7277
   -1.3481   0.3138
    0.9327   1.2199

 Canonical variate means
    0.9841   0.2797
    1.1805  -0.2632
   -2.1646  -0.0164
```