NAG 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,        &
                   LDCVM, E, LDE, NCV, CVX, LDCVX, TOL, IRANKX, WK, IWK,       &
                   IFAIL)

INTEGER            N, M, LDX, ISX(M), NX, ING(N), NG, NIG(NG), LDCVM, LDE,     &
                   NCV, LDCVX, IRANKX, IWK, IFAIL
REAL (KIND=nag_wp) X(LDX,M), WT(*), CVM(LDCVM,NX), E(LDE,6),                   &
                   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 then 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 $\left(n_g - 1\right)$ 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 nonzero elements of the diagonal matrix Δ, δ_i, for $i = 1, 2, \ldots, l$, are the l canonical correlations associated with the $l = \min\left(k, n_g - 1\right)$ canonical variates, where $l = \min\left(k, n_g\right)$.

The eigenvalues, λ_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 ith canonical variate. The values of the π_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 χ^2 statistic:

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

can be used. This is asymptotically distributed as a χ^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 ith 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.

WEIGHT = 'U'
: No weights are used.

WEIGHT = 'W' or 'V'
: Weights are used and must be supplied in WT.

If 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 nonzero weights.

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

2: N – INTEGER *Input*

On entry: n, the number of observations.

Constraint: $\mathrm{N} \geq \mathrm{NX} + \mathrm{NG}$.

3: M – INTEGER *Input*

On entry: m, the total number of variables.

Constraint: $\mathrm{M} \geq \mathrm{NX}$.

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

On entry: $\mathrm{X}(i,j)$ must contain the ith observation for the jth variable, for $i=1,2,\ldots,n$ and $j=1,2,\ldots,m$.

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 jth variable is to be included in the analysis.

If ISX(j) > 0, the variables contained in the jth 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 ≥ 1.

8: ING(N) – INTEGER array *Input*

On entry: ING(i) indicates which group the ith observation is in, for $i = 1, 2, \ldots, n$. The effective number of groups is the number of groups with nonzero 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 ≥ 2.

10: WT($*$) – REAL (KIND=nag_wp) array *Input*

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

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

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

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

Constraints:

WT(i) ≥ 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 (KIND=nag_wp) array *Output*

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

13: LDCVM – INTEGER *Input*

On entry: the first 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 (KIND=nag_wp) array *Output*

On exit: the statistics of the canonical variate analysis.

$\mathrm{E}(i,1)$
The canonical correlations, δ_i, for $i = 1, 2, \dots, l$.

$\mathrm{E}(i,2)$
The eigenvalues of the within-group sum of squares matrix, λ_i^2, for $i = 1, 2, \dots, l$.

$\mathrm{E}(i,3)$
The proportion of variation explained by the ith canonical variate, for $i = 1, 2, \dots, l$.

$\mathrm{E}(i,4)$
The χ^2 statistic for the ith canonical variate, for $i = 1, 2, \dots, l$.

$\mathrm{E}(i,5)$
The degrees of freedom for χ^2 statistic for the ith canonical variate, for $i = 1, 2, \dots, l$.

$\mathrm{E}(i,6)$
The significance level for the χ^2 statistic for the ith canonical variate, for $i = 1, 2, \dots, l$.

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: $\mathrm{LDE} \geq \min(\mathrm{NX}, \mathrm{NG} - 1)$.

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 (KIND=nag_wp) array *Output*

On exit: the canonical variate loadings. $\mathrm{CVX}(i,j)$ contains the loading coefficient for the ith variable on the jth canonical variate, for $i = 1, 2, \dots, n_x$ and $j = 1, 2, \dots, l$; 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: $\mathrm{LDCVX} \geq \mathrm{NX}$.

19: TOL – REAL (KIND=nag_wp) *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, the square root of ***machine precision*** is used instead.

Constraint: $\mathrm{TOL} \geq 0.0$.

20: IRANKX – INTEGER *Output*

On exit: the rank of the dependent variables.

If the variables are of full rank then $\mathrm{IRANKX} = \mathrm{NX}$.

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 $\mathrm{TOL} \times (\text{largest singular value})$.

21: WK(IWK) – REAL (KIND=nag_wp) 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 $NX \geq NG - 1$, $IWK \geq N \times NX + \max(5 \times (NX - 1) + (NX + 1) \times NX, N) + 1$;
if $NX < NG - 1$, $IWK \geq N \times NX + \max(5 \times (NX - 1) + (NG - 1) \times NX, N) + 1$.

23: 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, $NX < 1$,
or $NG < 2$,
or $M < NX$,
or $N < NX + NG$,
or $LDX < N$,
or $LDCVX < NX$,
or $LDCVM < NG$,
or $LDE < \min(NX, NG - 1)$,
or $NX \geq NG - 1$ and $IWK < N \times NX + \max(5 \times (NX - 1) + (NX + 1) \times NX, N)$,
or $NX < NG - 1$ and $IWK < N \times NX + \max(5 \times (NX - 1) + (NG - 1) \times NX, N)$,
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 nonzero 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

This example uses a sample of nine observations, each consisting of three variables plus a group indicator. There are three groups. An unweighted canonical variate analysis is performed and the results printed.

9.1 Program Text

```
    Program g03acfe

!     G03ACF Example Program Text

!     Mark 24 Release. NAG Copyright 2012.

!     .. Use Statements ..
      Use nag_library, Only: g03acf, nag_wp, x04caf
!     .. Implicit None Statement ..
      Implicit None
!     .. Parameters ..
      Integer, Parameter                :: nin = 5, nout = 6
!     .. Local Scalars ..
      Real (Kind=nag_wp)                :: tol
      Integer                           :: i, ifail, irankx, iwk, ldcvm, ldcvx, &
                                           lde, ldx, lwt, m, n, ncv, ng, nx
      Character (1)                     :: weight
!     .. Local Arrays ..
      Real (Kind=nag_wp), Allocatable   :: cvm(:,:), cvx(:,:), e(:,:), wk(:),   &
                                           wt(:), x(:,:)
      Integer, Allocatable              :: ing(:), isx(:), nig(:)
!     .. Intrinsic Procedures ..
      Intrinsic                         :: max, min
!     .. Executable Statements ..
      Write (nout,*) 'G03ACF Example Program Results'
      Write (nout,*)

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

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

```
      Select Case (weight)
      Case ('W','w','V','v')
        lwt = n
      Case Default
        lwt = 0
      End Select
      ldx = n
      ldcvm = ng
      lde = min(nx,ng-1)
      ldcvx = nx
      If (nx>=ng-1) Then
        iwk = n*nx + max(5*(nx-1)+(nx+1)*nx,n) + 1
      Else
        iwk = n*nx + max(5*(nx-1)+(ng-1)*nx,n) + 1
      End If
      Allocate (x(ldx,m),isx(m),ing(n),wt(lwt),nig(ng),cvm(ldcvm,nx),e(lde,6), &
        cvx(ldcvx,nx),wk(iwk))

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

!     Use default tolerance
      tol = 0.0E0_nag_wp

!     Perform canonical variate analysis
      ifail = 0
      Call g03acf(weight,n,m,x,ldx,isx,nx,ing,ng,wt,nig,cvm,ldcvm,e,lde,ncv, &
        cvx,ldcvx,tol,irankx,wk,iwk,ifail)

!     Display results
      Write (nout,99999) 'Rank of X = ', irankx
      Write (nout,*)
      Write (nout,*) &
        'Canonical    Eigenvalues Percentage     CHISQ      DF     SIG'
      Write (nout,*) 'Correlations              Variation'
      Write (nout,99998)(e(i,1:6),i=1,ncv)
      Write (nout,*)
      Flush (nout)
      ifail = 0
      Call x04caf('General',' ',nx,ncv,cvx,ldcvx, &
        'Canonical Coefficients for X',ifail)
      Write (nout,*)
      Flush (nout)
      ifail = 0
      Call x04caf('General',' ',ng,ncv,cvm,ldcvm,'Canonical variate means', &
        ifail)

99999 Format (1X,A,I0)
99998 Format (1X,2F12.4,F11.4,F10.4,F8.1,F8.4)
    End Program g03acfe
```

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          2
1     -1.7070     0.7277
2     -1.3481     0.3138
3      0.9327     1.2199

Canonical variate means
            1          2
1      0.9841     0.2797
2      1.1805    -0.2632
3     -2.1646    -0.0164
```