NAG Fortran Library Routine Document

G03ADF

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

G03ADF performs canonical correlation analysis upon input data matrices.

2 Specification

```
 SUBROUTINE G03ADF(WEIGHT, N, M, Z, LDZ, ISZ, NX, NY, WT, E, LDE, NCV,
1                  CVX, LDCVX, MCV, CVY, LDCVY, TOL, WK, IWK, IFAIL)
 INTEGER           N, M, LDZ, ISZ(M), NX, NY, LDE, NCV, LDCVX, MCV,
1                  LDCVY, IWK, IFAIL
 real              Z(LDZ,M), WT(*), E(LDE,6), CVX(LDCVX,MCV),
1                  CVY(LDCVY,MCV), TOL, WK(IWK)
 CHARACTER*1       WEIGHT
```

3 Description

Let there be two sets of variables, x and y. For a sample of n observations on n_x variables in a data matrix X and n_y variables in a data matrix Y, canonical correlation analysis seeks to find a small number of linear combinations of each set of variables in order to explain or summarise the relationships between them. The variables thus formed are known as canonical variates.

Let the variance-covariance of the two data sets be

$$\begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix}$$

and let

$$\Sigma = S_{yy}^{-1} S_{yx} S_{xx}^{-1} S_{xy}$$

then the canonical correlations can be calculated from the eigenvalues of the matrix Σ. However, G03ADF calculates the canonical correlations by means of a singular value decomposition (SVD) of a matrix V. If the rank of the data matrix X is k_x and the rank of the data matrix Y is k_y and both X and Y have had variable (column) means subtracted then the k_x by k_y matrix V is given by:

$$V = Q_x^{\mathrm{T}} Q_y,$$

where Q_x is the first k_x rows of the orthogonal matrix Q either from the QR decomposition of X if X is of full column rank, i.e., $k_x = n_x$:

$$X = Q_x R_x$$

or from the SVD of X if $k_x < n_x$:

$$X = Q_x D_x P_x^{\mathrm{T}}.$$

Similarly Q_y is the first k_y rows of the orthogonal matrix Q either from the QR decomposition of Y if Y is of full column rank, i.e., $k_y = n_y$:

$$Y = Q_y R_y$$

or from the SVD of Y if $k_y < n_y$:

$$Y = Q_y D_y P_y^{\mathrm{T}}.$$

Let the SVD of V be:

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

then the non-zero elements of the diagonal matrix Δ, δ_i, for $i = 1, 2, \ldots, l$, are the l canonical correlations associated with the l canonical variates, where $l = \min(k_x, k_y)$.

The eigenvalues, λ_i^2, of the matrix Σ are given by:

$$\lambda_i^2 = \frac{\delta_i^2}{1 + \delta_i^2}.$$

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:

$$\left(n - \tfrac{1}{2}(k_x + k_y + 3)\right) \sum_{j=i+1}^{l} \log(1 + \lambda_j^2)$$

can be used. This is asymptotically distributed as a χ^2 distribution with $(k_x - i)(k_y - i)$ degrees of freedom. If the test for $i = k_{\mathrm{min}}$ is not significant, then the remaining tests for $i > k_{\mathrm{min}}$ should be ignored.

The loadings for the canonical variates are calculated from the matrices U_x and U_y respectively. These matrices are scaled so that the canonical variates have unit variance.

4 References

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

Kendall M G and Stuart A (1976) *The Advanced Theory of Statistics (Volume 3)* (3rd Edition) Griffin

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.

If WEIGHT = 'U' (Unweighted), no weights are used.

If WEIGHT = 'W' (Weighted), weights are used and must be supplied in WT.

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

2: N – INTEGER *Input*

On entry: the number of observations, n.

Constraint: N > NX + NY.

3: M – INTEGER *Input*

On entry: the total number of variables, m.

Constraint: M ≥ NX + NY.

4: Z(LDZ,M) – ***real*** array *Input*

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

Both x and y variables are to be included in Z, the indicator array, ISZ, being used to assign the variables in Z to the x or y sets as appropriate.

5: LDZ – INTEGER *Input*

On entry: the first dimension of the array Z as declared in the (sub)program from which G03ADF is called.

Constraint: LDZ $\geq$ N.

6: ISZ(M) – INTEGER array *Input*

On entry: ISZ(j) indicates whether or not the jth variable is included in the analysis and to which set of variables it belongs.

If ISZ(j) > 0, then the variable contained in the jth column of Z is included as an x variable in the analysis.

If ISZ(j) < 0, then the variable contained in the jth column of Z is included as a y variable in the analysis.

If ISZ(j) $= 0$, then the variable contained in the jth column of Z is not included in the analysis.

Constraint: only NX elements of ISZ can be > 0 and only NY elements of ISZ can be < 0.

7: NX – INTEGER *Input*

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

Constraint: NX ≥ 1.

8: NY – INTEGER *Input*

On entry: the number of y variables in the analysis, n_y.

Constraint: NY ≥ 1.

9: WT(*) – ***real*** array *Input*

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

If WT(i) $= 0.0$, then the ith observation is not included in the analysis. The effective number of observations is the sum of weights.

If WEIGHT = 'U', then WT is not referenced and the effective number of observations is n.

Constraint: WT(i) ≥ 0.0, for $i = 1, 2, \ldots, n$ and the sum of weights $\geq$ NX + NY + 1.

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

On exit: the statistics of the canonical variate analysis.

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

E($i, 2$), the eigenvalues of Σ, λ_i^2, for $i = 1, 2, \ldots, l$.

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

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

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

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

11: LDE – INTEGER *Input*

On entry: the first dimension of the array E as declared in the (sub)program from which G03ADF is called.

Constraint: LDE $\geq \min(\mathrm{NX}, \mathrm{NY})$.

12: NCV – INTEGER *Output*

On exit: the number of canonical correlations, l. This will be the minimum of the rank of X and the rank of Y.

13: CVX(LDCVX,MCV) – ***real*** array *Output*

On exit: the canonical variate loadings for the x variables. CVX(i,j) contains the loading coefficient for the ith x variable on the jth canonical variate.

14: LDCVX – INTEGER *Input*

On entry: the first dimension of the array CVX as declared in the (sub)program from which G03ADF is called.

Constraint: LDCVX $\geq$ NX.

15: MCV – INTEGER *Input*

On entry: an upper limit to the number of canonical variates.

Constraint: MCV $\geq \min(\mathrm{NX}, \mathrm{NY})$.

16: CVY(LDCVY,MCV) – ***real*** array *Output*

On exit: the canonical variate loadings for the y variables. CVY(i,j) contains the loading coefficient for the ith y variable on the jth canonical variate.

17: LDCVY – INTEGER *Input*

On entry: the first dimension of the array CVY as declared in the (sub)program from which G03ADF is called.

Constraint: LDCVY $\geq$ NY.

18: 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 ≥ 0.0.

19: WK(IWK) – ***real*** array *Workspace*
20: IWK – INTEGER *Input*

On entry: the dimension of the array WK as declared in the (sub)program from which G03ADF is called.

Constraints:

if $\mathrm{NX} \geq \mathrm{NY}$, then
$\mathrm{IWK} \geq \mathrm{N} \times \mathrm{NX} + \mathrm{NX} + \mathrm{NY} + \max((5 \times (\mathrm{NX} - 1) + \mathrm{NX} \times \mathrm{NX}), \mathrm{N} \times \mathrm{NY})$,
if $\mathrm{NX} < \mathrm{NY}$, then
$\mathrm{IWK} \geq \mathrm{N} \times \mathrm{NY} + \mathrm{NX} + \mathrm{NY} + \max((5 \times (\mathrm{NY} - 1) + \mathrm{NY} \times \mathrm{NY}), \mathrm{N} \times \mathrm{NX})$.

21: 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, $\mathrm{NX} < 1$,
or $\mathrm{NY} < 1$,
or $\mathrm{M} < \mathrm{NX} + \mathrm{NY}$,
or $\mathrm{N} \leq \mathrm{NX} + \mathrm{NY}$,
or $\mathrm{MCV} < \min(\mathrm{NX}, \mathrm{NY})$,
or $\mathrm{LDZ} < \mathrm{N}$,
or $\mathrm{LDCVX} < \mathrm{NX}$,
or $\mathrm{LDCVY} < \mathrm{NY}$,
or $\mathrm{LDE} < \min(\mathrm{NX}, \mathrm{NY})$,
or $\mathrm{NX} \geq \mathrm{NY}$ and
$\mathrm{IWK} < \mathrm{N} \times \mathrm{NX} + \mathrm{NX} + \mathrm{NY} + \max((5 \times (\mathrm{NX} - 1) + \mathrm{NX} \times \mathrm{NX}), \mathrm{N} \times \mathrm{NY})$,
or $\mathrm{NX} < \mathrm{NY}$ and
$\mathrm{IWK} < \mathrm{N} \times \mathrm{NY} + \mathrm{NX} + \mathrm{NY} + \max((5 \times (\mathrm{NY} - 1) + \mathrm{NY} \times \mathrm{NY}), \mathrm{N} \times \mathrm{NX})$,
or WEIGHT $\neq$ 'U' or 'W',
or $\mathrm{TOL} < 0.0$.

IFAIL $= 2$

On entry, a WEIGHT = 'W' and value of $\mathrm{WT} < 0.0$.

IFAIL $= 3$

On entry, the number of x variables to be included in the analysis as indicated by ISZ is not equal to NX.
or the number of y variables to be included in the analysis as indicated by ISZ is not equal to NY.

IFAIL $= 4$

On entry, the effective number of observations is less than $\mathrm{NX} + \mathrm{NY} + 1$.

IFAIL $= 5$

A singular value decomposition has failed to converge. See F02WEF or F02WUF. This is an unlikely error exit.

IFAIL $= 6$

A canonical correlation is equal to 1. This will happen if the x and y variables are perfectly correlated.

IFAIL $= 7$

On entry, the rank of the X matrix or the rank of the Y matrix is 0. This will happen if all the x or y 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, G03ADF should be less affected by ill-conditioned problems.

8 Further Comments

None.

9 Example

A sample of nine observations with two variables in each set is read in. The second and third variables are x variables while the first and last are y variables. 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.

```
*     G03ADF Example Program Text
*     Mark 14 Release.  NAG Copyright 1989.
*     .. Parameters ..
      INTEGER          NMAX, IMAX, IWKMAX
      PARAMETER        (NMAX=9,IMAX=2,IWKMAX=40)
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
*     .. Local Scalars ..
      real             TOL
      INTEGER          I, IFAIL, IX, IY, J, M, N, NCV, NX, NY
      CHARACTER        WEIGHT
*     .. Local Arrays ..
      real             CVX(IMAX,IMAX), CVY(IMAX,IMAX), E(IMAX,6),
     +                 WK(IWKMAX), WT(NMAX), Z(NMAX,2*IMAX)
      INTEGER          ISZ(2*IMAX)
*     .. External Subroutines ..
      EXTERNAL         G03ADF
*     .. Executable Statements ..
      WRITE (NOUT,*) 'G03ADF Example Program Results'
*     Skip heading in data file
      READ (NIN,*)
      READ (NIN,*) N, M, IX, IY, WEIGHT
      IF (N.LE.NMAX .AND. IX.LE.IMAX .AND. IY.LE.IMAX) THEN
         IF (WEIGHT.EQ.'W' .OR. WEIGHT.EQ.'w') THEN
            DO 20 I = 1, N
               READ (NIN,*) (Z(I,J),J=1,M), WT(I)
   20       CONTINUE
         ELSE
            DO 40 I = 1, N
               READ (NIN,*) (Z(I,J),J=1,M)
   40       CONTINUE
         END IF
         READ (5,*) (ISZ(J),J=1,M)
         TOL = 0.000001e0
         NX = IX
         NY = IY
         IFAIL = 0
*
         CALL G03ADF(WEIGHT,N,M,Z,NMAX,ISZ,NX,NY,WT,E,IMAX,NCV,CVX,IMAX,
     +               IMAX,CVY,IMAX,TOL,WK,IWKMAX,IFAIL)
*
         WRITE (NOUT,*)
         WRITE (NOUT,99999) 'Rank of X = ', NX, ' Rank of Y = ', NY
         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, IX
             WRITE (NOUT,99997) (CVX(I,J),J=1,NCV)
   80     CONTINUE
          WRITE (NOUT,*)
          WRITE (NOUT,*) 'Canonical coefficients for Y'
          DO 100 I = 1, IY
             WRITE (NOUT,99997) (CVY(I,J),J=1,NCV)
  100     CONTINUE
      END IF
      STOP
*
99999 FORMAT (1X,A,I2,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

```
G03ADF Example Program Data
 9 4 2 2 'U'
 80.0 58.4 14.0 21.0
 75.0 59.2 15.0 27.0
 78.0 60.3 15.0 27.0
 75.0 57.4 13.0 22.0
 79.0 59.5 14.0 26.0
 78.0 58.1 14.5 26.0
 75.0 58.0 12.5 23.0
 64.0 55.5 11.0 22.0
 80.0 59.2 12.5 22.0
 -1    1    1    -1
```

9.3 Program Results

```
 G03ADF Example Program Results

 Rank of X =  2 Rank of Y =  2

 Canonical    Eigenvalues Percentage     Chisq      DF     Sig
 correlations              variation
       0.9570     10.8916     0.9863   14.3914     4.0  0.0061
       0.3624      0.1512     0.0137    0.7744     1.0  0.3789

 Canonical coefficients for X
   -0.4261   1.0337
   -0.3444  -1.1136

 Canonical coefficients for Y
   -0.1415   0.1504
   -0.2384  -0.3424
```