nag_mv_canon_corr (g03adc)

1. Purpose

nag_mv_canon_corr (g03adc) performs canonical correlation analysis upon input data matrices.

2. Specification

```
#include <nag.h>
#include <nagg03.h>

void nag_mv_canon_corr(Integer n, Integer m,  double z[], Integer tdz,
             Integer isz[], Integer nx, Integer ny, double wt[], double e[],
             Integer tde, Integer *ncv, double cvx[], Integer tdcvx,
             double cvy[], Integer tdcvy, double tol, NagError *fail)
```

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 matrix 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, nag_mv_canon_corr (g03adc) 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:

$$V = Q_x^T Q_y,$$

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 $k_x < n_x$:

$$X = Q_x D_x P_x^T.$$

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 $k_y < n_y$:

$$Y = Q_y D_y P_y^T.$$

Let the SVD of V be:

$$V = U_x \Delta U_y^T$$

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 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:

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

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 $i = k_{\text{min}}$ is not significant, then the remaining tests for $i > k_{\text{min}}$ 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. Parameters

n

Input: the number of observations, n.
Constraint: **n** > **nx** + **ny**.

m

Input: the total number of variables, m.
Constraint: **m** ≥ **nx** + **ny**.

z[n][tdx]

Input: **z**$[i-1][j-1]$ 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.

tdz

Input: the last dimension of the array **z** as declared in the calling program.
Constraint: **tdz** ≥ **m**.

isz[m]

Input: **isz**$[j-1]$ indicates whether or not the jth variable is to be included in the analysis and to which set of variables it belongs.

If **isz**$[j-1] > 0$, then the variable contained in the jth column of **z** is included as an x variable in the analysis.
If **isz**$[j-1] < 0$, then the variable contained in the jth column of **z** is included as a y variable in the analysis.
If **isz**$[j-1] = 0$, 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.

nx

Input: the number of x variables in the analysis, n_x.
Constraint: **nx** ≥ 1.

ny

Input: the number of y variables in the analysis, n_y.
Constraint: **ny** ≥ 1.

wt[n]

Input: the elements of **wt** must contain the weights to be used in the analysis. The effective number of observations is the sum of the weights. If $\mathbf{wt}[i-1] = 0.0$ then the ith observation is not included in the analysis.

Constraints:

$\mathbf{wt}[i-1] \geq 0.0$, for $i = 1, 2, \ldots, n$,
$\sum_{i=1}^{n} \mathbf{wt}[i-1] \geq \mathbf{nx} + \mathbf{ny} + 1$.

Note: If **wt** is set to the null pointer **NULL**, i.e., (double *)0, then **wt** is not referenced and the effective number of observations is n.

e[min(nx,ny)][tde]

Output: the statistics of the canonical variate analysis.

$\mathbf{e}[i-1][0]$, the canonical correlations, δ_i, for $i = 1, 2, \ldots, l$.

$\mathbf{e}[i-1][1]$, the eigenvalues of Σ, λ_i^2, for $i = 1, 2, \ldots, l$.

$\mathbf{e}[i-1][2]$, the proportion of variation explained by the ith canonical variate, for $i = 1, 2, \ldots, l$.

$\mathbf{e}[i-1][3]$, the χ^2 statistic for the ith canonical variate, for $i = 1, 2, \ldots, l$.

$\mathbf{e}[i-1][4]$, the degrees of freedom for χ^2 statistic for the ith canonical variate, for $i = 1, 2, \ldots, l$.

$\mathbf{e}[i-1][5]$, the significance level for the χ^2 statistic for the ith canonical variate, for $i = 1, 2, \ldots, l$.

tde

Input: the last dimension of the array **e** as declared in the calling program.

Constraint: $\mathbf{tde} \geq 6$.

ncv

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

cvx[nx][tdcvx]

Output: the canonical variate loadings for the x variables. $\mathbf{cvx}[i-1][j-1]$ contains the loading coefficient for the ith x variable on the jth canonical variate.

tdcvx

Input: the last dimension of the array **cvx** as declared in the calling program.

Constraint: **tdcvx** $\geq$ min**(nx,ny).**

cvy[ny][tdcvy]

Output: the canonical variate loadings for the y variables. $\mathbf{cvy}[i-1][j-1]$ contains the loading coefficient for the ith y variable on the jth canonical variate.

tdcvy

Input: the last dimension of the array **cvy** as declared in the calling program.

Constraint: **tdcvy** $\geq$ min**(nx,ny).**

tol

Input: 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: $\mathbf{tol} \geq 0.0$.

fail

The NAG error parameter, see the Essential Introduction to the NAG C Library.

5. Error Indications and Warnings

NE_INT_ARG_LT

On entry, **nx** must not be less than 1: **nx** = ⟨*value*⟩.
On entry, **ny** must not be less than 1: **ny** = ⟨*value*⟩.
On entry, **tde** must not be less than 6: **tde** = ⟨*value*⟩.

NE_REAL_ARG_LT

On entry, **tol** must not be less than 0.0: **tol** = ⟨*value*⟩.

NE_2_INT_ARG_LT

On entry, **tdz** = ⟨*value*⟩ while **m** = ⟨*value*⟩.
These parameters must satisfy **tdz** ≥ **m**.

NE_3_INT_ARG_CONS

On entry, **m** = ⟨*value*⟩, **nx** = ⟨*value*⟩ and **ny** = ⟨*value*⟩.
These parameters must satisfy **m** ≥ **nx** + **ny**.
On entry, **n** = ⟨*value*⟩, **nx** = ⟨*value*⟩ and **ny** = ⟨*value*⟩.
These parameters must satisfy **n** > **nx** + **ny**.
On entry, **tdcvx** = ⟨*value*⟩, **nx** = ⟨*value*⟩ and **ny** = ⟨*value*⟩.
These parameters must satisfy **tdcvx** ≥ min(**nx**,**ny**).
On entry, **tdcvy** = ⟨*value*⟩, **nx** = ⟨*value*⟩ and **ny** = ⟨*value*⟩.
These parameters must satisfy **tdcvy** ≥ min(**nx**,**ny**).

NE_NEG_WEIGHT_ELEMENT

On entry, **wt**[⟨*value*⟩] = ⟨*value*⟩.
Constraint: When referenced, all elements of **wt** must be non-negative.

NE_VAR_INCL_INDICATED

The number of variables, **nx** in the analysis = ⟨*value*⟩, while the number of **x** variables included in the analysis via array **isz** = ⟨*value*⟩.
Constraint: these two numbers must be the same.

The number of variables, **ny** in the analysis = ⟨*value*⟩, while the number of **y** variables included in the analysis via array **isz** = ⟨*value*⟩.
Constraint: these two numbers must be the same.

NE_OBSERV_LT_VAR

With weighted data, the effective number of observations given by the sum of weights = ⟨*value*⟩, while number of variables included in the analysis, **nx** + **ny** = ⟨*value*⟩.
Constraint: Effective number of observations ≥ **nx** + **ny** + 1.

NE_SVD_NOT_CONV

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

NE_CANON_CORR_1

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

NE_MAT_RANK_ZERO

The rank of the **x** matrix or the rank of the **y** matrix is zero.
This will happen if all the x and y variables are constants.

NE_ALLOC_FAIL

Memory allocation failed.

NE_INTERNAL_ERROR

An internal error has occurred in this function. Check the function call and any array sizes. If the call is correct then please consult NAG for assistance.

6. Further Comments

6.1. 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, nag_mv_canon_corr should be less affected by ill conditioned problems.

6.2. 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)* Griffin (3rd Edition).

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

7. See Also

None.

8. 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.

8.1. Program Text

```
/* nag_mv_canon_corr (g03adc) Example Program.
 *
 * Copyright 1998 Numerical Algorithms Group.
 *
 * Mark 5, 1998.
 *
 */
#include <nag.h>
#include <stdio.h>
#include <nag_stdlib.h>
#include <nagg03.h>

#define NMAX 9
#define IMAX 2
#define MMAX 2*IMAX

main()
{
  double e[IMAX][6];
  double z[NMAX][MMAX];
  double wt[NMAX];
  double tol, cvx[IMAX][IMAX], cvy[IMAX][IMAX];
  double *wtptr;

  Integer i, j, m, n;
  Integer ix, iy;
  Integer nx, ny;
  Integer ncv;
  Integer isz[4];
  Integer tdz=MMAX, tde=6, tdc=IMAX;

  char weight[2];

  Vprintf("g03adc Example Program Results\n\n");

  /* Skip heading in data file */
  Vscanf("%*[^\n]");

  Vscanf("%ld",&n);
  Vscanf("%ld",&m);
  Vscanf("%ld",&ix);
  Vscanf("%ld",&iy);
  Vscanf("%s",weight);

  if (n <= NMAX && ix <= IMAX && iy <= IMAX)
    {
      if (*weight == 'W')
        {
          for (i = 0; i < n; ++i)
            {
              for (j = 0; j < m; ++j)
                Vscanf("%lf",&z[i][j]);
              Vscanf("%lf",&wt[i]);
            }
          wtptr = wt;
        }
      else
        {
          for (i = 0; i < n; ++i)
            {
              for (j = 0; j < m; ++j)
                Vscanf("%lf",&z[i][j]);
            }
          wtptr = 0;
        }
      for (j = 0; j < m; ++j)
        Vscanf("%ld",&isz[j]);
      tol = 1e-6;
      nx = ix;
      ny = iy;

      g03adc(n, m, (double *)z, tdz, isz, nx, ny, wtptr, (double *)e, tde,
             &ncv, (double *)cvx, tdc, (double *)cvy, tdc, tol, NAGERR_DEFAULT);

      Vprintf("\n%s%2ld%s%2ld\n\n","Rank of x = ",nx," Rank of y = ",ny);
      Vprintf("    Canonical    Eigenvalues Percentage      Chisq \
         DF           Sig\n");
      Vprintf("    correlations              variation\n");

      for (i = 0; i < ncv; ++i)
        {
          for (j = 0; j < 6; ++j)
            Vprintf("%12.4f",e[i][j]);
          Vprintf("\n");
        }
      Vprintf("\nCanonical coefficients for x\n");
      for (i = 0; i < ix; ++i)
        {
          for (j = 0; j < ncv; ++j)
            Vprintf("%9.4f",cvx[i][j]);
          Vprintf("\n");
        }
      Vprintf("\nCanonical coefficients for y\n");
      for (i = 0; i < iy; ++i)
        {
          for (j = 0; j < ncv; ++j)
            Vprintf("%9.4f",cvy[i][j]);
          Vprintf("\n");
        }
      exit(EXIT_SUCCESS);
    }
  else
    {
      Vprintf("Incorrect input value of n or ix or iy.\n");
      exit(EXIT_FAILURE);
    }
}
```

8.2. Program Data

```
g03adc 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
```

8.3. Program Results

```
g03adc 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      4.0000      0.0061
      0.3624      0.1512      0.0137      0.7744      1.0000      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
```