NAG Toolbox

nag_mv_canon_corr (g03ad)

1 Purpose

nag_mv_canon_corr (g03ad) performs canonical correlation analysis upon input data matrices.

2 Syntax

```
[e, ncv, cvx, cvy, ifail] = nag_mv_canon_corr(z, isz, nx, ny, mcv, tol, 'n', n,
'm', m, 'wt', wt)
```

```
[e, ncv, cvx, cvy, ifail] = g03ad(z, isz, nx, ny, mcv, tol, 'n', n, 'm', m, 'wt',
wt)
```

Note: the interface to this routine has changed since earlier releases of the toolbox:

At Mark 24: *weight* was removed from the interface; **wt** was made optional

At Mark 22: **n** was made optional.

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 datasets 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 (g03ad) 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 columns 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 columns 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 nonzero 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:

$$\lambda_i^2 = \delta_i^2.$$

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:

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

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

Hastings N A J and Peacock J B (1975) *Statistical Distributions* Butterworth

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

5.1 Compulsory Input Parameters

1: **z**(ldz, **m**) – REAL (KIND=nag_wp) array

ldz, the first dimension of the array, must satisfy the constraint $ldz \geq$ **n**.

z(i, j) must contain the ith observation for the jth variable, for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$.

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.

2: **isz**(**m**) – INTEGER array

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

isz(j) > 0
The variable contained in the jth column of **z** is included as an x variable in the analysis.

isz(j) < 0
The variable contained in the jth column of **z** is included as a y variable in the analysis.

isz$(j) = 0$
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.

3: **nx** – INTEGER

The number of x variables in the analysis, n_x.

Constraint: **nx** ≥ 1.

4: **ny** – INTEGER

The number of y variables in the analysis, n_y.

Constraint: **ny** ≥ 1.

5: **mcv** – INTEGER

An upper limit to the number of canonical variates.

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

6: **tol** – REAL (KIND=nag_wp)

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: **tol** ≥ 0.0.

5.2 Optional Input Parameters

1: **n** – INTEGER

Default: the dimension of the array **wt** and the first dimension of the array **z**. (An error is raised if these dimensions are not equal.)

n, the number of observations.

Constraint: **n** $>$ **nx** $+$ **ny**.

2: **m** – INTEGER

Default: the dimension of the array **isz** and the second dimension of the array **z**. (An error is raised if these dimensions are not equal.)

m, the total number of variables.

Constraint: **m** $\geq$ **nx** $+$ **ny**.

3: **wt**(:) – REAL (KIND=nag_wp) array

The dimension of the array **wt** must be at least **n** if $weight =$ 'W', and at least 1 otherwise

If $weight =$ 'W', 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. The effective number of observations is the sum of weights.

If $weight =$ 'U', **wt** is not referenced and the effective number of observations is n.

Constraints:

wt$(i) \geq 0.0$, for $i = 1, 2, \ldots, n$;
the sum of weights $\geq$ **nx** $+$ **ny** $+ 1$.

5.3 Output Parameters

1: **e**(lde, **6**) – REAL (KIND=nag_wp) array

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 $i = 1, 2, \ldots, l$.

2: **ncv** – INTEGER

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

3: **cvx**($ldcvx$, **mcv**) – REAL (KIND=nag_wp) array

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.

4: **cvy**($ldcvy$, **mcv**) – REAL (KIND=nag_wp) array

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.

5: **ifail** – INTEGER

ifail $= 0$ unless the function detects an error (see Section 5).

6 Error Indicators and Warnings

Errors or warnings detected by the function:

ifail $= 1$

On entry, $\mathbf{nx} < 1$,
or $\mathbf{ny} < 1$,
or $\mathbf{m} < \mathbf{nx} + \mathbf{ny}$,
or $\mathbf{n} \leq \mathbf{nx} + \mathbf{ny}$,
or $\mathbf{mcv} < \min(\mathbf{nx}, \mathbf{ny})$,
or $ldz < \mathbf{n}$,
or $ldcvx < \mathbf{nx}$,
or $ldcvy < \mathbf{ny}$,
or $lde < \min(\mathbf{nx}, \mathbf{ny})$,
or $\mathbf{nx} \geq \mathbf{ny}$ and
$iwk < \mathbf{n} \times \mathbf{nx} + \mathbf{nx} + \mathbf{ny} + \max((5 \times (\mathbf{nx} - 1) + \mathbf{nx} \times \mathbf{nx}), \mathbf{n} \times \mathbf{ny})$,
or $\mathbf{nx} < \mathbf{ny}$ and
$iwk < \mathbf{n} \times \mathbf{ny} + \mathbf{nx} + \mathbf{ny} + \max((5 \times (\mathbf{ny} - 1) + \mathbf{ny} \times \mathbf{ny}), \mathbf{n} \times \mathbf{nx})$,
or $weight \neq$ 'U' or 'W',
or $\mathbf{tol} < 0.0$.

ifail $= 2$

On entry, a *weight* = 'W' and value of **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 **nx** + **ny** + 1.

ifail $= 5$

A singular value decomposition has failed to converge. See nag_eigen_real_triang_svd (f02wu). This is an unlikely error exit.

ifail $= 6$ (*warning*)

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

ifail $= 7$ (*warning*)

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.

ifail $= -99$

An unexpected error has been triggered by this routine. Please contact NAG.

ifail $= -399$

Your licence key may have expired or may not have been installed correctly.

ifail $= -999$

Dynamic memory allocation failed.

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, nag_mv_canon_corr (g03ad) should be less affected by ill-conditioned problems.

8 Further Comments

None.

9 Example

This example has nine observations and two variables in each set of the four variables read in, the second and third are x variables while the first and last are y variables. Canonical variate analysis is performed and the results printed.

9.1 Program Text

```
     function g03ad_example

fprintf('g03ad example results\n\n');

z   = [80, 58.4, 14.0, 21;
       75, 59.2, 15.0, 27;
       78, 60.3, 15.0, 27;
       75, 57.4, 13.0, 22;
       79, 59.5, 14.0, 26;
       78, 58.1, 14.5, 26;
       75, 58.0, 12.5, 23;
       64, 55.5, 11.0, 22;
       80, 59.2, 12.5, 22];
isz = [nag_int(-1);1;1;-1];
nx  = nag_int(2);
ny  = nx;
mcv = nx;
tol = 1e-06;

[e, ncv, cvx, cvy, ifail] = ...
  g03ad( ...
  z, isz, nx, ny, mcv, tol);

fprintf('Rank of x = %d, Rank of y = %d\n\n', nx, ny);
fprintf('Canonical    Eigenvalues Percentage     Chisq      DF     Sig\n');
fprintf('correlations              variation\n');
fprintf('%11.4f%12.4f%12.4f%10.4f%8.1f%8.4f\n',e');
fprintf('\n');

mtitle = 'Canonical Coefficients for x';
matrix = 'General';
diag   = ' ';
[ifail] = x04ca( ...
                 matrix, diag, cvx, mtitle);

fprintf('\n');
mtitle = 'Canonical Coefficients for y';
[ifail] = x04ca( ...
                 matrix, diag, cvy, mtitle);
```

9.2 Program Results

```
     g03ad example results

Rank of x = 2, Rank of y = 2

Canonical    Eigenvalues Percentage     Chisq      DF     Sig
correlations             variation
     0.9570      0.9159      0.8746   14.3914     4.0  0.0061
     0.3624      0.1313      0.1254    0.7744     1.0  0.3789

 Canonical Coefficients for x
            1          2
 1     -0.4261     1.0337
 2     -0.3444    -1.1136

 Canonical Coefficients for y
         1       2
 1  -0.1415  0.1504
 2  -0.2384 -0.3424
```