NAG Toolbox for MATLAB

g03ac

1 Purpose

g03ac performs a canonical variate (canonical discrimination) analysis.

2 Syntax

```
[nig, cvm, e, ncv, cvx, irankx, ifail] = g03ac(weight, x, isx, nx, ing,
ng, wt, tol, 'n', n, 'm', m)
```

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, g03ac 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 $l = \min(k, ng - 1)$ 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:

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

can be used. This is asymptotically distributed as a χ^2-distribution with $(k - i)\left(n_g - 1 - i\right)$ 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

5.1 Compulsory Input Parameters

1: **weight – string**

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: **x(ldx,m) – double array**

ldx, the first dimension of the array, must be at least **n**.

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

3: **isx(m) – int32 array**

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.

4: **nx – int32 scalar**

the number of variables in the analysis, n_x.

Constraint: **nx** ≥ 1.

5: **ing(n) – int32 array**

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

6: **ng – int32 scalar**

the number of groups, n_g.

Constraint: **ng** ≥ 2.

7: **wt(*) – double array**

Note: the dimension of the array **wt** must be at least **n** if **weight** = 'W' or 'V', and at least 1 otherwise.

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:

$$\mathbf{wt}(i) \geq 0.0, \text{ for } i = 1, 2, \ldots, n;$$

$$\sum_{1}^{n} \mathbf{wt}(i) \geq \mathbf{nx} + \text{effective number of groups.}$$

8: **tol – double scalar**

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 nonnegative 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 – int32 scalar**

Default: The dimension of the array **ing**.

n, the number of observations.

Constraint: **n** $\geq$ **nx** + **ng**.

2: **m – int32 scalar**

Default: The dimension of the arrays **isx**, **x**. (An error is raised if these dimensions are not equal.)

m, the total number of variables.

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

5.3 Input Parameters Omitted from the MATLAB Interface

ldx, ldcvm, lde, ldcvx, wk, iwk

5.4 Output Parameters

1: **nig(ng) – int32 array**

nig(j) gives the number of observations in group j, for $j = 1, 2, \ldots, n_g$.

2: **cvm(ldcvm,nx) – double array**

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.

3: **e(lde,6) – double 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 the within-group sum of squares matrix, 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 $i = 1,2,\ldots,l$.

4: **ncv – int32 scalar**

The number of canonical variates, l. This will be the minimum of $n_g - 1$ and the rank of **x**.

5: **cvx(ldcvx,ng − 1) – double array**

The canonical variate loadings. **cvx**(i,j) contains the loading coefficient for the ith variable on the jth canonical variate, for $i = 1,2,\ldots,n_x$ and $j = 1,2,\ldots,l$; the remaining columns, if any, are used as workspace.

6: **irankx – int32 scalar**

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** × (largest singular value).

7: **ifail – int32 scalar**

0 unless the function detects an error (see Section 6).

6 Error Indicators and Warnings

Errors or warnings detected by the function:

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** $< \min(\mathbf{nx}, \mathbf{ng} - 1)$,
or $\mathbf{nx} \geq \mathbf{ng} - 1$ and $\mathbf{iwk} < \mathbf{n} \times \mathbf{nx} + \max(5 \times (\mathbf{nx} - 1) + (\mathbf{nx} + 1) \times \mathbf{nx}, \mathbf{n})$,
or $\mathbf{nx} < \mathbf{ng} - 1$ and $\mathbf{iwk} < \mathbf{n} \times \mathbf{nx} + \max(5 \times (\mathbf{nx} - 1) + (\mathbf{ng} - 1) \times \mathbf{nx}, \mathbf{n})$,
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, g03ac should be less affected by ill-conditioned problems.

8 Further Comments

None.

9 Example

```
weight = ’U’;
x = [13.3, 10.6, 21.2;
     13.6, 10.2, 21;
     14.2, 10.7, 21.1;
     13.4, 9.4, 21;
     13.2, 9.6, 20.1;
     13.9, 10.4, 19.8;
     12.9, 10, 20.5;
     12.2, 9.9, 20.7;
     13.9, 11, 19.1];
isx = [int32(1);
     int32(1);
     int32(1)];
nx = int32(3);
ing = [int32(1);
     int32(2);
     int32(3);
     int32(1);
     int32(2);
     int32(3);
     int32(1);
     int32(2);
     int32(3)];
ng = int32(3);
wt = [];
tol = 1e-06;
[nig, cvm, e, ncv, cvx, irankx, ifail] = g03ac(weight, x, isx, nx, ing,
ng, wt, tol)
```

```
nig =
           3
           3
           3
cvm =
    0.9841    0.2797   -0.1653
    1.1805   -0.2632    0.0177
   -2.1646   -0.0164    0.1476
e =
    0.8826    3.5238    0.9795    7.9032    6.0000    0.2453
    0.2623    0.0739    0.0205    0.3564    2.0000    0.8368
ncv =
           2
cvx =
   -1.7070    0.7277
   -1.3481    0.3138
    0.9327    1.2199
irankx =
           3
ifail =
           0
```