NAG Toolbox

nag_mv_discrim (g03da)

1 Purpose

nag_mv_discrim (g03da) computes a test statistic for the equality of within-group covariance matrices and also computes matrices for use in discriminant analysis.

2 Syntax

```
[nig, gmn, det, gc, stat, df, sig, ifail] = nag_mv_discrim(x, isx, nvar, ing,
ng, 'n', n, 'm', m, 'wt', wt)
```

```
[nig, gmn, det, gc, stat, df, sig, ifail] = g03da(x, isx, nvar, ing, ng, '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.

3 Description

Let a sample of n observations on p variables come from n_g groups with n_j observations in the jth group and $\sum n_j = n$. If the data is assumed to follow a multivariate Normal distribution with the variance-covariance matrix of the jth group Σ_j, then to test for equality of the variance-covariance matrices between groups, that is, $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_{n_g} = \Sigma$, the following likelihood-ratio test statistic, G, can be used;

$$G = C\left\{\left(n - n_g\right)\log|S| - \sum_{j=1}^{n_g}\left(n_j - 1\right)\log\left|S_j\right|\right\},$$

where

$$C = 1 - \frac{2p^2 + 3p - 1}{6(p+1)\left(n_g - 1\right)}\left(\sum_{j=1}^{n_g}\frac{1}{\left(n_j - 1\right)} - \frac{1}{\left(n - n_g\right)}\right),$$

and S_j are the within-group variance-covariance matrices and S is the pooled variance-covariance matrix given by

$$S = \frac{\sum_{j=1}^{n_g}\left(n_j - 1\right)S_j}{\left(n - n_g\right)}.$$

For large n, G is approximately distributed as a χ^2 variable with $\frac{1}{2}p(p+1)\left(n_g - 1\right)$ degrees of freedom, see Morrison (1967) for further comments. If weights are used, then S and S_j are the weighted pooled and within-group variance-covariance matrices and n is the effective number of observations, that is, the sum of the weights.

Instead of calculating the within-group variance-covariance matrices and then computing their determinants in order to calculate the test statistic, nag_mv_discrim (g03da) uses a QR decomposition. The group means are subtracted from the data and then for each group, a QR decomposition is computed to give an upper triangular matrix R_j^*. This matrix can be scaled to give a matrix R_j such that $S_j = R_j^{\mathrm{T}} R_j$. The pooled R matrix is then computed from the R_j matrices. The values of $|S|$ and the $\left|S_j\right|$ can then be calculated from the diagonal elements of R and the R_j.

This approach means that the Mahalanobis squared distances for a vector observation x can be computed as $z^{\mathrm{T}}z$, where $R_j z = (x - \bar{x}_j)$, $\bar{x}_j$ being the vector of means of the jth group. These distances can be calculated by nag_mv_discrim_mahal (g03db). The distances are used in discriminant analysis and nag_mv_discrim_group (g03dc) uses the results of nag_mv_discrim (g03da) to perform several different types of discriminant analysis. The differences between the discriminant methods are, in part, due to whether or not the within-group variance-covariance matrices are equal.

4 References

Aitchison J and Dunsmore I R (1975) *Statistical Prediction Analysis* Cambridge

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

Krzanowski W J (1990) *Principles of Multivariate Analysis* Oxford University Press

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

5 Parameters

5.1 Compulsory Input Parameters

1: **x**$(ldx, \mathbf{m})$ – REAL (KIND=nag_wp) array

ldx, the first dimension of the array, must satisfy the constraint $ldx \geq \mathbf{n}$.

x(k, l) must contain the kth observation for the lth variable, for $k = 1, 2, \ldots, n$ and $l = 1, 2, \ldots, \mathbf{m}$.

2: **isx**$(\mathbf{m})$ – INTEGER array

isx(l) indicates whether or not the lth variable in **x** is to be included in the variance-covariance matrices.

If **isx**$(l) > 0$ the lth variable is included, for $l = 1, 2, \ldots, \mathbf{m}$; otherwise it is not referenced.

Constraint: **isx**$(l) > 0$ for **nvar** values of l.

3: **nvar** – INTEGER

p, the number of variables in the variance-covariance matrices.

Constraint: **nvar** ≥ 1.

4: **ing**$(\mathbf{n})$ – INTEGER array

ing(k) indicates to which group the kth observation belongs, for $k = 1, 2, \ldots, n$.

Constraint: $1 \leq$ **ing**$(k) \leq$ **ng**, for $k = 1, 2, \ldots, n$

The values of **ing** must be such that each group has at least **nvar** members.

5: **ng** – INTEGER

The number of groups, n_g.

Constraint: **ng** ≥ 2.

5.2 Optional Input Parameters

1: **n** – INTEGER

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

n, the number of observations.

Constraint: **n** ≥ 1.

2: **m** – INTEGER

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

The number of variables in the data array **x**.

Constraint: $\mathbf{m} \geq \mathbf{nvar}$.

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 and the effective number of observations for a group is the sum of the weights of the observations in that group. If $\mathbf{wt}(k) = 0.0$ the kth observation is excluded from the calculations.

If $weight =$ 'U', **wt** is not referenced and the effective number of observations for a group is the number of observations in that group.

Constraint: if $weight =$ 'W', $\mathbf{wt}(k) \geq 0.0$, for $k = 1, 2, \ldots, n$.

5.3 Output Parameters

1: **nig**(**ng**) – INTEGER array

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

2: **gmn**($ldgmn$, **nvar**) – REAL (KIND=nag_wp) array

The jth row of **gmn** contains the means of the p selected variables for the jth group, for $j = 1, 2, \ldots, n_g$.

3: **det**(**ng**) – REAL (KIND=nag_wp) array

The logarithm of the determinants of the within-group variance-covariance matrices.

4: **gc**($(\mathbf{ng}+\mathbf{1}) \times \mathbf{nvar} \times (\mathbf{nvar}+\mathbf{1})/\mathbf{2}$) – REAL (KIND=nag_wp) array

The first $p(p+1)/2$ elements of **gc** contain R and the remaining n_g blocks of $p(p+1)/2$ elements contain the R_j matrices. All are stored in packed form by columns.

5: **stat** – REAL (KIND=nag_wp)

The likelihood-ratio test statistic, G.

6: **df** – REAL (KIND=nag_wp)

The degrees of freedom for the distribution of G.

7: **sig** – REAL (KIND=nag_wp)

The significance level for G.

8: **ifail** – INTEGER

$\mathbf{ifail} = 0$ 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, **nvar** < 1,
or **n** < 1,
or **ng** < 2,
or **m** $<$ **nvar**,
or $ldx <$ **n**,
or $ldgmn <$ **ng**,
or $weight \neq$ 'U' or 'W'.

ifail $= 2$

On entry, $weight =$ 'W' and a value of **wt** < 0.0.

ifail $= 3$

On entry, there are not exactly **nvar** elements of **isx** > 0,
or a value of **ing** is not in the range 1 to **ng**,
or the effective number of observations for a group is less than 1,
or a group has less than **nvar** members.

ifail $= 4$

R or one of the R_j is not of full rank.

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

The accuracy is dependent on the accuracy of the computation of the QR decomposition. See nag_lapack_dgeqrf (f08ae) for further details.

8 Further Comments

The time taken will be approximately proportional to np^2.

9 Example

The data, taken from Aitchison and Dunsmore (1975), is concerned with the diagnosis of three ‘types’ of Cushing's syndrome. The variables are the logarithms of the urinary excretion rates (mg/24hr) of two steroid metabolites. Observations for a total of 21 patients are input and the statistics computed by nag_mv_discrim (g03da). The printed results show that there is evidence that the within-group variance-covariance matrices are not equal.

9.1 Program Text

```
     function g03da_example

fprintf('g03da example results\n\n');

x = [1.1314,  2.4596;
     1.0986,  0.2624;
     0.6419, -2.3026;
     1.3350, -3.2189;
     1.4110,  0.0953;
     0.6419, -0.9163;
     2.1163,  0.0000;
     1.3350, -1.6094;
     1.3610, -0.5108;
     2.0541,  0.1823;
     2.2083, -0.5108;
     2.7344,  1.2809;
     2.0412,  0.4700;
     1.8718, -0.9163;
     1.7405, -0.9163;
     2.6101,  0.4700;
     2.3224,  1.8563;
     2.2192,  2.0669;
     2.2618,  1.1314;
     3.9853,  0.9163;
     2.7600,  2.0281];
[n,m] = size(x);
isx  = ones(m,1,nag_int_name);
nvar = nag_int(m);
ing  = ones(n,1,nag_int_name);
ing(7:16) = nag_int(2);
ing(17:n) = nag_int(3);
ng        = nag_int(3);

[nig, gmean, det, gc, stat, df, sig, ifail] = ...
  g03da( ...
  x, isx, nvar, ing, ng);

mtitle = 'Group means';
matrix = 'General';
diag   = ' ';
[ifail] = x04ca( ...
                 matrix, diag, gmean, mtitle);
fprintf('\nLog of determinants\n\n');
fprintf('%10.4f%10.4f%10.4f\n\n', det);
fprintf(' Stat = %7.4f\n', stat);
fprintf('   DF = %7.4f\n', df);
fprintf('  SIG = %7.4f\n', sig);
```

9.2 Program Results

```
     g03da example results

 Group means
             1          2
 1      1.0433    -0.6034
 2      2.0073    -0.2060
 3      2.7097     1.5998

Log of determinants

   -0.8273   -3.0460   -2.2877

 Stat = 19.2410
   DF =  6.0000
  SIG =  0.0038
```