NAG Toolbox

nag_mv_discrim_mahal (g03db)

1 Purpose

nag_mv_discrim_mahal (g03db) computes Mahalanobis squared distances for group or pooled variance-covariance matrices. It is intended for use after nag_mv_discrim (g03da).

2 Syntax

```
[d, ifail] = nag_mv_discrim_mahal(equal, mode, gmn, gc, nobs, isx, x, 'nvar',
nvar, 'ng', ng, 'm', m)
```

```
[d, ifail] = g03db(equal, mode, gmn, gc, nobs, isx, x, 'nvar', nvar, 'ng', ng,
'm', m)
```

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

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

3 Description

Consider p variables observed on n_g populations or groups. Let $\bar{x}_j$ be the sample mean and S_j the within-group variance-covariance matrix for the jth group and let x_k be the kth sample point in a dataset. A measure of the distance of the point from the jth population or group is given by the Mahalanobis distance, D_{kj}:

$$D_{kj}^2 = \left(x_k - \bar{x}_j\right)^{\mathrm{T}} S_j^{-1} \left(x_k - \bar{x}_j\right).$$

If the pooled estimated of the variance-covariance matrix S is used rather than the within-group variance-covariance matrices, then the distance is:

$$D_{kj}^2 = \left(x_k - \bar{x}_j\right)^{\mathrm{T}} S^{-1} \left(x_k - \bar{x}_j\right).$$

Instead of using the variance-covariance matrices S and S_j, nag_mv_discrim_mahal (g03db) uses the upper triangular matrices R and R_j supplied by nag_mv_discrim (g03da) such that $S = R^{\mathrm{T}}R$ and $S_j = R_j^{\mathrm{T}}R_j$. D_{kj}^2 can then be calculated as $z^{\mathrm{T}}z$ where $R_jz = \left(x_k - \bar{x}_j\right)$ or $Rz = \left(x_k - \bar{x}_j\right)$ as appropriate.

A particular case is when the distance between the group or population means is to be estimated. The Mahalanobis squared distance between the ith and jth groups is:

$$D_{ij}^2 = \left(\bar{x}_i - \bar{x}_j\right)^{\mathrm{T}} S_j^{-1} \left(\bar{x}_i - \bar{x}_j\right)$$

or

$$D_{ij}^2 = \left(\bar{x}_i - \bar{x}_j\right)^{\mathrm{T}} S^{-1} \left(\bar{x}_i - \bar{x}_j\right).$$

Note: $D_{jj}^2 = 0$ and that in the case when the pooled variance-covariance matrix is used $D_{ij}^2 = D_{ji}^2$ so in this case only the lower triangular values of D_{ij}^2, $i > j$, are computed.

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

5 Parameters

5.1 Compulsory Input Parameters

1: **equal** – CHARACTER(1)

Indicates whether or not the within-group variance-covariance matrices are assumed to be equal and the pooled variance-covariance matrix used.

equal = 'E'
The within-group variance-covariance matrices are assumed equal and the matrix R stored in the first $p(p+1)/2$ elements of **gc** is used.

equal = 'U'
The within-group variance-covariance matrices are assumed to be unequal and the matrices R_j, for $j = 1, 2, \dots, n_g$, stored in the remainder of **gc** are used.

Constraint: **equal** = 'E' or 'U'.

2: **mode** – CHARACTER(1)

Indicates whether distances from sample points are to be calculated or distances between the group means.

mode = 'S'
The distances between the sample points given in **x** and the group means are calculated.

mode = 'M'
The distances between the group means will be calculated.

Constraint: **mode** = 'M' or 'S'.

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

$ldgmn$, the first dimension of the array, must satisfy the constraint $ldgmn \geq$ **ng**.

The jth row of **gmn** contains the means of the p selected variables for the jth group, for $j = 1, 2, \dots, n_g$. These are returned by nag_mv_discrim (g03da).

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** should contain the upper triangular matrix R and the next n_g blocks of $p(p+1)/2$ elements should contain the upper triangular matrices R_j. All matrices must be stored packed by column. These matrices are returned by nag_mv_discrim (g03da). If **equal** = 'E' only the first $p(p+1)/2$ elements are referenced, if **equal** = 'U' only the elements $p(p+1)/2+1$ to $(n_g+1)p(p+1)/2$ are referenced.

Constraints:

if **equal** = 'E', $R \neq 0.0$;
if **equal** = 'U', the diagonal elements of the $R_j \neq 0.0$, for $j = 1, 2, \dots,$ **ng**.

5: **nobs** – INTEGER

If **mode** = 'S', the number of sample points in **x** for which distances are to be calculated.

If **mode** = 'M', **nobs** is not referenced.

Constraint: if **nobs** ≥ 1, **mode** = 'S'.

6: **isx**(:) – INTEGER array

The dimension of the array **isx** must be at least $\max(1, \mathbf{m})$

If **mode** = 'S', **isx**(l) indicates if the lth variable in **x** is to be included in the distance calculations. If **isx**$(l) > 0$ the lth variable is included, for $l = 1, 2, \dots, \mathbf{m}$; otherwise the lth variable is not referenced.

If **mode** $=$ 'M', **isx** is not referenced.

Constraint: if **mode** $=$ 'S', **isx**$(l) > 0$ for **nvar** values of l.

7: **x**$(ldx,:)$ – REAL (KIND=nag_wp) array

The first dimension, ldx, of the array **x** must satisfy

if **mode** $=$ 'S', $ldx \geq$ **nobs**;
otherwise 1.

The second dimension of the array **x** must be at least $\max(1, \mathbf{m})$.

If **mode** $=$ 'S' the kth row of **x** must contain x_k. That is **x**(k,l) must contain the kth sample value for the lth variable, for $k = 1, 2, \ldots,$ **nobs** and $l = 1, 2, \ldots,$ **m**. Otherwise **x** is not referenced.

5.2 Optional Input Parameters

1: **nvar** – INTEGER

Default: the second dimension of the array **gmn**.

p, the number of variables in the variance-covariance matrices as specified to nag_mv_discrim (g03da).

Constraint: **nvar** ≥ 1.

2: **ng** – INTEGER

Default: the first dimension of the array **gmn**.

The number of groups, n_g.

Constraint: **ng** ≥ 2.

3: **m** – INTEGER

Default: the dimension of the arrays **isx**, **x**.

If **mode** $=$ 'S', the number of variables in the data array **x**.

If **mode** $=$ 'M', **m** is not referenced.

Constraint: if **m** $\geq$ **nvar**, **mode** $=$ 'S'.

5.3 Output Parameters

1: **d**$(ldd, \mathbf{ng})$ – REAL (KIND=nag_wp) array

The squared distances.

If **mode** $=$ 'S', **d**(k,j) contains the squared distance of the kth sample point from the jth group mean, D_{kj}^2, for $k = 1, 2, \ldots,$ **nobs** and $j = 1, 2, \ldots, n_g$.

If **mode** $=$ 'M' and **equal** $=$ 'U', **d**(i,j) contains the squared distance between the ith mean and the jth mean, D_{ij}^2, for $i = 1, 2, \ldots, n_g$ and $j = 1, 2, \ldots, i-1, i+1, \ldots, n_g$. The elements **d**$(i,i)$ are not referenced, for $i = 1, 2, \ldots, n_g$.

If **mode** $=$ 'M' and **equal** $=$ 'E', **d**(i,j) contains the squared distance between the ith mean and the jth mean, D_{ij}^2, for $i = 1, 2, \ldots, n_g$ and $j = 1, 2, \ldots, i-1$. Since $D_{ij} = D_{ji}$ the elements **d**(i,j) are not referenced, for $i = 1, 2, \ldots, n_g$ and $j = i+1, \ldots, n_g$.

2: **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, **nvar** < 1,
or **ng** < 2,
or $ldgmn <$ **ng**,
or **mode** $=$ 'S' and **nobs** < 1,
or **mode** $=$ 'S' and **m** $<$ **nvar**,
or **mode** $=$ 'S' and $ldx <$ **nobs**,
or **mode** $=$ 'S' and $ldd <$ **nobs**,
or **mode** $=$ 'M' and $ldd <$ **ng**,
or **equal** $\neq$ 'E' or 'U',
or **mode** $\neq$ 'M' or 'S'.

ifail $= 2$

On entry, **mode** $=$ 'S' and the number of variables indicated by **isx** is not equal to **nvar**,
or **equal** $=$ 'E' and a diagonal element of R is zero,
or **equal** $=$ 'U' and a diagonal element of R_j for some j is zero.

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 will depend upon the accuracy of the input R or R_j matrices.

8 Further Comments

If the distances are to be used for discrimination, see also nag_mv_discrim_group (g03dc).

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 group means and R matrices are computed by nag_mv_discrim (g03da). A further six observations of unknown type are input, and the distances from the group means of the 21 patients of known type are computed under the assumption that the within-group variance-covariance matrices are not equal. These results are printed and indicate that the first four are close to one of the groups while observations 5 and 6 are some distance from any group.

9.1 Program Text

```
     function g03db_example

fprintf('g03db 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);

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

equal = 'U';
mode = 'Sample points';
nobs = nag_int(6);

% Data from which to compute distances
x = [1.6292, -0.9163;
     2.5572, 1.6094;
     2.5649, -0.2231;
     0.9555, -2.3026;
     3.4012, -2.3026;
     3.0204, -0.2231];

% Compute distances
[d, ifail] = g03db( ...
     equal, mode, gmean, gc, nobs, isx, x);

mtitle = 'Distances';
matrix = 'General';
diag   = ' ';
[ifail] = x04ca( ...
                 matrix, diag, d, mtitle);
```

9.2 Program Results

```
     g03db example results

 Distances
             1           2           3
 1      3.3393      0.7521     50.9283
 2     20.7771      5.6559      0.0597
 3     21.3631      4.8411     19.4978
 4      0.7184      6.2803    124.7323
 5     55.0003     88.8604     71.7852
 6     36.1703     15.7849     15.7489
```