NAG Toolbox

nag_mv_discrim_group (g03dc)

1 Purpose

nag_mv_discrim_group (g03dc) allocates observations to groups according to selected rules. It is intended for use after nag_mv_discrim (g03da).

2 Syntax

```
[prior, p, iag, ati, ifail] = nag_mv_discrim_group(typ, equal, priors, nig, gmn,
gc, det, isx, x, prior, atiq, 'nvar', nvar, 'ng', ng, 'nobs', nobs, 'm', m)
```

```
[prior, p, iag, ati, ifail] = g03dc(typ, equal, priors, nig, gmn, gc, det, isx,
x, prior, atiq, 'nvar', nvar, 'ng', ng, 'nobs', nobs, 'm', m)
```

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

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

3 Description

Discriminant analysis is concerned with the allocation of observations to groups using information from other observations whose group membership is known, X_t; these are called the training set. 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; these are calculated from a training set of n observations with n_j observations in the jth group, and let x_k be the kth observation from the set of observations to be allocated to the n_g groups. The observation can be allocated to a group according to a selected rule. The allocation rule or discriminant function will be based on the distance of the observation from an estimate of the location of the groups, usually the group means. A measure of the distance of the observation from the jth group mean is given by the Mahalanobis distance, D_{kj}:

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

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

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

Instead of using the variance-covariance matrices S and S_j, nag_mv_discrim_group (g03dc) 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 $R^{\mathrm{T}}{}_j z = \left(x_k - x_j\right)$ or $R^{\mathrm{T}}z = \left(x_k - x\right)$ as appropriate.

In addition to the distances, a set of prior probabilities of group membership, π_j, for $j = 1, 2, \ldots, n_g$, may be used, with $\sum \pi_j = 1$. The prior probabilities reflect your view as to the likelihood of the observations coming from the different groups. Two common cases for prior probabilities are $\pi_1 = \pi_2 = \cdots = \pi_{n_g}$, that is, equal prior probabilities, and $\pi_j = n_j/n$, for $j = 1, 2, \ldots, n_g$, that is, prior probabilities proportional to the number of observations in the groups in the training set.

nag_mv_discrim_group (g03dc) uses one of four allocation rules. In all four rules the p variables are assumed to follow a multivariate Normal distribution with mean μ_j and variance-covariance matrix Σ_j if the observation comes from the jth group. The different rules depend on whether or not the within-group variance-covariance matrices are assumed equal, i.e., $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_{n_g}$, and whether a predictive or estimative approach is used. If $p\left(x_k \mid \mu_j, \Sigma_j\right)$ is the probability of observing the observation x_k from group j, then the posterior probability of belonging to group j is:

$$p\big(j \mid x_k, \mu_j, \Sigma_j\big) \propto p\big(x_k \mid \mu_j, \Sigma_j\big)\pi_j. \tag{3}$$

In the estimative approach, the arguments μ_j and Σ_j in (3) are replaced by their estimates calculated from X_t. In the predictive approach, a non-informative prior distribution is used for the arguments and a posterior distribution for the arguments, $p\big(\mu_j, \Sigma_j \mid X_t\big)$, is found. A predictive distribution is then obtained by integrating $p\big(j \mid x_k, \mu_j, \Sigma_j\big)p\big(\mu_j, \Sigma_j \mid X\big)$ over the argument space. This predictive distribution then replaces $p\big(x_k \mid \mu_j, \Sigma_j\big)$ in (3). See Aitchison and Dunsmore (1975), Aitchison *et al.* (1977) and Moran and Murphy (1979) for further details.

The observation is allocated to the group with the highest posterior probability. Denoting the posterior probabilities, $p\big(j \mid x_k, \mu_j, \Sigma_j\big)$, by q_j, the four allocation rules are:

(i) Estimative with equal variance-covariance matrices – Linear Discrimination

$$\log q_j \propto -\tfrac{1}{2}D_{kj}^2 + \log \pi_j$$

(ii) Estimative with unequal variance-covariance matrices – Quadratic Discrimination

$$\log q_j \propto -\tfrac{1}{2}D_{kj}^2 + \log \pi_j - \tfrac{1}{2}\log\left|S_j\right|$$

(iii) Predictive with equal variance-covariance matrices

$$q_j^{-1} \propto \big((n_j+1)/n_j\big)^{p/2}\Big\{1 + \big[n_j/\big((n-n_g)\big(n_j+1\big)\big)\big]D_{kj}^2\Big\}^{\left(n+1-n_g\right)/2}$$

(iv) Predictive with unequal variance-covariance matrices

$$q_j^{-1} \propto C\Big\{\Big(\Big(n_j^2-1\Big)/n_j\Big)\left|S_j\right|\Big\}^{p/2}\Big\{1 + \Big(n_j/\Big(n_j^2-1\Big)\Big)D_{kj}^2\Big\}^{n_j/2},$$

where

$$C = \frac{\Gamma\big(\frac{1}{2}(n_j-p)\big)}{\Gamma\big(\frac{1}{2}n_j\big)}.$$

In the above the appropriate value of D_{kj}^2 from (1) or (2) is used. The values of the q_j are standardized so that,

$$\sum_{j=1}^{n_g} q_j = 1.$$

Moran and Murphy (1979) show the similarity between the predictive methods and methods based upon likelihood ratio tests.

In addition to allocating the observation to a group, nag_mv_discrim_group (g03dc) computes an atypicality index, $I_j(x_k)$. The predictive atypicality index is returned, irrespective of the value of the parameter **typ**. This represents the probability of obtaining an observation more typical of group j than the observed x_k (see Aitchison and Dunsmore (1975) and Aitchison *et al.* (1977)). The atypicality index is computed for unequal within-group variance-covariance matrices as:

$$I_j(x_k) = P\big(B \leq z : \tfrac{1}{2}p, \tfrac{1}{2}(n_j-p)\big)$$

where $P(B \leq \beta : a, b)$ is the lower tail probability from a beta distribution and

$$z = D_{kj}^2/\Big(D_{kj}^2 + \Big(n_j^2-1\Big)/n_j\Big),$$

and for equal within-group variance-covariance matrices as:

$$I_j(x_k) = P\big(B \leq z : \tfrac{1}{2}p, \tfrac{1}{2}(n-n_g-p+1)\big),$$

with

$$z = D_{kj}^2/\Big(D_{kj}^2 + (n-n_g)\big(n_j+1\big)/n_j\Big).$$

If $I_j(x_k)$ is close to 1 for all groups it indicates that the observation may come from a grouping not represented in the training set. Moran and Murphy (1979) provide a frequentist interpretation of $I_j(x_k)$.

4 References

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

Aitchison J, Habbema J D F and Kay J W (1977) A critical comparison of two methods of statistical discrimination *Appl. Statist.* **26** 15–25

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

Moran M A and Murphy B J (1979) A closer look at two alternative methods of statistical discrimination *Appl. Statist.* **28** 223–232

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

5 Parameters

5.1 Compulsory Input Parameters

1: **typ** – CHARACTER(1)

Whether the estimative or predictive approach is used.

typ = 'E'
The estimative approach is used.

typ = 'P'
The predictive approach is used.

Constraint: **typ** = 'E' or 'P'.

2: **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_i, for $i = 1, 2, \ldots, n_g$, stored in the remainder of **gc** are used.

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

3: **priors** – CHARACTER(1)

Indicates the form of the prior probabilities to be used.

priors = 'E'
Equal prior probabilities are used.

priors = 'P'
Prior probabilities proportional to the group sizes in the training set, n_j, are used.

priors = 'I'
The prior probabilities are input in **prior**.

Constraint: **priors** = 'E', 'I' or 'P'.

4: **nig**(**ng**) – INTEGER array

The number of observations in each group in the training set, n_j.

Constraints:

if **equal** = 'E', **nig**$(j) > 0$ and $\sum_{j=1}^{n_g}$**nig**$(j) >$ **ng** + **nvar**, for $j = 1, 2, \ldots, n_g$;
if **equal** = 'U', **nig**$(j) >$ **nvar**, for $j = 1, 2, \ldots, n_g$.

5: **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 variables for the jth group, for $j = 1, 2, \ldots, n_j$. These are returned by nag_mv_discrim (g03da).

6: **gc**((**ng** + **1**) × **nvar** × (**nvar** + **1**)/**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 $\left(n_g+1\right)p(p+1)/2$ are referenced.

Constraints:

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

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

If **equal** = 'U'. the logarithms of the determinants of the within-group variance-covariance matrices as returned by nag_mv_discrim (g03da). Otherwise **det** is not referenced.

8: **isx**(**m**) – INTEGER array

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 $l = 1, 2, \ldots,$ **m**; otherwise the lth variable is not referenced.

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

9: **x**(ldx, **m**) – REAL (KIND=nag_wp) array

ldx, the first dimension of the array, must satisfy the constraint $ldx \geq$ **nobs**.

x(k, l) must contain the kth observation for the lth variable, for $k = 1, 2, \ldots,$ **nobs** and $l = 1, 2, \ldots,$ **m**.

10: **prior**(**ng**) – REAL (KIND=nag_wp) array

If **priors** = 'I', the prior probabilities for the n_g groups.

Constraint: if **priors** = 'I', **prior**$(j) > 0.0$ and $\left|1 - \sum_{j=1}^{n_g}\textbf{prior}(j)\right| \leq 10 \times$ ***machine precision***, for $j = 1, 2, \ldots, n_g$.

11: **atiq** – LOGICAL

atiq must be *true* if atypicality indices are required. If **atiq** is *false* the array **ati** is not set.

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.

Constraint: **nvar** ≥ 1.

2: **ng** – INTEGER

Default: the dimension of the arrays **nig**, **det**, **prior** and the first dimension of the array **gmn**. (An error is raised if these dimensions are not equal.)

The number of groups, n_g.

Constraint: **ng** ≥ 2.

3: **nobs** – INTEGER

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

The number of observations in **x** which are to be allocated.

Constraint: **nobs** ≥ 1.

4: **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: **m** $\geq$ **nvar**.

5.3 Output Parameters

1: **prior**(**ng**) – REAL (KIND=nag_wp) array

If **priors** = 'P', the computed prior probabilities in proportion to group sizes for the n_g groups.

If **priors** = 'I', the input prior probabilities will be unchanged.

If **priors** = 'E', **prior** is not set.

2: **p**(ldp, **ng**) – REAL (KIND=nag_wp) array

p(k, j) contains the posterior probability p_{kj} for allocating the kth observation to the jth group, for $k = 1, 2, \ldots,$ **nobs** and $j = 1, 2, \ldots, n_g$.

3: **iag**(**nobs**) – INTEGER array

The groups to which the observations have been allocated.

4: **ati**(ldp, :) – REAL (KIND=nag_wp) array

The first dimension of the array **ati** will be **nobs**.

The second dimension of the array **ati** will be **ng** if **atiq** $= true$ and 1 otherwise.

If **atiq** is *true*, **ati**(k, j) will contain the predictive atypicality index for the kth observation with respect to the jth group, for $k = 1, 2, \ldots,$ **nobs** and $j = 1, 2, \ldots, n_g$.

If **atiq** is *false*, **ati** is not set.

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, **nvar** < 1,
or **ng** < 2,
or **nobs** < 1,
or **m** $<$ **nvar**,
or $ldgmn <$ **ng**,
or $ldx <$ **nobs**,
or $ldp <$ **nobs**,
or **typ** $\neq$ 'E' or '**p**',
or **equal** $\neq$ 'E' or 'U',
or **priors** $\neq$ 'E', 'I' or '**p**'.

ifail $= 2$

On entry, the number of variables indicated by **isx** is not equal to **nvar**,
or **equal** = 'E' and **nig**$(j) \leq 0$, for some j,
or **equal** = 'E' and $\sum_{j=1}^{n_g}$**nig**$(j) \leq$ **ng** + **nvar**,
or **equal** = 'U' and **nig**$(j) \leq$ **nvar** for some j.

ifail $= 3$

On entry, **priors** = 'I' and **prior**$(j) \leq 0.0$ for some j,
or **priors** = 'I' and $\sum_{j=1}^{n_g}$**prior**(j) is not within $10 \times$ ***machine precision*** of 1.

ifail $= 4$

On entry, **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 of the returned posterior probabilities will depend on the accuracy of the input R or R_j matrices. The atypicality index should be accurate to four significant places.

8 Further Comments

The distances D_{kj}^2 can be computed using nag_mv_discrim_mahal (g03db) if other forms of discrimination are required.

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 allocations made using the predictive approach and under the assumption that the within-group covariance matrices are not equal. The posterior probabilities of group membership, q_j, and the atypicality index are printed along with the allocated group. The atypicality index shows that observations 5 and 6 do not seem to be typical of the three types present in the initial 21 observations.

9.1 Program Text

```
     function g03dc_example

fprintf('g03dc 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);

% Data to group
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];

% Grouping parameters
typ    = 'P';
equal  = 'U';
```

```
priors = 'Equal priors';
prior  = zeros(3, 1);
atiq   = true;

[prior, p, iag, ati, ifail] = ...
  g03dc( ...
  typ, equal, priors, nig, gmean, gc, det, isx, x, prior, atiq);

fprintf('   Obs       Posterior        Allocated     Atypicality\n');
fprintf('             probabilities    to group      index\n');
for i=1:6
  fprintf('%6d     ', i);
  fprintf('%6.3f', p(i,:));
  fprintf('%6d     ', iag(i));
  fprintf('%6.3f', ati(i,:));
  fprintf('\n');
 end
```

9.2 Program Results

```
     g03dc example results

   Obs       Posterior        Allocated     Atypicality
             probabilities    to group      index
     1      0.094 0.905 0.002     2      0.596 0.254 0.975
     2      0.005 0.168 0.827     3      0.952 0.836 0.018
     3      0.019 0.920 0.062     2      0.954 0.797 0.912
     4      0.697 0.303 0.000     1      0.207 0.860 0.993
     5      0.317 0.013 0.670     3      0.991 1.000 0.984
     6      0.032 0.366 0.601     3      0.981 0.978 0.887
```