NAG Toolbox for MATLAB

g03dc

1 Purpose

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

2 Syntax

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

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, ${D_{kj}}^2$:

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

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, g03dc uses the upper triangular matrices R and R_j supplied by 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.

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.

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:

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

In the estimative approach, the parameters 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 parameters and a posterior distribution for the parameters, $p\left(\mu_j, \Sigma_j \mid X_t\right)$, is found. A predictive distribution is then obtained by integrating $p\left(j \mid x_k, \mu_j, \Sigma_j\right)p\left(\mu_j, \Sigma_j \mid X\right)$ over the parameter space. This predictive distribution then replaces $p\left(x_k \mid \mu_j, \Sigma_j\right)$ 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, $p\left(j \mid x_k, \mu_j, \Sigma_j\right)$, 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

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

(iv) Predictive with unequal variance-covariance matrices

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

where

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

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, g03dc computes an atypicality index, $I_j(x_k)$. 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:

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

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

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

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

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

with

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

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 – string**

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 – string**

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 – string**

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) – int32 array**

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

Constraints:

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

5: **gmn(ldgmn,nvar) – double array**

ldgmn, the first dimension of the array, must be at least **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 g03da.

6: **gc((ng + 1) × nvar × (nvar + 1)/2) – double 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 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) – double array**

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

8: **nobs – int32 scalar**

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

Constraint: $\mathbf{nobs} \geq 1$.

9: **isx(m) – int32 array**

$\mathbf{isx}(l)$ indicates if the lth variable in **x** is to be included in the distance calculations.

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

Constraint: $\mathbf{isx}(l) > 0$ for **nvar** values of l.

10: **x(ldx,m) – double array**

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

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

11: **prior(ng) – double array**

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

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

12: **atiq – logical scalar**

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 – int32 scalar**

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

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

Constraint: **nvar** ≥ 1.

2: **ng – int32 scalar**

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

the number of groups, n_g.

Constraint: **ng** ≥ 2.

3: **m – int32 scalar**

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

p, the number of variables in the data array **x**.

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

5.3 Input Parameters Omitted from the MATLAB Interface

ldgmn, ldx, ldp, wk

5.4 Output Parameters

1: **prior(ng) – double 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) – double 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) – int32 array**

The groups to which the observations have been allocated.

4: **ati(ldp,$*$) – double array**

The first dimension of the array **ati** must be at least **nobs**

The second dimension of the array must be at least **ng** if **atiq** = **true**, and at least 1 otherwise

If **atiq** is **true**, **ati**(k,j) will contain the 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 – 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, **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 $\sum_{j=1}^{n_g} \mathbf{nig}(j) \leq \mathbf{ng} + \mathbf{nvar}$,
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 $\sum_{j=1}^{n_g} \mathbf{prior}(j)$ 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.

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 g03db if other forms of discrimination are required.

9 Example

```
typ = 'P';
equal = 'U';
priors = 'Equal priors';
nig = [int32(6);
      int32(10);
      int32(5)];
gmean = [1.0433, -0.6034166666666667;
      2.00727, -0.20604;
      2.70974, 1.5998];
gc = [-0.5099642881287538;
      -0.279705472386133;
```

```
     -1.217327847040481;
     -0.3326727521153484;
     -0.3723518779712077;
     -1.987589395382754;
     -0.4603014906920608;
     -0.7041634974247672;
     0.4737334252803499;
     0.7451327720614629;
     -0.3251057349548681;
     -0.4275545007358186];
det = [-0.8273469064608421;
     -3.045968198109008;
     -2.287732741158105];
nobs = int32(6);
isx = [int32(1);
     int32(1)];
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];
prior = zeros(3, 1);
atiq = true;
[priorOut, p, iag, ati, ifail] = ...
     g03dc(typ, equal, priors, nig, gmean, gc, det, nobs, isx, x, prior,
atiq)
```

```
priorOut =
     0
     0
     0
p =
    0.0939    0.9046    0.0015
    0.0047    0.1682    0.8270
    0.0186    0.9196    0.0618
    0.6969    0.3026    0.0005
    0.3174    0.0130    0.6696
    0.0323    0.3664    0.6013
iag =
           2
           3
           2
           1
           3
           3
ati =
    0.5956    0.2539    0.9747
    0.9519    0.8360    0.0184
    0.9540    0.7966    0.9122
    0.2073    0.8599    0.9929
    0.9908    0.9999    0.9843
    0.9807    0.9779    0.8871
ifail =
           0
```