NAG Toolbox for MATLAB

g03da

1 Purpose

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] = g03da(weight, x, isx, nvar,
ing, ng, wt, 'n', n, 'm', m)
```

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, 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 $R_j z = \left(x - \bar{x}_j\right)$, $\bar{x}_j$ being the vector of means of the jth group. These distances can be calculated by g03db. The distances are used in discriminant analysis and g03dc uses the results of 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: **weight – string**

Indicates if weights are to be used.

weight = 'U'

No weights are used.

weight = 'W'

Weights are to be used and must be supplied in **wt**.

Constraint: **weight** = 'U' or 'W'.

2: **x(ldx,m) – double array**

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

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

3: **isx(m) – int32 array**

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

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

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

4: **nvar – int32 scalar**

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

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

5: **ing(n) – int32 array**

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

Constraint: $1 \leq \mathbf{ing}(k) \leq \mathbf{ng}$, for $k = 1,2,\ldots,n$.

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

6: **ng – int32 scalar**

the number of groups, n_g.

Constraint: $\mathbf{ng} \geq 2$.

7: **wt(*) – double array**

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

The effective number of observations for each group must be greater than 1.

5.2 Optional Input Parameters

1: **n – int32 scalar**

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

n, the number of observations.

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

2: **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: $\mathbf{m} \geq \mathbf{nvar}$.

5.3 Input Parameters Omitted from the MATLAB Interface

ldx, ldgmn, wk, iwk

5.4 Output Parameters

1: **nig**(**ng**) **– int32 array**

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

2: **gmn**(**ldgmn,nvar**) **– double 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**) **– double array**

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

4: **gc**((**ng** + **1**) × **nvar** × (**nvar** + **1**)/**2**) **– double 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 – double scalar**

The likelihood-ratio test statistic, G.

6: **df – double scalar**

The degrees of freedom for the distribution of G.

7: **sig – double scalar**

The significance level for G.

8: **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 **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.

7 Accuracy

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

8 Further Comments

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

9 Example

```
weight = 'U';
x = [1.1314, 2.4596;
     1.0986, 0.2624;
     0.6419, -2.3026;
     1.335, -3.2189;
     1.411, 0.0953;
     0.6419, -0.9163;
     2.1163, 0;
     1.335, -1.6094;
     1.361, -0.5108;
     2.0541, 0.1823;
     2.2083, -0.5108;
     2.7344, 1.2809;
     2.0412, 0.47;
     1.8718, -0.9163;
     1.7405, -0.9163;
     2.6101, 0.47;
     2.3224, 1.8563;
     2.2192, 2.0669;
     2.2618, 1.1314;
     3.9853, 0.9163;
```

```
     2.76, 2.0281];
isx = [int32(1);
     int32(1)];
nvar = int32(2);
ing = [int32(1);
     int32(1);
     int32(1);
     int32(1);
     int32(1);
     int32(1);
     int32(2);
     int32(2);
     int32(2);
     int32(2);
     int32(2);
     int32(2);
     int32(2);
     int32(2);
     int32(2);
     int32(2);
     int32(3);
     int32(3);
     int32(3);
     int32(3);
     int32(3)];
ng = int32(3);
wt = [0];
[nig, gmean, det, gc, stat, df, sig, ifail] = ...
    g03da(weight, x, isx, nvar, ing, ng, wt)
```

```
nig =
           6
          10
           5
gmean =
    1.0433   -0.6034
    2.0073   -0.2060
    2.7097    1.5998
det =
   -0.8273
   -3.0460
   -2.2877
gc =
   -0.5100
   -0.2797
   -1.2173
   -0.3327
   -0.3724
   -1.9876
   -0.4603
   -0.7042
    0.4737
    0.7451
   -0.3251
   -0.4276
stat =
   19.2410
df =
     6
sig =
    0.0038
ifail =
           0
```