NAG Toolbox for MATLAB

g03db

1 Purpose

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

2 Syntax

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

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 data set. A measure of the distance of the point from the jth population or group is given by the Mahalanobis distance, ${D_{kj}}^2$:

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

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, g03db 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 $R_j z = \left(x_k - \bar{x}_j\right)$ 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 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 – 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_j, for $j = 1, 2, \dots, n_g$, stored in the remainder of **gc** are used.

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

2: **mode – string**

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: **ng – int32 scalar**

the number of groups, n_g.

Constraint: **ng** ≥ 2.

4: **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 selected variables for the jth group, for $j = 1, 2, \dots, n_g$. These are returned by g03da.

5: **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 $(n_g+1)p(p+1)/2$ 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**.

6: **nobs – int32 scalar**

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'.

7: **isx(*) – int32 array**

Note: the dimension of the array **isx** must be at least max(1, **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 $l = 1, 2, \dots,$ **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.

8: **x(ldx,$*$) – double array**

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

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

The second dimension of the array must be at least max$(1,$ **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 – int32 scalar**

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

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

Constraint: **nvar** ≥ 1.

2: **m – int32 scalar**

Default: The dimension of the array **isx** The second dimension of the array **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 Input Parameters Omitted from the MATLAB Interface

ldgmn, ldx, ldd, wk

5.4 Output Parameters

1: **d(ldd,ng) – double 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 $j = i,i+1,\ldots,n_g$.

2: **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 **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** ≠ 'E' or 'U',
or **mode** ≠ '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.

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 g03dc.

9 Example

```
equal = 'U';
mode = 'Sample points';
ng = int32(3);
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];
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];
[d, ifail] = g03db(equal, mode, ng, gmean, gc, nobs, isx, x)
```

```
d =
    3.3393    0.7521   50.9283
   20.7771    5.6559    0.0597
   21.3631    4.8411   19.4978
    0.7184    6.2803  124.7323
   55.0003   88.8604   71.7852
   36.1703   15.7849   15.7489
ifail =
           0
```