nag_mv_discrim (g03dac)

1. Purpose

nag_mv_discrim (g03dac) computes a test statistic for the equality of within-group covariance matrices and also computes matrices for use in discriminant analysis.

2. Specification

```
#include <nag.h>
#include <nagg03.h>

void nag_mv_discrim(Integer n, Integer m,  double x[],
           Integer tdx, Integer isx[], Integer nvar, Integer ing[],
           Integer ng, double wt[],  Integer nig[], double gmean[],
           Integer tdg, double det[], double gc[], double *stat,
           double *df, double *sig, NagError *fail)
```

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, $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_{n_g} = \Sigma$, the following likelihood-ratio test statistic, G, can be used;

$$G = C\left\{(n-n_g)\log|S| - \sum_{j=1}^{n_g}(n_j-1)\log|S_j|\right\},$$

where

$$C = 1 - \frac{2p^2+3p-1}{6(p+1)(n_g-1)}\left(\sum_{j=1}^{n_g}\frac{1}{(n_j-1)} - \frac{1}{(n-n_g)}\right),$$

and S_j are the within-group variance-covariance matrices and S is the pooled variance-covariance matrix given by

$$S = \frac{\sum_{j=1}^{n_g}(n_j-1)S_j}{(n-n_g)}.$$

For large n, G is approximately distributed as a χ^2 variable with $\frac{1}{2}p(p+1)(n_g-1)$ 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, nag_mv_discrim 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 $S_j = R_j^T R_j$. The pooled R matrix is then computed from the R_j matrices. The values of $|S|$ and the $|S_j|$ 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 $z^T z$, where $R_j z = (x - \bar{x}_j)$, $\bar{x}_j$ being the vector of means of the jth group. These distances can be calculated by nag_mv_discrim_mahaldist (g03dbc). The distances are used in discriminant analysis and nag_mv_discrim_group (g03dcc) uses the results of nag_mv_discrim 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. Parameters

n

Input: the number of observations, n.

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

m

Input: the number of variables in the data array **x**.

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

x[n][tdx]

Input: $\mathbf{x}[k-1][l-1]$ must contain the kth observation for the lth variable, for $k = 1, 2, \ldots, n$; $l = 1, 2, \ldots,$ **m** .

tdx

Input: the last dimension of the array **x** as declared in the calling program.
Constraint: $\mathbf{tdx} \geq \mathbf{m}$.

isx[m]

Input: $\mathbf{isx}[l-1]$ indicates whether or not the lth variable in **x** is to be included in the variance-covariance matrices.

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

Constraint: $\mathbf{isx}[l-1] > 0$ for **nvar** values of l.

nvar

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

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

ing[n]

Input: $\mathbf{ing}[k-1]$ indicates to which group the kth observation belongs, for $k = 1, 2, \ldots, n$.

Constraint: $1 \leq \mathbf{ing}[k-1] \leq \mathbf{ng}$ for $k = 1, 2, \ldots, n$ and the values of **ing** must be such that each group has at least **nvar** members.

ng

Input: the number of groups, n_g.

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

wt[n]

Input: the 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 $\mathbf{wt}[k-1] = 0.0$ then the kth observation is excluded from the calculations.

Constraint: $\mathbf{wt}[k-1] \geq 0.0$ for $k = 1, 2, \ldots, n$ and the effective number of observations for each group must be greater than 1.

Note: If **wt** is set to the null pointer **NULL**, i.e., (double *)0, then **wt** is not referenced and the effective number of observations for a group is the number of observations in that group.

nig[ng]

Output: $\mathbf{nig}[j-1]$ contains the number of observations in the jth group, for $j = 1, 2, \ldots, n_g$.

gmean[ng][tdg]

Output: the jth row of **gmean** contains the means of the p selected variables for the jth group, for $j = 1, 2, \ldots, n_g$.

tdg

Input: the last dimension of the array **gmean** as declared in the calling program.
Constraint: $\mathbf{tdg} \geq \mathbf{nvar}$.

det[ng]

Output: the logarithm of the determinants of the within-group variance-covariance matrices.

gc[(ng+1)∗nvar∗(nvar+1)/2]

Output: 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.

stat

Output: the likelihood-ratio test static, G.

df

Output: the degrees of freedom for the distribution of G.

sig

Output: the significance level for G.

fail

The NAG error parameter, see the Essential Introduction to the NAG C Library.

5. Error Indications and Warnings

NE_INT_ARG_LT

On entry, **n** must not be less than 1: **n** = ⟨*value*⟩.
On entry, **nvar** must not be less than 1: **nvar** = ⟨*value*⟩.
On entry, **ng** must not be less than 2: **ng** = ⟨*value*⟩.

NE_2_INT_ARG_LT

On entry, **m** = ⟨*value*⟩ while **nvar** = ⟨*value*⟩.
These parameters must satisfy **m** ≥ **nvar**.
On entry, **tdx** = ⟨*value*⟩ while **m** = ⟨*value*⟩.
These parameters must satisfy **tdx** ≥ **m**.
On entry, **tdg** = ⟨*value*⟩ while **nvar** = ⟨*value*⟩.
These parameters must satisfy **tdg** ≥ **nvar**.

NE_INTARR_INT

On entry, **ing**[⟨*value*⟩] = ⟨*value*⟩, **ng** = ⟨*value*⟩.
Constraint: $1 \leq$ **ing**$[i-1] \leq$ **ng**, $i = 1, 2, \ldots, n$.

NE_NEG_WEIGHT_ELEMENT

On entry, **wt**[⟨*value*⟩] = ⟨*value*⟩.
Constraint: when referenced, all elements of **wt** must be non-negative.

NE_VAR_INCL_INDICATED

The number of variables, **nvar** in the analysis = ⟨*value*⟩, while number of variables included in the analysis via array **isx** = ⟨*value*⟩.
Constraint: these two numbers must be the same.

NE_GROUP_OBSERV

On entry, group ⟨*value*⟩ has ⟨*value*⟩ effective observations.
Constraint: in each group the effective number of observations must be ≥ 1.

NE_GROUP_VAR

On entry, group ⟨*value*⟩ has ⟨*value*⟩ members, while **nvar** = ⟨*value*⟩.
Constraint: number of members in each group ≥ **nvar**

NE_GROUP_VAR_RANK

The variables in group ⟨*value*⟩ are not of full rank.

NE_VAR_RANK

The variables are not of full rank.

NE_ALLOC_FAIL

Memory allocation failed.

NE_INTERNAL_ERROR

An internal error has occurred in this function.
Check the function call and any array sizes. If the call is correct then please consult NAG for assistance.

6. Further Comments

The time will be approximately proportional to np^2.

6.1. Accuracy

The accuracy is dependent on the accuracy of the computation of the QR decomposition.

6.2. 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)* Griffin (3rd Edition).

Krzanowski W J (1990) *Principles of Multivariate Analysis* Oxford University Press.

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

7. See Also

nag_mv_discrim_mahaldist (g03dbc)
nag_mv_discrim_group (g03dcc)

8. 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 statistics computed by nag_mv_discrim (g03dac). The printed results show that there is evidence that the within-group variance-covariance matrices are not equal.

8.1. Program Text

```
/* nag_mv_discrim (g03dac) Example Program.
 *
 * Copyright 1998 Numerical Algorithms Group.
 *
 * Mark 5, 1998.
 *
 */
#include <nag.h>
#include <stdio.h>
#include <nag_stdlib.h>
#include <nagg03.h>

#define NMAX 21
#define MMAX 2
#define GPMAX 3

main()
{
  double stat;
  double det[GPMAX], gc[(GPMAX+1)*MMAX*(MMAX+1)/2],
  gmean[GPMAX][MMAX], wt[NMAX],
  x[NMAX][MMAX];
  double df;
  double *wtptr = 0;
  double sig;

  Integer nvar;
  Integer i, j, m, n;
  Integer ing[NMAX], isx[MMAX], nig[GPMAX];
  Integer ng;
  Integer tdgmean=MMAX, tdx= MMAX;

  char weight[2];

  Vprintf("g03dac Example Program Results\n\n");

  /*  Skip headings in data file */
  Vscanf("%*[^\n]");
  Vscanf("%ld",&n);
  Vscanf("%ld",&m);
  Vscanf("%ld",&nvar);
  Vscanf("%ld",&ng);
  Vscanf("%s",weight);
  if (n <= NMAX && m <= MMAX)
    {
      if (*weight == 'W')
        {
          for (i = 0; i < n; ++i)
            {
              for (j = 0; j < m; ++j)
                Vscanf("%lf",&x[i][j]);
              Vscanf("%ld",&ing[i]);
              Vscanf("%lf",&wt[i]);
            }
          wtptr = wt;
        }
      else
        {
          for (i = 0; i < n; ++i)
            {
              for (j = 0; j < m; ++j)
                Vscanf("%lf",&x[i][j]);
              Vscanf("%ld",&ing[i]);
            }
        }
      for (j = 0; j < m; ++j)
        Vscanf("%ld",&isx[j]);

      g03dac(n, m, (double *)x, tdx, isx, nvar, ing, ng, wtptr, nig,
             (double *)gmean, tdgmean, det, gc, &stat, &df, &sig, NAGERR_DEFAULT);

      Vprintf("\nGroup means\n\n");
      for (i = 0; i < ng; ++i)
        {
          for (j = 0; j < nvar; ++j)
            Vprintf("%10.4f",gmean[i][j]);
          Vprintf("\n");
        }
      Vprintf("\nLOG of determinants\n\n");
      for (j = 0; j < ng; ++j)
        Vprintf("%10.4f",det[j]);
      Vprintf("\n\n%s%7.4f\n","stat = ",stat);
      Vprintf("%s%7.4f\n","  df = ",df);
      Vprintf("%s%7.4f\n"," sig = ",sig);
      exit(EXIT_SUCCESS);
    }
  else
    {
      Vprintf("Incorrect input value of n or m.\n");
      exit(EXIT_FAILURE);
    }
}
```

8.2. Program Data

```
g03dac Example Program Data
  21 2 2 3 U
  1.1314    2.4596   1
  1.0986    0.2624   1
  0.6419   -2.3026   1
  1.3350   -3.2189   1
  1.4110    0.0953   1
  0.6419   -0.9163   1
  2.1163    0.0000   2
  1.3350   -1.6094   2
  1.3610   -0.5108   2
  2.0541    0.1823   2
  2.2083   -0.5108   2
  2.7344    1.2809   2
  2.0412    0.4700   2
  1.8718   -0.9163   2
  1.7405   -0.9163   2
  2.6101    0.4700   2
  2.3224    1.8563   3
  2.2192    2.0669   3
  2.2618    1.1314   3
  3.9853    0.9163   3
  2.7600    2.0281   3
  1     1
```

8.3. Program Results

```
g03dac Example Program Results


Group means

    1.0433   -0.6034
    2.0073   -0.2060
    2.7097    1.5998

LOG of determinants

   -0.8273   -3.0460   -2.2877

stat = 19.2410
  df =  6.0000
 sig =  0.0038
```