NAG Toolbox for MATLAB

g02hb

1 Purpose

g02hb finds, for a real matrix X of full column rank, a lower triangular matrix A such that $\left(A^{\mathrm{T}}A\right)^{-1}$ is proportional to a robust estimate of the covariance of the variables. g02hb is intended for the calculation of weights of bounded influence regression using g02hd.

2 Syntax

```
[a, z, nit, ifail] = g02hb(ucv, n, x, a, tol, maxit, nitmon, 'm', m, 'bl', bl, 'bd', bd)
```

3 Description

In fitting the linear regression model

$$y = X\theta + \epsilon,$$

where y is a vector of length n of the dependent variable,

X is an n by m matrix of independent variables,

θ is a vector of length m of unknown parameters,

and ϵ is a vector of length n of unknown errors,

it may be desirable to bound the influence of rows of the X matrix. This can be achieved by calculating a weight for each observation. Several schemes for calculating weights have been proposed (see Hampel *et al.* 1986 and Marazzi 1987a). As the different independent variables may be measured on different scales one group of proposed weights aims to bound a standardized measure of influence. To obtain such weights the matrix A has to be found such that

$$\frac{1}{n}\sum_{i=1}^{n} u\left(\|z_i\|_2\right) z_i z_i^{\mathrm{T}} = I, \qquad (I \text{ is the identity matrix})$$

and

$$z_i = Ax_i,$$

where x_i is a vector of length m containing the elements of the ith row of X,

A is an m by m lower triangular matrix,

z_i is a vector of length m,

and u is a suitable function.

The weights for use with g02hd may then be computed using

$$w_i = f\left(\|z_i\|_2\right)$$

for a suitable user function f.

g02hb finds A using the iterative procedure

$$A_k = (S_k + I)A_{k-1},$$

where $S_k = \left(s_{jl}\right)$, for j and $l = 1, 2, \ldots, m$ is a lower triangular matrix such that

$$s_{jl} = \begin{cases} -\min\left[\max\left(h_{jl}/n, -BL\right), BL\right], & j > l \\ -\min\left[\max\left(\frac{1}{2}\left(h_{jj}/n - 1\right), -BD\right), BD\right], & j = l \end{cases}$$

$$h_{jl} = \sum_{i=1}^{n} u\left(\|z_i\|_2\right) z_{ij} z_{il}$$

and BD and BL are suitable bounds.

In addition the values of $\|z_i\|_2$, for $i = 1, 2, \ldots, n$, are calculated.

g02hb is based on routines in ROBETH; see Marazzi 1987a.

4 References

Hampel F R, Ronchetti E M, Rousseeuw P J and Stahel W A 1986 *Robust Statistics. The Approach Based on Influence Functions* Wiley

Huber P J 1981 *Robust Statistics* Wiley

Marazzi A 1987a Weights for bounded influence regression in ROBETH *Cah. Rech. Doc. IUMSP, No. 3 ROB 3* Institut Universitaire de Médecine Sociale et Préventive, Lausanne

5 Parameters

5.1 Compulsory Input Parameters

1: **ucv – string containing name of m-file**

ucv must return the value of the function u for a given value of its argument. The value of u must be nonnegative.

Its specification is:

```
[result] = ucv(t)
```

Input Parameters

1: **t – double scalar**

The argument for which **ucv** must be evaluated.

Output Parameters

1: **result – double scalar**

The result of the function.

2: **n – int32 scalar**

n, the number of observations.

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

3: **x(ldx,m) – double array**

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

The real matrix X, i.e., the independent variables. $\mathbf{x}(i,j)$ must contain the ijth element of **x**, for $i = 1, 2, \ldots n$, $j = 1, 2, \ldots, m$.

4: **a(m × (m + 1)/2) – double array**

An initial estimate of the lower triangular real matrix A. Only the lower triangular elements must be given and these should be stored row-wise in the array.

The diagonal elements must be $\neq 0$, although in practice will usually be > 0. If the magnitudes of the columns of X are of the same order the identity matrix will often provide a suitable initial value for A. If the columns of X are of different magnitudes, the diagonal elements of the initial value of A should be approximately inversely proportional to the magnitude of the columns of X.

5: **tol – double scalar**

The relative precision for the final value of A. Iteration will stop when the maximum value of $|s_{jl}|$ is less than **tol**.

Constraint: **tol** > 0.0.

6: **maxit – int32 scalar**

The maximum number of iterations that will be used during the calculation of A.

A value of **maxit** $= 50$ will often be adequate.

Constraint: **maxit** > 0.

7: **nitmon – int32 scalar**

Determines the amount of information that is printed on each iteration.

nitmon > 0

The value of A and the maximum value of $|s_{jl}|$ will be printed at the first and every **nitmon** iterations.

nitmon ≤ 0

No iteration monitoring is printed.

When printing occurs the output is directed to the current advisory message unit (see x04ab).

5.2 Optional Input Parameters

1: **m – int32 scalar**

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

m, the number of independent variables.

Constraint: $1 \leq \mathbf{m} \leq \mathbf{n}$.

2: **bl – double scalar**

The magnitude of the bound for the off-diagonal elements of S_k.

Suggested value: **bl** $= 0.9$.

Default: 0.9

Constraint: **bl** > 0.

3: **bd – double scalar**

The magnitude of the bound for the diagonal elements of S_k.

Suggested value: **bd** $= 0.9$.

Default: 0.9

Constraint: **bd** > 0.

5.3 Input Parameters Omitted from the MATLAB Interface

ldx, wk

5.4 Output Parameters

1: **a**$(\mathbf{m} \times (\mathbf{m}+\mathbf{1})/\mathbf{2})$ **– double array**

The lower triangular elements of the matrix A, stored row-wise.

2: **z(n) – double array**

The value $\|z_i\|_2$, for $i = 1, 2, \ldots, n$.

3: **nit – int32 scalar**

The number of iterations performed.

4: **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, $\mathbf{n} \leq 1$,
or $\mathbf{m} < 1$,
or $\mathbf{n} < \mathbf{m}$,
or $\mathbf{ldx} < \mathbf{n}$.

ifail $= 2$

On entry, $\mathbf{tol} \leq 0$,
or $\mathbf{maxit} \leq 0$,
or diagonal element of $\mathbf{a} = 0$,
or $\mathbf{bl} \leq 0$,
or $\mathbf{bd} \leq 0$.

ifail $= 3$

Value returned by $\mathbf{ucv} < 0$.

ifail $= 4$

The function has failed to converge in **maxit** iterations.

7 Accuracy

On successful exit the accuracy of the results is related to the value of **tol**; see Section 5.

8 Further Comments

The existence of A will depend upon the function u; (see Hampel *et al.* 1986 and Marazzi 1987a), also if X is not of full rank a value of A will not be found. If the columns of X are almost linearly related then convergence will be slow.

9 Example

g02hb_ucv.m

```
function [result] = ucv(t)
  ucvc = 2.5;
  result = 1;
  if (t ~= 0.0)
     q = ucvc/t;
     q2 = q*q;
     [pc, ifail] = s15ab(q);
     l = x02ak();
```

```
      if (q2 < -log(l))
         pd = exp(-q2/2.0)/sqrt(pi*2.0);
      else
         pd = 0.0;
      end
      result = (2.0*pc-1.0)*(1.0-q2) + q2 - 2.0*q*pd;
   end
```

```
n = int32(5);
x = [1, -1, -1;
     1, -1, 1;
     1, 1, -1;
     1, 1, 1;
     1, 0, 3];
a = [1;
     0;
     1;
     0;
     0;
     1];
tol = 5e-05;
maxit = int32(50);
nitmon = int32(0);
[aOut, z, nit, ifail] = g02hb('g02hb_ucv', n, x, a, tol, maxit, nitmon)
```

```
aOut =
    1.3208
    0.0000
    1.4518
   -0.5753
    0.0000
    0.9340
z =
    2.4760
    1.9953
    2.4760
    1.9953
    2.5890
nit =
          16
ifail =
           0
```