

NAG Library Routine Document

G02BSF

Note: before using this routine, please read the Users' Note for your implementation to check the interpretation of ***bold italicised*** terms and other implementation-dependent details.

1 Purpose

G02BSF computes Kendall and/or Spearman nonparametric rank correlation coefficients for a set of data omitting cases with missing values from only those calculations involving the variables for which the values are missing; the data array is preserved, and the ranks of the observations are not available on exit from the routine.

2 Specification

```

SUBROUTINE G02BSF(N, M, X, LDX, MISS, XMISS, ITYPE, RR, LDRR, NCASES,
1          CNT, LDCNT, KWORKA, KWORKB, KWORKC, KWORKD, WORK1,
2          WORK2, IFAIL)

  INTEGER          N, M, LDX, MISS(M), ITYPE, LDRR, NCASES, LDCNT,
1          KWORKA(N), KWORKB(N), KWORKC(N), KWORKD(N), IFAIL
  double precision X(LDX,M), XMISS(M), RR(LDRR,M), CNT(LDCNT,M),
1          WORK1(N), WORK2(N)

```

3 Description

The input data consists of n observations for each of m variables, given as an array

$$[x_{ij}], \quad i = 1, 2, \dots, n \ (n \geq 2) \text{ and } j = 1, 2, \dots, m \ (m \geq 2),$$

where x_{ij} is the i th observation on the j th variable. In addition each of the m variables may optionally have associated with it a value which is to be considered as representing a missing observation for that variable; the missing value for the j th variable is denoted by xm_j . Missing values need not be specified for all variables.

Let $w_{ij} = 0$ if the i th observation for the j th variable is a missing value, i.e., if a missing value, xm_j , has been declared for the j th variable, and $x_{ij} = xm_j$ (see also Section 7); and $w_{ij} = 1$ otherwise, for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, m$.

The observations are first ranked, a pair of variables at a time as follows:

For a given pair of variables, j and l say, each of the observations x_{ij} for which the product $w_{ij}w_{il} = 1$, for $i = 1, 2, \dots, n$, has associated with it an additional number, the 'rank' of the observation, which indicates the magnitude of that observation relative to the magnitude of the other observations on variable j for which $w_{ij}w_{il} = 1$.

The smallest of these valid observations for variable j is assigned to rank 1, the second smallest valid observation for variable j the rank 2, the third smallest rank 3, and so on until the largest such observation is given the rank n_{jl} , where

$$n_{jl} = \sum_{i=1}^n w_{ij}w_{il}.$$

If a number of cases all have the same value for the variable j , then they are each given an 'average' rank, e.g., if in attempting to assign the rank $h + 1$, k observations for which $w_{ij}w_{il} = 1$ were found to have the same value, then instead of giving them the ranks

$$h + 1, h + 2, \dots, h + k,$$

all k observations would be assigned the rank

$$\frac{2h + k + 1}{2}$$

and the next value in ascending order would be assigned the rank

$$h + k + 1.$$

The variable l is then ranked in a similar way. The process is then repeated for all pairs of variables j and l , for $j = 1, 2, \dots, m$ and $l = j, \dots, m$. Let $y_{ij(l)}$ be the rank assigned to the observation x_{ij} when the j th and l th variables are being ranked, and $y_{il(j)}$ be the rank assigned to the observation x_{il} during the same process, for $i = 1, 2, \dots, n$, $j = 1, 2, \dots, m$ and $l = j, j + 1, \dots, m$.

The quantities calculated are:

(a) Kendall's tau rank correlation coefficients:

$$R_{jk} = \frac{\sum_{h=1}^n \sum_{i=1}^n w_{hj} w_{hk} w_{ij} w_{ik} \text{sign}(y_{hj(k)} - y_{ij(k)}) \text{sign}(y_{hk(j)} - y_{ik(j)})}{\sqrt{[n_{jk}(n_{jk} - 1) - T_{j(k)}][n_{jk}(n_{jk} - 1) - T_{k(j)}]}}, \quad j, k = 1, 2, \dots, m,$$

$$\text{where } n_{jk} = \sum_{i=1}^n w_{ij} w_{ik}$$

and $\text{sign } u = 1$ if $u > 0$

$\text{sign } u = 0$ if $u = 0$

$\text{sign } u = -1$ if $u < 0$

and $T_{j(k)} = \sum t_j(t_j - 1)$ where t_j is the number of ties of a particular value of variable j when the j th and k th variables are being ranked, and the summation is over all tied values of variable j .

(b) Spearman's rank correlation coefficients:

$$R_{jk}^* = \frac{n_{jk}(n_{jk}^2 - 1) - 6 \sum_{i=1}^n w_{ij} w_{ik} (y_{ij(k)} - y_{ik(j)})^2 - \frac{1}{2}(T_{j(k)}^* + T_{k(j)}^*)}{\sqrt{[n_{jk}(n_{jk}^2 - 1) - T_{j(k)}^*][n_{jk}(n_{jk}^2 - 1) - T_{k(j)}^*]}}, \quad j, k = 1, 2, \dots, m,$$

$$\text{where } n_{jk} = \sum_{i=1}^n w_{ij} w_{ik}$$

and $T_{j(k)}^* = \sum t_j(t_j^2 - 1)$, where t_j is the number of ties of a particular value of variable j when the j th and k th variables are being ranked, and the summation is over all tied values of variable j .

4 References

Siegel S (1956) *Non-parametric Statistics for the Behavioral Sciences* McGraw-Hill

5 Parameters

- 1: N – INTEGER Input
On entry: n , the number of observations or cases.
Constraint: $N \geq 2$.
- 2: M – INTEGER Input
On entry: m , the number of variables.
Constraint: $M \geq 2$.

- 3: X(LDX,M) – **double precision** array *Input*
On entry: X(*i*,*j*) must be set to x_{ij} , the value of the *i*th observation on the *j*th variable, for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, m$.
- 4: LDX – INTEGER *Input*
On entry: the first dimension of the array X as declared in the (sub)program from which G02BSF is called.
Constraint: $LDX \geq N$.
- 5: MISS(M) – INTEGER array *Input*
On entry: MISS(*j*) must be set equal to 1 if a missing value, xm_j , is to be specified for the *j*th variable in the array X, or set equal to 0 otherwise. Values of MISS must be given for all *m* variables in the array X.
- 6: XMISS(M) – **double precision** array *Input*
On entry: XMISS(*j*) must be set to the missing value, xm_j , to be associated with the *j*th variable in the array X, for those variables for which missing values are specified by means of the array MISS (see Section 7).
- 7: ITYPE – INTEGER *Input*
On entry: the type of correlation coefficients which are to be calculated.
 ITYPE = -1
 Only Kendall's tau coefficients are calculated.
 ITYPE = 0
 Both Kendall's tau and Spearman's coefficients are calculated.
 ITYPE = 1
 Only Spearman's coefficients are calculated.
Constraint: ITYPE = -1, 0 or 1.
- 8: RR(LDRR,M) – **double precision** array *Output*
On exit: the requested correlation coefficients.
 If only Kendall's tau coefficients are requested (ITYPE = -1), RR(*j*,*k*) contains Kendall's tau for the *j*th and *k*th variables.
 If only Spearman's coefficients are requested (ITYPE = 1), RR(*j*,*k*) contains Spearman's rank correlation coefficient for the *j*th and *k*th variables.
 If both Kendall's tau and Spearman's coefficients are requested (ITYPE = 0), the upper triangle of RR contains the Spearman coefficients and the lower triangle the Kendall coefficients. That is, for the *j*th and *k*th variables, where *j* is less than *k*, RR(*j*,*k*) contains the Spearman rank correlation coefficient, and RR(*k*,*j*) contains Kendall's tau, for $j, k = 1, 2, \dots, m$.
 (Diagonal terms, RR(*j*,*j*), are unity for all three values of ITYPE.)
- 9: LDRR – INTEGER *Input*
On entry: the first dimension of the array RR as declared in the (sub)program from which G02BSF is called.
Constraint: $LDRR \geq M$.
- 10: NCASES – INTEGER *Output*
On exit: the minimum number of cases used in the calculation of any of the correlation coefficients (when cases involving missing values have been eliminated).

- 11: CNT(LDCNT,M) – *double precision* array *Output*
On exit: the number of cases, n_{jk} , actually used in the calculation of the rank correlation coefficient for the j th and k th variables, for $j, k = 1, 2, \dots, m$.
- 12: LDCNT – INTEGER *Input*
On entry: the first dimension of the array CNT as declared in the (sub)program from which G02BSF is called.
Constraint: LDCNT $\geq$ M.
- 13: KWORKA(N) – INTEGER array *Workspace*
 14: KWORKB(N) – INTEGER array *Workspace*
 15: KWORKC(N) – INTEGER array *Workspace*
 16: KWORKD(N) – INTEGER array *Workspace*
 17: WORK1(N) – *double precision* array *Workspace*
 18: WORK2(N) – *double precision* array *Workspace*
- 19: IFAIL – INTEGER *Input/Output*
On entry: IFAIL must be set to 0, -1 or 1. If you are unfamiliar with this parameter you should refer to Section 3.3 in the Essential Introduction for details.
On exit: IFAIL = 0 unless the routine detects an error (see Section 6).
 For environments where it might be inappropriate to halt program execution when an error is detected, the value -1 or 1 is recommended. If the output of error messages is undesirable, then the value 1 is recommended. Otherwise, because for this routine the values of the output parameters may be useful even if IFAIL $\neq$ 0 on exit, the recommended value is -1. **When the value -1 or 1 is used it is essential to test the value of IFAIL on exit.**

6 Error Indicators and Warnings

If on entry IFAIL = 0 or -1, explanatory error messages are output on the current error message unit (as defined by X04AAF).

Note: G02BSF may return useful information for one or more of the following detected errors or warnings.

Errors or warnings detected by the routine:

IFAIL = 1

On entry, $N < 2$.

IFAIL = 2

On entry, $M < 2$.

IFAIL = 3

On entry, LDX $<$ N,
 or LDRR $<$ M,
 or LDCNT $<$ M.

IFAIL = 4

On entry, ITYPE $<$ -1,
 or ITYPE $>$ 1.

IFAIL = 5

After observations with missing values were omitted, fewer than two cases remained for at least one pair of variables. (The pairs of variables involved can be determined by examination of the contents

of the array CNT.) All correlation coefficients based on two or more cases are returned by the routine even if IFAIL = 5.

7 Accuracy

You are warned of the need to exercise extreme care in your selection of missing values. G02BSF treats all values in the inclusive range $\left(1 \pm 0.1^{(X02BEF-2)}\right) \times x_{mj}$, where x_{mj} is the missing value for variable j specified by you as missing values for variable j .

You must therefore ensure that the missing value chosen for each variable is sufficiently different from all valid values for that variable so that none of the valid values fall within the range indicated above.

8 Further Comments

The time taken by G02BSF depends on n and m , and the occurrence of missing values.

9 Example

This example reads in a set of data consisting of nine observations on each of three variables. Missing values of 0.99, 9.0 and 0.0 are declared for the first, second and third variables respectively. The program then calculates and prints both Kendall's tau and Spearman's rank correlation coefficients for all three variables, omitting cases with missing values from only those calculations involving the variables for which the values are missing. The program therefore eliminates cases 4, 5, 7 and 9 in calculating and correlation between the first and second variables, cases 5, 8 and 9 for the first and third variables, and cases 4, 7 and 8 for the second and third variables.

9.1 Program Text

```
*      G02BSF Example Program Text
*      Mark 14 Revised. NAG Copyright 1989.
*      .. Parameters ..
      INTEGER          M, N, LDX, LDRR, LDCNT
      PARAMETER        (M=3,N=9,LDX=N,LDRR=M,LDCNT=M)
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
*      .. Local Scalars ..
      INTEGER          I, IFAIL, ITYPE, J, NCASES
*      .. Local Arrays ..
      DOUBLE PRECISION CNT(LDCNT,M), RR(LDRR,M), WORK1(N), WORK2(N),
+      X(LDX,M), XMISS(M)
      INTEGER          KWORKA(N), KWORKB(N), KWORKC(N), KWORKD(N),
+      MISS(M)
*      .. External Subroutines ..
      EXTERNAL         G02BSF
*      .. Executable Statements ..
      WRITE (NOUT,*) 'G02BSF Example Program Results'
*      Skip heading in data file
      READ (NIN,*)
      READ (NIN,*) ((X(I,J),J=1,M),I=1,N)
      WRITE (NOUT,*)
      WRITE (NOUT,99999) 'Number of variables (columns) =', M
      WRITE (NOUT,99999) 'Number of cases      (rows)    =', N
      WRITE (NOUT,*)
      WRITE (NOUT,*) 'Data matrix is:-'
      WRITE (NOUT,*)
      WRITE (NOUT,99998) (J,J=1,M)
      WRITE (NOUT,99997) (I,(X(I,J),J=1,M),I=1,N)
      WRITE (NOUT,*)
*
*      Set up missing values before calling routine
*
      MISS(1) = 1
      MISS(2) = 1
      MISS(3) = 1
```

```

XMISS(1) = 0.99D0
XMISS(2) = 9.0D0
XMISS(3) = 0.00D0
ITYPE = 0
IFAIL = 1
*
CALL G02BSF(N,M,X,LDX,MISS,XMISS,ITYPE,RR,LDRR,NCASES,CNT,LDCNT,
+          KWORKA,KWORKB,KWORKC,KWORKD,WORK1,WORK2,IFAIL)
*
IF (IFAIL.LT.0) THEN
  WRITE (NOUT,*)
  WRITE (NOUT,99999) ' ** G02BSF returned with IFAIL = ', IFAIL
ELSE
  IF (IFAIL.NE.0) THEN
    WRITE (NOUT,99999) 'Routine fails, IFAIL =', IFAIL
  ELSE
    WRITE (NOUT,*) 'Matrix of rank correlation coefficients:'
    WRITE (NOUT,*) 'Upper triangle -- Spearman''s'
    WRITE (NOUT,*) 'Lower triangle -- Kendall''s tau'
    WRITE (NOUT,*)
    WRITE (NOUT,99998) (I,I=1,M)
    WRITE (NOUT,99997) (I,(RR(I,J),J=1,M),I=1,M)
    WRITE (NOUT,*)
    WRITE (NOUT,99999)
+      'Minimum number of cases used for any pair of variables:',
+      NCASES
    WRITE (NOUT,*)
    WRITE (NOUT,*) 'Numbers used for each pair are:'
    WRITE (NOUT,99998) (I,I=1,M)
    WRITE (NOUT,99997) (I,(CNT(I,J),J=1,M),I=1,M)
  END IF
END IF
*
99999 FORMAT (1X,A,I5)
99998 FORMAT (1X,3I12)
99997 FORMAT (1X,I3,3F12.4)
END

```

9.2 Program Data

G02BSF Example Program Data

1.70	1.00	0.50
2.80	4.00	3.00
0.60	6.00	2.50
1.80	9.00	6.00
0.99	4.00	2.50
1.40	2.00	5.50
1.80	9.00	7.50
2.50	7.00	0.00
0.99	5.00	3.00

9.3 Program Results

G02BSF Example Program Results

```

Number of variables (columns) = 3
Number of cases      (rows)   = 9

```

Data matrix is:-

	1	2	3
1	1.7000	1.0000	0.5000
2	2.8000	4.0000	3.0000
3	0.6000	6.0000	2.5000
4	1.8000	9.0000	6.0000
5	0.9900	4.0000	2.5000
6	1.4000	2.0000	5.5000
7	1.8000	9.0000	7.5000
8	2.5000	7.0000	0.0000
9	0.9900	5.0000	3.0000

Matrix of rank correlation coefficients:

Upper triangle -- Spearman's

Lower triangle -- Kendall's tau

	1	2	3
1	1.0000	0.1000	0.4058
2	0.0000	1.0000	0.0896
3	0.2760	0.0000	1.0000

Minimum number of cases used for any pair of variables: 5

Numbers used for each pair are:

	1	2	3
1	7.0000	5.0000	6.0000
2	5.0000	7.0000	6.0000
3	6.0000	6.0000	8.0000
