NAG Library Routine Document

G02BXF

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

G02BXF calculates the sample means, the standard deviations, the variance-covariance matrix, and the matrix of Pearson product-moment correlation coefficients for a set of data. Weights may be used.

2 Specification

```
SUBROUTINE G02BXF (WEIGHT, N, M, X, LDX, WT, XBAR, STD, V, LDV, R,          &
                   IFAIL)

INTEGER             N, M, LDX, LDV, IFAIL
REAL (KIND=nag_wp) X(LDX,M), WT(*), XBAR(M), STD(M), V(LDV,M),              &
                   R(LDV,M)
CHARACTER(1)        WEIGHT
```

3 Description

For n observations on m variables the one-pass algorithm of West (1979) as implemented in G02BUF is used to compute the means, the standard deviations, the variance-covariance matrix, and the Pearson product-moment correlation matrix for p selected variables. Suitables weights may be used to indicate multiple observations and to remove missing values. The quantities are defined by:

(a) The means

$$\bar{x}_j = \frac{\sum_{i=1}^{n} w_i x_{ij}}{\sum_{i=1}^{n} w_i} \quad j = 1, \ldots, p$$

(b) The variance-covariance matrix

$$C_{jk} = \frac{\sum_{i=1}^{n} w_i \left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right)}{\sum_{i=1}^{n} w_i - 1} \quad j, k = 1, \ldots, p$$

(c) The standard deviations

$$s_j = \sqrt{C_{jj}} \quad j = 1, \ldots, p$$

(d) The Pearson product-moment correlation coefficients

$$R_{jk} = \frac{C_{jk}}{\sqrt{C_{jj} C_{kk}}} \quad j, k = 1, \ldots, p$$

where x_{ij} is the value of the ith observation on the jth variable and w_i is the weight for the ith observation which will be 1 in the unweighted case.

Note that the denominator for the variance-covariance is $\sum_{i=1}^{n} w_i - 1$, so the weights should be scaled so that the sum of weights reflects the true sample size.

4 References

Chan T F, Golub G H and Leveque R J (1982) *Updating Formulae and a Pairwise Algorithm for Computing Sample Variances* Compstat, Physica-Verlag

West D H D (1979) Updating mean and variance estimates: An improved method *Comm. ACM* **22** 532–555

5 Parameters

1: WEIGHT – CHARACTER(1) *Input*

On entry: indicates whether weights are to be used.

WEIGHT = 'U'
Weights are not used and unit weights are assumed.

WEIGHT = 'W' or 'V'
Weights are used and must be supplied in WT. The only difference between WEIGHT = 'W' or WEIGHT = 'V' is in computing the variance. If WEIGHT = 'W' the divisor for the variance is the sum of the weights minus one and if WEIGHT = 'V' the divisor is the number of observations with nonzero weights minus one. The former is useful if the weights represent the frequency of the observed values.

Constraint: WEIGHT = 'U', 'V' or 'W'.

2: N – INTEGER *Input*

On entry: the number of data observations in the sample.

Constraint: $\mathrm{N} > 1$.

3: M – INTEGER *Input*

On entry: the number of variables.

Constraint: $\mathrm{M} \geq 1$.

4: X(LDX, M) – REAL (KIND=nag_wp) array *Input*

On entry: $\mathrm{X}(i,j)$ must contain the ith observation for the jth variable, for $i = 1,2,\ldots,\mathrm{N}$ and $j = 1,2,\ldots,\mathrm{M}$.

5: LDX – INTEGER *Input*

On entry: the first dimension of the array X as declared in the (sub)program from which G02BXF is called.

Constraint: $\mathrm{LDX} \geq \mathrm{N}$.

6: WT($*$) – REAL (KIND=nag_wp) array *Input*

Note: the dimension of the array WT must be at least N if WEIGHT = 'W' or 'V', and at least 1 otherwise.

On entry: w, the optional frequency weighting for each observation, with $\mathrm{WT}(i) = w_i$. Usually w_i will be an integral value corresponding to the number of observations associated with the ith data value, or zero if the ith data value is to be ignored. If WEIGHT = 'U', w_i is set to 1 for all i and WT is not referenced.

Constraint: if WEIGHT = 'W' or 'V', $\displaystyle\sum_{i=1}^{\mathrm{N}} \mathrm{WT}(i) > 1.0$, $\mathrm{WT}(i) \geq 0.0$, for $i = 1,2,\ldots,\mathrm{N}$.

7: XBAR(M) – REAL (KIND=nag_wp) array *Output*

On exit: the sample means. XBAR(j) contains the mean of the jth variable.

8: STD(M) – REAL (KIND=nag_wp) array *Output*

On exit: the standard deviations. STD(j) contains the standard deviation for the jth variable.

9: V(LDV, M) – REAL (KIND=nag_wp) array *Output*

On exit: the variance-covariance matrix. V(j, k) contains the covariance between variables j and k, for $j = 1, 2, \ldots, \mathrm{M}$ and $k = 1, 2, \ldots, \mathrm{M}$.

10: LDV – INTEGER *Input*

On entry: the first dimension of the arrays R and V as declared in the (sub)program from which G02BXF is called.

Constraint: LDV $\geq$ M.

11: R(LDV, M) – REAL (KIND=nag_wp) array *Output*

On exit: the matrix of Pearson product-moment correlation coefficients. R(j, k) contains the correlation coefficient between variables j and k.

12: 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.

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.**

On exit: IFAIL $= 0$ unless the routine detects an error or a warning has been flagged (see Section 6).

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: G02BXF 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, M < 1,
or N ≤ 1,
or LDX $<$ N,
or LDV $<$ M.

IFAIL $= 2$

On entry, WEIGHT $\neq$ 'U', 'V' or 'W'.

IFAIL $= 3$

On entry, WEIGHT $=$ 'W' or 'V' and a value of WT < 0.0.

IFAIL $= 4$

WEIGHT = 'W' and the sum of weights is not greater than 1.0, or WEIGHT = 'V' and fewer than 2 observations have nonzero weights.

IFAIL $= 5$

A variable has a zero variance. In this case V and STD are returned as calculated but R will contain zero for any correlation involving a variable with zero variance.

IFAIL $= -99$

An unexpected error has been triggered by this routine. Please contact NAG.

See Section 3.8 in the Essential Introduction for further information.

IFAIL $= -399$

Your licence key may have expired or may not have been installed correctly.

See Section 3.7 in the Essential Introduction for further information.

IFAIL $= -999$

Dynamic memory allocation failed.

See Section 3.6 in the Essential Introduction for further information.

7 Accuracy

For a discussion of the accuracy of the one pass algorithm see Chan *et al.* (1982) and West (1979).

8 Parallelism and Performance

G02BXF is not threaded by NAG in any implementation.

G02BXF makes calls to BLAS and/or LAPACK routines, which may be threaded within the vendor library used by this implementation. Consult the documentation for the vendor library for further information.

Please consult the X06 Chapter Introduction for information on how to control and interrogate the OpenMP environment used within this routine. Please also consult the Users' Note for your implementation for any additional implementation-specific information.

9 Further Comments

None.

10 Example

The data are some of the results from 1988 Olympic Decathlon. They are the times (in seconds) for the 100m and 400m races and the distances (in metres) for the long jump, high jump and shot. Twenty observations are input and the correlation matrix is computed and printed.

10.1 Program Text

```
    Program g02bxfe

!     G02BXF Example Program Text

!     Mark 25 Release. NAG Copyright 2014.

!     .. Use Statements ..
      Use nag_library, Only: g02bxf, nag_wp, x04caf
```

```
!     .. Implicit None Statement ..
      Implicit None
!     .. Parameters ..
      Integer, Parameter               :: nin = 5, nout = 6
!     .. Local Scalars ..
      Integer                          :: i, ifail, ldv, ldx, lwt, m, n
      Logical                          :: zero_var
      Character (1)                    :: weight
!     .. Local Arrays ..
      Real (Kind=nag_wp), Allocatable  :: r(:,:), std(:), v(:,:), wt(:),       &
                                          x(:,:), xbar(:)
!     .. Executable Statements ..
      Write (nout,*) 'G02BXF Example Program Results'
      Write (nout,*)

!     Skip heading in data file
      Read (nin,*)

!     Read in problem size
      Read (nin,*) weight, n, m
      If (weight=='W' .Or. weight=='w') Then
        lwt = n
      Else
        lwt = 0
      End If
      ldx = n
      ldv = m
      Allocate (x(ldx,m),wt(lwt),xbar(m),std(m),v(ldv,m),r(ldv,m))

!     Read in data
      If (lwt>0) Then
        Read (nin,*)(x(i,1:m),wt(i),i=1,n)
      Else
        Read (nin,*)(x(i,1:m),i=1,n)
      End If

!     Calculate summary statistics
      ifail = -1
      Call g02bxf(weight,n,m,x,ldx,wt,xbar,std,v,ldv,r,ifail)
      If (ifail/=0) Then
        If (ifail==5) Then
          zero_var = .True.
        Else
          Go To 100
        End If
      Else
        zero_var = .False.
      End If

!     Display results
      Write (nout,*) '     Means'
      Write (nout,*)
      Write (nout,99999)(xbar(i),i=1,m)
      Write (nout,*)
      Write (nout,*) '     Standard deviations'
      Write (nout,*)
      Write (nout,99999)(std(i),i=1,m)
      Write (nout,*)
      Flush (nout)
      ifail = 0
      Call x04caf('Upper','Non-unit',m,m,r,ldv,'     Correlation matrix', &
        ifail)
      If (zero_var) Then
        Write (nout,*) ' NOTE: some variances are zero'
      End If

100   Continue

99999 Format (1X,10F13.4)
    End Program g02bxfe
```

10.2 Program Data

```
G02BXF Example Program Data
 'u'     20       5
11.25 48.9 7.43 2.270 15.48
10.87 47.7 7.45 1.971 14.97
11.18 48.2 7.44 1.979 14.20
10.62 49.0 7.38 2.026 15.02
11.02 47.4 7.43 1.974 12.92
10.83 48.3 7.72 2.124 13.58
11.18 49.3 7.05 2.064 14.12
11.05 48.2 6.95 2.001 15.34
11.15 49.1 7.12 2.035 14.52
11.23 48.6 7.28 1.970 15.25
10.94 49.9 7.45 1.974 15.34
11.18 49.0 7.34 1.942 14.48
11.02 48.2 7.29 2.063 12.92
10.99 47.8 7.37 1.973 13.61
11.03 48.9 7.45 1.974 14.20
11.09 48.8 7.08 2.039 14.51
11.46 51.2 6.75 2.008 16.07
11.57 49.8 7.00 1.944 16.60
11.07 47.9 7.04 1.947 13.41
10.89 49.6 7.07 1.798 15.84
```

10.3 Program Results

```
 G02BXF Example Program Results

      Means

       11.0810      48.7900       7.2545       2.0038      14.6190

      Standard deviations

        0.2132       0.9002       0.2349       0.0902       1.0249

      Correlation matrix
          1       2       3       4       5
 1   1.0000  0.4416 -0.5427  0.0696  0.3912
 2           1.0000 -0.5058 -0.0678  0.7057
 3                   1.0000  0.2768 -0.4352
 4                           1.0000 -0.1494
 5                                   1.0000
```