

G03AAF – NAG Fortran Library Routine Document

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

G03AAF performs a principal component analysis on a data matrix; both the principal component loadings and the principal component scores are returned.

2 Specification

```

SUBROUTINE G03AAF(MATRIX, STD, WEIGHT, N, M, X, LDX, ISX, S, WT,
1              NVAR, E, LDE, P, LDP, V, LDV, WK, IFAIL)
  INTEGER      N, M, LDX, ISX(M), NVAR, LDE, LDP, LDV, IFAIL
  real        X(LDX,M), S(M), WT(*), E(LDE,6), P(LDP,NVAR),
1              V(LDV,NVAR), WK(NVAR*NVAR+5*(NVAR-1))
  CHARACTER*1  MATRIX, STD, WEIGHT

```

3 Description

Let X be an n by p data matrix of n observations on p variables x_1, x_2, \dots, x_p and let the p by p variance-covariance matrix of x_1, x_2, \dots, x_p be S . A vector a_1 of length p is found such that:

$$a_1^T S a_1 \text{ is maximized subject to } a_1^T a_1 = 1.$$

The variable $z_1 = \sum_{i=1}^p a_{1i} x_i$ is known as the first principal component and gives the linear combination of

the variables that gives the maximum variation. A second principal component, $z_2 = \sum_{i=1}^p a_{2i} x_i$, is found such that:

$$a_2^T S a_2 \text{ is maximized subject to } a_2^T a_2 = 1 \text{ and } a_2^T a_1 = 0.$$

This gives the linear combination of variables that is orthogonal to the first principal component that gives the maximum variation. Further principal components are derived in a similar way.

The vectors a_1, a_2, \dots, a_p , are the eigenvectors of the matrix S and associated with each eigenvector is the eigenvalue, λ_i^2 . The value of $\lambda_i^2 / \sum \lambda_i^2$ gives the proportion of variation explained by the i th principal component. Alternatively the a_i 's can be considered as the right singular vectors in a singular value decomposition with singular values λ_i of the data matrix centred about its mean and scaled by $1/\sqrt{(n-1)}$, X_s . This latter approach is used in G03AAF, with

$$X_s = V \Lambda P'$$

where Λ is a diagonal matrix with elements λ_i , P' is the p by p matrix with columns a_i and V is an n by p matrix with $V'V = I$, which gives the principal component scores.

Principal component analysis is often used to reduce the dimension of a data set, replacing a large number of correlated variables with a smaller number of orthogonal variables that still contain most of the information in the original data set.

The choice of the number of dimensions required is usually based on the amount of variation accounted for by the leading principal components. If k principal components are selected then a test of the equality of the remaining $p - k$ eigenvalues is

$$(n - (2p + 5)/6) \left\{ - \sum_{i=k+1}^p \log(\lambda_i^2) + (p - k) \log \left(\sum_{i=k+1}^p \lambda_i^2 / (p - k) \right) \right\}$$

which has, asymptotically, a χ^2 distribution with $\frac{1}{2}(p - k - 1)(p - k + 2)$ degrees of freedom.

Equality of the remaining eigenvalues indicates that if any more principal components are to be considered then they all should be considered.

Instead of the variance-covariance matrix the correlation matrix, the sums of squares and cross-products matrix or a standardised sums of squares and cross-products matrix may be used. In the last case S is replaced by $\sigma^{-\frac{1}{2}}S\sigma^{-\frac{1}{2}}$ for a diagonal matrix σ with positive elements. If the correlation matrix is used the χ^2 approximation for the statistic given above is not valid.

The principal component scores, F , are the values of the principal component variables for the observations. These can be standardised so that the variance of these scores for each principal component is 1.0 or equal to the corresponding eigenvalue.

Weights can be used with the analysis, in which case the matrix X is first centred about the weighted means then each row is scaled by an amount $\sqrt{w_i}$, where w_i is the weight for the i th observation.

4 References

- [1] Chatfield C and Collins A J (1980) *Introduction to Multivariate Analysis* Chapman and Hall
- [2] Cooley W C and Lohnes P R (1971) *Multivariate Data Analysis* Wiley
- [3] Hammarling S (1985) The singular value decomposition in multivariate statistics *SIGNUM Newsl.* **20** (3) 2–25
- [4] Kendall M G and Stuart A (1969) *The Advanced Theory of Statistics (Volume 1)* Griffin (3rd Edition)
- [5] Morrison D F (1967) *Multivariate Statistical Methods* McGraw–Hill

5 Parameters

1: MATRIX — CHARACTER*1 *Input*

On entry: indicates for which type of matrix the principal component analysis is to be carried out.

If MATRIX = 'C', then it is for the correlation matrix.

If MATRIX = 'S', then it is for a standardised matrix, with standardisations given by S.

If MATRIX = 'U', then it is for the sums of squares and cross-products matrix.

If MATRIX = 'V', then it is for the variance-covariance matrix.

Constraint: MATRIX = 'C', 'S', 'U' or 'V'.

2: STD — CHARACTER*1 *Input*

On entry: indicates if the principal component scores are to be standardised.

If STD = 'S', then the principal component scores are standardised so that $F'F = I$, i.e., $F = X_s P \Lambda^{-1} = V$.

If STD = 'U', then the principal component scores are unstandardised, i.e., $F = X_s P = V \Lambda$.

If STD = 'Z', then the principal component scores are standardised so that they have unit variance.

If STD = 'E', then the principal component scores are standardised so that they have variance equal to the corresponding eigenvalue.

Constraint: STD = 'E', 'S', 'U' or 'Z'.

3: WEIGHT — CHARACTER*1 *Input*

On entry: indicates if weights are to be used.

If WEIGHT = 'U' (Unweighted), then no weights are used.

If WEIGHT = 'W' (Weighted), then weights are used and must be supplied in WT.

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

- 4:** N — INTEGER *Input*
On entry: the number of observations, n .
Constraint: $N \geq 2$.
- 5:** M — INTEGER *Input*
On entry: the number of variables in the data matrix, m .
Constraint: $M \geq 1$.
- 6:** X(LDX,M) — *real* array *Input*
On entry: X(i , j) must contain the i th observation for the j th variable, for $i = 1, 2, \dots, n$;
 $j = 1, 2, \dots, m$.
- 7:** LDX — INTEGER *Input*
On entry: the first dimension of the array X as declared in the (sub)program from which G03AAF is called.
Constraint: $LDX \geq N$.
- 8:** ISX(M) — INTEGER array *Input*
On entry: ISX(j) indicates whether or not the j th variable is to be included in the analysis.
 If ISX(j) > 0, then the variable contained in the j th column of X is included in the principal component analysis, for $j = 1, 2, \dots, m$.
Constraint: ISX(j) > 0 for NVAR values of j .
- 9:** S(M) — *real* array *Input/Output*
On entry: the standardisations to be used, if any.
 If MATRIX = 'S', then the first m elements of S must contain the standardisation coefficients, the diagonal elements of σ .
Constraint: if ISX(j) > 0, then S(j) > 0.0, for $j = 1, 2, \dots, m$.
On exit:
 If MATRIX = 'S', then S is unchanged on exit.
 If MATRIX = 'C', then S contains the variances of the selected variables. S(j) contains the variance of the variable in the j th column of X if ISX(j) > 0.
 If MATRIX = 'U' or 'V', then S is not referenced.
- 10:** WT(*) — *real* array *Input*
On entry: if WEIGHT = 'W', then the first n elements of WT must contain the weights to be used in the principal component analysis.
 If WT(i) = 0.0, then the i th observation is not included in the analysis. The effective number of observations is the sum of the weights.
 If WEIGHT = 'U', then WT is not referenced and the effective number of observations is n .
Constraint: WT(i) \geq 0.0, for $i = 1, 2, \dots, n$ and the sum of weights \geq NVAR + 1.
- 11:** NVAR — INTEGER *Input*
On entry: the number of variables in the principal component analysis, p .
Constraint: $1 \leq \text{NVAR} \leq \min(N-1, M)$.

- 12:** E(LDE,6) — *real* array *Output*
On exit: the statistics of the principal component analysis.
 E(i,1), the eigenvalues associated with the i th principal component, λ_i^2 , for $i = 1, 2, \dots, p$.
 E(i,2), the proportion of variation explained by the i th principal component, for $i = 1, 2, \dots, p$.
 E(i,3), the cumulative proportion of variation explained by the first i th principal components, for $i = 1, 2, \dots, p$.
 E(i,4), the χ^2 statistics, for $i = 1, 2, \dots, p$.
 E(i,5), the degrees of freedom for the χ^2 statistics, for $i = 1, 2, \dots, p$.
 If MATRIX \neq 'C', then E(i,6) contains significance level for the χ^2 statistic, for $i = 1, 2, \dots, p$.
 If MATRIX = 'C', then E(i,6) is returned as zero.
- 13:** LDE — INTEGER *Input*
On entry: the first dimension of the array E as declared in the (sub)program from which G03AAF is called.
Constraint: LDE \geq NVAR.
- 14:** P(LDP,NVAR) — *real* array *Output*
On exit: the first NVAR columns of P contain the principal component loadings, a_i . The j th column of P contains the NVAR coefficients for the j th principal component.
- 15:** LDP — INTEGER *Input*
On entry: the first dimension of the array P as declared in the (sub)program from which G03AAF is called.
Constraint: LDP \geq NVAR.
- 16:** V(LDV,NVAR) — *real* array *Output*
On exit: the first NVAR columns of V contain the principal component scores. The j th column of V contains the N scores for the j th principal component.
 If WEIGHT = 'W', then any rows for which WT(i) is zero will be set to zero.
- 17:** LDV — INTEGER *Input*
On entry: the first dimension of the array V as declared in the (sub)program from which G03AAF is called.
Constraint: LDV \geq N.
- 18:** WK(NVAR*NVAR+5*(NVAR-1)) — *real* array *Workspace*
- 19:** IFAIL — INTEGER *Input/Output*
On entry: IFAIL must be set to 0, -1 or 1. For users not familiar with this parameter (described in Chapter P01) the recommended value is 0.
On exit: IFAIL = 0 unless the routine detects an error (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`).

Errors detected by the routine:

`IFAIL = 1`

On entry, `M < 1`,
 or `N < 2`,
 or `NVAR < 1`,
 or `NVAR > M`,
 or `NVAR ≥ N`,
 or `LDX < N`,
 or `LDV < N`,
 or `LDP < NVAR`,
 or `LDE < NVAR`,
 or `MATRIX ≠ 'C', 'S', 'U' or 'V'`,
 or `STD ≠ 'S', 'U', 'Z' or 'E'`,
 or `WEIGHT ≠ 'U' or 'W'`.

`IFAIL = 2`

On entry, `WEIGHT = 'W'` and a value of `WT < 0.0`.

`IFAIL = 3`

On entry, there are not `NVAR` values of `ISX > 0`,
 or `WEIGHT = 'W'` and the effective number of observations is less than `NVAR + 1`.

`IFAIL = 4`

On entry, `S(j) ≤ 0.0` for some $j = 1, 2, \dots, m$, when `MATRIX = 'S'` and `ISX(j) > 0`.

`IFAIL = 5`

The singular value decomposition has failed to converge. See `F02WEF`. This is an unlikely error exit.

`IFAIL = 6`

All eigenvalues/singular values are zero. This will be caused by all the variables being constant.

7 Accuracy

As `G03AAF` uses a singular value decomposition of the data matrix, it will be less affected by ill-conditioned problems than traditional methods using the eigenvalue decomposition of the variance-covariance matrix.

8 Further Comments

None.

9 Example

A data set is taken from Cooley and Lohnes [2], it consists of ten observations on three variables. The unweighted principal components based on the variance-covariance matrix are computed and unstandardised principal component scores requested.

9.1 Program Text

Note. The listing of the example program presented below uses bold italicised terms to denote precision-dependent details. Please read the Users' Note for your implementation to check the interpretation of these terms. As explained in the Essential Introduction to this manual, the results produced may not be identical for all implementations.

```

*   G03AAF Example Program Text
*   Mark 17 Revised.  NAG Copyright 1995.
*   .. Parameters ..
INTEGER          NMAX, MMAX
PARAMETER       (NMAX=12,MMAX=3)
INTEGER          NIN, NOUT
PARAMETER       (NIN=5,NOUT=6)
*   .. Local Scalars ..
INTEGER          I, IFAIL, J, M, N, NVAR
CHARACTER       MATRIX, STD, WEIGHT
*   .. Local Arrays ..
real           E(MMAX,6), P(MMAX,MMAX), S(MMAX), V(NMAX,MMAX),
+              WK(MMAX*MMAX+5*(MMAX-1)), WT(NMAX), X(NMAX,MMAX)
INTEGER          ISX(MMAX)
*   .. External Subroutines ..
EXTERNAL        G03AAF
*   .. Executable Statements ..
WRITE (NOUT,*) 'G03AAF Example Program Results'
*   Skip heading in data file
READ (NIN,*)
READ (NIN,*) MATRIX, STD, WEIGHT, N, M
IF (N.LE.NMAX .AND. M.LE.MMAX) THEN
    IF (WEIGHT.EQ.'U' .OR. WEIGHT.EQ.'u') THEN
        DO 20 I = 1, N
            READ (NIN,*) (X(I,J),J=1,M)
20    CONTINUE
        ELSE
            DO 40 I = 1, N
                READ (NIN,*) (X(I,J),J=1,M), WT(I)
40    CONTINUE
        END IF
        READ (NIN,*) (ISX(J),J=1,M), NVAR
        IF (MATRIX.EQ.'S' .OR. MATRIX.EQ.'s') READ (NIN,*) (S(J),J=1,M)
        IFAIL = 0
*
        CALL G03AAF(MATRIX,STD,WEIGHT,N,M,X,NMAX,ISX,S,WT,NVAR,E,MMAX,
+              P,MMAX,V,NMAX,WK,IFAIL)
*
        WRITE (NOUT,*)
        WRITE (NOUT,*)
+   'Eigenvalues Percentage Cumulative      Chisq      DF      Sig'
        WRITE (NOUT,*) '          variation  variation'
        WRITE (NOUT,*)
        DO 60 I = 1, NVAR
            WRITE (NOUT,99999) (E(I,J),J=1,6)
60    CONTINUE
        WRITE (NOUT,*)
        WRITE (NOUT,*) 'Eigenvalues'
        WRITE (NOUT,*)
        DO 80 I = 1, NVAR
            WRITE (NOUT,99998) (P(I,J),J=1,NVAR)
80    CONTINUE
        WRITE (NOUT,*)
        WRITE (NOUT,*) 'Principal component scores'

```

```

        WRITE (NOUT,*)
        DO 100 I = 1, N
            WRITE (NOUT,99997) I, (V(I,J),J=1,NVAR)
100     CONTINUE
        END IF
        STOP
*
99999 FORMAT (1X,F11.4,2F12.4,F10.4,F8.1,F8.4)
99998 FORMAT (1X,8F9.4)
99997 FORMAT (1X,I2,(8F9.3))
        END

```

9.2 Program Data

G03AAF Example Program Data

```

'V' 'E' 'U' 10 3
7.0 4.0 3.0
4.0 1.0 8.0
6.0 3.0 5.0
8.0 6.0 1.0
8.0 5.0 7.0
7.0 2.0 9.0
5.0 3.0 3.0
9.0 5.0 8.0
7.0 4.0 5.0
8.0 2.0 2.0
1 1 1 3

```

9.3 Program Results

G03AAF Example Program Results

Eigenvalues	Percentage variation	Cumulative variation	Chisq	DF	Sig
8.2739	0.6515	0.6515	8.6127	5.0	0.1255
3.6761	0.2895	0.9410	4.1183	2.0	0.1276
0.7499	0.0590	1.0000	0.0000	0.0	0.0000

Eigenvalues

0.1376	0.6990	0.7017
0.2505	0.6609	-0.7075
-0.9583	0.2731	-0.0842

Principal component scores

1	2.151	-0.173	-0.107
2	-3.804	-2.887	-0.510
3	-0.153	-0.987	-0.269
4	4.707	1.302	-0.652
5	-1.294	2.279	-0.449
6	-4.099	0.144	0.803
7	1.626	-2.232	-0.803
8	-2.114	3.251	0.168
9	0.235	0.373	-0.275
10	2.746	-1.069	2.094
