1 Purpose

g03aa performs a principal component analysis on a data matrix; both the principal component loadings and the principal component scores are returned.

2 Syntax

\[ [s, e, p, v, ifail] = g03aa(matrix, std, weight, x, isx, s, wt, nvar, 'n', n, 'm', m) \]

Note: The interface to this routine has changed at this mark from earlier releases of the toolbox: \( n \) has been made optional.

3 Description

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

\[ a_1^T S a_1 \]

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:

\[ a_2^T S a_2 \]

is maximized subject to \( a_2^T a_2 = 1 \) 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 \( a_1, a_2, \ldots, a_p \) are the eigenvectors of the matrix \( S \) and associated with each eigenvector is the eigenvalue, \( \lambda_i \). The value of \( \lambda_i / \sum \lambda_i \) 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 \( \lambda_i \) of the data matrix centred about its mean and scaled by

\[ \frac{1}{\sqrt{n}} \frac{1}{\sqrt{p}} \]

which has, asymptotically, a \( \chi^2 \)-distribution with \( \frac{1}{2}(p-k)(p-k+2) \) 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 standardized sums of squares and cross-products matrix may be used. In the last case $S$ is replaced by $\sigma^{-\frac{1}{2}}SS\sigma^{-\frac{1}{2}}$ for a diagonal matrix $\sigma$ with positive elements. If the correlation matrix is used, the $\chi^2$ 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 standardized 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

5 Parameters
5.1 Compulsory Input Parameters
1: matrix – string
   Indicates for which type of matrix the principal component analysis is to be carried out.
   matrix = 'C'
   It is for the correlation matrix.
   matrix = 'S'
   It is for a standardized matrix, with standardizations given by $s$.
   matrix = 'U'
   It is for the sums of squares and cross-products matrix.
   matrix = 'V'
   It is for the variance-covariance matrix.
   Constraint: matrix = 'C', 'S', 'U' or 'V'.

2: std – string
   Indicates if the principal component scores are to be standardized.
   std = 'S'
   The principal component scores are standardized so that $F'F = I$, i.e., $F = X_sP\Lambda^{-1} = V$.
   std = 'U'
   The principal component scores are unstandardized, i.e., $F = X_sP = VA$.
   std = 'Z'
   The principal component scores are standardized so that they have unit variance.
The principal component scores are standardized so that they have variance equal to the corresponding eigenvalue.

Constraint: \( \text{std} = 'E', 'S', 'U' \) or 'Z'.

3: weight – string
Indicates if weights are to be used.

weight = 'U'
No weights are used.

weight = 'W'
Weights are used and must be supplied in \( \text{wt} \).

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

4: \( x(ldx,m) \) – double array

\( ldx \), the first dimension of the array, must satisfy the constraint \( ldx \geq n \).

\( x(i,j) \) must contain the \( i \)th observation for the \( j \)th variable, for \( i = 1, 2, \ldots, n \) and \( j = 1, 2, \ldots, m \).

5: \( \text{isx}(m) \) – int32 array

\( m \), the dimension of the array, must satisfy the constraint \( m \geq 1 \).

\( \text{isx}(j) \) indicates whether or not the \( j \)th variable is to be included in the analysis.

If \( \text{isx}(j) > 0 \), the variable contained in the \( j \)th column of \( x \) is included in the principal component analysis, for \( j = 1, 2, \ldots, m \).

Constraint: \( \text{isx}(j) > 0 \) for \( \text{nvar} \) values of \( j \).

6: \( s(m) \) – double array

\( m \), the dimension of the array, must satisfy the constraint \( m \geq 1 \).

The standardizations to be used, if any.

If \( \text{matrix} = 'S' \), the first \( m \) elements of \( s \) must contain the standardization coefficients, the diagonal elements of \( \sigma \).

Constraint: if \( \text{isx}(j) > 0 \), \( s(j) > 0.0 \), for \( j = 1, 2, \ldots, m \).

7: \( \text{wt}(:) \) – double array

Note: the dimension of the array \( \text{wt} \) must be at least \( n \) if \( \text{weight} = 'W' \), and at least 1 otherwise.

If \( \text{weight} = 'W' \), the first \( n \) elements of \( \text{wt} \) must contain the weights to be used in the principal component analysis.

If \( \text{wt}(i) = 0.0 \), the \( i \)th observation is not included in the analysis. The effective number of observations is the sum of the weights.

If \( \text{weight} = 'U' \), \( \text{wt} \) is not referenced and the effective number of observations is \( n \).

Constraint: \( \text{wt}(i) \geq 0.0 \), for \( i = 1, 2, \ldots, n \) and the sum of weights \( \geq n_{\text{var}} + 1 \).

8: \( \text{nvar} \) – int32 scalar

\( p \), the number of variables in the principal component analysis.

Constraint: \( 1 \leq \text{nvar} \leq \min(n - 1, m) \).
5.2 Optional Input Parameters

1: \( n \) – int32 scalar

*Default:* The first dimension of the array \( x \).

\( n \), the number of observations.

*Constraint:* \( n \geq 2 \).

2: \( m \) – int32 scalar

*Default:* The dimension of the arrays \( isx \), \( s \) and the second dimension of the array \( x \). (An error is raised if these dimensions are not equal.)

\( m \), the number of variables in the data matrix.

*Constraint:* \( m \geq 1 \).

5.3 Input Parameters Omitted from the MATLAB Interface

\( ldx \), \( lde \), \( ldp \), \( ldv \), \( wk \)

5.4 Output Parameters

1: \( s(m) \) – double array

If \( \text{matrix} = 'S' \), \( s \) is unchanged on exit.

If \( \text{matrix} = 'C' \), \( s \) 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 \( \text{matrix} = 'U' \) or \( 'V' \), \( s \) is not referenced.

2: \( e(lde,6) \) – double array

The statistics of the principal component analysis.

\( e(i,1) \)

The eigenvalues associated with the \( i \)th principal component, \( \lambda_i^2 \), for \( i = 1,2,\ldots,p \).

\( e(i,2) \)

The proportion of variation explained by the \( i \)th principal component, for \( i = 1,2,\ldots,p \).

\( e(i,3) \)

The cumulative proportion of variation explained by the first \( i \)th principal components, for \( i = 1,2,\ldots,p \).

\( e(i,4) \)

The \( \chi^2 \) statistics, for \( i = 1,2,\ldots,p \).

\( e(i,5) \)

The degrees of freedom for the \( \chi^2 \) statistics, for \( i = 1,2,\ldots,p \).

If \( \text{matrix} \neq 'C' \), \( e(i,6) \) contains significance level for the \( \chi^2 \) statistic, for \( i = 1,2,\ldots,p \).

If \( \text{matrix} = 'C' \), \( e(i,6) \) is returned as zero.

3: \( p(ldp,nvar) \) – double array

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.

4: \( v(ldv,nvar) \) – double array

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 \( \text{weight} = 'W' \), any rows for which \( \text{wt}(i) \) is zero will be set to zero.

5: \[ \text{ifail} \rightarrow \text{int32 scalar} \]
    ifail = 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, \( m < 1 \),
  - or \( n < 2 \),
  - or \( \text{nvar} < 1 \),
  - or \( \text{nvar} > m \),
  - or \( \text{nvar} \geq n \),
  - or \( \text{ldx} < n \),
  - or \( \text{ldv} < n \),
  - or \( \text{ldp} < \text{nvar} \),
  - or \( \text{ldf} < \text{nvar} \),
  - or \( \text{matrix} \neq 'C', 'S', 'U' \) or 'V',
  - or \( \text{std} \neq 'S', 'U', 'Z' \) or 'E',
  - or \( \text{weight} \neq 'U' \) or 'W'.

- ifail = 2
  - On entry, weight = 'W' and a value of \( \text{wt} < 0.0 \).

- ifail = 3
  - On entry, there are not \( \text{nvar} \) values of \( \text{isx} > 0 \),
  - or weight = 'W' and the effective number of observations is less than \( \text{nvar} + 1 \).

- ifail = 4
  - On entry, \( \text{s}(j) \leq 0.0 \) for some \( j = 1, 2, \ldots, m \), when matrix = 'S' and isx\((j) > 0 \).

- ifail = 5
  - The singular value decomposition has failed to converge. 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 g03aa 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**

```plaintext
matrix = 'V';
std = 'E';
weight = 'U';
x = [7, 4, 3];
```
4, 1, 8;
6, 3, 5;
8, 6, 1;
8, 5, 7;
7, 2, 9;
5, 3, 3;
9, 5, 8;
7, 4, 5;
8, 2, 2;
isx = [int32(1);1;1];
s = [-5.04677090184712e-39;
    -5.04512289241806e-39;
    -1.790699005126953];
wt = [0];
nvar = int32(3);
[sOut, e, p, v, ifail] = g03aa(matrix, std, weight, x, isx, s, wt, nvar)
sOut =
   -0.0000
   -0.0000
   -1.7907
e =
3.6761  0.2895  0.9410  4.1183  2.0000  0.1276
  0.7499  0.0590  1.0000  0.0000  0.0000  0.0000
p =
-0.1376  0.6990  -0.7017
-0.2505  0.6609  0.7075
  0.9583  0.2731  0.0842
v =
-2.1514  -0.1731  0.1068
  3.8042  -2.8875  0.5104
  0.1532  -0.9869  0.2694
-4.7065   1.3015  0.6517
  1.2938   2.2791  0.4492
  4.0993   0.1436  -0.8031
-1.6258  -2.2321  0.8028
  2.1145   3.2512  -0.1684
-0.2348   0.3730  0.2751
-2.7464  -1.0689  -2.0940
ifail =
   0