NAG Toolbox

nag_mv_prin_comp (g03aa)

1 Purpose

nag_mv_prin_comp (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] = nag_mv_prin_comp(matrix, std, x, isx, s, nvar, 'n', n, 'm',
m, 'wt', wt)
```

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

Note: the interface to this routine has changed since earlier releases of the toolbox:

At Mark 24: *weight* was removed from the interface; **wt** was made optional

At Mark 22: **n** was 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:

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

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:

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

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, λ_i^2. The value of $\lambda_i^2 / \sum \lambda_i^2$ gives the proportion of variation explained by the ith 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 nag_mv_prin_comp (g03aa), 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 dataset, replacing a large number of correlated variables with a smaller number of orthogonal variables that still contain most of the information in the original dataset.

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

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

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 standardized 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 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 ith observation.

4 References

Chatfield C and Collins A J (1980) *Introduction to Multivariate Analysis* Chapman and Hall

Cooley W C and Lohnes P R (1971) *Multivariate Data Analysis* Wiley

Hammarling S (1985) The singular value decomposition in multivariate statistics *SIGNUM Newsl.* **20(3)** 2–25

Kendall M G and Stuart A (1969) *The Advanced Theory of Statistics (Volume 1)* (3rd Edition) Griffin

Morrison D F (1967) *Multivariate Statistical Methods* McGraw–Hill

5 Parameters

5.1 Compulsory Input Parameters

1: **matrix** – CHARACTER(1)

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** – CHARACTER(1)

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., $F = X_sP = V\Lambda$.

std = 'Z'

The principal component scores are standardized so that they have unit variance.

std = 'E'

The principal component scores are standardized so that they have variance equal to the corresponding eigenvalue.

Constraint: **std** = 'E', 'S', 'U' or 'Z'.

3: **x**$(ldx,$ **m**$)$ – REAL (KIND=nag_wp) array

ldx, the first dimension of the array, must satisfy the constraint $ldx \geq$ **n**.

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

4: **isx**(**m**) – INTEGER array

isx(j) indicates whether or not the jth variable is to be included in the analysis.

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

Constraint: **isx**$(j) > 0$ for **nvar** values of j.

5: **s**(**m**) – REAL (KIND=nag_wp) array

The standardizations to be used, if any.

If **matrix** = 'S', the first m elements of **s** must contain the standardization coefficients, the diagonal elements of σ.

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

6: **nvar** – INTEGER

p, the number of variables in the principal component analysis.

Constraint: $1 \leq$ **nvar** $\leq \min($**n** $- 1,$ **m**$)$.

5.2 Optional Input Parameters

1: **n** – INTEGER

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

n, the number of observations.

Constraint: **n** ≥ 2.

2: **m** – INTEGER

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** ≥ 1.

3: **wt**(:) – REAL (KIND=nag_wp) array

The dimension of the array **wt** must be at least **n** if $weight$ = 'W', and at least 1 otherwise

If $weight$ = 'W', the first n elements of **wt** must contain the weights to be used in the principal component analysis.

If **wt**$(i) = 0.0$, the ith observation is not included in the analysis. The effective number of observations is the sum of the weights.

If *weight* = 'U', **wt** is not referenced and the effective number of observations is n.

Constraints:

$\mathbf{wt}(i) \geq 0.0$, for $i = 1, 2, \ldots, n$;
the sum of weights $\geq \mathbf{nvar} + 1$.

5.3 Output Parameters

1: **s**(**m**) – REAL (KIND=nag_wp) array

If **matrix** = 'S', **s** is unchanged on exit.

If **matrix** = 'C', **s** contains the variances of the selected variables. $\mathbf{s}(j)$ contains the variance of the variable in the jth column of **x** if $\mathbf{isx}(j) > 0$.

If **matrix** = 'U' or 'V', **s** is not referenced.

2: $\mathbf{e}(lde, \mathbf{6})$ – REAL (KIND=nag_wp) array

The statistics of the principal component analysis.

$\mathbf{e}(i, 1)$

The eigenvalues associated with the ith principal component, λ_i^2, for $i = 1, 2, \ldots, p$.

$\mathbf{e}(i, 2)$

The proportion of variation explained by the ith principal component, for $i = 1, 2, \ldots, p$.

$\mathbf{e}(i, 3)$

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

$\mathbf{e}(i, 4)$

The χ^2 statistics, for $i = 1, 2, \ldots, p$.

$\mathbf{e}(i, 5)$

The degrees of freedom for the χ^2 statistics, for $i = 1, 2, \ldots, p$.

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

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

3: $\mathbf{p}(ldp, \mathbf{nvar})$ – REAL (KIND=nag_wp) array

The first **nvar** columns of **p** contain the principal component loadings, a_i. The jth column of **p** contains the **nvar** coefficients for the jth principal component.

4: $\mathbf{v}(ldv, \mathbf{nvar})$ – REAL (KIND=nag_wp) array

The first **nvar** columns of **v** contain the principal component scores. The jth column of **v** contains the **n** scores for the jth principal component.

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

5: **ifail** – INTEGER

ifail = 0 unless the function detects an error (see Section 5).

6 Error Indicators and Warnings

Errors or warnings detected by the function:

ifail = 1

On entry, $\mathbf{m} < 1$,
or $\mathbf{n} < 2$,
or $\mathbf{nvar} < 1$,

or **nvar** $>$ **m**,
or **nvar** $\geq$ **n**,
or $ldx <$ **n**,
or $ldv <$ **n**,
or $ldp <$ **nvar**,
or $lde <$ **nvar**,
or **matrix** $\neq$ 'C', 'S', 'U' or 'V',
or **std** $\neq$ 'S', 'U', 'Z' or 'E',
or $weight \neq$ '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**$(j) \leq 0.0$ for some $j = 1, 2, \dots, m$, when **matrix** = 'S' and **isx**$(j) > 0$.

ifail $= 5$

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

ifail $= 6$ (*warning*)

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

ifail $= -99$

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

ifail $= -399$

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

ifail $= -999$

Dynamic memory allocation failed.

7 Accuracy

As nag_mv_prin_comp (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

A dataset is taken from Cooley and Lohnes (1971), it consists of ten observations on three variables. The unweighted principal components based on the variance-covariance matrix are computed and the principal component scores requested. The principal component scores are standardized so that they have variance equal to the corresponding eigenvalue.

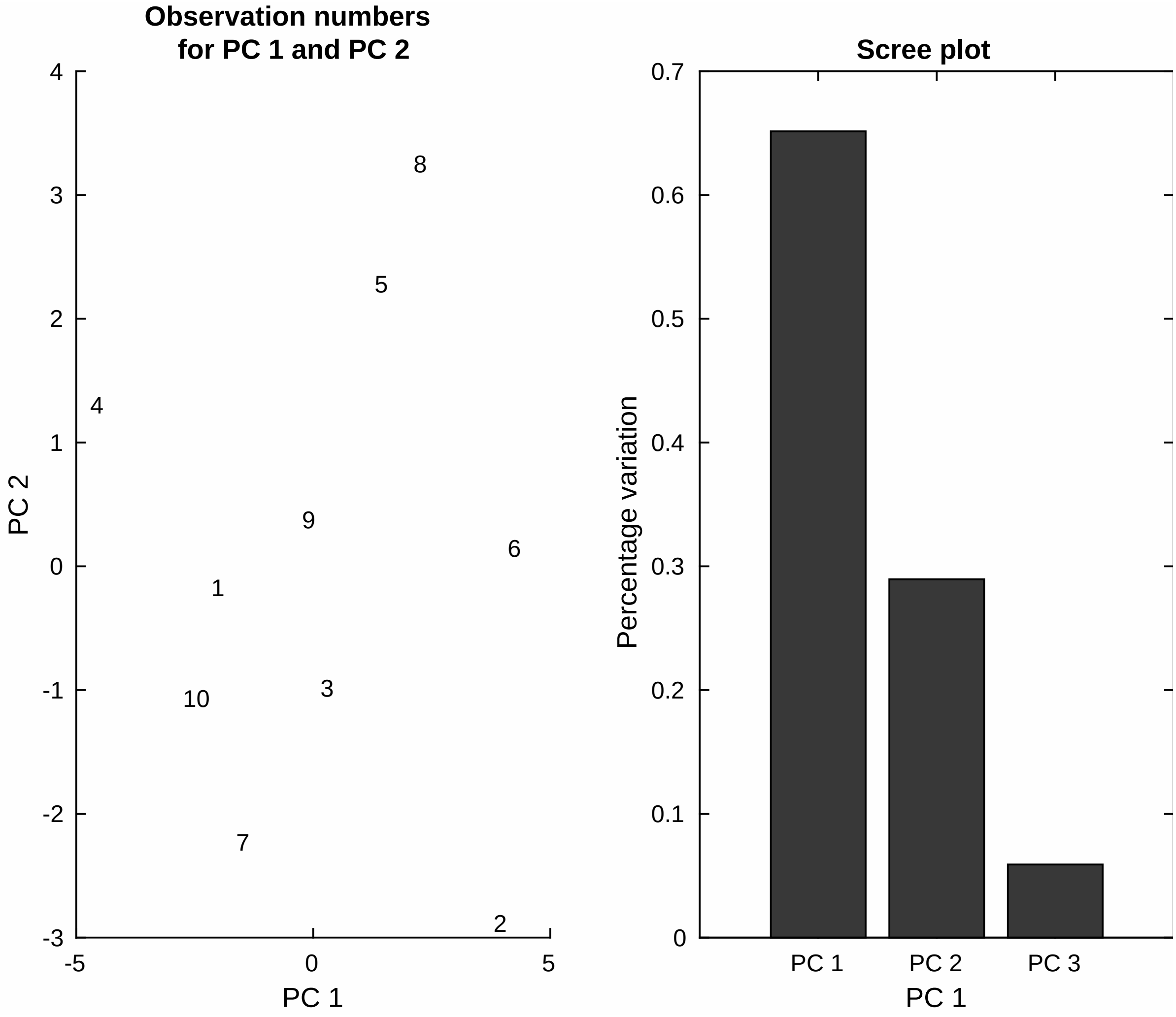
9.1 Program Text

```
     function g03aa_example

fprintf('g03aa example results\n\n');

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];
n    = size(x,2);

matrix = 'V';
std  = 'E';
isx  = ones(n,1,nag_int_name);
s    = zeros(n,1);
nvar = nag_int(n);

[s, e, p, v, ifail] = g03aa( ...
      matrix, std, x, isx, s, nvar);

fprintf('Eigenvalues  Percentage  Cumulative     Chisq      DF     Sig\n');
fprintf('              variation   variation\n\n');
fprintf('%11.4f%12.4f%12.4f%10.4f%8.1f%8.4f\n',e');
fprintf('\n');

mtitle = 'Principal component loadings';
matrix = 'General';
diag   = ' ';

[ifail] = x04ca( ...
                 matrix, diag, p, mtitle);

fprintf('\n');
mtitle = 'Principal component scores';
[ifail] = x04ca( ...
                 matrix, diag, v, mtitle);

fig1 = figure;
subplot(1,2,1);
xlabel('PC 1');
ylabel('PC 2');
title({'Observation numbers', 'for PC 1 and PC 2'});
axis([-5 5 -3 4]);
for j = 1:size(x,1)
  ch = sprintf('%d',j);
  text(v(j,1),v(j,2),ch);
end
subplot(1,2,2);
bar(e(:,2));
ax = gca;
ax.XTickLabel = {'PC 1','PC 2','PC 3'};
xlabel('PC 1');
ylabel('Percentage variation');
title('Scree plot');
```

9.2 Program Results

```
     g03aa example results

Eigenvalues  Percentage  Cumulative     Chisq      DF     Sig
              variation   variation

     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

 Principal component loadings
          1       2       3
 1  -0.1376  0.6990 -0.7017
 2  -0.2505  0.6609  0.7075
 3   0.9583  0.2731  0.0842

 Principal component scores
              1          2          3
  1     -2.1514    -0.1731     0.1068
  2      3.8042    -2.8875     0.5104
  3      0.1532    -0.9869     0.2694
  4     -4.7065     1.3015     0.6517
  5      1.2938     2.2791     0.4492
  6      4.0993     0.1436    -0.8031
  7     -1.6258    -2.2321     0.8028
  8      2.1145     3.2512    -0.1684
  9     -0.2348     0.3730     0.2751
 10     -2.7464    -1.0689    -2.0940
```


Observation numbers
for PC 1 and PC 2
PC 2
PC 1
Scree plot
Percentage variation
PC 1
PC 2
PC 3
PC 1