Principal components

Principal components are uncorrelated linear combinations of observed variables that account for maximum variance in the observed variables. For further explanation of principal components and their interpretation, we refer to Mardia, Kent, & Bibby (1980, Chapter 8), Reyment & Jöreskog (1993), or any other book covering principal components. For an explanation of the difference between principal components and factor analysis, see Jöreskog (1979).

Unlike factor analysis, which is scale independent in a certain sense, the results of principal components depend on the unit of measurement in the observed variables. For example, the analysis of the covariance matrix and the correlation matrix can give very different results, especially if the variances differ much across variables. This is the case in our example. Principal components are best applied to variables which are measured in the same units of measurements and which have approximately the same variances. In other situations, it is best to use the correlation matrix. Principal components are sometimes applied to the moment matrix instead of the covariance or correlation matrix, *i.e.*, the means are not subtracted from the observed scores before analysis. This too can give a different result compared with the covariance or correlation matrix.

Principal components may be obtained by including a PC command in a PRELIS, SIMPLIS, or LISREL command file. The syntax of the PC command is similar to that of the FA command described in the previous section. In PRELIS the syntax is

```
PC NC = k PS
```

where NC = k is used to specify that the first k principal components are to be computed. If NC = k is omitted, all principal components are computed, *i.e.*, $k = p$, where p is the number of observed variables. With PS on the PC command, the scores of the principal components are computed for each unit in the sample and saved in the file **input.PSC**, where input is the name of the PRELIS command file (without the suffix). This file can be read by PRELIS in free format.

In the LISREL command language the syntax is the same, except that PS is not allowed since individual component scores can only be computed from raw data.

In the SIMPLIS command language the syntax is

```
Principal Components with k Components
```

1. Principal Components of Five Meteorological Variables

Mardia, Kent, & Bibby (1980, p. 248) give the data on five meteorological variables presented in the table below.

Table: Raw data in five meteorological variables

Year	x_1	x_2	x_3	x_4	x_5
1920-21	87.9	19.6	1.0	1661	28.37
1921-22	89.9	15.2	90.1	968	23.77
1922-23	153.0	19.7	56.6	1353	26.04
1923-24	132.1	17.0	91.0	1293	25.74
1924-25	88.8	18.3	93.7	1153	26.68
1925-26	220.9	17.8	106.9	1286	24.29
1926-27	117.7	17.8	65.5	1104	28.00
1927-28	109.0	18.3	41.8	1574	28.37
1928-29	156.1	17.8	57.4	1222	24.96
1929-30	181.5	16.8	140.6	902	21.66
1930-31	181.4	17.0	74.3	1150	24.37

The corresponding unbiased covariance matrix is given below.

Table: Covariance matrix of five meteorological variables

x_1	x_2	x_3	x_4	x_5
1793.298				
-4.921	1.637			
799.564	-29.279	1346.859		
-2439.351	217.198	-6822.728	52914.656	
-57.214	1.735	-62.080	361.803	4.496

The variables are

x_1 = rainfall in November and December (in millimeters)

x_2 = average July temperature (in degrees Celsius)

x_3 = rainfall in July (in millimeters)

x_4 = radiation in July (in millimeters of alcohol)

x_5 = average harvest yield (in quintals per hectare)

We illustrate how the principal components can be obtained for the covariance matrix using a LISREL command file (see file **PCEX1.LIS**):

```
Example of Principal Components
DA NI = 5 NO = 11
CM
1793.298
-4.921 1.637
799.564 -29.279 1346.859
-2439.351 217.198 -6822.728 52914.656
-57.214 1.735 -62.080 361.803 4.496
PC
OU
```

The results are given in the output as follows.

```
Principal Component Analysis
```

Eigenvalues and Eigenvectors

	PC_1	PC_2	PC_3	PC_4	PC_5
Eigenvalue	53926.56	1839.57	293.11	1.22	0.48
StandError	22994.36	784.39	124.98	0.52	0.21
% Variance	96.19	3.28	0.52	0.00	0.00
Cum. % Var	96.19	99.47	100.00	100.00	100.00
VAR 1	-0.048	0.943	-0.328	0.014	-0.016
VAR 2	0.004	0.003	-0.009	0.550	0.835
VAR 3	-0.129	0.319	0.939	0.014	0.001
VAR 4	0.990	0.088	0.107	-0.005	-0.000
VAR 5	0.007	-0.023	-0.003	0.835	-0.550

Correlations between Variables and Principal Components

	PC_1	PC_2	PC_3	PC_4	PC_5
VAR 1	-0.264	0.955	-0.133	0.000	-0.000
VAR 2	0.738	0.091	-0.125	0.476	0.453
VAR 3	-0.818	0.373	0.438	0.000	0.000
VAR 4	1.000	0.016	0.008	-0.000	-0.000
VAR 5	0.750	-0.464	-0.027	0.436	-0.180

Variance Contributions

	PC_1	PC_2	PC_3	PC_4	PC_5
VAR 1	0.070	0.913	0.018	0.000	0.000
VAR 2	0.544	0.008	0.016	0.227	0.205
VAR 3	0.669	0.139	0.192	0.000	0.000
VAR 4	1.000	0.000	0.000	0.000	0.000
VAR 5	0.562	0.215	0.001	0.190	0.033

The first line gives the eigenvalues of the covariance matrix. These are the variances of the principal components. The second line gives the eigenvalues in percentage of the total variance and the third line gives these percentages cumulatively. The next five columns give the eigenvectors of the covariance matrix normalized so that their sum of squares is one and such that the largest value is positive. These eigenvectors are the coefficients (weights) in the linear combinations defining the principal components.

The next set of five columns shows the correlations between the observed variables and the principal components and the last set of five columns gives the variance contribution of each principal component to the variance of each observed variable. If all principal components are computed, as is the case here, these variance contributions sum to one row-wise.

The first two principal components account for 99.47% of the variance. The first component is dominated by the radiation variable x_4. The second component is a rainfall variable dominated by x_1 and x_3.

To analyze the correlation matrix instead of the covariance matrix of the five meteorological variables, just add MA = KM on the DA command in **PCEX1.LIS** (see file **PCEX2.LIS**). We leave it to the reader to verify that this gives a very different result and interpretation.

One can also compute the component scores, *i.e.*, the scores on the linear combinations. For this, one must use PRELIS and the raw data. File **PCEX3.PRL** illustrates how to obtain the component scores of the first two principal components:

```
Principal Components with Components Scores
DA NI=5
RA=PCEX.RAW
CO ALL
PC NC=2 PS OU MA=CM
```

The eigenvalues, eigenvectors and correlations between variables and principal components are given in the output file.

Principal Component Analysis

 Eigenvalues and Eigenvectors

	PC_1	PC_2
Eigenvalue	53927.94	1999.96
StandError	22994.95	852.79
% Variance	95.89	3.56
Cum. % Var	95.89	99.44
VAR 1	-0.048	0.954
VAR 2	0.004	0.003
VAR 3	-0.129	0.288
VAR 4	0.990	0.084

```
    VAR 5        0.007      -0.021
```

Correlations between Variables and Principal Components

```
                 PC_1        PC_2
                --------    --------
    VAR 1       -0.254       0.960
    VAR 2        0.738       0.091
    VAR 3       -0.818       0.351
    VAR 4        1.000       0.016
    VAR 5        0.750      -0.443
```

Variance Contributions

```
                 PC_1        PC_2
                --------    --------
    VAR 1        0.064       0.922
    VAR 2        0.544       0.008
    VAR 3        0.669       0.123
    VAR 4        1.000       0.000
    VAR 5        0.562       0.196
```

The component scores are given in the file **PCEX3.PSC**:

```
     423.256      -6.692
    -275.908     -12.942
     111.016       8.089
      49.359      20.667
     -86.970     -38.177
      38.374      66.545
    -131.071    -101.073
     335.221      -4.458
     -20.796       2.797
    -348.384      55.395
     -94.095       9.848
```