Principal components: analysis of meteorological data

Contents

1. Introduction

PCA is used in exploratory data analysis and for making predictive models. It is commonly used for dimensionality reduction by projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible.

The principal components are eigenvectors of the data's covariance matrix. Thus, the principal components are often computed by eigen decomposition of the data covariance matrix or singular value decomposition of the data matrix. PCA is the simplest of the true eigenvector-based multivariate analyses and is closely related to factor analysis. Factor analysis typically incorporates more domain specific assumptions about the underlying structure and solves eigenvectors of a slightly different matrix. It should also be noted that factor analysis is a model which can be tested.

In this example, data from Mardia, Kent, and Bibby (1980) is used to illustrate how principal component analysis can be performed using LISREL. The data consists of annual measurements over a period of 11 years of five meteorological characteristics:

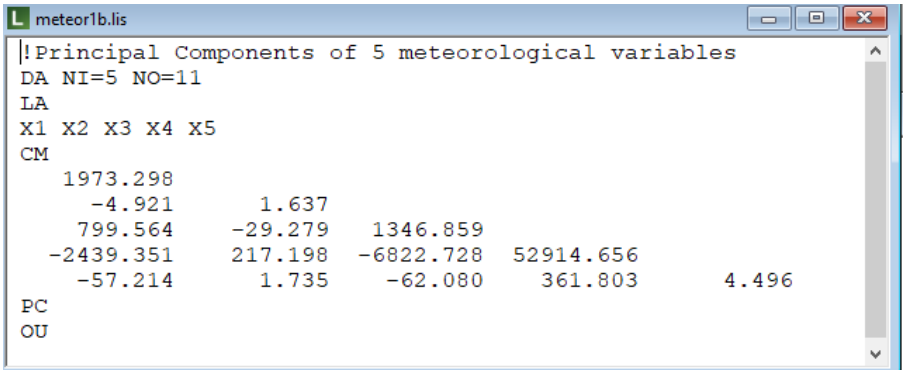
- x_1 : the rainfall, measured in millimeters, during November and December
- x_2 : the average July temperature (in degrees Celsius)
- x_3 : the rainfall during July, again measured in millimeters
- x_4 : the radiation in July (in millimeters of alcohol)
- x_5 : the average harvest yield (in quintals per hectare, a quintal being a unit of weight equal to 100 kg.)

The covariance matrix of these five variables is used as input for the principal component analysis and is shown below.

$$\begin{bmatrix} 1973.298 & & & & \\ -4.921 & 1.637 & & & \\ 799.564 & -29.279 & 1346.859 & & \\ -2439.351 & 217.198 & -6822.728 & 52914.656 & \\ -57.214 & 1.735 & -62.080 & 361.803 & 4.496 \end{bmatrix}$$

1. PCA based on the covariance matrix

The PC command is used to request a principal component analysis, as shown in the syntax file below. Note that all files used here can be found in the **MVABOOK\Chapter5** folder. The number of variables and number of observations are specified on the DA line using the NI and NO keywords.



In the first section of output obtained, the smallest and largest eigenvalues are reported.

```
Total Variance = 56240.946 Generalized Variance = 0.220269D+11

 Largest Eigenvalue = 53927.942 Smallest Eigenvalue = 0.521

 Condition Number = 321.814
```

In the next section, the first information reported is the eigenvalues of the covariance matrix., along with their standard errors. Note that, due to our small sample size of 11 here, these standard errors are not reliable. The second part of the table contains the eigenvectors of the covariance matrix, normalized so that their sum of squares is equal to 1 and also so that the largest value is positive. These eigenvectors serve as the coefficients in the linear components that define the principal components.

```
 Eigenvalues and Eigenvectors

                PC_1        PC_2        PC_3        PC_4        PC_5
             --------    --------    --------    --------    --------
Eigenvalue   53927.94     1999.96      311.26        1.26        0.52
StandError   22994.95      852.79      132.72        0.54        0.22
% Variance      95.89        3.56        0.55        0.00        0.00
Cum. % Var      95.89       99.44      100.00      100.00      100.00
             --------    --------    --------    --------    --------
       X1      -0.048       0.954      -0.296       0.013      -0.013
       X2       0.004       0.003      -0.008       0.508       0.861
       X3      -0.129       0.288       0.949       0.016      -0.001
```

```
X4      0.990      0.084      0.109     -0.005     -0.001
X5      0.007     -0.021     -0.008      0.861     -0.508
```

This is followed by the correlations between the principal components and the observed variables.

```
Correlations between Variables and Principal Components

            PC_1       PC_2       PC_3       PC_4       PC_5
        --------   --------   --------   --------   --------
X1      -0.254      0.960     -0.118      0.000     -0.000
X2       0.738      0.091     -0.115      0.446      0.486
X3      -0.818      0.351      0.456      0.000     -0.000
X4       1.000      0.016      0.008     -0.000     -0.000
X5       0.750     -0.443     -0.065      0.456     -0.173
```

The last section of output gives the variance contributions of each principal component to the variance of each of the observed variables. We see that the last 3 principal components do not contribute much to the variance in the observed variables, in contrast with the first two principal components. The first principal component contributes greatly to the variance in x_4, while the second contributes greatly to the variance in x_1 and to a lesser extent to the variance in x_3. Recall that these two variables were measures of rainfall, so we see that the second principal component as a "rainfall" component.

```
Variance Contributions

            PC_1       PC_2       PC_3       PC_4       PC_5
        --------   --------   --------   --------   --------
X1      0.064      0.922      0.014      0.000      0.000
X2      0.544      0.008      0.013      0.199      0.236
X3      0.669      0.123      0.208      0.000      0.000
X4      1.000      0.000      0.000      0.000      0.000
X5      0.562      0.196      0.004      0.208      0.030
```

2. PCA based on the correlation matrix

The principal components analysis can also be based on the correlation matrix. The correlation matrix and syntax for this analysis is shown below. Note the addition of the keyword MA = KM on the DA command.

meteor2b.lis

```
!PComponents of 5 meteorological variables
DA NI=5 NO=11 MA=KM
LA
X1 X2 X3 X4 X5
CM
   1973.298
     -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 output for this analysis is as follows:

```
Total Variance = 5.000 Generalized Variance = 0.0213
```

Largest Eigenvalue = 3.399 Smallest Eigenvalue = 0.130

Condition Number = 5.110

Principal Component Analysis

Eigenvalues and Eigenvectors

	PC_1	PC_2	PC_3	PC_4	PC_5
Eigenvalue	3.40	1.02	0.29	0.16	0.13
StandError	1.45	0.43	0.13	0.07	0.06
% Variance	67.97	20.31	5.90	3.21	2.60
Cum. % Var	67.97	88.29	94.18	97.40	100.00
X1	-0.293	0.809	-0.231	0.173	-0.419
X2	0.423	0.482	0.677	-0.338	0.123
X3	-0.500	0.040	0.505	0.585	0.389
X4	0.483	0.286	-0.433	0.365	0.604
X5	0.502	-0.170	0.214	0.617	-0.541

Correlations between Variables and Principal Components

	PC_1	PC_2	PC_3	PC_4	PC_5
X1	-0.540	0.816	-0.125	0.070	-0.151
X2	0.780	0.486	0.368	-0.136	0.044
X3	-0.921	0.040	0.274	0.234	0.140
X4	0.890	0.288	-0.235	0.146	0.218
X5	0.926	-0.171	0.116	0.247	-0.195

When we compare the correlations between variables and principal components over the two analyses, we note that the fourth and fifth principal components have a larger correlation with x_2 in the case of the covariance matrix analysis. While the second principal component is still highly correlation with x_1, the second analysis shows the next highest correlation for this component to be with x_2, in contrast with x_3 previously. The first principal component can be seen as a contrast between x_1 and x_3 on the one hand and the other three variables on the other. It is clear that using a covariance matrix vs a correlation matrix in this analysis delivers very different results.