Principal components: stock market 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, we use data on stock market prices (Johnson and Wichern, 2002). Data are available for five oil companies trading on the New York Stock Exchange, these being Allied Chemical, Du Pont, Union Carbide, Exxon and Texaco. For each of these companies the percentage weekly rate of returns for a period of 100 weeks are available. Data are available in the file **smp.lsf**. Data can be found in the **MVABOOK\Chapter 5** folder. The first 20 lines of this file are shown below.

smp.lsf

	AllChem	DuPoint	UnCarbid	Exxon	Texaco
1	0.00	0.00	0.00	3.95	0.00
2	2.70	-4.49	-0.30	-1.45	4.35
3	12.28	6.08	8.81	8.62	7.81
4	5.70	2.99	6.68	1.35	1.95
5	6.37	-0.38	-3.98	-1.86	-2.42
6	0.35	5.08	8.29	7.43	4.95
7	-4.56	-3.30	0.26	-0.96	-2.83
8	5.88	4.17	8.14	-1.46	1.46
9	0.00	-1.94	0.24	0.16	-2.87
10	0.69	-2.60	0.70	-4.11	-2.46
11	1.03	0.64	8.39	1.03	0.00
12	-3.07	2.02	-4.09	-3.90	-5.05
13	-0.35	11.88	8.97	6.01	2.13
14	6.01	7.96	2.88	3.67	2.60
15	-0.33	-0.10	2.80	2.89	-1.02
16	5.56	9.13	4.28	5.94	-1.58
17	5.13	-0.75	-4.14	-1.63	5.85
18	-6.10	-4.36	2.36	0.46	-1.51
19	-3.57	1.82	-2.11	-0.76	-1.02
20	0.00	-2.16	-0.78	8.85	8.25

2. Principal component analysis

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. We use PRELIS syntax for this analysis. Note that we use the covariance matrix as input (MA = CM) and do not specify the number of components.

pcnpv1.prl

```
SY=pasteur_npv.lsf
PC
OU MA=CM
```

```
 Univariate Summary Statistics for Continuous Variables

 Variable      Mean  St. Dev.   Skewness  Kurtosis  Minimum Freq.  Maximum Freq.
 --------      ----  --------   -------   --------  ------- -----  ------- -----
  AllChem     0.543     4.037      0.206     0.027   -9.665     1   12.281     1
  DuPoint     0.483     3.506      0.564     0.499   -7.576     1   11.881     1
UnCarbid      0.565     3.945      0.157     0.107   -9.146     1   10.262     1
   Exxon      0.629     2.833      0.631     0.520   -5.313     1    8.855     1
  Texaco      0.371     2.755      0.519     0.332   -5.050     1    8.247     1

 Test of Univariate Normality for Continuous Variables

              Skewness          Kurtosis       Skewness and Kurtosis

 Variable Z-Score P-Value   Z-Score P-Value   Chi-Square P-Value

  AllChem    0.876   0.381     0.245   0.806        0.827   0.661
  DuPoint    2.282   0.022     1.106   0.269        6.433   0.040
UnCarbid     0.667   0.505     0.414   0.679        0.616   0.735
```

Exxon	2.524	0.012	1.138	0.255	7.664	0.022
Texaco	2.118	0.034	0.833	0.405	5.180	0.075

The univariate statistics is followed by eigenvalues and eigenvectors of the components. The first component contributes the most by far at 57.13 %. Using Kaiser's rule of only retaining those components with an eigenvalue larger than 1, this indicates that we should only retain the first principal component.

Eigenvalues and Eigenvectors

	PC_1	PC_2	PC_3	PC_4	PC_5
Eigenvalue	2.86	0.81	0.54	0.45	0.34
StandError	0.40	0.11	0.08	0.06	0.05
% Variance	57.13	16.18	10.80	9.03	6.86
Cum. % Var	57.13	73.31	84.11	93.14	100.00
AllChem	0.464	-0.241	0.613	-0.381	-0.453
DuPoint	0.457	-0.509	-0.178	-0.211	0.675
UnCarbid	0.470	-0.261	-0.337	0.664	-0.396
Exxon	0.422	0.525	-0.539	-0.473	-0.179
Texaco	0.421	0.582	0.434	0.381	0.387

CALL PMSE,FMT1 (1H ,A8,10F11.3)

Correlations between Variables and Principal Components

	PC_1	PC_2	PC_3	PC_4	PC_5
AllChem	0.783	-0.217	0.451	-0.256	-0.265
DuPoint	0.773	-0.458	-0.131	-0.142	0.395
UnCarbid	0.794	-0.234	-0.248	0.446	-0.232
Exxon	0.713	0.472	-0.396	-0.318	-0.105
Texaco	0.712	0.524	0.319	0.256	0.227

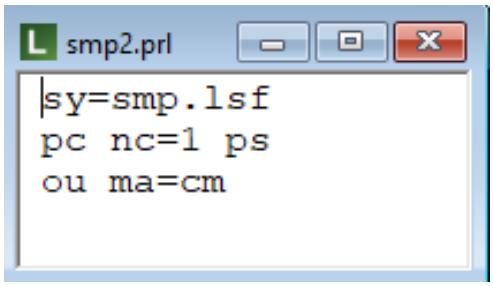
CALL PMSE,FMT1 (1H ,A8,10F11.3)

Variance Contributions

	PC_1	PC_2	PC_3	PC_4	PC_5
AllChem	0.614	0.047	0.203	0.066	0.070
DuPoint	0.597	0.210	0.017	0.020	0.156
UnCarbid	0.631	0.055	0.061	0.199	0.054
Exxon	0.508	0.223	0.157	0.101	0.011
Texaco	0.507	0.274	0.102	0.066	0.051

3. Principal scores

The modified syntax file shown below requests the estimation of only one principal component (NC = 1) and also the calculation of the principal scores via the use of the PS option.

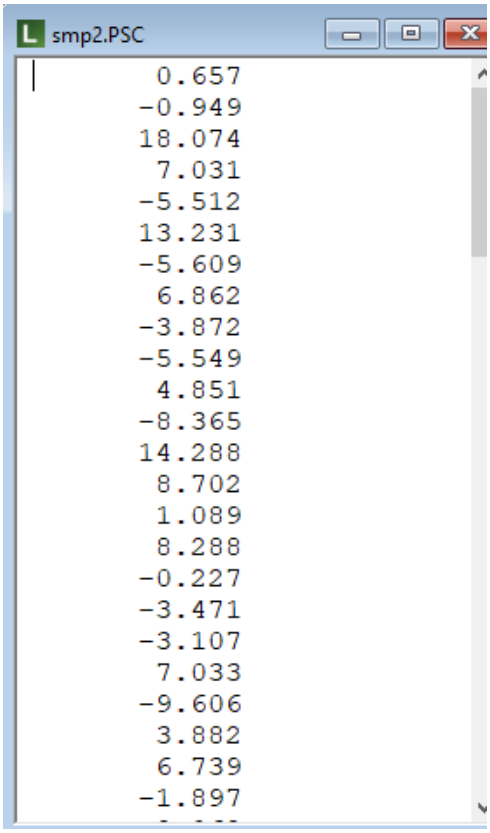


The standard LISREL output file reports the eigenvalue and contributions of the principal component as:

```
Eigenvalues and Eigenvectors

               PC_1
            --------
Eigenvalue     35.95
StandError      5.08
% Variance     60.16
Cum. % Var     60.16
            --------
  AllChem      0.561
  DuPoint      0.470
 UnCarbid      0.547
    Exxon      0.291
   Texaco      0.284
 CALL PMSE,FMT1    (1H ,A8,10F11.3)

 Correlations between Variables and Principal Components

               PC_1
            --------
  AllChem      0.833
  DuPoint      0.804
 UnCarbid      0.832
    Exxon      0.616
   Texaco      0.619
 CALL PMSE,FMT1    (1H ,A8,10F11.3)

 Variance Contributions

               PC_1
            --------
  AllChem      0.693
  DuPoint      0.646
 UnCarbid      0.692
    Exxon      0.379
   Texaco      0.383
```

In addition, the use of the PS option invokes the creation of a text file containing the principal component scores. The filename of the principal score file will correspond to that of the syntax file, and the file extension for principal scores files

is *.**fsc**. The first few lines of **smp2.fsc** are shown below. As this is a simple text file, importing it into an LSF file for further use in LISREL is easy.



For information on how to use these scores to compute a price index for the five oil companies we have data for, the reader is referred to the *Multivariate Analysis with LISREL* text.