Principal components: nine psychological variables

Contents

1. Introduction

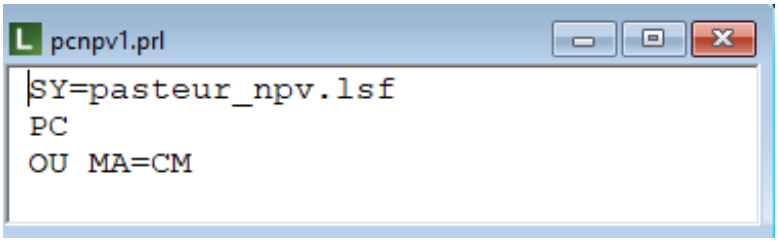
PCA is used in exploratory data analysis and for making predictive models. It is commonly used for dimensionality reduction by projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible.

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 nine psychological variables for students from the Pasteur school data. Data are available on nine selected tests. The nine tests are Visual Perception, Cubes, Paper Form Board, General Information, Sentence Completion, Word Classification, Figure Recognition, Object-Number, and Number-Figure. Given the number of variables of interest here, it is more convenient to use an LSF file if the data are available in this format. In this case, data are available in the file **Pasteur_npv.lsf**.

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. We use PRELIS syntax for this analysis.



Basic descriptive statistics for the nine tests are given first.

Univariate Summary Statistics for Continuous Variables

Variable	Mean	St. Dev.	Skewness	Kurtosis	Minimum	Freq.	Maximum	Freq.
VISPERC	29.647	7.110	-0.378	0.736	4.000	1	45.000	2
CUBES	23.936	4.921	0.695	0.255	14.000	1	37.000	3
LOZENGES	19.897	9.311	0.156	-1.074	2.000	1	36.000	5
PARCOMP	8.468	3.457	0.221	-0.052	0.000	1	18.000	1
SENCOMP	15.981	5.244	-0.132	-0.839	4.000	1	27.000	1
WORDMEAN	13.455	6.932	1.009	2.034	1.000	2	43.000	1
ADDITION	101.942	24.958	0.303	-0.395	47.000	1	171.000	1
COUNTDOT	111.263	19.577	0.362	0.091	70.000	1	166.000	1
SCCAPS	195.038	35.708	0.226	0.194	100.000	1	310.000	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
VISPERC	-1.934	0.053	1.682	0.093	6.568	0.037
CUBES	3.360	0.001	0.782	0.434	11.900	0.003
LOZENGES	0.818	0.414	-5.966	0.000	36.257	0.000
PARCOMP	1.153	0.249	0.018	0.985	1.329	0.514
SENCOMP	-0.691	0.489	-3.585	0.000	13.332	0.001
WORDMEAN	4.555	0.000	3.207	0.001	31.028	0.000
ADDITION	1.567	0.117	-1.131	0.258	3.736	0.154
COUNTDOT	1.856	0.064	0.398	0.691	3.601	0.165
SCCAPS	1.176	0.240	0.645	0.519	1.798	0.407

This is followed by the covariance matrix that forms the basis of the Principal Component Analysis. We note that the variable SCCAPS has the largest variation at 1275, followed by ADDITION with a variance of 622.881. The smallest variance is for the variable PARCOMP at 11.954.

Covariance Matrix

	VISPERC	CUBES	LOZENGES	PARCOMP	SENCOMP	WORDMEAN
VISPERC	50.552					
CUBES	9.713	24.215				
LOZENGES	29.957	15.355	86.686			
PARCOMP	10.289	1.069	3.977	11.954		
SENCOMP	11.400	2.296	2.140	13.041	27.503	
WORDMEAN	21.032	5.468	12.518	15.960	26.318	48.056
ADDITION	6.412	-18.875	-4.574	22.330	15.825	35.504

COUNTDOT	19.951	3.101	24.414	8.876	12.199	28.976
SCCAPS	76.336	31.693	96.256	16.866	30.207	47.221

Covariance Matrix

	ADDITION	COUNTDOT	SCCAPS
ADDITION	622.881		
COUNTDOT	197.267	383.253	
SCCAPS	239.357	255.383	1275.082

The first principal components account for only 57.6% of the total variance. Even when the first three components are looked at together, they only account for 91% of the total variance. These results are not particularly informative, especially if we note that the first component is essentially equal to the variable with the largest variance (SCCAPS) and the last with the variable with the smallest variance (PARCOMP). The rest of the components follow in this pattern.

Eigenvalues and Eigenvectors

	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6
Eigenvalue	1457.41	587.13	259.49	102.66	63.99	26.54
StandError	165.02	66.48	29.38	11.62	7.25	3.01
% Variance	57.60	23.21	10.26	4.06	2.53	1.05
Cum. % Var	57.60	80.81	91.06	95.12	97.65	98.70
VISPERC	0.057	-0.038	0.035	0.487	-0.119	0.850
CUBES	0.018	-0.052	0.031	0.173	0.048	0.023
LOZENGES	0.069	-0.072	0.068	0.679	0.621	-0.333
PARCOMP	0.019	0.026	-0.008	0.175	-0.239	-0.042
SENCOMP	0.027	0.010	0.008	0.225	-0.471	-0.256
WORDMEAN	0.046	0.036	0.033	0.411	-0.559	-0.313
ADDITION	0.326	0.853	-0.395	0.047	0.073	0.018
COUNTDOT	0.278	0.317	0.902	-0.089	0.007	0.010
SCCAPS	0.897	-0.401	-0.145	-0.107	-0.022	-0.014

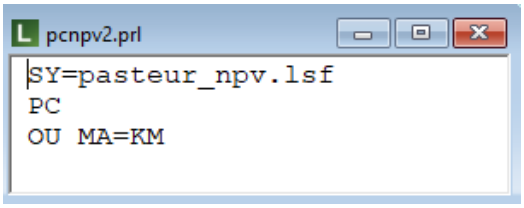
Eigenvalues and Eigenvectors

	PC_7	PC_8	PC_9
Eigenvalue	19.50	9.55	3.91
StandError	2.21	1.08	0.44
% Variance	0.77	0.38	0.15
Cum. % Var	99.47	99.85	100.00
VISPERC	-0.102	0.006	-0.096
CUBES	0.979	0.069	0.024
LOZENGES	-0.154	0.058	-0.013
PARCOMP	-0.062	0.322	0.895
SENCOMP	-0.048	0.687	-0.433
WORDMEAN	0.008	-0.645	-0.013
ADDITION	0.039	0.008	-0.024
COUNTDOT	-0.003	0.016	0.007
SCCAPS	-0.012	-0.008	0.008

In this case, it would be better to use the correlation matrix as input instead.

2. PCA based on the correlation matrix

The syntax for the principal components analysis based on the correlation matrix remains almost the same, except that the keyword MA = KM on the DA command replaces the previously used MA = CM.



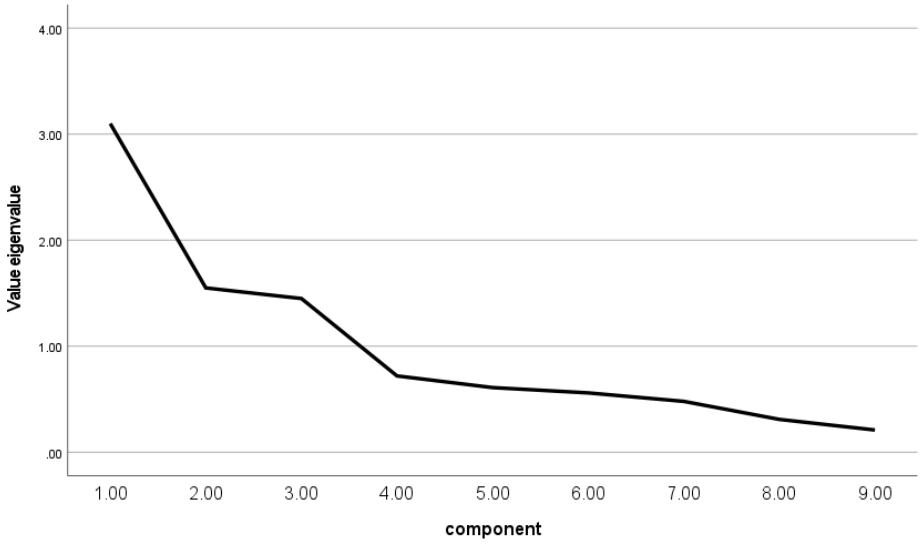
The output for this analysis is as follows:

```
Eigenvalues and Eigenvectors

                PC_1        PC_2        PC_3        PC_4        PC_5        PC_6
             --------    --------    --------    --------    --------    --------
Eigenvalue       3.10        1.55        1.45        0.72        0.61        0.56
StandError       0.35        0.18        0.16        0.08        0.07        0.06
% Variance      34.40       17.25       16.15        7.99        6.79        6.25
Cum. % Var      34.40       51.65       67.80       75.79       82.58       88.83
             --------    --------    --------    --------    --------    --------
  VISPERC        0.380       0.301      -0.114      -0.428       0.004      -0.062
    CUBES        0.168       0.511      -0.175       0.711      -0.233       0.317
 LOZENGES        0.238       0.546       0.018      -0.433      -0.262       0.010
  PARCOMP        0.448      -0.280      -0.200      -0.085      -0.046       0.135
  SENCOMP        0.429      -0.302      -0.265       0.185       0.176      -0.121
 WORDMEAN        0.470      -0.183      -0.179       0.086      -0.081      -0.098
 ADDITION        0.197      -0.289       0.549      -0.090      -0.286       0.659
 COUNTDOT        0.229      -0.006       0.576       0.239      -0.341      -0.642
   SCCAPS        0.272       0.252       0.425       0.100       0.800       0.082

 Eigenvalues and Eigenvectors

                PC_7        PC_8        PC_9
             --------    --------    --------
Eigenvalue       0.48        0.31        0.21
StandError       0.05        0.04        0.02
% Variance       5.33        3.47        2.37
Cum. % Var      94.16       97.63      100.00
             --------    --------    --------
  VISPERC        0.722       0.104       0.181
    CUBES        0.127      -0.061       0.009
 LOZENGES       -0.614      -0.076       0.080
  PARCOMP        0.022      -0.662      -0.463
  SENCOMP       -0.214      -0.077       0.724
 WORDMEAN       -0.154       0.706      -0.410
 ADDITION        0.042       0.123       0.191
 COUNTDOT        0.084      -0.149      -0.004
   SCCAPS       -0.084      -0.013      -0.132
```

When we compare the correlations between variables and principal components over the two analyses, we note that we again have no single component contributing the most to the total variance. The question now becomes how many of these components we should retain, and what to base that decision on.

3. Number of components

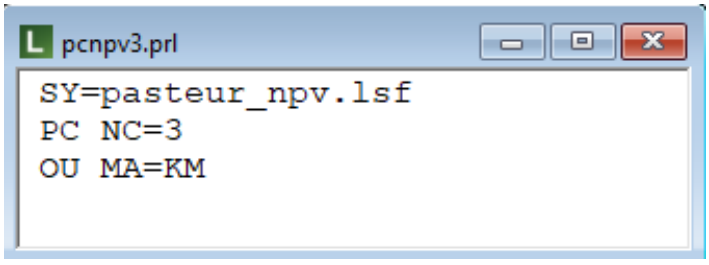
When we plot the eigenvalues associated with each component in order of size, the following graph is obtained. We hope to use this as a visual aid to retain only those principal components which are statistically different. In other words, components for which the largest eigenvalues are statistically different.



The graph shows no "elbow' in the plot that might suggest a good cutting off point for the number of components. We thus opt for another approach: retaining only components for which the eigenvalues of the correlation matrix are larger than one. The rationale here is that principal components with a variance smaller than 1 will explain less variance than any variable, as all variables have a variance of 1 in a correlation matrix.

Revisiting the results obtained in the previous section, we note that according to this rule, only the first three principal components qualify for selection.

We amend our syntax file to request the estimation of only three principal components by adding the keyword NC = 3 on the PC line:



For this analysis, the following output is obtained:

Eigenvalues and Eigenvectors

	PC_1	PC_2	PC_3
Eigenvalue	3.10	1.55	1.45
StandError	0.35	0.18	0.16
% Variance	34.40	17.25	16.15
Cum. % Var	34.40	51.65	67.80
VISPERC	0.380	0.301	-0.114
CUBES	0.168	0.511	-0.175
LOZENGES	0.238	0.546	0.018
PARCOMP	0.448	-0.280	-0.200
SENCOMP	0.429	-0.302	-0.265
WORDMEAN	0.470	-0.183	-0.179
ADDITION	0.197	-0.289	0.549
COUNTDOT	0.229	-0.006	0.576
SCCAPS	0.272	0.252	0.425

Correlations between Variables and Principal Components

	PC_1	PC_2	PC_3
VISPERC	0.668	0.375	-0.137
CUBES	0.296	0.637	-0.212
LOZENGES	0.419	0.681	0.021
PARCOMP	0.788	-0.348	-0.241
SENCOMP	0.755	-0.376	-0.319
WORDMEAN	0.827	-0.228	-0.216
ADDITION	0.347	-0.360	0.661
COUNTDOT	0.404	-0.007	0.695
SCCAPS	0.479	0.314	0.513

Variance Contributions

	PC_1	PC_2	PC_3
VISPERC	0.446	0.140	0.019
CUBES	0.088	0.406	0.045
LOZENGES	0.176	0.464	0.000
PARCOMP	0.621	0.121	0.058
SENCOMP	0.569	0.142	0.102
WORDMEAN	0.684	0.052	0.047
ADDITION	0.120	0.129	0.437
COUNTDOT	0.163	0.000	0.482
SCCAPS	0.229	0.099	0.263