Submitted by-Aditya Gautam

Roll No-12

M.sc. 3rd semester

Date of Assignment-24/12/2020

Date of Submission-05/01/2021

Experiment No-10

Topic- Principle Component Analysis

Problem- A census provided information on 5 socio-economic variables for a place in the United States, which is as shown below-

SL no	Total population('000)	Median school years	Total employment('000)	Health service employment	Median home value (in 1000$)
1	5.935	14.2	2.265	2.27	2.91
2	1.583	13.1	0.597	0.75	2.62
3	2.599	12.7	1.237	1.11	1.72
4	4.009	15.2	1.649	0.81	3.02
5	4.687	14.7	2.312	2.50	2.22
6	8.044	15.6	3.641	4.51	2.36
7	2.766	13.3	1.294	1.03	1.97
8	6.538	17.0	2.618	2.39	1.85
9	6.451	12.9	3.147	5.52	2.01
10	3.314	12.2	1.606	2.18	1.82
11	3.777	13.2	2.119	2.82	1.80
12	1.530	13.8	0.798	0.84	4.25
13	2.768	13.6	1.336	1.75	2.64
14	6.585	14.9	2.763	1.91	3.7

Can the sample variance be summarized by two or more principal components?

Theory-

Principal component analysis is a data reduction technique, which is concerned with the variance covariance structure of a set of large no of variables through a few linear combination of these variables.

Often, much of the variability in the data on p-variables, can be accounted for by a small no k(k<p)

of the principal components. In such a case, the k-components contains as much information as there is in the original data of n measurements on p-variables. Thus, the data set is reduced to one containing n measurements on k principal components. Algebraically, principal components are particular linear combinations of the p random variable $X_1,X_2,\ldots\ldots..X_p$.

Suppose ,we have the data on p variables ($X'=(X_1,X_2,\ldots\ldots..X_p)$) from some p-dimensional population with mean vector $\underset{\sim}{\mu}$ & covariance matrix $\sum$. Then, we have from the sample of n-values, the sample variance covariance matrix $S=\left(s_{ij}\right)_{pxp}$, where,

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_{ik}-\bar{X}_i\right)\left(X_{jk}-\bar{X}_j\right) \qquad \text{i,j=1,2,.....p.}$$

Consider, the linear combinations $y_i = \underset{\sim i}{a'}\underset{\sim}{X} = \begin{pmatrix} a_{i1} & a_{i2}\cdots & a_{ip}\end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$

$$= a_{i1}x_1 + a_{i2}x_2 + \ldots\ldots\ldots + a_{ip}x_p\,, \quad \text{i=1,2,.....p}$$

The sample principal components are defined as those uncorrelated which have the maximum sample variance i.e which maximizes var $\left(\underset{\sim i}{a'},\underset{\sim i}{x}\right)$

Further , suppose that the Eigen value of S are $\hat{\lambda}_1,\ldots\ldots,\hat{\lambda}_p$. Then, the proportion of variation explained by the i^{th} sample principal component is given by $\dfrac{\hat{\lambda}_i}{\sum_{i=1}^{p}\hat{\lambda}_i}$,i=1,2…,p

If most of the total population variance say 80-90% can be attributed to the 1st random principal components $(1 \le r \le p)$, then these random components can replace the original p-variables without much loss of information. Therefore, the principal components to be retained becomes-

$$\underset{\sim}{y}_i = \underset{\sim}{e}_i^{/} \underset{\sim}{X} = \begin{pmatrix} \hat{e}_{i1} & \hat{e}_{i2} & \dots & \hat{e}_{ip} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$$

where $\underset{\sim}{e}_i^{/}$ is the Eigen vector corresponding to the Eigen value $\hat{\lambda}_i$.

Calculation-

The following R-program is used to obtain a solution of the given problem-

```
x1=c(5.935,1.583,2.599,4.009,4.687,8.044,2.766,6.538,6.451,3.314,3.777,1.530,2.768,6.585)
x2=c(14.2,13.1,12.7,15.2,14.7,15.6,13.3,17.0,12.9,12.2,13.2,13.8,13.6,14.9)
x3=c(2.265,0.597,1.237,1.649,2.312,3.641,1.294,2.618,3.147,1.606,2.119,0.798,1.336,2.763)
x4=c(2.27,0.75,1.11,0.81,2.50,4.51,1.03,2.39,5.52,2.18,2.82,0.84,1.75,1.91)
x5=c(2.91,2.62,1.72,3.02,2.22,2.36,1.97,1.85,2.01,1.82,1.80,4.25,2.64,3.7)
x=array(c(x1,x2,x3,x4,x5),dim=c(14,5))
s=mat.or.vec(5,5)
for(i in 1:5){
for(j in 1:5){
s[i,j]=cov(x[,i],x[,j])}}
s
e_val=eigen(s)$values
e_val
e_vec=eigen(s)$vectors
e_vec
p=mat.or.vec(5,1)
```

```
for(i in 1:5){
p[i]=e_val[i]/sum(e_val)}
p
prop=cumsum(p)
prop
```

Conclusion-

Proportion of variation explained by the 1st sample principal component $= \frac{\lambda_1}{\sum \lambda_i} = 0.735376708$

Proportion of variation explained by the 2nd sample principal component $= \frac{\lambda_2}{\sum \lambda_i} = 0.190082087$

Since, the 1st two sample principal components can very well summarize the total sample variance (about 92%) of the total variance, a reduction in the data from the 14 observations on 5 variables to 14 observations on 2 principal components is reasonable.

The sample principal components are–

$$y_1 = \begin{pmatrix} 0.7809593 & 0.3056211 & 0.3343107 & 0.4280031 & -0.0418096 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix}$$

$$y_2 = \begin{pmatrix} -0.07598093 & -0.74792486 & 0.07982963 & 0.58087464 & -0.30174026 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix}$$