

Chapter X**Principal Component Analysis****Susheel Kumar Sarkar**ICAR- Indian Agricultural Statistics Research Institute
New Delhi-110012Email: Susheel.sarkar@icar.gov.in

Most of the times the variables under study are highly correlated and as such they are effectively “saying the same thing”. To examine the relationships among a set of p correlated variables, it may be useful to transform the original set of variables to a new set of uncorrelated variables called principal components. These new variables are linear combinations of original variables and are derived in decreasing order of importance in terms of accounts for variation. Principal component analysis is an advantageous statistical procedure that has found applications in various fields. It helps in reducing the dimensionality of the data set and identification of patterns in the data through graphical illustration. In nutshell, we can say that principal component analysis is a data reduction procedure.

The purpose of principal component analysis is to derive a small number of linear combinations (principal components) of a set of variables that retain as much information in the original variables as possible. Often a small number of principal components can be used in place of the original variables for plotting, regression, clustering and so on. Principal component analysis can also be viewed as a technique to remove multicollinearity in the data. In this technique, we transform the original set of variables to a new set of uncorrelated random variables, called principal components. These new variables are linear combinations of the originals variables and are derived in decreasing order of importance so that the first principal component accounts for as much as possible of the variation in the original data. The researcher often hopes that the first few components will account for most of the variation in the original data so that the effective dimensionality of the data can be reduced (Abdi and Williams, 2010).

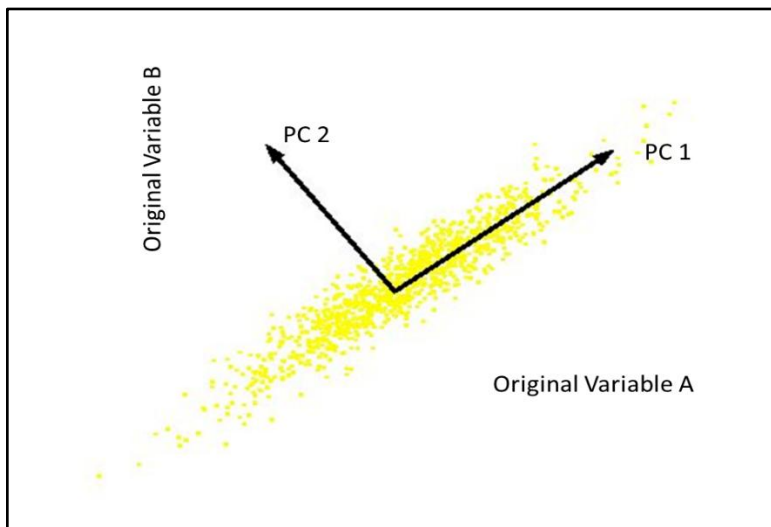
Principal component analysis often reveals relationships that were not previously suspected and thereby allows interpretations that would not ordinarily result. Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system with $x_1, x_2, \dots, x_p$ as the coordinate axes. The new axes represent the directions with maximum variability and provide a simple and more parsimonious description of the covariance structure.

Application of PCA

- Reduction of data
- Plotting of data in two or three dimensions
- Help in analysis where no. of variables is more than the number of observations
- Help in solve the problem due to multicollinearity

Orthogonal directions of greatest variance in data

Projections along PCs discriminate the data most along any one axis that ensures independence among variables.



Computing the Components

Data points are vectors in a multidimensional space. Projection of vector x onto an axis (dimension) u is $u \cdot x$. Direction of greatest variability is that in which the average square of the projection is

greatest i.e. u such that $E((u \cdot x)^2)$ over all x is maximized (we subtract the mean along each dimension, and center the original axis system at the centroid of all data points, for simplicity). This direction of u is the direction of the first Principal Component.

Geometry of PCA

S.No	X1	Mean corrected	X2	Mean corrected
1	16	8	8	5
2	12	4	10	7
3	13	5	6	3
4	11	3	2	-1
5	10	2	8	5
6	9	1	-1	-4
7	8	0	4	1
8	7	-1	6	3
9	5	-3	-3	-6
10	3	-5	-1	-4
11	2	-6	-3	-6
12	0	-8	0	-3
Mean	8	0	3	0
Variance	23.09	23.09	21.09	21.09091

From above table sum of square and cross product matrix can be calculated as under

$$SSCA = \begin{bmatrix} 254 & 181 \\ 181 & 232 \end{bmatrix}$$

If we divide all elements of this matrix by $12-1=11$ we get Variance co-variance matrix

$$S = \begin{bmatrix} 23.091 & 16.455 \\ 16.455 & 21.091 \end{bmatrix}$$

And finally, we can have correlation matrix

$$R = \begin{bmatrix} 1.000 & 0.746 \\ 0.746 & 1.000 \end{bmatrix}$$

Rotating 10 degree, we get

Observations	Mean corrected x_1^*	x_1^*
1	8	8.75
2	4	5.158
3	5	5.447
4	3	2.781
5	2	2.84
6	1	0.289
7	0	0.174
8	-1	-0.463
9	-3	-3.999
10	-5	-5.621
11	-6	-6.954
12	-8	-8.402
Mean	0	
Variance		28.659

Optimum rotation (43.261)

Angle with x_1	Total variance	variance of x_1^*	percent
0	44.18	23.09	52.26
10	44.18	28.65	64.86
20	44.18	33.43	75.67
30	44.18	36.84	83.38
40	44.18	38.46	87.06
43	44.18	38.57	87.31
50	44.18	38.12	86.28
60	44.18	35.84	81.12
70	44.18	31.90	72.20
80	44.18	26.77	60.61
90	44.18	21.09	47.73

Mean corrected values

	Mean corrected x_1	Mean corrected x_2	x_1^*	x_2^*
	8	5	9.25	-1.84
	4	7	7.71	2.36
	5	3	5.69	-1.24
	3	-1	1.50	-2.78
	2	5	4.88	2.27

	1	-4	-2.01	-3.59
	0	1	0.68	0.73
	-1	3	1.32	2.87
	-3	-6	-6.29	-2.31
	-5	-4	-6.38	0.51
	-6	-6	-8.48	-0.26
	-8	-3	-7.88	3.29
Mean	0	0	0	0
SS			424.33	61.67
Variance			38.576	5.606

$$SSCA = \begin{bmatrix} 424.334 & 0.000 \\ 0.000 & 61.666 \end{bmatrix}$$

$$S = \begin{bmatrix} 38.576 & 0.000 \\ 0.000 & 5.606 \end{bmatrix}$$

$$R = \begin{bmatrix} 1.000 & 0.000 \\ 0.000 & 1.000 \end{bmatrix}$$

Derivation of Principal Components

Suppose $\mathbf{x}' = [x_1, \dots, x_p]$ is a p -dimensional random variable with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = ((\sigma_{ij}))$. The problem is to find a new set of variables, say $z_1, z_2, \dots, z_p$, which are uncorrelated and whose variances decrease from first to last. Each z_j is taken to be linear combination of the x 's, so that

$$\begin{aligned} z_j &= a_{1j}x_1 + a_{2j}x_2 + \dots + a_{pj}x_p \\ &= \mathbf{a}'_j \mathbf{x} \end{aligned} \quad (1)$$

where $\mathbf{a}'_j = [a_{1j}, a_{2j}, \dots, a_{pj}]$ is a vector of constants. Equation (1) contains an arbitrary scale factor. We therefore impose the condition that $\mathbf{a}'_j \mathbf{a}_j = \sum_{k=1}^p a_{kj}^2 = 1$. We shall see that this particular normalization procedure ensures that the overall transformation is orthogonal in other words, that distance in p -space is preserved. The first principal component $z_1 = a_{11}x_1 + a_{12}x_2 + \dots + a_{1p}x_p$, is found by choosing $\mathbf{a}_1$ such variance of z_1 is as large as possible subject to the condition that $a_{11}^2 + a_{12}^2 + \dots + a_{1p}^2 = 1$ or $\mathbf{a}'_1 \mathbf{a}_1 = 1$.

This constraint is introduced because if this is not done, then $\text{Var}(z_1)$ can be increased simply by multiplying any a_{1j} 's by a constant factor.

The second principal component $z_2 = a_{21}x_1 + a_{22}x_2 + \dots + a_{2p}x_p$, is found by choosing $\mathbf{a}_2$ such that $\text{Var}(z_2)$ is as large as possible next to $\text{Var}(z_1)$ subject to the constraint that

$a_{21}^2 + a_{22}^2 + \dots + a_{2p}^2 = 1$ or $\mathbf{a}'_2 \mathbf{a}_2 = 1$ and $\text{Cov}(z_1, z_2) = 0$. Similarly, remaining $p-2$ principal components $z_3, \dots, z_p$ can be derived so as to be uncorrelated with the first two principal components and among themselves and to have decreasing variance.

We begin by finding the first principal component. We want to choose $\mathbf{a}_1$ so as to maximize the variance of z_1 subject to the normalization constraint that $\mathbf{a}'_1 \mathbf{a}_1 = 1$.

$$\begin{aligned} \text{Var}(z_1) &= \text{Var}(\mathbf{a}'_1 X) \\ &= \mathbf{a}'_1 \Sigma \mathbf{a}_1 \end{aligned} \quad (2)$$

Thus, we take $\mathbf{a}'_1 \Sigma \mathbf{a}_1$ as our objective function.

Contribution of Variability Through Spectral Decomposition of Σ

Spectral decomposition of Σ is given by

$$\Sigma \mathbf{A} \mathbf{A}' = \lambda_1 \mathbf{a}_1 \mathbf{a}_1' + \lambda_2 \mathbf{a}_2 \mathbf{a}_2' + \cdots + \lambda_p \mathbf{a}_p \mathbf{a}_p'.$$

This clearly shows that the contribution of the first principal component to Σ is $\lambda_1 \mathbf{a}_1 \mathbf{a}_1'$ when that is removed, the residual matrix of variances and covariance is

$$\Sigma - \lambda_1 \mathbf{a}_1 \mathbf{a}_1' = \sum_{j=2}^p \lambda_j \mathbf{a}_j \mathbf{a}_j'$$

The second principal component contributes $\lambda_2 \mathbf{a}_2 \mathbf{a}_2'$ and so on (Wall *et al.*, 2003).

Identification of Important Components

The usual procedure is to look at the first few components which, hopefully, account for a large proportion of the total variance. In order to do this, it is necessary to decide which eigenvalues are 'large' and which are 'small' so that components corresponding to the later may be disregarded. When analyzing a correlation matrix where the sum of the eigenvalues is p , many researchers use the rule that eigenvalues less than 1 may be disregarded. This arbitrary policy is a useful rule of thumb but has no theoretical justification. It may be better to look at the pattern of eigenvalues and see if there is a natural breakpoint. The fact that there is no objective way of deciding how many components to retain is a serious drawback of the method.

Estimating the Principal Components

The above derivation of the principal components of $\mathbf{x}$ assumes that Σ is known. Generally we deal with sample data and Σ is unknown. In such situations, Σ is replaced by $\mathbf{S}$, the sample covariance matrix. The derivation of the principal components of $\mathbf{x}$ using the sample variances and covariance $\mathbf{S}$ is same as before. Let us denote the eigenvalues of $\mathbf{S}$ in descending order of size by $\lambda_1, \lambda_2, \dots, \lambda_p$ and the corresponding eigenvectors by $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p$. Since $\mathbf{S}$ is positive semi-definite, the eigenvalues are all non-

negative and represent the estimated variances of the different components (Wold *et al.*, 1987).

If our sample of ‘individuals’ a random sample from a larger population, then $\{\lambda_i\}$ and $\{\mathbf{a}_i\}$ may be regarded as estimates of the eigenvalues and vectors of Σ , giving us estimates of the principal components of $\mathbf{x}$. But no assumptions have been made about the underlying population, and without such assumptions it is impossible to derive the sampling properties of the estimates. If we are prepared to assume that the observations are taken from a multivariate normal distribution, then some sampling theory is available (Johnson and Wichern, 2002). But this theory is of limited principal value, partly because many of the results are for the asymptotic case (as $n \rightarrow \infty$), and partly because the normality assumption is often questionable. In any case, the ‘sample’ may be observations for a complete population. Thus the modern tendency is to view Principal Component Analysis as a mathematical technique with no underlying statistical model. The principal components obtained from the sample covariance matrix $\mathbf{S}$ are seen as the principal components and not as estimates of the corresponding quantities obtained from Σ . The ‘hats’ over λ_i and $\mathbf{a}_i$ are often omitted. Indeed, it is not even necessary to regard $\mathbf{x}$ and $\mathbf{z}$ as random variables (Jolliffe and Cadima, 2016).

Example: The Iris dataset is a classic dataset in the field of machine learning and statistics. It was introduced by the British biologist and statistician Ronald A. Fisher in 1936. The dataset consists of measurements from three species of iris flowers, namely Iris setosa, Iris versicolor, and Iris virginica. Each species has 50 samples, making a total of 150 samples in the dataset.

Features:

The dataset contains four features (attributes) for each flower sample:

- Sepal Length (cm): The length of the sepal, which is the outermost whorl of a flower.
- Sepal Width (cm): The width of the sepal.
- Petal Length (cm): The length of the petal, which is the inner whorl of a flower.

- Petal Width (cm): The width of the petal.

Target Variable:

The target variable (label) is the species of the iris flower.

There are three species in the dataset:

- Iris setosa
- Iris versicolor
- Iris virginica

```
data("iris") # call inbuilt data "iris"
```

```
iris # print data and understand
```

```
set.seed(555) # random value for unbiased sampling
```

```
ind <- sample(2,nrow(iris),replace=T,prob = c(0.8,0.2)) # making  
two random samples one for training and another for testing,
```

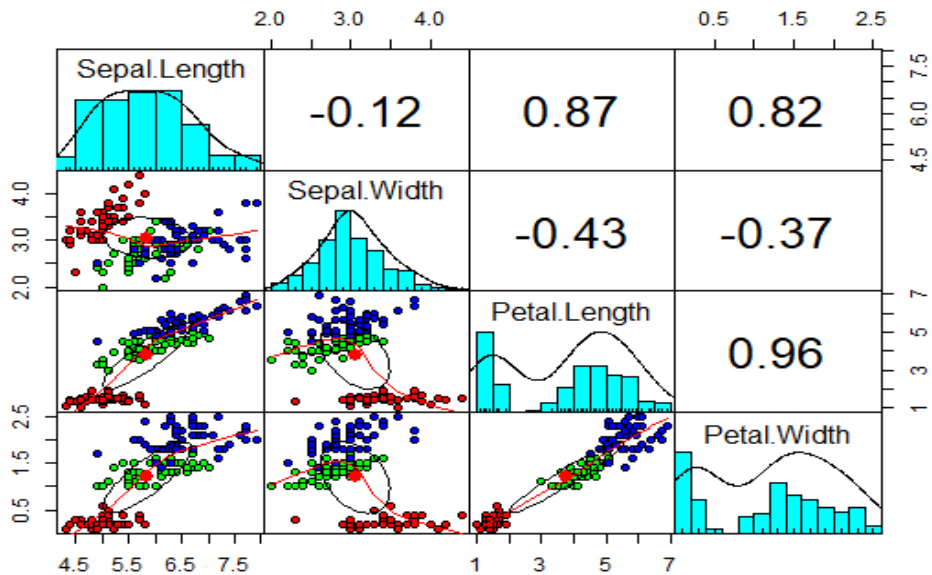
```
training <- iris[ind==1,] #training dataset
```

```
testing <- iris[ind==2,] #testing dataset
```

let's have a pictorial look of data

```
library(psych)
```

```
pairs.panels(iris[1:4],gap=0,bg=c("red", "green",  
"blue")[iris$Species], pch=21)
```



Red colour depicts setosa; green colour depicts versicolor and blue colour depicts virginica

R programme for Principal Component Analysis is as follows

```
mydata = read.table(file.choose(), header=T) #gives an option to  
select data file
```

```
fit = princomp(mydata, cor=TRUE) # here we are using correlation  
matrix for PCA
```

```
summary(fit) # print variance accounted for
```

```
loadings(fit) # pc loadings
```

```
plot(fit, type="lines") # scree plot
```

```
fit$scores # the principal components
```

```
> pc <- prcomp(training[, -5], center = T, scale. = T)
```

```
> attributes(pc)
```

```
$names
```

```
[1] "sdev" "rotation" "center" "scale" "x"
```

```
$class
```

```
[1] "prcomp"
```

```
> pc$center
```

```
Sepal.Length Sepal.Width Petal.Length Petal.Width  
5.822881 3.055085 3.688983 1.167797
```

```
> pc$scale
```

```
Sepal.Length Sepal.Width Petal.Length Petal.Width  
0.8556910 0.4452211 1.7942350 0.7639172
```

```
> print(pc)
```

```
Standard deviations (1, ..., p=4):
```

```
[1] 1.7110469 0.9598531 0.3613169 0.1430058
```

```
Rotation (n x k) = (4 x 4):
```

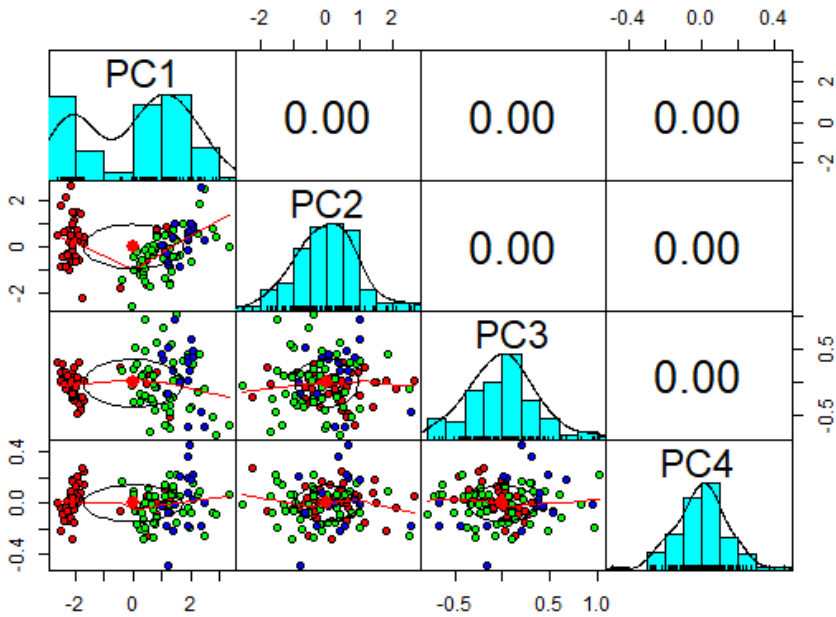
	PC1	PC2	PC3	PC4
Sepal.Length	0.5246028	0.37200075	-0.7054364	0.2979376
Sepal.Width	-0.2627864	0.92602023	0.2388251	-0.1280329
Petal.Length	0.5799268	0.02610399	0.1039516	-0.8075875
Petal.Width	0.5651726	0.05848559	0.6591783	0.4925883

```
> summary(pc)
```

```
Importance of components:
```

	PC1	PC2	PC3	PC4
Standard deviation	1.7110	0.9599	0.36132	0.14301
Proportion of Variance	0.7319	0.2303	0.03264	0.00511
Cumulative Proportion	0.7319	0.9623	0.99489	1.00000

```
> pairs.panels(pc$x,gap=0,bg=c("red", "green",  
"blue"))[iris$Species], pch=21)
```



Proportionate contribution by each PC's

	PC1	PC2	PC3	PC4
1	-2.16	0.50	-0.13	0.03
3	-2.26	-0.30	0.03	0.02
4	-2.20	-0.54	0.07	-0.08
5	-2.28	0.67	0.00	-0.04
6	-1.97	1.49	0.02	0.01
8	-2.13	0.25	-0.10	-0.02
9	-2.24	-1.05	0.12	-0.04
12	-2.22	0.17	0.07	-0.14
13	-2.12	-0.67	-0.24	0.00
14	-2.53	-0.90	0.15	-0.04
16	-2.14	2.65	0.03	0.06
17	-2.10	1.48	0.00	0.19

	PC1	PC2	PC3	PC4
18	-2.09	0.51	-0.05	0.09
19	-1.80	1.40	-0.36	0.08
21	-1.82	0.43	-0.42	0.02
22	-2.10	0.94	0.15	0.05
23	-2.66	0.49	0.31	0.00
24	-1.72	0.12	0.04	0.14
25	-2.12	0.17	0.09	-0.27
26	-1.86	-0.58	-0.31	0.05
27	-1.95	0.27	0.08	0.06
29	-2.04	0.34	-0.27	0.09
30	-2.17	-0.29	0.05	-0.12
31	-2.04	-0.46	-0.09	-0.05
32	-1.74	0.44	-0.26	0.24
33	-2.50	1.79	0.03	-0.22
34	-2.33	2.13	-0.09	-0.04
35	-2.02	-0.41	-0.18	0.03
36	-2.11	-0.17	-0.22	0.17
37	-1.95	0.68	-0.47	0.21
38	-2.42	0.62	0.00	-0.14
39	-2.33	-0.84	0.17	-0.03
40	-2.07	0.30	-0.18	0.01
41	-2.18	0.47	0.03	0.10
42	-1.78	-2.25	-0.20	0.27
43	-2.45	-0.43	0.28	-0.09
44	-1.86	0.49	0.31	0.16
45	-2.03	1.15	0.23	-0.16
46	-1.98	-0.66	-0.07	0.13
48	-2.29	-0.34	0.12	-0.06
49	-2.12	1.01	-0.18	-0.01
50	-2.10	0.04	-0.16	0.05
51	1.13	0.85	-0.63	0.06
52	0.78	0.59	-0.06	0.01
53	1.27	0.60	-0.51	0.03

	PC1	PC2	PC3	PC4
54	0.45	-1.70	-0.01	0.05
55	1.11	-0.20	-0.36	0.11
56	0.44	-0.56	0.13	-0.25
57	0.79	0.76	0.17	-0.08
58	-0.43	-1.78	0.24	-0.07
60	0.07	-0.99	0.54	-0.06
61	-0.07	-2.57	-0.04	-0.01
62	0.49	-0.05	0.22	0.03
63	0.59	-1.71	-0.73	0.06
64	0.76	-0.17	-0.05	-0.16
65	0.02	-0.41	0.21	0.09
67	0.40	-0.17	0.49	-0.21
68	0.20	-0.76	-0.29	-0.20
69	1.24	-1.58	-0.44	0.23
70	0.21	-1.25	-0.16	-0.06
71	0.79	0.40	0.62	-0.11
72	0.52	-0.40	-0.23	0.12
73	1.26	-0.90	-0.33	-0.01
74	0.67	-0.39	-0.28	-0.26
75	0.74	-0.05	-0.41	0.06
77	1.28	-0.07	-0.68	0.06
78	1.39	0.33	-0.22	0.07
79	0.71	-0.21	0.10	-0.04
80	0.01	-1.02	-0.30	0.06
87	1.08	0.51	-0.35	0.05
88	1.07	-1.34	-0.64	0.15
89	0.13	-0.20	0.29	-0.16
90	0.33	-1.28	0.10	-0.01
91	0.32	-1.07	0.09	-0.28
92	0.67	0.04	0.00	-0.15
93	0.38	-0.95	-0.18	0.00
95	0.34	-0.82	0.14	-0.12
98	0.62	-0.14	-0.24	-0.01

	PC1	PC2	PC3	PC4
99	-0.39	-1.48	0.20	0.17
100	0.31	-0.57	0.10	-0.07
103	2.22	0.54	-0.15	0.07
105	1.89	0.29	0.43	-0.03
106	2.75	0.77	-0.52	-0.07
107	0.42	-1.50	0.97	-0.18
109	2.02	-0.69	-0.35	-0.08
110	2.29	1.87	0.45	0.10
111	1.40	0.68	0.32	0.10
112	1.63	-0.41	0.06	0.05
114	1.29	-1.13	0.60	0.06
115	1.50	-0.43	1.03	0.22
117	1.50	0.25	0.06	-0.16
118	2.45	2.49	-0.08	-0.25
119	3.29	0.00	-0.63	0.07
120	1.28	-1.66	-0.24	-0.07
121	2.06	0.89	0.28	0.16
122	1.02	-0.55	0.84	-0.01
123	2.89	0.39	-0.79	-0.09
124	1.36	-0.47	0.03	0.13
125	1.73	0.99	0.33	-0.07
126	1.97	0.98	-0.38	-0.19
128	1.06	0.07	0.36	-0.03
130	1.88	0.55	-0.67	-0.18
131	2.44	0.25	-0.67	0.01
132	2.33	2.56	-0.44	-0.17
133	1.89	-0.17	0.39	0.08
134	1.14	-0.28	-0.16	-0.18
135	1.23	-0.78	-0.16	-0.48
138	1.38	0.42	0.20	-0.22
139	0.97	0.03	0.43	-0.01
140	1.88	0.66	0.04	0.19
141	2.04	0.60	0.47	0.23

	PC1	PC2	PC3	PC4
142	1.93	0.67	0.19	0.46
143	1.19	-0.67	0.54	-0.07
144	2.07	0.85	0.38	0.03
145	2.03	1.02	0.67	0.19
146	1.90	0.38	0.31	0.37
147	1.59	-0.87	0.02	0.21
149	1.42	0.99	0.95	-0.01

Applications of Principal Component Analysis

- An important benefit of Principal Component Analysis is that it provides a quick way of assessing the effective dimensionality of a set of data. If the first few components account for most of the variation in the original data, it is often a good idea to use these first few component scores in subsequent analysis.
- Plotting of data becomes difficult with more than three variables. But if the first two components account for a large proportion of the total variation, then it will often be useful to plot the values of the first two component scores for each individual. In other words, Principal Component Analysis enables us to plot the data in two dimensions. This procedure is known as Biplot. A biplot is a scatter plot that graphically displays both the row factors and the column factors of a two-way data. The concept of biplots were first developed by Gabriel (1971). Since then, biplot has been used in data visualization and pattern analysis in various research fields, from psychology to economics to agronomy. Currently there are over 50000 web pages containing the keyword 'biplot'. This technique has extensively been used in the analysis of multi-environment trials. For generating a biplot, we take the matrix representing the effects of two factors. This matrix is then subjected to singular value decomposition.
- In particular, one can then look for outliers or for groups or 'clusters' of individuals. This is an important use of Principal Component Analysis and often groupings of variables which would not be found by other means.
- Reduction in dimensionality can help in discriminant analysis when a large number of correlated variables (p) are observed on

n observation ($n < p$). As the number of observations is less than the number of variables, there will be unpleasant singularity problems unless the dimensionality is drastically reduced. Hopefully, Principal Component Analysis does that.

- Multiple regression can be dangerous if the so-called independent variables are in fact highly correlated. Various techniques, such as ridge regression, have been developed to overcome this problem. An alternative approach, which may well be more fruitful, is to regress, not on the original variables, but on those components which are most highly correlated with the dependent variable. This is called principal components regression.

It is quite likely that first few principal components account for most of the variability in the original data. If so, these few principal components can then replace the initial p variables in subsequent analysis, thus reducing the effective dimensionality of the problem. An analysis of principal components often reveals relationships that were not previously suspected and thereby allows interpretation that would not ordinarily result.

The Principal Components Analysis is more of a mean to an end rather than end in itself because this frequently serves as intermediate steps in much larger investigations by reducing the dimensionality of the problem and providing easier interpretation. It is a mathematical technique, which does not require user to specify the statistical model or assumption about distribution of original variates. It may also be mentioned that principal components are artificial variables and often it is not possible to assign physical meaning to them. Further, since Principal Components Analysis transforms original set of variables to new set of uncorrelated variables. It is worth stressing that if the original variables are uncorrelated, then there is no point in carrying out the Principal Components Analysis. It is important to note here that the principal components depend on the scale of measurement. Conventional way of getting rid of this problem is to use the standardized variables with unit variances.

References

- Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4), 433-459.

- Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3), 453-467.
- Johnson, R. A., & Wichern, D. W. (2002). Applied multivariate statistical analysis.
- Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202.
- Wall, M. E., Rechtsteiner, A., & Rocha, L. M. (2003). Singular value decomposition and principal component analysis *A practical approach to microarray data analysis* (pp. 91-109): Springer.
- Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3), 37-52.

How to Cite: Sarkar, S. K. (2023). Principal Component Analysis. In S. Rathod, B. Sailaja, N. Bandumula, S. Arun Kumar, P. A. Lakshmi Prasanna, P. Jeyakumar, A. Waris, P. Muthuraman, and R. M. Sundaram (Eds.), *Statistical Procedures for Analysing Agricultural Data using R* (pp. 139-156). ICAR - Indian Institute of Rice Research, Hyderabad.