Multivariate Analysis with R

Principal Component and Its Graphical Analysis in R

Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur

Principal Component Analysis: Example- Construction

```
eigen(sigma)
eigen() decomposition
$values
[1] 5.8284271 2.0000000 0.1715729

$vectors
           [,1] [,2]      [,3]
[1,] -0.3826834    0 0.9238795
[2,]  0.9238795    0 0.3826834
[3,]  0.0000000    1 0.0000000
```

$$\Sigma = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

The principal components are

$$Y_1 = \underline{e}_1'\underline{X} = (-0.38, 0.92, 0.00)(X_1, X_2, X_3)' = -0.38X_1 + 0.92X_2$$

$$Y_2 = \underline{e}_2'\underline{X} = (0, 0, 1)(X_1, X_2, X_3)' = X_3$$

$$Y_3 = \underline{e}_3'\underline{X} = (0.92, 0.38, 0.00)(X_1, X_2, X_3)' = 0.92X_1 + 0.38X_2$$

Principal Component Analysis: Example- Variance

```
eigen() decomposition
$values
[1] 5.8284271 2.0000000 0.1715729
```

The variance of first, second and third principal components are

$Var(Y_1) = Var(-0.38X_1 + 0.92X_2)$

$= (-0.38)^2 Var(X_1) + (0.92)^2 Var(X_2) + 2 \times (-0.38) \times (0.92) Cov(X_1, X_2)$

$= (-0.38)^2 (1) + (0.92)^2 Var(X_2) + 2 \times (-0.38) \times (0.92) \times (-2)$

$= 5.83 = \lambda_1$

$$\Sigma = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

$Var(Y_2) = 2.00 = \lambda_2$

$Var(Y_3) = 0.17 = \lambda_3$

Also $\sum_{i=1}^{3} Var(X_i) = \sigma_{11} + \sigma_{22} + \sigma_{33} = 1 + 5 + 2 = 8$

$\sum_{i=1}^{3} Var(Y_i) = \lambda_1 + \lambda_2 + +\lambda_3 = 5.83 + 2.00 + 0.17 = 8.$

Principal Component Analysis: Example- Proportion of Variance

The proportion of total variance accounted by

- First principal component is $\frac{\boldsymbol{\lambda_1}}{\boldsymbol{\lambda_1+\lambda_2+\lambda_3}} = \frac{\mathbf{5.83}}{\mathbf{8}} = \mathbf{0.73 = 73\%}$
- Second principal component is $\frac{\boldsymbol{\lambda_2}}{\boldsymbol{\lambda_1+\lambda_2+\lambda_3}} = \frac{\mathbf{2}}{\mathbf{8}} = \mathbf{0.25 = 25\%}$
- Third principal component is $\frac{\boldsymbol{\lambda_3}}{\boldsymbol{\lambda_1+\lambda_2+\lambda_3}} = \frac{\mathbf{0.17}}{\mathbf{8}} = \mathbf{0.02 = 2\%}$

First two principal components is $\frac{\boldsymbol{\lambda_1+\lambda_2}}{\boldsymbol{\lambda_1+\lambda_2+\lambda_3}} = \frac{\mathbf{5.83+2}}{\mathbf{8}} = \mathbf{0.98 = 98\%}$

So the first two principal components $\boldsymbol{Y_1}$ and $\boldsymbol{Y_2}$ could replace the three original variables $\boldsymbol{X_1}, \boldsymbol{X_2}$ and $\boldsymbol{X_3}$ with little loss of information.

Principal Component Analysis: Example- Correlation

The correlation coefficient between

- Y_1 and X_1 is $\rho(Y_1, X_1) = \frac{e_{11}\sqrt{\lambda_1}}{\sqrt{\sigma_{11}}} = \frac{0.383\times\sqrt{5.83}}{\sqrt{1}} = 0.925$

- Y_1 and X_2 is $\rho(Y_1, X_2) = \frac{e_{21}\sqrt{\lambda_1}}{\sqrt{\sigma_{22}}} = \frac{-0.924\times\sqrt{5.83}}{\sqrt{5}} = -0.998$

So X_1 and X_2 are, individually, about equally important to the first principal component.

Also $\rho(Y_2, X_1) = \rho(Y_2, X_2) = 0$,

$\rho(Y_2, X_3) = \frac{e_{32}\sqrt{\lambda_2}}{\sqrt{\sigma_{33}}} = \frac{1\times\sqrt{2}}{\sqrt{2}} = 1$ (As expected)

Other correlations can be ignored since the third component is unimportant.

Principal Component Analysis in R:

Description

Performs a principal components analysis on the given data matrix and returns the results as an object of class prcomp.

Usage `prcomp(x, ...)`

```
prcomp(formula, data = NULL, subset,...)

prcomp(x, retx = TRUE, center = TRUE, scale. =
FALSE, tol = NULL, rank. = NULL, ...)
```

Arguments

`formula` a formula with no response variable, referring only to numeric variables.

`data` an optional data frame containing the variables in `formula`.

`x` a numeric or complex matrix (or data frame) which provides the data for the principal components analysis.

Principal Component Analysis in R:

Arguments

`retx` a logical value indicating whether the rotated variables should be returned.

`center` a logical value indicating whether the variables should be shifted to be zero centered.

`scale.` a logical value indicating whether the variables should be scaled to have unit variance before the analysis takes place.

`tol` a value indicating the magnitude below which components should be omitted.

`rank.` optionally, a number specifying the maximal rank, i.e., the maximal number of principal components to be used.

`newdata` An optional data frame or matrix in which to look for variables with which to predict. If omitted, the scores are used.

Principal Component Analysis in R:

Value

`prcomp` returns a list containing the following components:

`sdev` the standard deviations of the principal components.

`rotation` the matrix of variable loadings (i.e., a matrix whose columns contain the eigenvectors).

`x` if `retx` is true the value of the rotated data (the centred (and scaled if requested) data multiplied by the rotation matrix) is returned.

`center, scale` the centering and scaling used, or `FALSE`.

Principal Component Analysis in R: Example

Let $\Sigma = \begin{pmatrix} 11 & -12 & 6 \\ -12 & 15 & 4 \\ 6 & 4 & 12 \end{pmatrix}$ be the a (3×3) symmetric and positive definite covariance matrix of random vector $\underline{X} = (X_1, X_2, X_3)'$.

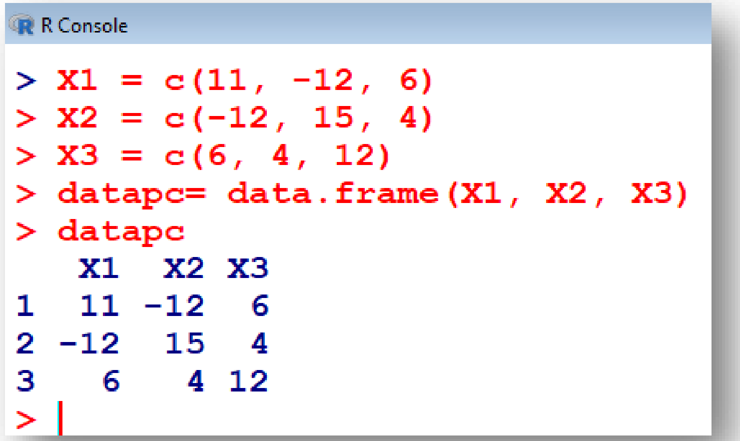
```
X1 = c(11, -12, 6)

X2 = c(-12, 15, 4)

X3 = c(6, 4, 12)

datapc= data.frame(X1, X2, X3)

datapc
    X1  X2 X3
1   11 -12  6
2  -12  15  4
3    6   4 12
```



Principal Component Analysis in R: Raw Data

```
pcdataout = prcomp(datapc)
pcdataout
Standard deviations (1, .., p=3):
[1] 1.783783e+01 5.460014e+00 3.787415e-16

Rotation (n x k) = (3 x 3):
          PC1       PC2        PC3
X1 -0.6619978 0.4807521  0.5750098
X2  0.7454701 0.5018419  0.4386673
X3 -0.0776738 0.7190494 -0.6906046
```

The principal components are

$Y_1 = \underline{e}_1'\underline{X} = -0.66X_1 + 0.74X_2 - 0.07X_3$ with $Var(Y_1) = 17.8^2$,

$Y_2 = \underline{e}_2'\underline{X} = 0.48X_1 + 0.50X_2 + 0.72X_3$ with $Var(Y_2) = 5.46^2$,

$Y_3 = \underline{e}_3'\underline{X} = 0.58X_1 + 0.44X_2 - 0.69X_3$ with $Var(Y_3) = \left(3.78 \times 10^{-16}\right)^2$

Principal Component Analysis in R: Raw Data

```
> summary(pcdataout)

Importance of components:

                           PC1     PC2       PC3
Standard deviation     17.8378 5.46001 2.272e-15
Proportion of Variance  0.9143 0.08567 0.000e+00
Cumulative Proportion   0.9143 1.00000 1.000e+00
```

Principal Component Analysis in R: Raw Data

R Console

```
> X1 = c(11, -12, 6)
> X2 = c(-12, 15, 4)
> X3 = c(6, 4, 12)
> datapc= data.frame(X1, X2, X3)
> pcdataout = prcomp(datapc)
> pcdataout
Standard deviations (1, .., p=3):
[1] 1.783783e+01 5.460014e+00 2.272449e-15

Rotation (n x k) = (3 x 3):
          PC1       PC2        PC3
X1 -0.6619978 0.4807521  0.5750098
X2  0.7454701 0.5018419  0.4386673
X3 -0.0776738 0.7190494 -0.6906046
> summary(pcdataout)
Importance of components:
                           PC1     PC2       PC3
Standard deviation     17.8378 5.46001 2.272e-15
Proportion of Variance  0.9143 0.08567 0.000e+00
Cumulative Proportion   0.9143 1.00000 1.000e+00
> 
```

Principal Component Analysis: Biplot

Description : Plot a biplot on the current graphics device.

Biplot helps in visualizing the similarities and dissimilarities between the samples, and shows the impact of each attribute on each of the principal components.

Usage `biplot(x, ...)`

```
biplot(x, y, var.axes = TRUE, col, cex =
rep(par("cex"), 2), xlabs = NULL, ylabs = NULL,
xlim  = NULL, ylim  = NULL, arrow.len = 0.1, main
= NULL, xlab = NULL, ylab = NULL, ...)
```

Arguments

`x` **The biplot, a fitted object.**

`y` **The second set of points (a two-column matrix), usually associated with variables.**

Principal Component Analysis: Biplot

Arguments

`var.axes` If `TRUE` the second set of points have arrows representing them as (unscaled) axes.

`col` A vector of length 2 giving the colours for the first and second set of points respectively (and the corresponding axes).

`cex` The character expansion factor used for labelling the points.

`arrow.len` The length of the arrow heads on the axes plotted in `var.axes` is true. The arrow head can be suppressed by `arrow.len = 0.`

`xlim, ylim` Limits for the x and y axes in the units of the first set of variables.

Principal Component Analysis: Biplot

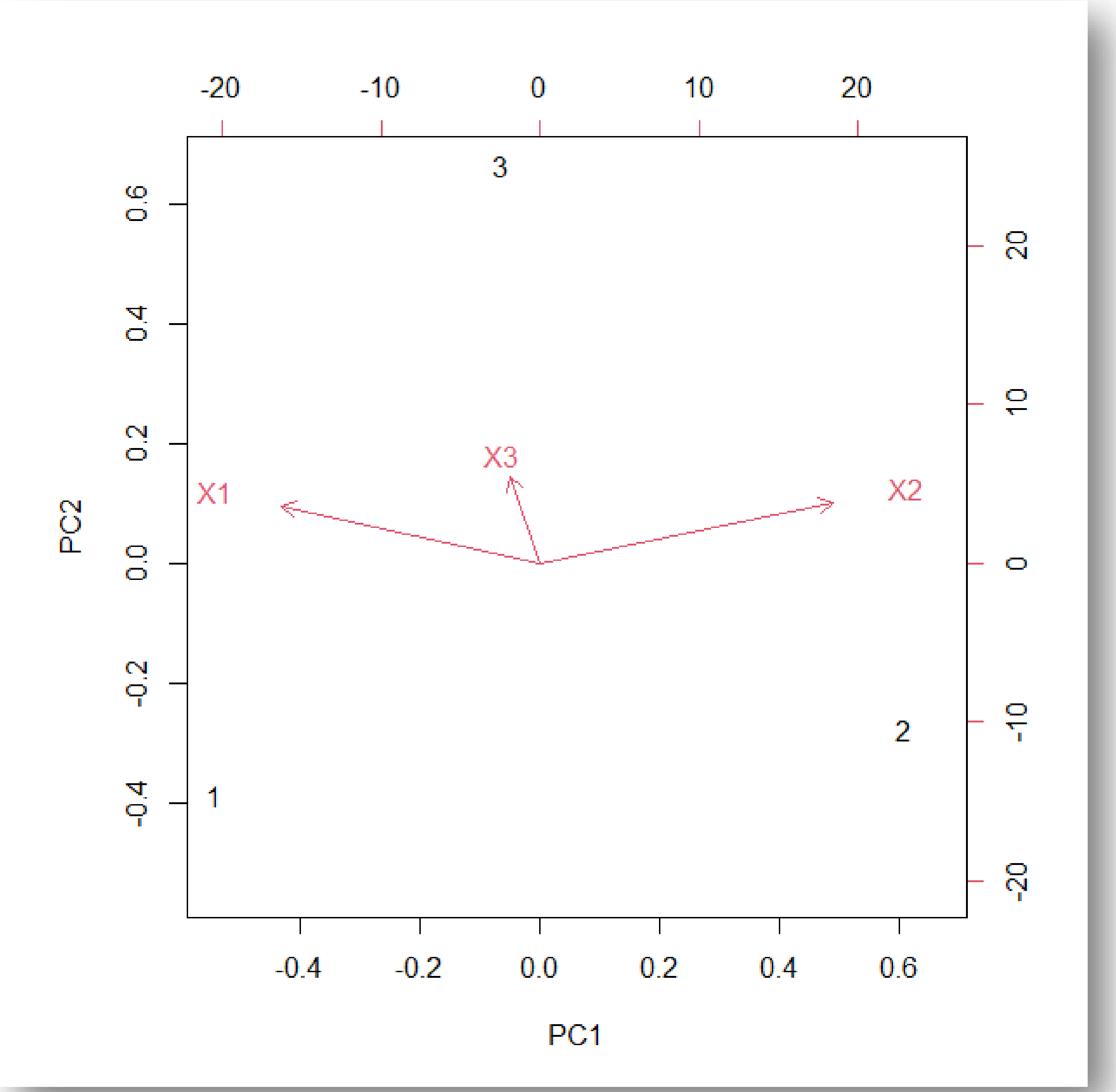
- The biplot of the principal component shows both PC scores of samples (dots) and loadings of variables (vectors).
- The further away these vectors are from a PC origin, the more influence they have on that PC.
- Representation of angle degrees between two variables :
  - 90^0 indicates no correlation between the two variables,
  - less than 90^0 indicates a positive correlation between two variables and
  - more than 90^0 indicates a negative correlation between two variable.

Principal Component Analysis: Biplot

- **The vector line length represents the variance level of the variable:**
  - **the longer the vector line, the greater the variance.**
  - **the shorter the line, the less the variance.**

Principal Component Analysis: Biplot of Raw Data

```
biplot(pcdataout)
```



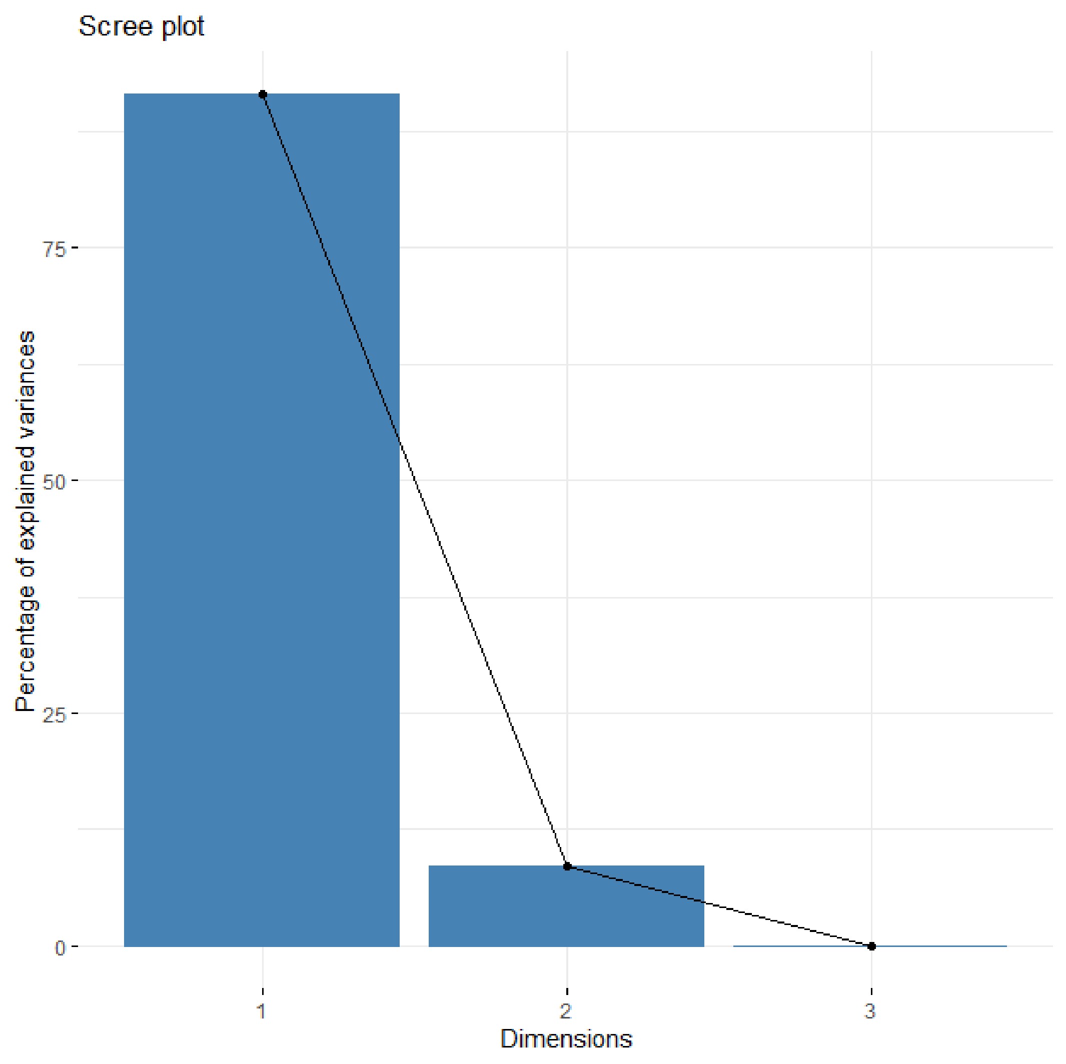
Principal Component Analysis: Scree Plot

- **Scree Plot is used to visualize the importance of each principal component. A scree plot is a diagnostic tool to check whether PCA works well on the data or not.**
- **A scree plot shows how much variation each principal component captures from the data. It is used to determine the number of principal components to be retained.**
- **The scree plot has eigenvalues on *y*-axis and the number of principal components on *x*-axis.**
- **An ideal curve should be steep, then bends at an "elbow" — this is your cutting-off point — and after that flattens out.**
- **It is generated with the function `fviz_eig()`**

Principal Component Analysis: Scree Plot

```
library(factoextra)

fviz_eig(pcdataout)
```



Principal Component Analysis: Contribution of Each Variable

Cos2 visualization is used to determine how much each variable is represented in a given component. It is computed using the function

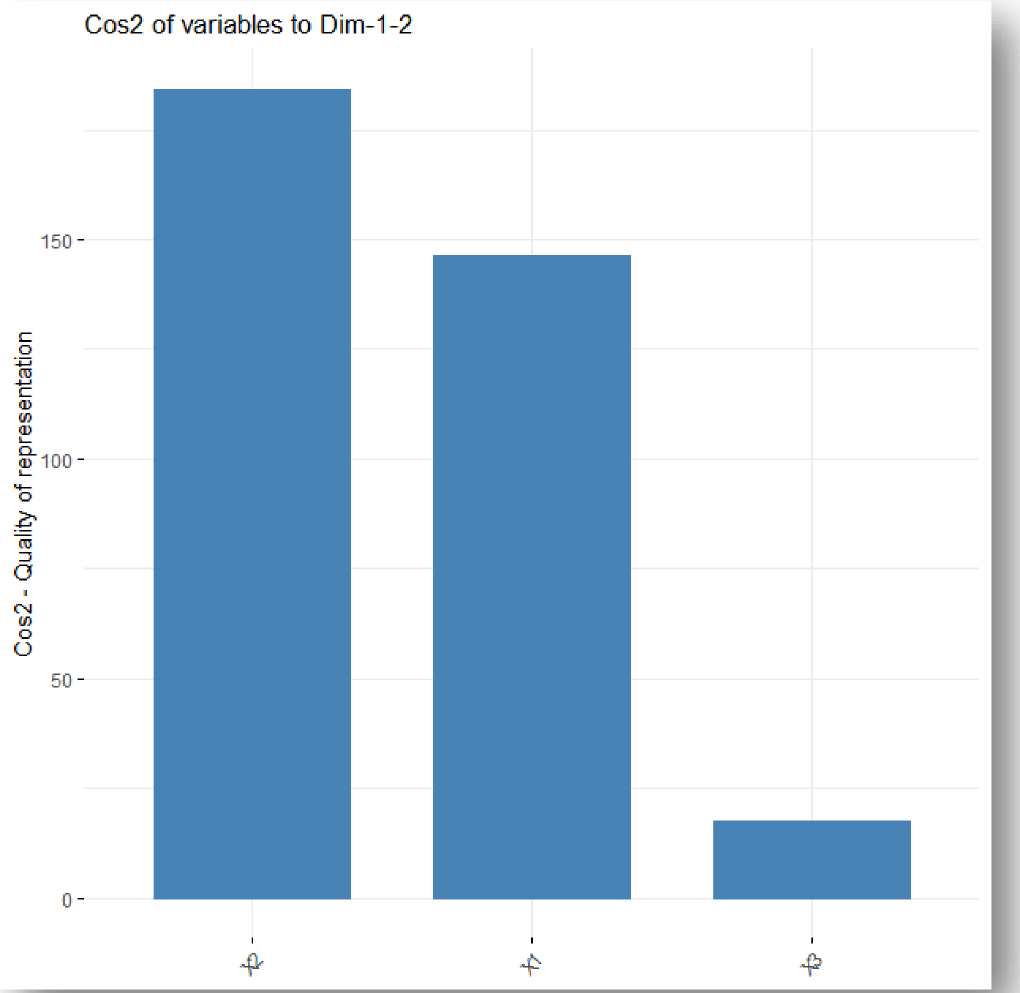
`fviz_cos2`

A high value indicates a good representation of the variable on that component.

A low value indicates that that component does not perfectly represent the variable.

Principal Component Analysis: Contribution of Each Variable

```
library(factoextra)
fviz_cos2(pcdataout, choice = "var", axes = 1:2)
```



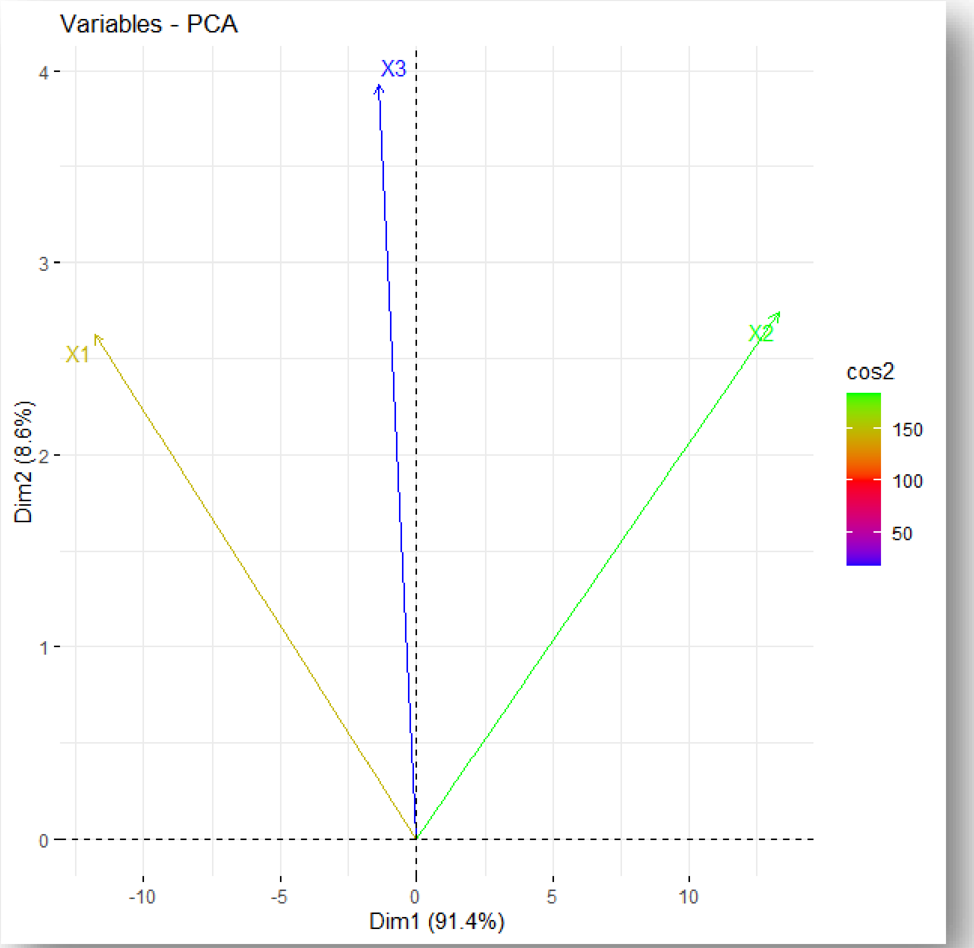
Principal Component Analysis: Biplot combined with cos2

Biplot and attribute importance can be combined to create a single biplot, where attributes with similar cos2 scores will have similar colours.

This is achieved by the `fviz_pca_var` function.

Principal Component Analysis: Biplot combined with cos2

```
fviz_pca_var(pcdataout, col.var = "cos2",
gradient.cols = c("blue", "red", "green"),
repel = TRUE)
```



Principal Component Analysis: Scaled data

```
pcdataoutcs = prcomp(datapc, center = TRUE, scale. =
TRUE)
> pcdataoutcs
Standard deviations (1, .., p=3):
[1] 1.455088e+00 9.395320e-01 1.316260e-16

Rotation (n x k) = (3 x 3):
          PC1         PC2        PC3
X1 -0.6859507 -0.06526322 -0.7247153
X2  0.6085311  0.49461231 -0.6205228
X3 -0.3989504  0.86665990  0.2995650
```

The principal components are

$Y_1 = \underline{e}_1'\underline{X} = -0.69X_1 + 0.61X_2 - 0.39X_3$ with $Var(Y_1) = 1.45^2$,

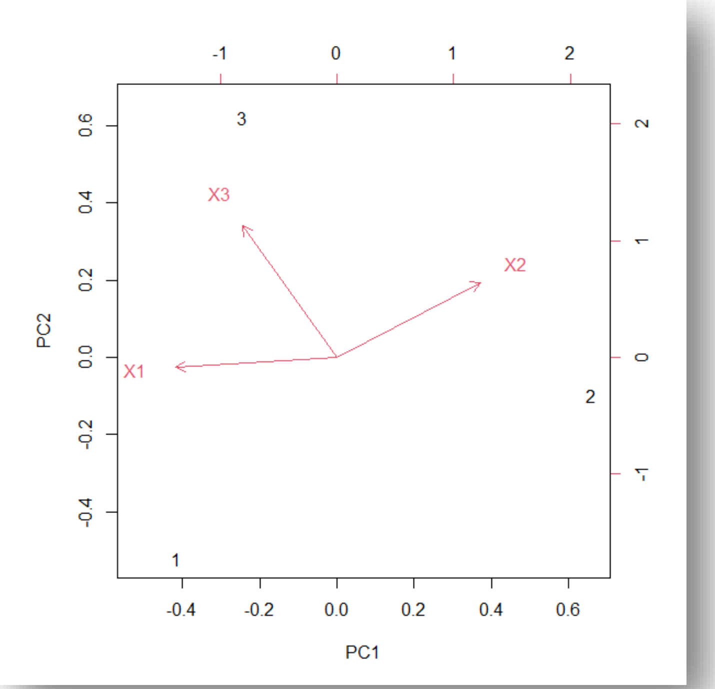
$Y_2 = \underline{e}_2'\underline{X} = -0.07X_1 + 0.49X_2 + 0.86X_3$ with $Var(Y_2) = 0.939^2$,

$Y_3 = \underline{e}_3'\underline{X} = -0.72X_1 - 0.62X_2 + 0.29X_3$ with $Var(Y_3) = \left(1.31 \times 10^{-16}\right)^2$

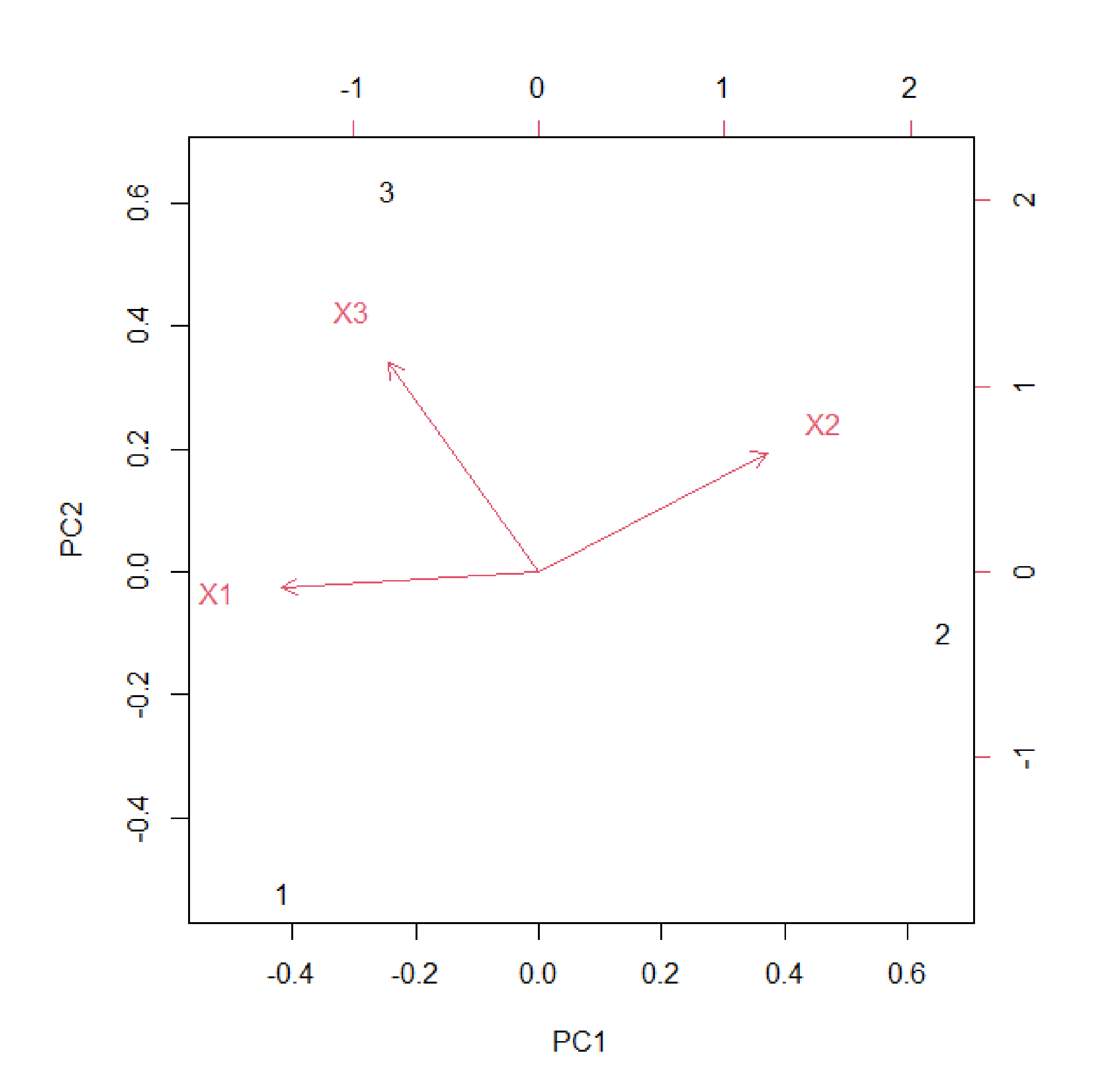
Principal Component Analysis: Scaled data

```
> summary(pcdataoutcs)
Importance of components:
                          PC1     PC2       PC3
Standard deviation     17.8378 5.46001 2.272e-15
Proportion of Variance  0.9143 0.08567 0.000e+00
Cumulative Proportion   0.9143 1.00000 1.000e+00
```

```
> biplot(pcdataoutcs)
```



Principal Component Analysis: Biplot with Scaled Data



Principal Component Analysis: Scaled Data

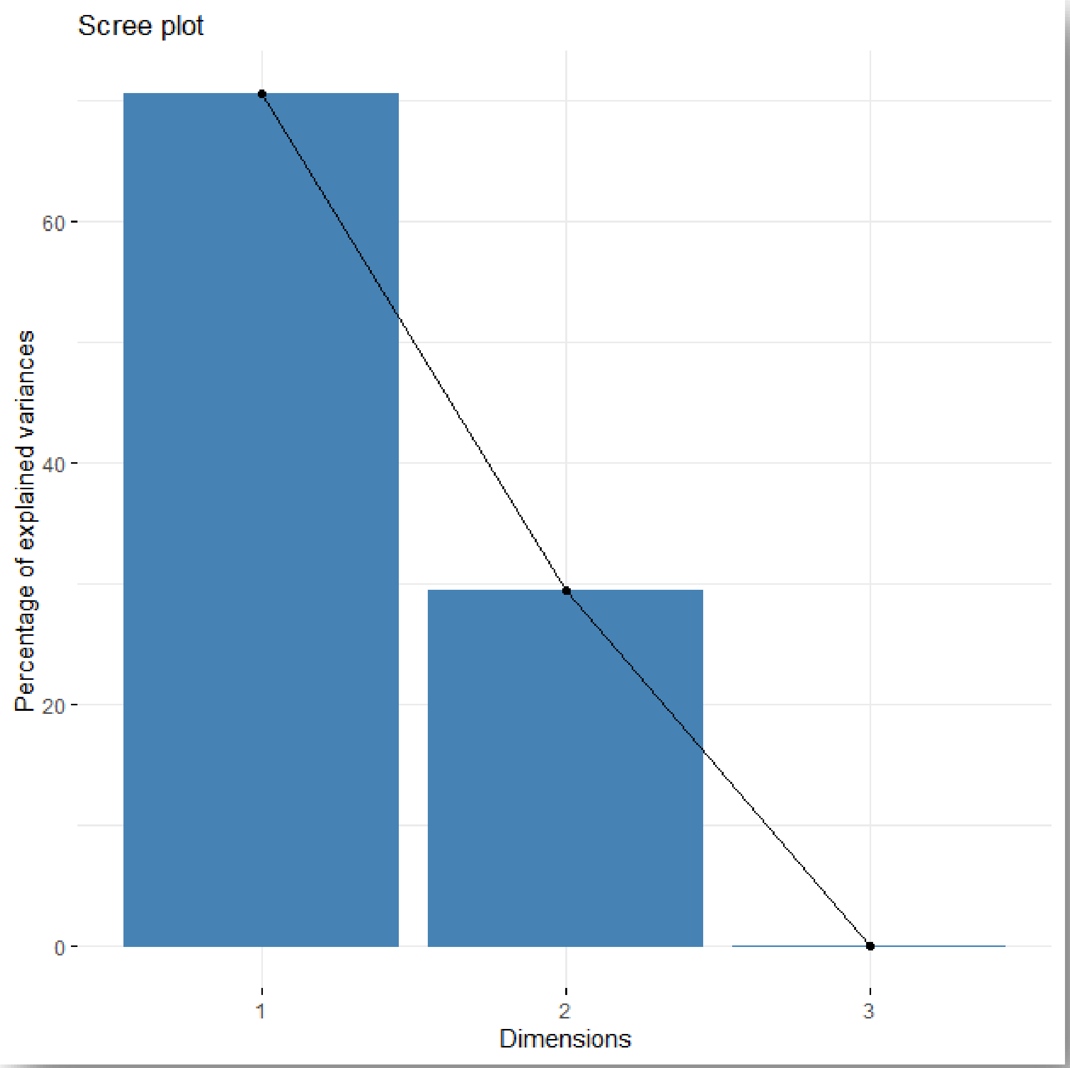
R Console

```
> X1 = c(11, -12, 6)
> X2 = c(-12, 15, 4)
> X3 = c(6, 4, 12)
> datapc= data.frame(X1, X2, X3)
> pcdataoutcs = prcomp(datapc, center = TRUE, scale. = TRUE)
> pcdataoutcs
Standard deviations (1, .., p=3):
[1] 1.455088e+00 9.395320e-01 1.233994e-16

Rotation (n x k) = (3 x 3):
          PC1         PC2        PC3
X1 -0.6859507 -0.06526322  0.7247153
X2  0.6085311  0.49461231  0.6205228
X3 -0.3989504  0.86665990 -0.2995650
> summary(pcdataoutcs)
Importance of components:
                          PC1    PC2       PC3
Standard deviation     1.4551 0.9395 1.234e-16
Proportion of Variance 0.7058 0.2942 0.000e+00
Cumulative Proportion  0.7058 1.0000 1.000e+00
> biplot(pcdataoutcs)
> 
```

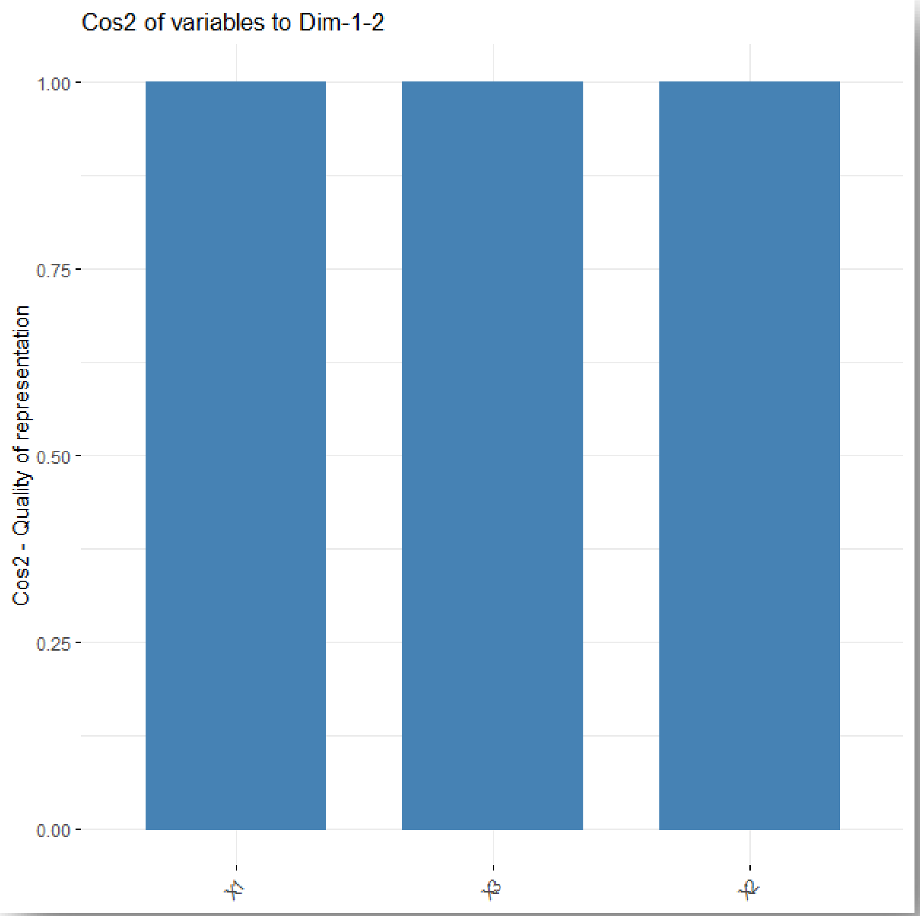
Principal Component Analysis: Scree plot of Scaled Data

```
fviz_eig(pcdataoutcs)
```



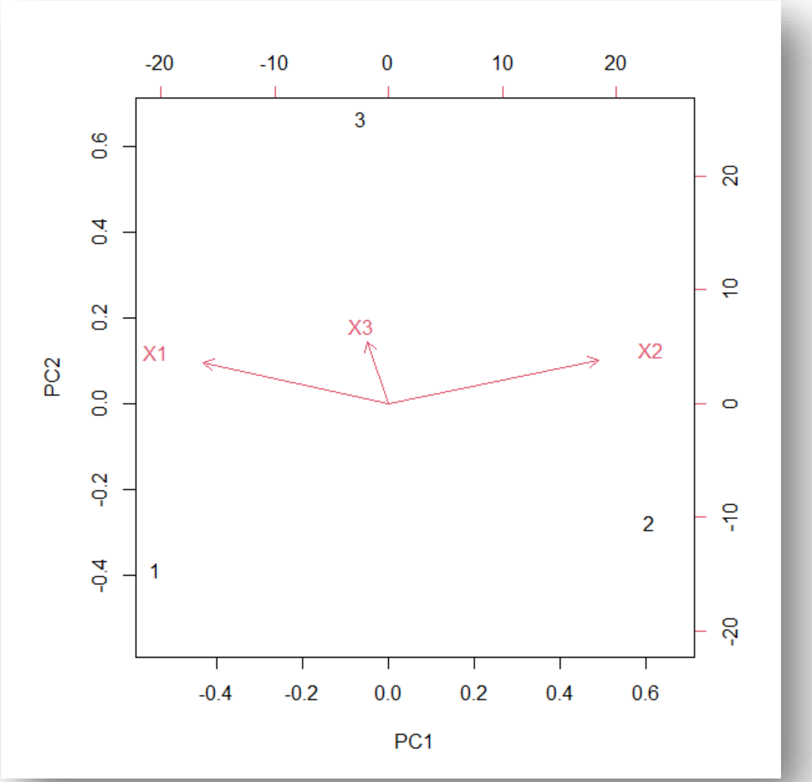
Principal Component Analysis: Contribution of Each Variable with Scaled Data

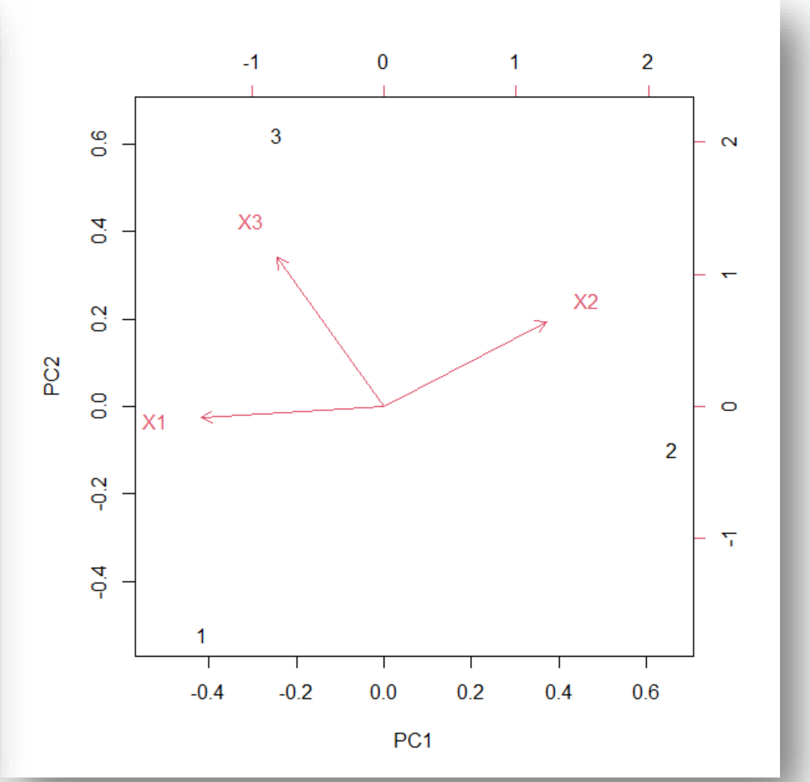
```
fviz_cos2(pcdataoutcs, choice = "var", axes = 
1:2)
```



Principal Component Analysis : Example- Comparison of Biplots of Scaled and Unscaled Data

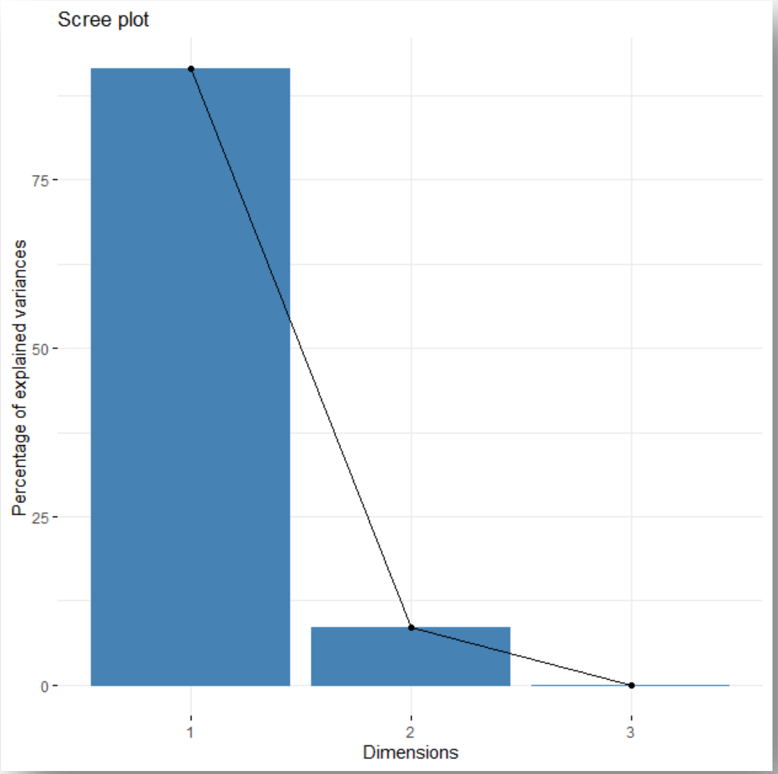
Unscaled Data	Scaled data

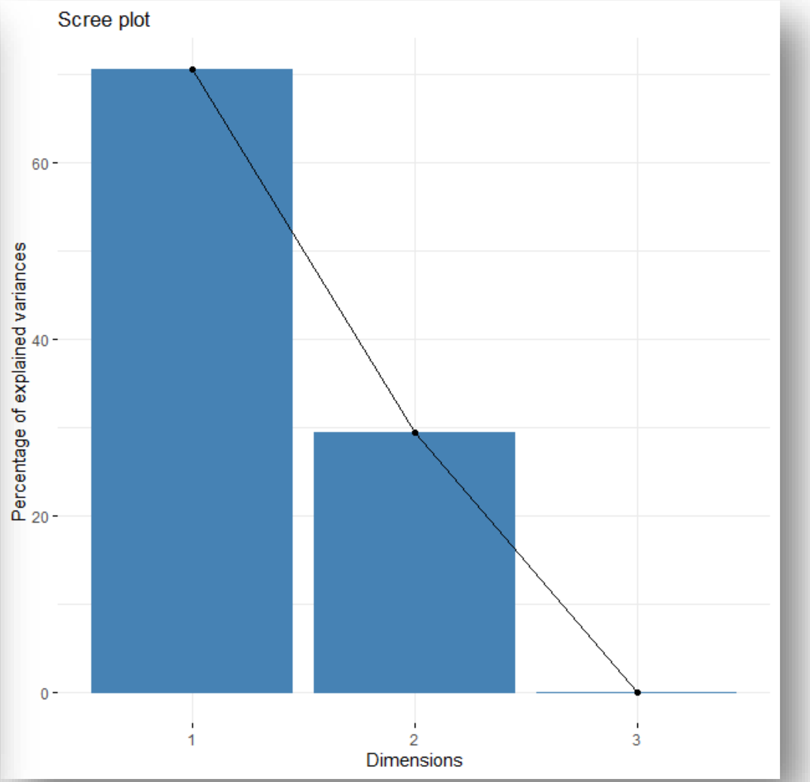




Principal Component Analysis : Example- Comparison of Scree Plots of Scaled and Unscaled Data

Unscaled Data	Scaled data





Principal Component Analysis: Example with USArrest Data

Use the built in data `USArrests` in R software

Violent Crime Rates by US State

Description : This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

Usage: `USArrests`

Format: A data frame with 50 observations on 4 variables.

[,1]	Murder	numeric	Murder arrests (per 100,000)
[,2]	Assault	numeric	Assault arrests (per 100,000)
[,3]	UrbanPop	numeric	Percent urban population
[,4]	Rape	numeric	Rape arrests (per 100,000)

Note: `USArrests` contains the data as in McNeil's monograph. For the UrbanPop percentages, a review of the table (No. 21) in the Statistical Abstracts 1975 reveals a transcription error for Maryland (and that McNeil used the same "round to even" rule that R's round() uses), as found by Daniel S Coven (Arizona).

Source: World Almanac and Book of facts 1975. (Crime rates).

References: McNeil, D. R. (1977) Interactive Data Analysis. New York: Wiley.

Principal Component Analysis: Example with USArrest Data

Use the built in data **USArrests**.

R Console

```
> datausa
               Murder Assault UrbanPop Rape
Alabama          13.2     236       58 21.2
Alaska           10.0     263       48 44.5
Arizona           8.1     294       80 31.0
Arkansas          8.8     190       50 19.5
California        9.0     276       91 40.6
Colorado          7.9     204       78 38.7
Connecticut       3.3     110       77 11.1
Delaware          5.9     238       72 15.8
Florida          15.4     335       80 31.9
Georgia          17.4     211       60 25.8
Hawaii            5.3      46       83 20.2
Idaho             2.6     120       54 14.2
Illinois         10.4     249       83 24.0
Indiana           7.2     113       65 21.0
Iowa              2.2      56       57 11.3
Kansas            6.0     115       66 18.0
Kentucky          9.7     109       52 16.3
Louisiana        15.4     249       66 22.2
Maine             2.1      83       51  7.8
Maryland         11.3     300       67 27.8
Massachusetts     4.4     149       85 16.3
Michigan         12.1     255       74 35.1
Minnesota         2.7      72       66 14.9
Mississippi      16.1     259       44 17.1
Missouri          9.0     178       70 28.2
Montana           6.0     109       53 16.4
Nebraska          4.3     102       62 16.5
Nevada           12.2     252       81 46.0
New Hampshire     2.1      57       56  9.5
New Jersey        7.4     159       89 18.8
New Mexico       11.4     285       70 32.1
New York         11.1     254       86 26.1
North Carolina   13.0     337       45 16.1
North Dakota      0.8      45       44  7.3
Ohio              7.3     120       75 21.4
Oklahoma          6.6     151       68 20.0
Oregon            4.9     159       67 29.3
Pennsylvania      6.3     106       72 14.9
Rhode Island      3.4     174       87  8.3
South Carolina   14.4     279       48 22.5
South Dakota      3.8      86       45 12.8
Tennessee        13.2     188       59 26.9
Texas            12.7     201       80 25.5
Utah              3.2     120       80 22.9
Vermont           2.2      48       32 11.2
Virginia          8.5     156       63 20.7
Washington        4.0     145       73 26.2
West Virginia     5.7      81       39  9.3
Wisconsin         2.6      53       66 10.8
Wyoming           6.8     161       60 15.6
> datausa <- na.omit(datausa)
>
```

Principal Component Analysis: Example with USArrest Data

```
data("USArrests") #load data

head(USArrests) #view first six rows of data
           Murder Assault UrbanPop Rape
Alabama      13.2     236       58 21.2
Alaska       10.0     263       48 44.5
Arizona       8.1     294       80 31.0
Arkansas      8.8     190       50 19.5
California    9.0     276       91 40.6
Colorado      7.9     204       78 38.7
```

Principal Component Analysis: Example with USArrest Data

```
pcdataUSarrest = prcomp(USArrests)
> pcdataUSarrest
Standard deviations (1, .., p=4):
[1] 83.732400 14.212402  6.489426  2.482790

Rotation (n x k) = (4 x 4):
                PC1         PC2         PC3         PC4
Murder   0.04170432 -0.04482166  0.07989066 -0.99492173
Assault  0.99522128 -0.05876003 -0.06756974  0.03893830
UrbanPop 0.04633575  0.97685748 -0.20054629 -0.05816914
Rape     0.07515550  0.20071807  0.97408059  0.07232502
```

The principal components are

$$Y_1 = 0.04\,\mathbf{Murder} + 0.99\,\mathbf{Assult} + 0.05\,\mathbf{UrbanPop} + 0.08\,\mathbf{Rape}$$

$$Y_2 = -0.04\,\mathbf{Murder} - 0.05\,\mathbf{Assult} + 0.97\,\mathbf{UrbanPop} + 0.20\,\mathbf{Rape}$$

$$Y_3 = 0.08\,\mathbf{Murder} - 0.07\,\mathbf{Assult} - 0.20\,\mathbf{UrbanPop} + 0.97\,\mathbf{Rape}$$

$$Y_4 = -0.99\,\mathbf{Murder} + 0.04\,\mathbf{Assult} - 0.06\,\mathbf{UrbanPop} + 0.07\,\mathbf{Rape}$$

Principal Component Analysis: Example with USArrest Data

```
> summary(pcdataUSarrest)

Importance of components:

                           PC1      PC2    PC3     PC4
Standard deviation     83.7324 14.21240 6.4894 2.48279
Proportion of Variance  0.9655  0.02782 0.0058 0.00085
Cumulative Proportion   0.9655  0.99335 0.9991 1.00000
```

Principal Component Analysis: Example with USArrest Data

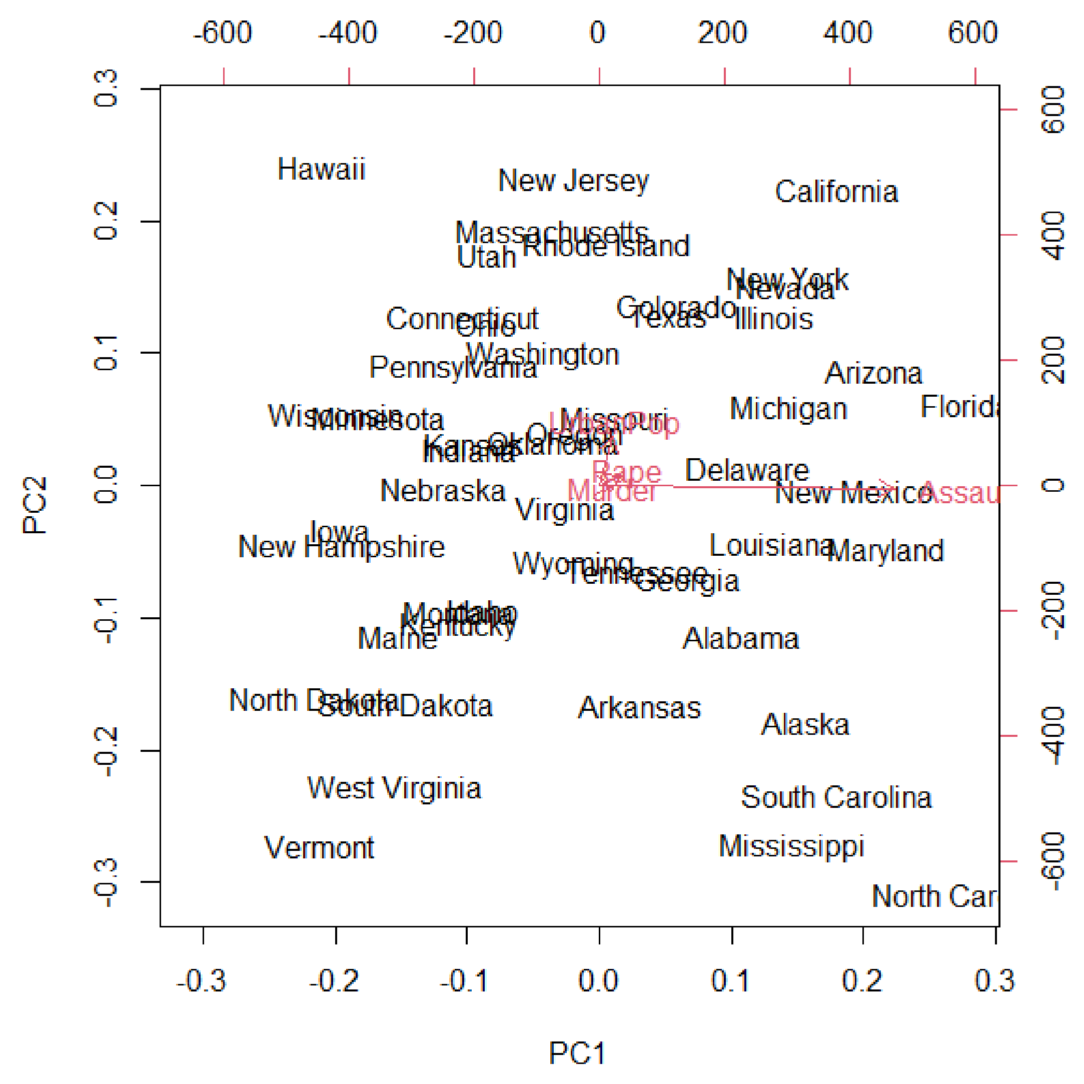
R Console

```
> data("USArrests") #load data
> head(USArrests) #view first six rows of data
           Murder Assault UrbanPop Rape
Alabama      13.2     236       58 21.2
Alaska       10.0     263       48 44.5
Arizona       8.1     294       80 31.0
Arkansas      8.8     190       50 19.5
California    9.0     276       91 40.6
Colorado      7.9     204       78 38.7
> pcdataUSarrest = prcomp(USArrests)
> pcdataUSarrest
Standard deviations (1, .., p=4):
[1] 83.732400 14.212402  6.489426  2.482790

Rotation (n x k) = (4 x 4):
                PC1         PC2         PC3         PC4
Murder   0.04170432 -0.04482166  0.07989066 -0.99492173
Assault  0.99522128 -0.05876003 -0.06756974  0.03893830
UrbanPop 0.04633575  0.97685748 -0.20054629 -0.05816914
Rape     0.07515550  0.20071807  0.97408059  0.07232502
> summary(pcdataUSarrest)
Importance of components:
                           PC1      PC2    PC3     PC4
Standard deviation     83.7324 14.21240 6.4894 2.48279
Proportion of Variance  0.9655  0.02782 0.0058 0.00085
Cumulative Proportion   0.9655  0.99335 0.9991 1.00000
> 
```

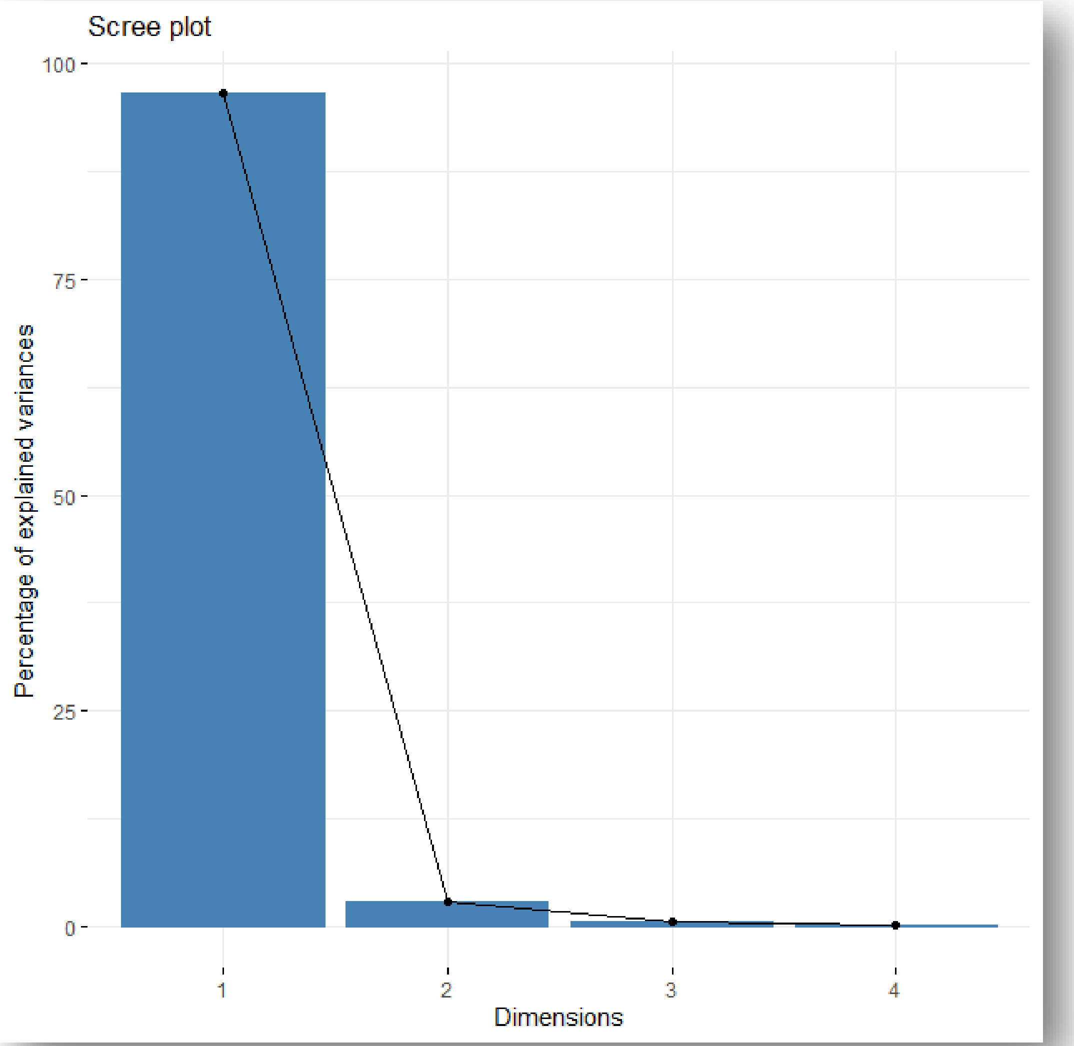
Principal Component Analysis: Biplot with USArrest Data

```
> biplot(pcdataUSarrest)
```



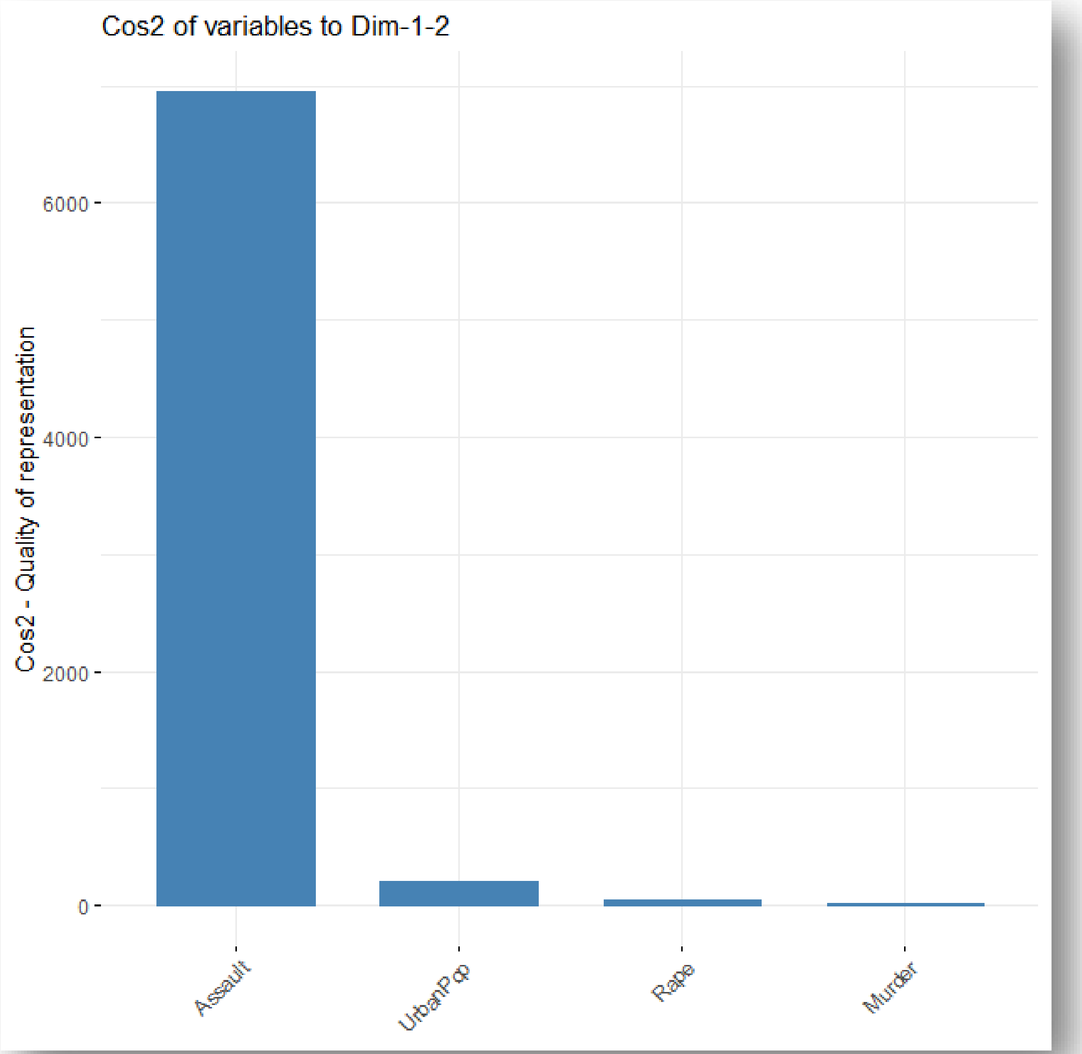
Principal Component Analysis: Scree Plot with USArrest Data

```
fviz_eig(pcdataUSarrest)
```



Principal Component Analysis: Cos Plot with USArrest Data

```
fviz_cos2(pcdataUSarrest, choice = "var", axes =
1:2)
```



Principal Component Analysis: Scree and Cos Plots with USArrest Data

```
fviz_pca_var(pcdataUSarrest, col.var = "cos2",
gradient.cols = c("blue", "red", "green",
"magenta"), repel = TRUE)
```