ADVANCED BIOSTATICS ABSTAT17

Principal Component Analysis

Carina Silva
(*carina.silva@estesl.ipl.pt*)
Higher School of Technologies and Health of Lisbon &
Center of Statistics and Applications, University of Lisbon

IGC, April, 10–13, 2017

What happens when a data set has too many variables

Here are few possible situations which you might come across:

- You find that most of the variables are correlated.

What happens when a data set has too many variables

Here are few possible situations which you might come across:

- You find that most of the variables are correlated.
- You lose patience and decide to run a model on whole data. This returns poor accuracy.

What happens when a data set has too many variables

Here are few possible situations which you might come across:

- You find that most of the variables are correlated.
- You lose patience and decide to run a model on whole data. This returns poor accuracy.
- You become indecisive about what to do.

What happens when a data set has too many variables

Here are few possible situations which you might come across:

- You find that most of the variables are correlated.
- You lose patience and decide to run a model on whole data. This returns poor accuracy.
- You become indecisive about what to do.
- You start thinking of some strategic method to find few important variables.

Introduction

Multivariate statistical analyses shows that most techniques fall into one of the following categories:

- Data reduction or structural simplification.
- Sorting and grouping.
- Investigation of the dependence among variables.
- Prediction.
- ...

Data Organization

Multivariate data are a collection of observations (or measurements) of:

- p variables ($k = 1, ..., p$)
- n cases ($j = 1, ..., n$)

Data Organization

- x_{jk} measurement of the k^{th} variable on the j^{th} case.

	Variable 1	Variable 2	...	Variable k	...	Variable p
1:	x_{11}	x_{12}	...	x_{1k}	...	x_{1p}
2:	x_{21}	x_{22}	...	x_{2k}	...	x_{2p}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
j:	x_{j1}	x_{j2}	...	x_{jk}	...	x_{jp}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
n:	x_{n1}	x_{n2}	...	x_{nk}	...	x_{np}

- The first subscript (j) represents the ROW location in the data array.
- The second subscript (k) represents the COLUMN location in the data array.

Descriptive Statistics Review

- When we have a large amount of data, it is often hard to get a manageable description of the nature of the variables under study.
- Such descriptive statistics include:
  - Means
  - Variances
  - Covariances
  - Correlations

Sample Mean

- For the k^{th} variable, the sample mean is: $\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk}$
- An array of the means for all p variables then looks like this (which we will come to know as the mean vector):

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \\ \bar{x}_4 \end{bmatrix}$$

Sample Variance

- For the k^{th} variable, the sample variance is:
  $s_k^2 = s_{kk} = \frac{1}{n}\sum_{j=1}^{n}(x_{jk} - \bar{x}_k)^2$
- Note the "kk" subscript, this will be important because the equation that produces the variance for a single variable is a derivation of the equation of the covariance for a pair of variables.
- For a pair of variables, i and k, the sample covariance is:
  $s_{ik} = \frac{1}{n}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$

Sample Covariance Matrix

- Making a matrix of all sample covariances give us:

$$\mathbf{S}_n = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{bmatrix}$$

Sample Correlation

- Sample covariances are dependent upon the scale of the variables under study.
- For this reason, the correlation is often used to describe the association between two variables.
- For a pair of variables, i and k, the sample correlation is found by dividing the sample covariance by the product of the standard deviation of the variables:
  $r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}}$
- The sample correlation:
  - Ranges from -1 to 1.
  - Measures linear association.
  - Is invariant under linear transformations of i and k.
  - Is a biased estimator.

Sample Correlation Matrix

- Making a matrix of all sample correlations give us:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}$$

Dimension Reduction

Biological research studies use more and more high-throughput biotechnologies generating large datasets with many dimensions.

Dealing with multidimensionality has been challenging to researchers, due to the difficulty in humans comprehending more than three dimensions to discover meaningful features such as linear relationships, outliers, clusters, gaps, and so on.

Several data decomposition techniques are available for the purpose of reducing dimensionality. One of the most used technique is:

- Principal Component Analysis (PCA)

What is Principal Component Analysis?

- PCA is a method of extracting important variables (in form of components) from a large set of variables available in a data set.
- With fewer variables, visualization also becomes much more meaningful.

What is PCA?

- Let's say we have a data set of dimension $n = 300$ observations and $p = 50$ variables (predictors).

What is PCA?

- Let's say we have a data set of dimension $n = 300$ observations and $p = 50$ variables (predictors).
- Since we have a large $p = 50$, there can be $p(p\text{-}1)/2$ scatter plots, i.e, more than 1000 plots possible to analyze the variable relationship.

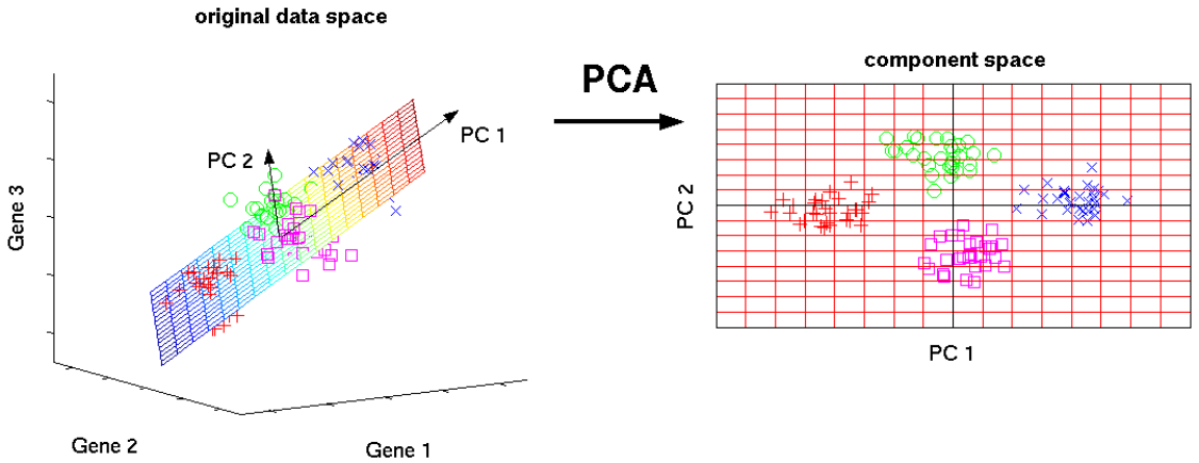


Figura: Font: Matthias Scholz, Ph.D. thesis

The three original variables (genes) are reduced to a lower number of two new variables termed principal components (PC). Such two-dimensional visualization of the samples allow us to draw qualitative conclusions about the separability of experimental conditions (marked by different colors).

PCA

There are two primary reasons for using PCA:

- Data Reduction: PCA is most commonly used to condense the information contained in a large number of original variables into a smaller set of new composite dimensions, with a minimum loss of information.
- Interpretation: PCA can be used to discover important features of a large data set. It often reveals relationships that were previously unsuspected, thereby allowing interpretations that would not ordinarily result

PCA is typically used as an intermediate step in data analysis when the number of input variables is otherwise too large for useful analysis

Introduction

- The primary goal of the PCA is to describe the variability in a multivariate data of p correlated variables by a smaller set of derived variables that are uncorrelated.

[1] linear combination is an expression constructed from a set of terms by multiplying each term by a constant and adding the results (e.g. a linear combination of x and y would be any expression of the form ax + by, where a and b are constants).

Introduction

- The primary goal of the PCA is to describe the variability in a multivariate data of p correlated variables by a smaller set of derived variables that are uncorrelated.
- The derived variables, called principal components, are linear combinations[1] of the original variables, and are listed in the order of their respective importance.

[1] linear combination is an expression constructed from a set of terms by multiplying each term by a constant and adding the results (e.g. a linear combination of x and y would be any expression of the form ax + by, where a and b are constants).

Introduction

- The primary goal of the PCA is to describe the variability in a multivariate data of p correlated variables by a smaller set of derived variables that are uncorrelated.
- The derived variables, called principal components, are linear combinations[1] of the original variables, and are listed in the order of their respective importance.
- The aim is to discard subsequent principal components after a large percentage of the variation has been explained by the first k principal component, where $k << p$.

[1] linear combination is an expression constructed from a set of terms by multiplying each term by a constant and adding the results (e.g. a linear combination of x and y would be any expression of the form ax + by, where a and b are constants).

Some algebra - Vectors

- Vectors are a sequence of numbers corresponding to measurements along various dimensions.

Some algebra - Vectors

- Vectors are a sequence of numbers corresponding to measurements along various dimensions.
- The numbers comprising the vector are called *components*, and the number equals the dimensionality of the vector.

Some algebra - Vectors

- Vectors are a sequence of numbers corresponding to measurements along various dimensions.
- The numbers comprising the vector are called *components*, and the number equals the dimensionality of the vector.
- $\overrightarrow{x} = [8, 6, 7, 5, 3]$ is a vector with dimension 5 and $x_1 = 8$ is a component in position 1 of the vector.

Some algebra - Vector Terminology

- Vector length (or norm): $|\overrightarrow{v}| = \sqrt{\sum_{i=1}^{n} x_i^2}$. E.g., if $\overrightarrow{v} = [4, 11, 8, 10]$, then $|\overrightarrow{v}| = \sqrt{4^2 + 11^2 + 8^2 + 10^2} = 17.35$

Some algebra - Vector Terminology

- Vector length (or norm): $|\overrightarrow{v}| = \sqrt{\sum_{i=1}^{n} x_i^2}$. E.g., if $\overrightarrow{v} = [4, 11, 8, 10]$, then $|\overrightarrow{v}| = \sqrt{4^2 + 11^2 + 8^2 + 10^2} = 17.35$
- Vector addition: $[3, 2, 1, -2] + [2, -1, 4, 1] = [(3 + 2), (2 - 1), (1 + 4), (-2 + 1)] = [5, 1, 5, -1]$

Some algebra - Vector Terminology

- Vector length (or norm): $|\overrightarrow{v}| = \sqrt{\sum_{i=1}^{n} x_i^2}$. E.g., if $\overrightarrow{v} = [4, 11, 8, 10]$, then $|\overrightarrow{v}| = \sqrt{4^2 + 11^2 + 8^2 + 10^2} = 17.35$
- Vector addition: $[3, 2, 1, -2] + [2, -1, 4, 1] = [(3 + 2), (2 - 1), (1 + 4), (-2 + 1)] = [5, 1, 5, -1]$
- Scalar multiplication: A scalar is a real number which is multiplied by each component of a vector. E.g., $1.5 * [3, 6, 8, 4] = [1.5 * 3, 1.5 * 6, 1.5 * 8, 1.5 * 4] = [4.5, 9, 12, 6]$.

Some algebra - Vector Terminology

- Inner product: (dot product or scalar product), is a multiplication of vectors of the same dimension: $(\overrightarrow{x}, \overrightarrow{y}) = \overrightarrow{x}.\overrightarrow{y} = \sum_{i=1}^{n} x_i y_i$

Some algebra - Vector Terminology

- Inner product: (dot product or scalar product), is a multiplication of vectors of the same dimension: $(\overrightarrow{x}, \overrightarrow{y}) = \overrightarrow{x}.\overrightarrow{y} = \sum_{i=1}^{n} x_i y_i$
- Orthognonality: two vectors are *orthogonal* to each other if their inner product equals zero.

Some algebra - Vector Terminology

- Inner product: (dot product or scalar product), is a multiplication of vectors of the same dimension: $(\overrightarrow{x}, \overrightarrow{y}) = \overrightarrow{x}.\overrightarrow{y} = \sum_{i=1}^{n} x_i y_i$
- Orthognonality: two vectors are *orthogonal* to each other if their inner product equals zero.
- Normal vector: is a vector of length (norm) 1. Any vector with length>1 can be normalized by dividing each component by the vector's length. E.g., if $\overrightarrow{v} = [2, 4, 1, 2]$, then $|\overrightarrow{v}| = \sqrt{2^2 + 4^2 + 1^2 + 2^2} = 5$. Then $\overrightarrow{u} = [2/5, 4/5, 1/5, 2/5]$ is a normal vector because $|\overrightarrow{u}| = 1$.

Some algebra - Vector Terminology

- Inner product: (dot product or scalar product), is a multiplication of vectors of the same dimension: $(\overrightarrow{x}, \overrightarrow{y}) = \overrightarrow{x}.\overrightarrow{y} = \sum_{i=1}^{n} x_i y_i$
- Orthognonality: two vectors are *orthogonal* to each other if their inner product equals zero.
- Normal vector: is a vector of length (norm) 1. Any vector with length>1 can be normalized by dividing each component by the vector's length. E.g., if $\overrightarrow{v} = [2, 4, 1, 2]$, then $|\overrightarrow{v}| = \sqrt{2^2 + 4^2 + 1^2 + 2^2} = 5$. Then $\overrightarrow{u} = [2/5, 4/5, 1/5, 2/5]$ is a normal vector because $|\overrightarrow{u}| = 1$.
- Orthonormal vectors: vectors of unit length that are orthogonal to each other.

Some algebra - Identity matrix

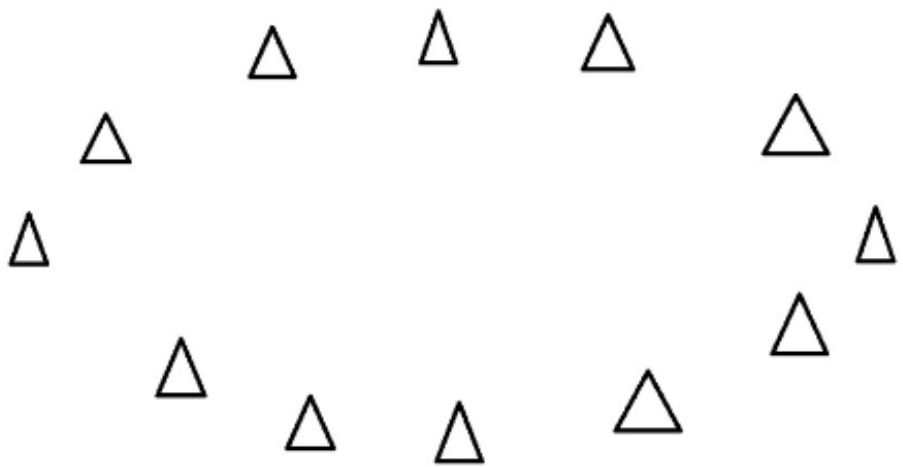
The identity matrix is a square matrix with entries on the diagonal equal 1 and all other entries equal zero. The diagonal is all entries $a_{ii}, i = 1, ..., n$. Usually this matrix is denoted by I and is a particular case of a diagonal matrix E.g.,

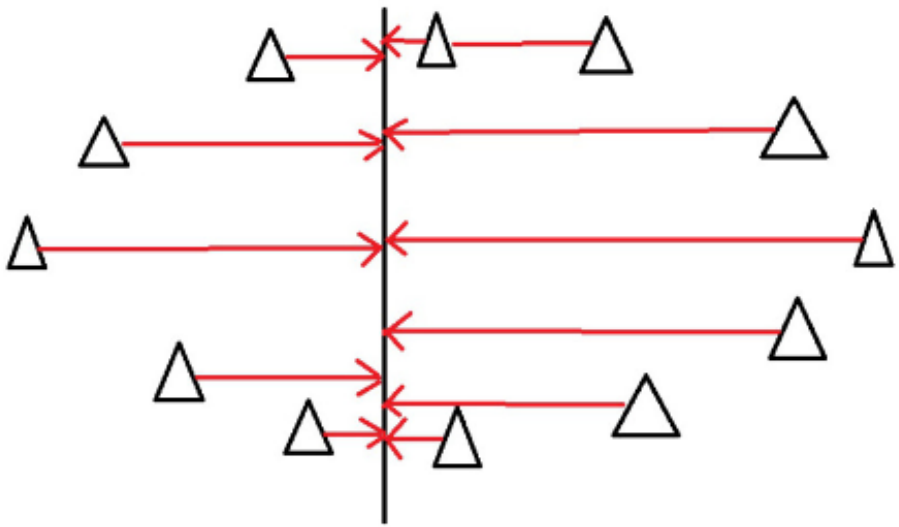
$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

What are principal components?

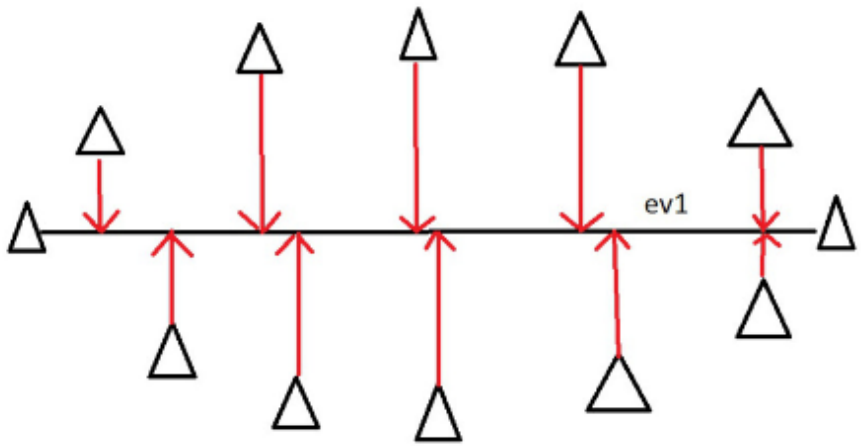
- They're the underlying structure in the data.
- They are the directions where there is the most variance, the directions where the data is most spread out

- Imagine that the triangles are points of data. To find the direction where there is most variance, find the straight line where the data is most spread out when projected onto it.





- The data isn't very spread out here, therefore it doesn't have a large variance. It is probably not the principal component.



- On this line the data is way more spread out, it has a large variance. In fact there isn't a straight line you can draw that has a larger variance than a horizontal one. A horizontal line is therefore the principal component in this example.

Let's say we have a random vector $X_1, X_2, ..., X_p$.

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

with population variance-covariance matrix:

$$var(\mathbf{X}) = \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p \end{pmatrix}$$

Variance and Covariance

To understand how PCA works, we need to recall the concepts of variance and covariance.

- Variance of a sample is given by
  $s^2 = \frac{\sum(x_i - \bar{x})}{n-1}$
- Covariance between two variables, x and y, as
  $cov(x, y) = \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$,
  when x varies with y, this expression will tend do accumulate positive terms; when they are independent, the covariance will be zero, and will be negative if they are anti-correlated. Note that the variance of a variable is just the covariance of that variable with itself.

$$\xi_1 = \phi_{11}X_1 + \phi_{12}X_2 + \ldots + \phi_{1p}X_p$$
$$\xi_2 = \phi_{21}X_1 + \phi_{22}X_2 + \ldots + \phi_{2p}X_p$$
$$\cdots$$
$$\xi_p = \phi_{p1}X_1 + \phi_{p2}X_2 + \ldots + \phi_{pp}X_p$$

where:

- ξ_i are the principal components, where $cov(\xi_i, \xi_j) = 0$
- ϕ_{ij} are the coefficients, where they are collected into the vectors:

$$\phi_i = \begin{pmatrix} \phi_{i1} \\ \phi_{i2} \\ \vdots \\ \phi_{ip} \end{pmatrix}$$

these vectors are called eigenvectors.

How do we find the coefficients ϕ_{ij} for a principal component?

- The solution involves the eigenvalues and eigenvectors of the variance-covariance matrix Σ.
- The variance for the ith principal component is equal to the ith eigenvalue:

$$var(\xi_i) = var(\phi_{i1}X_1 + \phi_{i2}X_2 + \ldots + \phi_{ip}X_p) = \lambda_i$$

- The eigenvectors are normal vectors: $|\phi_i| = 1$
- Where the eigenvalues are ordered so that λ_1 is the largest value and λ_p the smallest: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$
- The principal components are uncorrelated with one another.
- $trace(\Sigma) = \sigma_1^2 + \sigma_2^2 + ... + \sigma_p^2 = \lambda_1 + \lambda_2 + ... + \lambda_p$
  This will give us an interpretation of the components in terms of the amount of the full variation explained by each component.

How to start

We assume that the multi-dimensional data have been collected in a data matrix, in which the rows are associated with the cases and the columns with the variables.

Preparing Data Analysis

- Start From Correlation Matrix or Covariance Matrix
  - The correlation matrix is simply the covariance matrix, standardized by setting all variances equal to one.
  - When scales of variables are similar, the covariance matrix is always preferred, as the correlation matrix will lose information when standardizing the variance.
  - The correlation matrix is recommended when variables are measured in different scales. (if most of the correlation coefficients are smaller than 0.3, PCA will not help.)

Eigenvectors and Eigenvalues

Eigenvectors and values exist in pairs: every eigenvector has a corresponding eigenvalue.The number of eigenvectors/values that exist equals the number of dimensions the data set has (p).

- An eigenvector is a direction, in the example above the eigenvector was the direction of the line (vertical, horizontal, 45 degrees etc.)
- An eigenvalue is a number, telling you how much variance there is in the data in that direction, in the example above the eigenvalue is a number telling us how spread out the data is on the line.

- The data covariance matrix Σ ($p \times p$ matrix) is transformed into a diagonal matrix Λ ($p \times p$ matrix), whose diagonal elements λ_{ii} are the variances of the new variables

- The data covariance matrix Σ ($p \times p$ matrix) is transformed into a diagonal matrix Λ ($p \times p$ matrix), whose diagonal elements λ_{ii} are the variances of the new variables
- This diagonal matrix is constructed such that the diagonal elements are in descending order.

- The data covariance matrix Σ ($p \times p$ matrix) is transformed into a diagonal matrix Λ ($p \times p$ matrix), whose diagonal elements λ_{ii} are the variances of the new variables
- This diagonal matrix is constructed such that the diagonal elements are in descending order.

- Hence λ_{11} represents the largest amount of variation in the data and λ_{pp} the least amount.

- The data covariance matrix Σ ($p \times p$ matrix) is transformed into a diagonal matrix Λ ($p \times p$ matrix), whose diagonal elements λ_{ii} are the variances of the new variables
- This diagonal matrix is constructed such that the diagonal elements are in descending order.

- Hence λ_{11} represents the largest amount of variation in the data and λ_{pp} the least amount.

- The proportion of the total variation contributed by each component ξ_i is given by $\frac{\lambda_{ii}}{\sum \lambda_{ii}}$.

- The data covariance matrix Σ ($p \times p$ matrix) is transformed into a diagonal matrix Λ ($p \times p$ matrix), whose diagonal elements λ_{ii} are the variances of the new variables
- This diagonal matrix is constructed such that the diagonal elements are in descending order.

- Hence λ_{11} represents the largest amount of variation in the data and λ_{pp} the least amount.

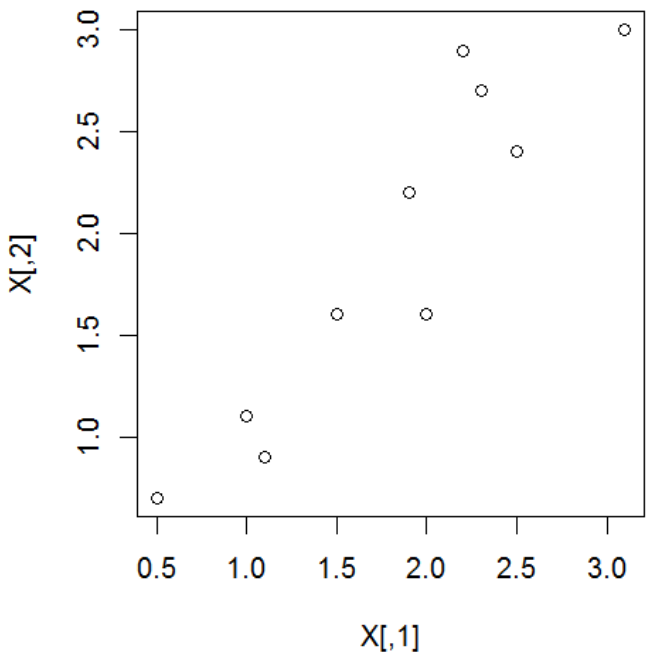
- The proportion of the total variation contributed by each component ξ_i is given by $\frac{\lambda_{ii}}{\sum \lambda_{ii}}$.

- We retain only the components which together represent a certain percentage of the total variation, say, 90%.

Example in R

```
X<-matrix(c(2.5, 2.4,
0.5, 0.7,
2.2, 2.9,
1.9, 2.2,
3.1, 3.0,
2.3, 2.7,
2, 1.6,
1, 1.1,
1.5, 1.6,
1.1, 0.9),ncol=2,byrow=T)

> X
      [,1] [,2]
 [1,]  2.5  2.4
 [2,]  0.5  0.7
 [3,]  2.2  2.9
 [4,]  1.9  2.2
 [5,]  3.1  3.0
 [6,]  2.3  2.7
 [7,]  2.0  1.6
 [8,]  1.0  1.1
 [9,]  1.5  1.6
[10,]  1.1  0.9
```

```
plot(X)    #plot the data
```



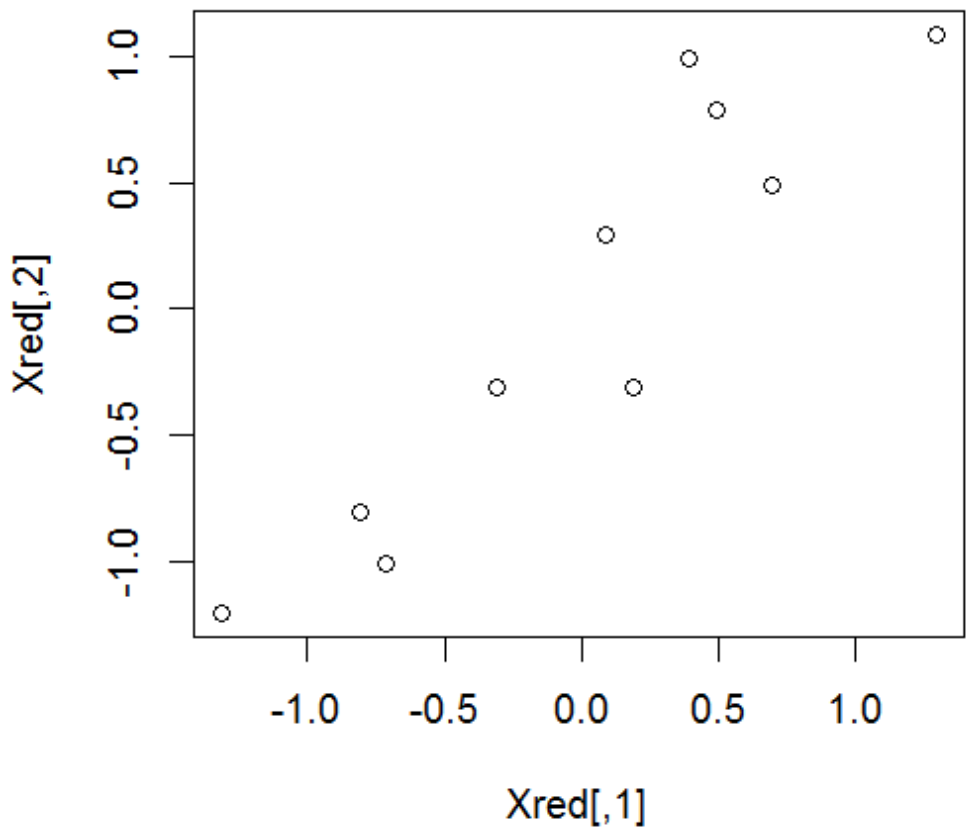
The first thing you need to do is center or standardize the data. Since the data is in the same scale, we just center the data.

```
#center the data

Xred<-matrix(0,nrow=nrow(X),ncol=ncol(X))
#subtract the mean

Xred[,1]<-X[,1]-mean(X[,1])
Xred[,2]<-X[,2]-mean(X[,2])
```

Note: If the variables in the data set are not in the same scale, you will need to divide them by the standard deviation to get them standardized, and then you have to use the correlation matrix instead of the covariance matrix.

Plot the standardized data

`plot(Xred)`



Get the covariance matrix, eigenvalues and eigenvectors.

```
> S<-cov(Xred)
> S
          [,1]      [,2]
[1,] 0.6165556 0.6154444
[2,] 0.6154444 0.7165556

> #compute the eigenvalues and eigenvectors
> ev<-eigen(S)

> ev
$values
[1] 1.2840277 0.0490834

$vectors
          [PC1]       [PC2]
[Var 1] 0.6778734 -0.7351787
[Var 2] 0.7351787  0.6778734

>sum(ev$values) #sum of the eigenvalues represents the total of the variance
[1] 1.333111

# in the data set
>var(Xred[,1])+var(Xred[,2])
[1] 1.333111
#the eigenvectors have norm 1
> sum(ev$vector[,1]^2)
[1] 1
> sum(ev$vector[,2]^2)
[1} 1
```

```
> ev
$values
[1] 1.2840277 0.0490834

$vectors
           [,1]       [,2]
[1,] 0.6778734 -0.7351787
[2,] 0.7351787  0.6778734
```

loading

PC1 **PC2**

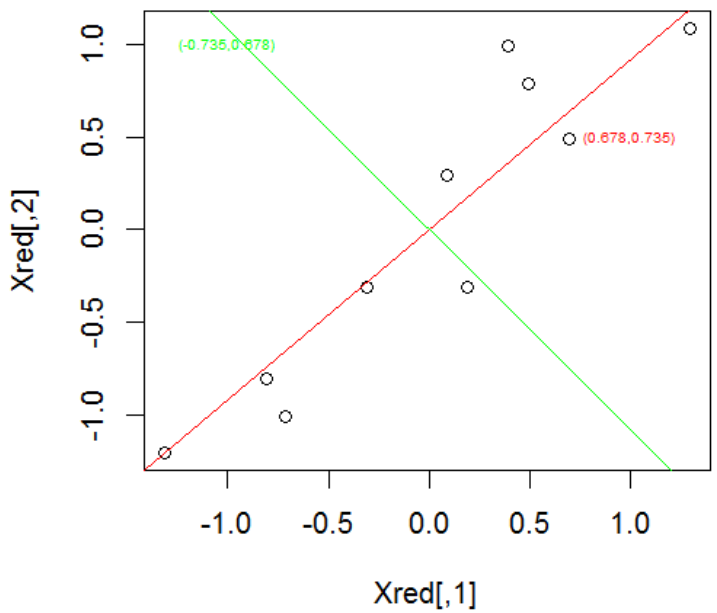
eigenvectors

PC1=0.678Variable1+0.735Variable

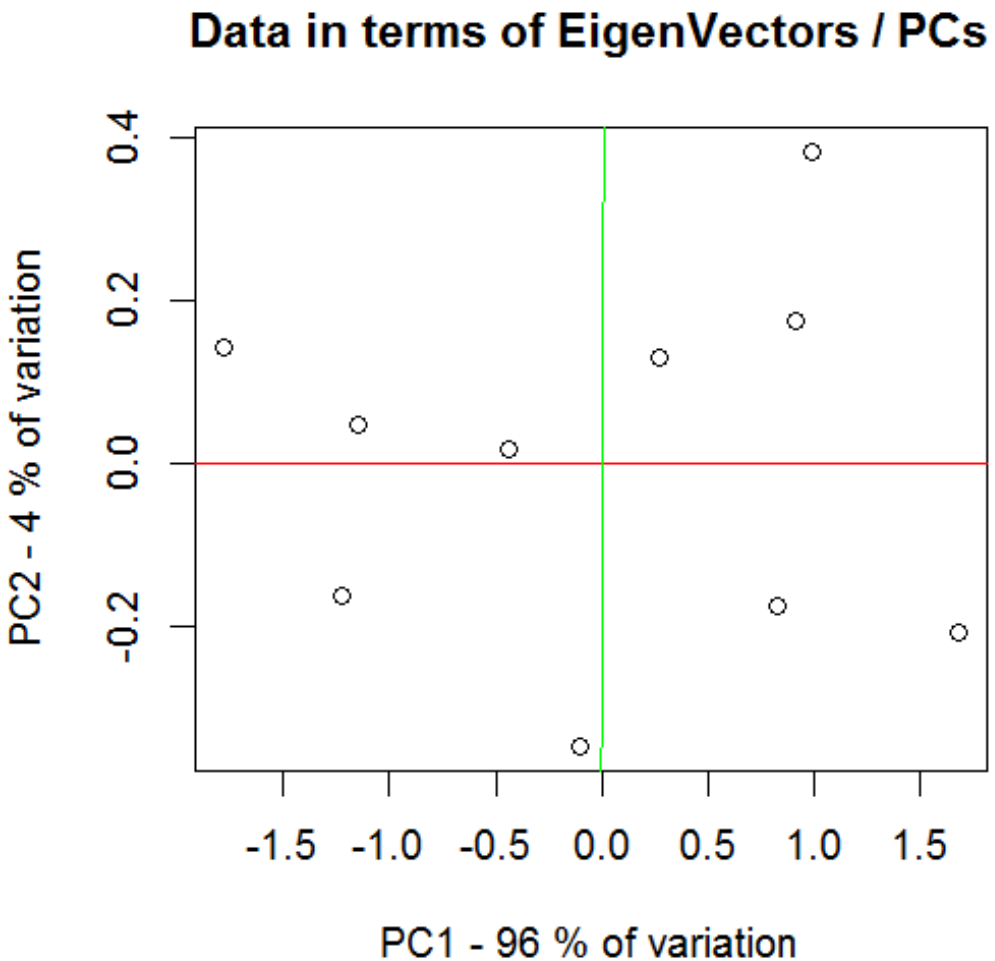
Indicates the association of the principal component with each variable

The eigenvectors are the principal components.

```
loadings = ev$vectors

# Let's plot them on the standardized plot

pc1.slope = ev$vectors[1,1]/ev$vectors[2,1]
pc2.slope = ev$vectors[1,2]/ev$vectors[2,2]

abline(0,pc1.slope,col="red")
abline(0,pc2.slope,col="green")

text(1,0.5,"(0.678,0.735)"
text(-1,1,"(-0.735,0.678)"
```



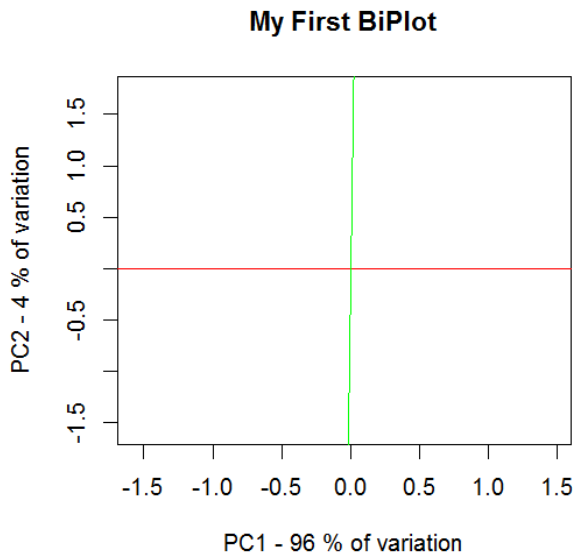
```
# See how much variation each eigenvector accounts for

pc1.var = 100*round(ev$values[1]/sum(ev$values),digits=2)
[1] 96
pc2.var = 100*round(ev$values[2]/sum(ev$values),digits=2)
[1] 4

# Multiply the scaled data by the eigen vectors (principal components)
scores<-t(ev$vector)%*%t(Xred) #data expressed in terms in principal components
t(scores)

             [,1]        [,2]
 [1,]  0.82797019 -0.17511531
 [2,] -1.77758033  0.14285723
 [3,]  0.99219749  0.38437499
 [4,]  0.27421042  0.13041721
 [5,]  1.67580142 -0.20949846
 [6,]  0.91294910  0.17528244
 [7,] -0.09910944 -0.34982470
 [8,] -1.14457216  0.04641726
 [9,] -0.43804614  0.01776463
[10,] -1.22382056 -0.16267529

xlab=paste("PC1 - ",pc1.var," % of variation",sep="")
ylab=paste("PC2 - ",pc2.var," % of variation",sep="")
plot(t(scores),main="Data in terms of EigenVectors / PC",xlab=xlab,ylab=ylab)
abline(0,0,col="red")
abline(0,90,col="green")
```


Data in terms of EigenVectors / PCs
PC2 - 4 % of variation
0.4
0.2
0.0
-0.2
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
PC1 - 96 % of variation

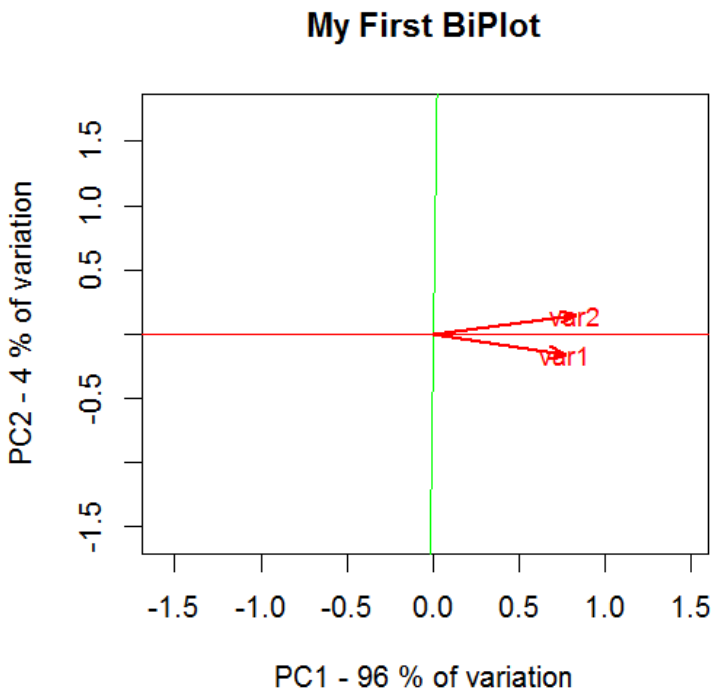
Biplot

It plots the scores (points) and the variables (vectors) on the same graph

```
plot(t(scores)[,1]/sd[1],t(scores)[,2]/sd[2],main="My First BiPlot",xlab=xlab,ylab=ylab,type="n")
abline(0,0,col="red")
abline(0,90,col="green")

# First plot the variables as vectors
arrows(0,0,loadings[,1]*sd[1],loadings[,2]*sd[2],length=0.1, lwd=2,angle=20, col="red")
text(loadings[,1]*sd[1],loadings[,2]*sd[2],c("var1","var2"), col="red", cex=0.9)
```



Biplot


My First BiPlot
PC2 - 4 % of variation
PC1 - 96 % of variation
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
var2
var1

Exercise 2

Consider the iris dataset (included with `R`) which gives the petal width, petal length, sepal width, sepal length and species for 150 irises. To view more information about the dataset, enter `help(iris)`.

Perform a PCA analysis to the `iris` data.