Principal Components Analysis (PCA)

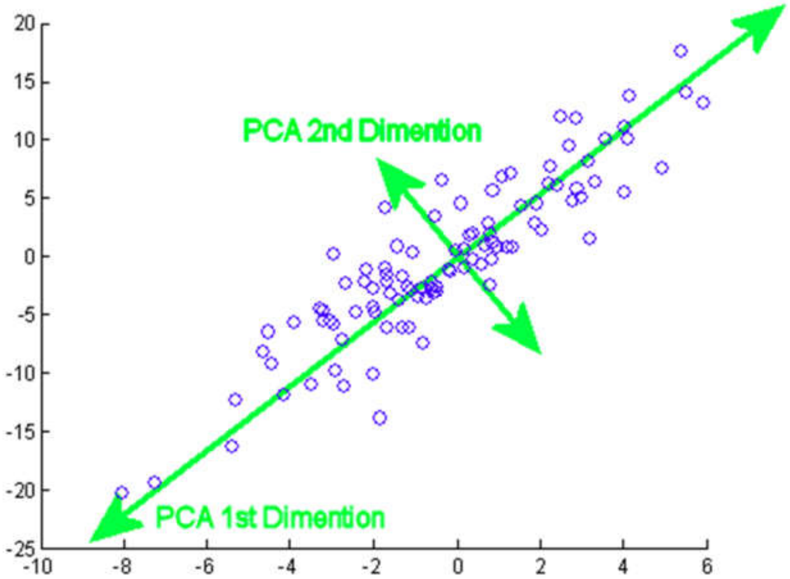
PCA is a method used to reduce number of variables (dimension) in a data. It reduces the dimension of the data with the aim of retaining as much information as possible.

It is a mathematical procedure that transforms a number of (possibly) correlated variables of a data into a (smaller) number of uncorrelated variables called **principal components**.

PCA has various names in various fields, so it is also known as the Karhunen-Lo`eve transformation, the Hoteling transformation, the method of empirical orthogonal functions, singular value decomposition or factor analysis.

Mathematics behind PCA:

There are several equivalent ways of deriving the principal components mathematically. **The simplest one is by finding the directions of projections of data-vector which maximize the variance.** The first principal component is the direction in feature space along which projections have the largest variance. The second principal component is the direction which maximizes variance among all directions orthogonal to the first. The k^{th} component is the variance-maximizing direction orthogonal to the previous ($k - 1$) components.



As PCA tries to "retain as much information as possible", **we should to look for the projection with the smallest average (mean-squared) distance between the original vectors and their projections on to the principal components.** It will be proved that this is equivalent to maximizing the variance.

Now, let us assume that the data have been "centered", so that every feature dimension has mean 0. The centered data is represented by matrix **X**, where rows are objects and columns are feature dimensions or simply ***feature***s.

For example:

Data =

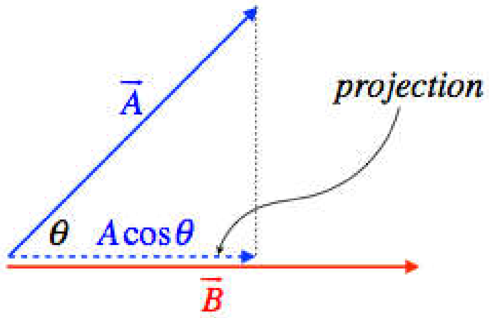
x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

a) Original data

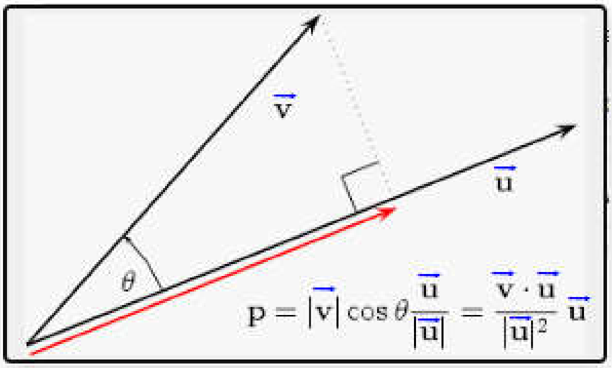
DataAdjust =

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

b) mean subtracted or "centered" data after adjustment



$$\frac{\vec{A}\cdot\vec{B}}{|\vec{B}|} = A\cos\theta$$


$\vec{v}$
$\vec{u}$
θ
$p = |\vec{v}|\cos\theta\frac{\vec{u}}{|\vec{u}|} = \frac{\vec{v}\cdot\vec{u}}{|\vec{u}|^2}\vec{u}$

Now, if we project the a data vector $\overrightarrow{x_i}$ on $\overrightarrow{w_i}$, where we assume that $\overrightarrow{w_i}$ is in the direction a principal component, then the error in projection,

$$\begin{aligned} \|\vec{x_i} - (\vec{w}\cdot\vec{x_i})\vec{w}\|^2 &= \|\vec{x_i}\|^2 - 2(\vec{w}\cdot\vec{x_i})(\vec{w}\cdot\vec{x_i}) + \|\vec{w}\|^2 \\ &= \|\vec{x_i}\|^2 - 2(\vec{w}\cdot\vec{x_i})^2 + 1 \end{aligned}$$

summing,

$$= \left(n + \sum_{i=1}^{n}\|\vec{x_i}\|^2\right) - 2\sum_{i=1}^{n}(\vec{w}\cdot\vec{x_i})^2$$

The first term in the big parenthesis doesn't depend on $\overrightarrow{w_l}$, so it doesn't matter for trying to minimize the sum-of-square error. To make the error small, what we must do is to make the second sum big, i.e., we want to maximize,

$$\sum_{i=1}^{n} (\vec{w} \cdot \vec{x_i})^2$$

Equivalently, since n doesn't depend on $\vec{w}$, we want to maximize

$$\frac{1}{n}\sum_{i=1}^{n} (\vec{w} \cdot \vec{x_i})^2$$

Now, the mean of a square is equal to the square of the mean plus the variance:

$$\frac{1}{n}\sum_{i=1}^{n} (\vec{w} \cdot \vec{x_i})^2 = \left(\frac{1}{n}\sum_{i=1}^{n} \vec{x_i} \cdot \vec{w}\right)^2 + \mathrm{Var}\left[\vec{w} \cdot \vec{x_i}\right]$$

But, as the data is centered, the mean of the projections is zero. Hence, it is proved that minimizing the residual sum of squares is equivalent to maximizing the variance of the projections.

Method for Maximizing Variance

If our *p* dimensional data vectors are stacked into an *n × p* matrix, **X**, where n is the number of vectors or number of data, then the projections are given by ***Xw***, which is an ***n × 1*** matrix. The variance is,

$$\sigma^2_{\vec{w}} = \frac{1}{n}\sum_{i} (\vec{x_i} \cdot \vec{w})^2$$

$$
\begin{aligned}
&= \frac{1}{n}(\mathbf{Xw})^T(\mathbf{Xw}) \\
&= \frac{1}{n}\mathbf{w}^T\mathbf{X}^T\mathbf{Xw} \\
&= \mathbf{w}^T\frac{\mathbf{X}^T\mathbf{X}}{n}\mathbf{w} \\
&= \mathbf{w}^T\mathbf{Vw}
\end{aligned}
$$

V is the covariance matrix of X.

So, the problem is to find a set of $\overrightarrow{w_i}$ to maximize the variance.

To do this, we need to make sure that we only look at unit vectors because we want to make the Principal components uncorrelated to each other, then we can avoid the redundancy of the data. Hence, it is a problem of constrained maximization (optimization). The constraint is that, $w^Tw = 1$.

Let a function *f(w)* that we want to maximize. Here, that function is $\mathbf{w}^T\mathbf{V}\mathbf{w}$. We also have an equality constraint, *g(w) = c*. Here, *g(w)* = $\mathbf{w}^T\mathbf{w}$ and *c* = 1. We re-arrange the constraint equation so its right hand side is zero, *g(w) − c = 0*. Now we can add an extra variable to the problem, called the **Lagrange multiplier λ**, and consider

u(w, λ) = f(w)−λ(g(w)−c) as our new **objective function**,

so we differentiate with respect to both arguments and set the derivatives equal to zero:

$$
\begin{aligned}
\frac{\partial u}{\partial w} &= 0 = \frac{\partial f}{\partial w} - \lambda\frac{\partial g}{\partial w} \\
\frac{\partial u}{\partial \lambda} &= 0 = -(g(w) - c)
\end{aligned}
$$

For our problem,

$$\begin{aligned} u &= \mathbf{w}^T\mathbf{V}\mathbf{w} - \lambda(\mathbf{w}^T\mathbf{w} - 1) \\ \frac{\partial u}{\partial \mathbf{w}} &= 2\mathbf{V}\mathbf{w} - 2\lambda\mathbf{w} = 0 \\ \mathbf{V}\mathbf{w} &= \lambda\mathbf{w} \end{aligned}$$

Hence, $w^TVw=w^T\lambda w=\lambda\ (w^Tw)= \lambda$

Thus, desired vector w is an eigenvector of the covariance matrix V, and the maximizing vector will be the one associated with the largest eigenvalue λ.

Remember: V is a symmetric matrix.

Some Proofs:

1. Prove that, a square matrix A is *orthogonally diagonalizable* if there exists an orthogonal matrix Q such that $Q^T AQ = D$ is a diagonal matrix.

Also Prove that, in that case A is a symmetric matrix.

Proof: If we have $Q^TAQ = D$ then times Q on the left, and Q^T on the right gives

$A = QDQ^T$ (since $Q^T = Q^{-1}$).

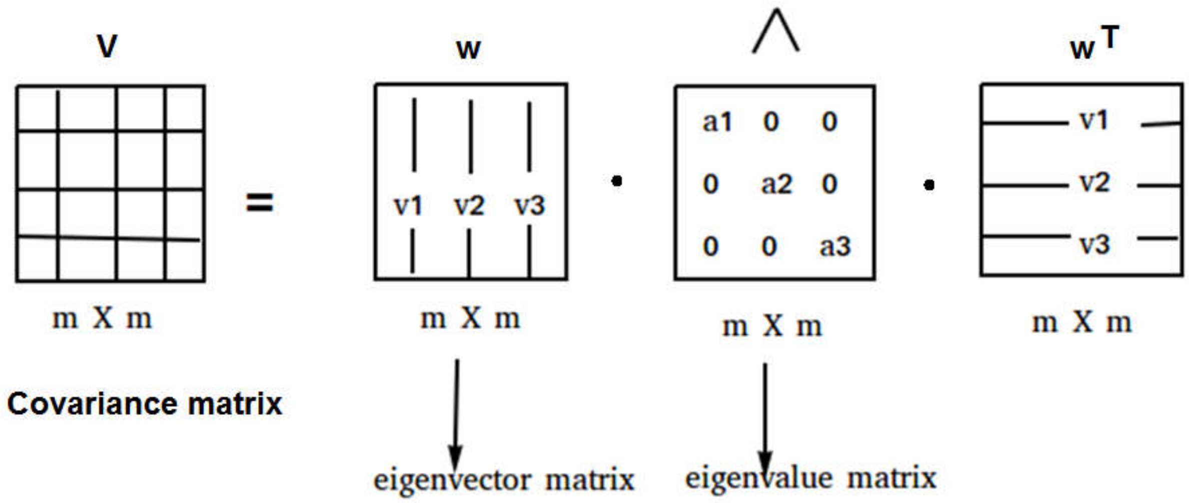
Then $A^T = (QDQ^T)^T = (Q^T)^T D^T Q^T = QDQ^T = A$, so A is symmetric.

Prove that, Eigenvectors of a symmetric matrix corresponding to different eigenvalues are orthogonal.

Proof. Let $A^T = A$ have eigenvectors $\vec{v}_1$ and $\vec{v}_2$ for eigenvalues $\lambda_1 \neq \lambda_2$. We compute the dot product $(A\vec{v}_1)\cdot\vec{v}_2 = (\lambda_1\vec{v}_1)\cdot\vec{v}_2 = \lambda_1(\vec{v}_1\cdot\vec{v}_2)$. On the other hand, the left-hand side can be written as a matrix product:

$(A\vec{v}_1)\cdot\vec{v}_2 = \vec{v}_1\,(A\vec{v}_2) = \vec{v}_1\cdot(\lambda_2\vec{v}_2) = \lambda_2(\vec{v}_1\cdot\vec{v}_2)$. Thus, $\lambda_1(\vec{v}_1\cdot\vec{v}_2) = \lambda_2(\vec{v}_1\cdot\vec{v}_2)$. Since $\lambda_1 \neq \lambda_2$ we must have $\vec{v}_1\cdot\vec{v}_2 = 0$.

Note that, $\mathbf{V}\vec{w_i}=\lambda\vec{w_i} \Rightarrow \mathbf{Vw=w\Lambda} \Rightarrow \mathbf{V= w\Lambda w^T}$



Summarizing the steps of PCA:

Step 1: Get the data, say a *nXp* matrix.

Step 2: Centre the data, i.e., subtract mean for each dimension *p* and get the centered data X of dimension *nXp*

Data =

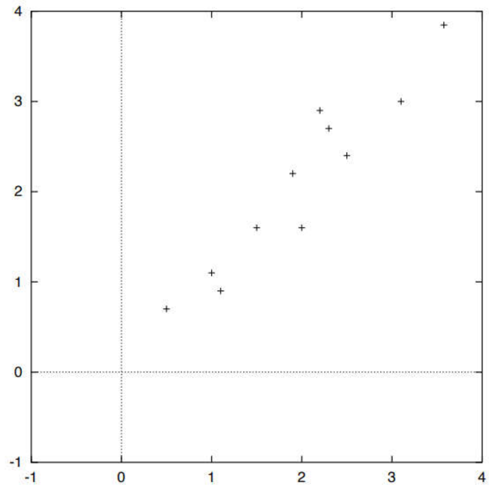
x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

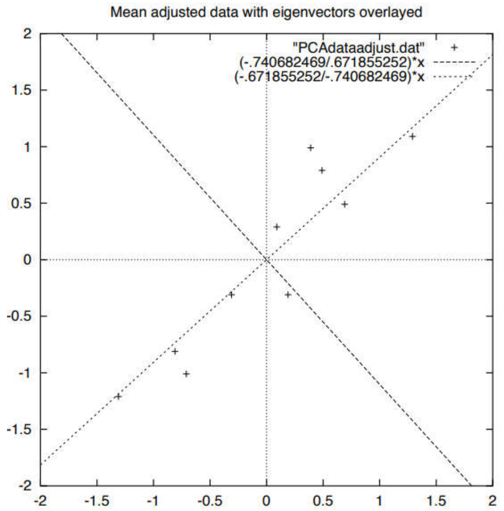
DataAdjust =

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

a) Original data

b) mean subtracted or “centered” data after adjustment





Step 3: Calculate the covariance matrix (V) of centered data X

The dimension of V will be ***pXp***

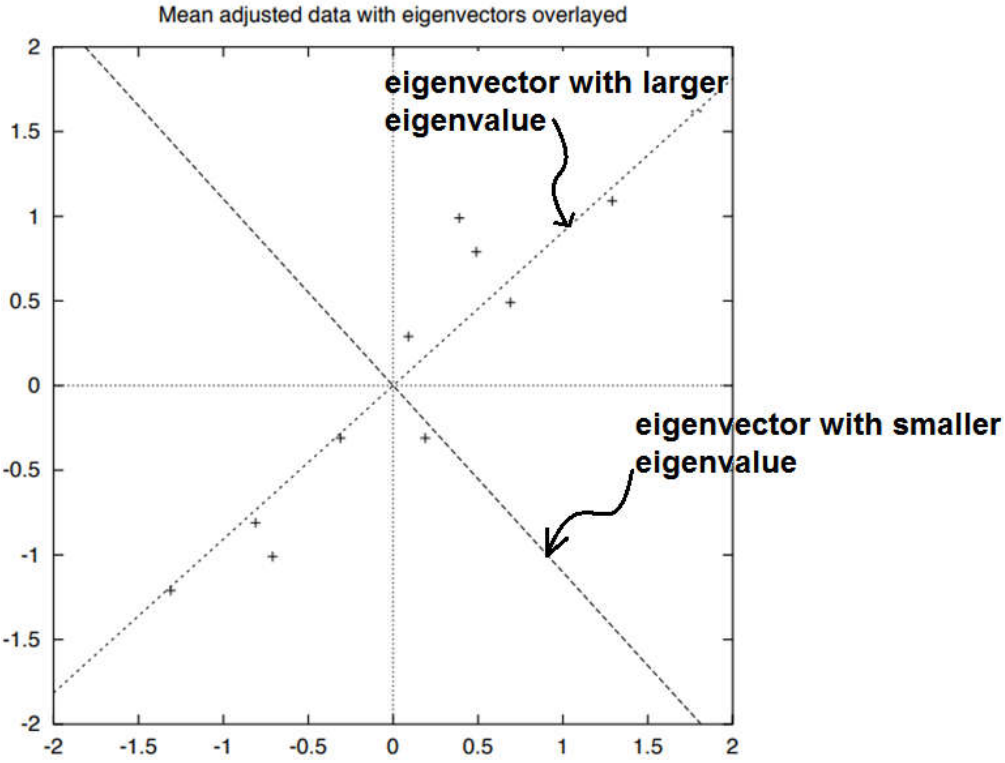
$$cov = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

Step 4: Calculate the eigenvectors and eigenvalues of the covariance matrix

The number of eigenvalues and eigenvector will be *p (i.e., the dimension of the data)*

$$eigenvalues = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

$$eigenvectors = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$

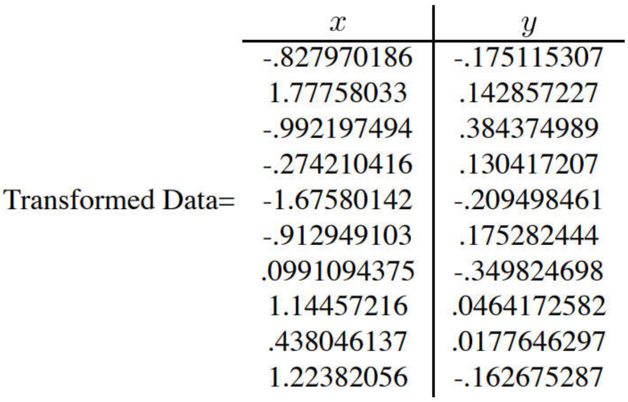


Step 5: Eigenvector with large eigenvalue indicates high variance ,i.e., more significant direction for projection

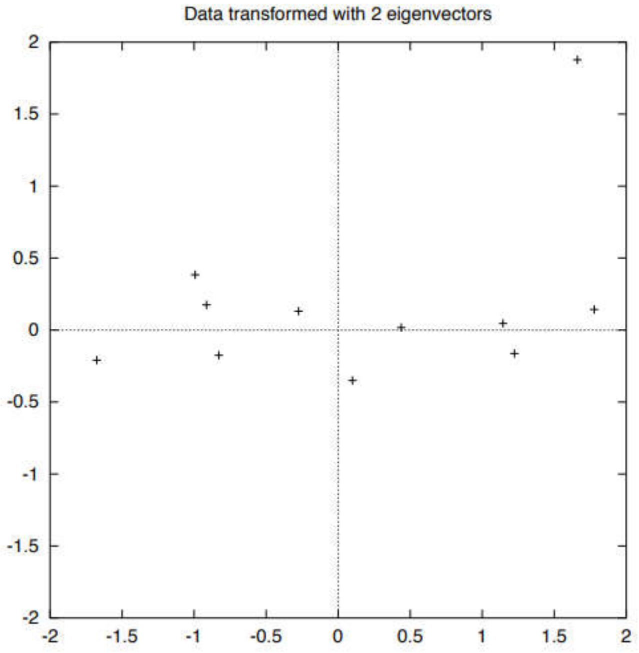
Note that, the new axes are not the earlier ones, the data is presented wrt new axes (i.e, new bases).

With all the eigenvectors the new data will be represented along new bases i.e, new transformed axes which are nothing but in the directions of the eigenvectors.

$[X_{new}]^{T}{}_{pXn}$=$[\Lambda]_{pXp}$•$[X]^{T}{}_{pXn}$



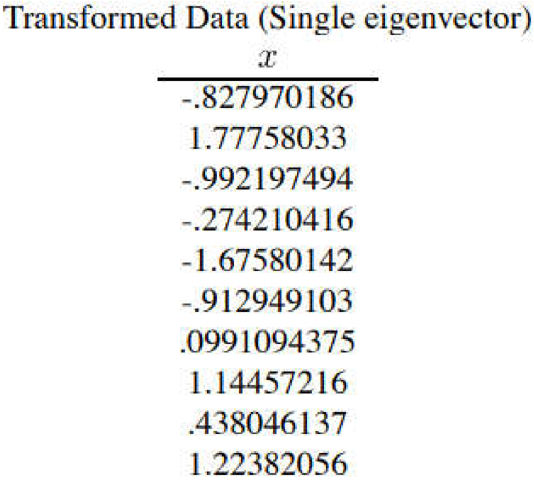
	x	y
	-.827970186	-.175115307
	1.77758033	.142857227
	-.992197494	.384374989
	-.274210416	.130417207
Transformed Data=	-1.67580142	-.209498461
	-.912949103	.175282444
	.0991094375	-.349824698
	1.14457216	.0464172582
	.438046137	.0177646297
	1.22382056	-.162675287


Data transformed with 2 eigenvectors

If we consider the only the larger eigenvector,

$$\begin{pmatrix} -.677873399 \\ -.735178656 \end{pmatrix}$$

Then that data dimension is reduced,



Transformed Data (Single eigenvector)

x
-.827970186
1.77758033
-.992197494
-.274210416
-1.67580142
-.912949103
.0991094375
1.14457216
.438046137
1.22382056

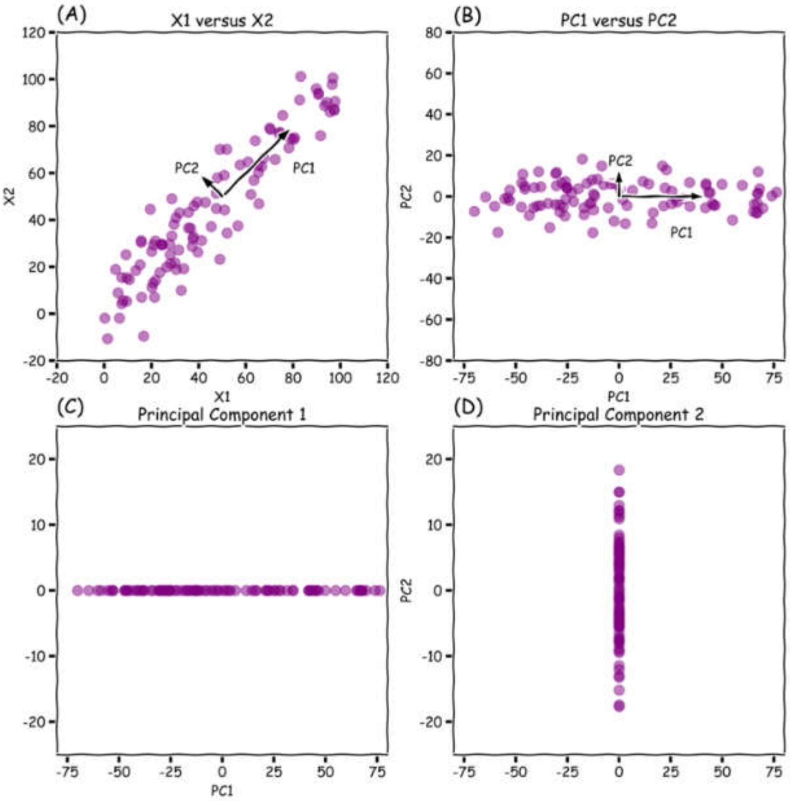
Therefore, if we choose eigenvectors with the first *k* largest eigenvalues,

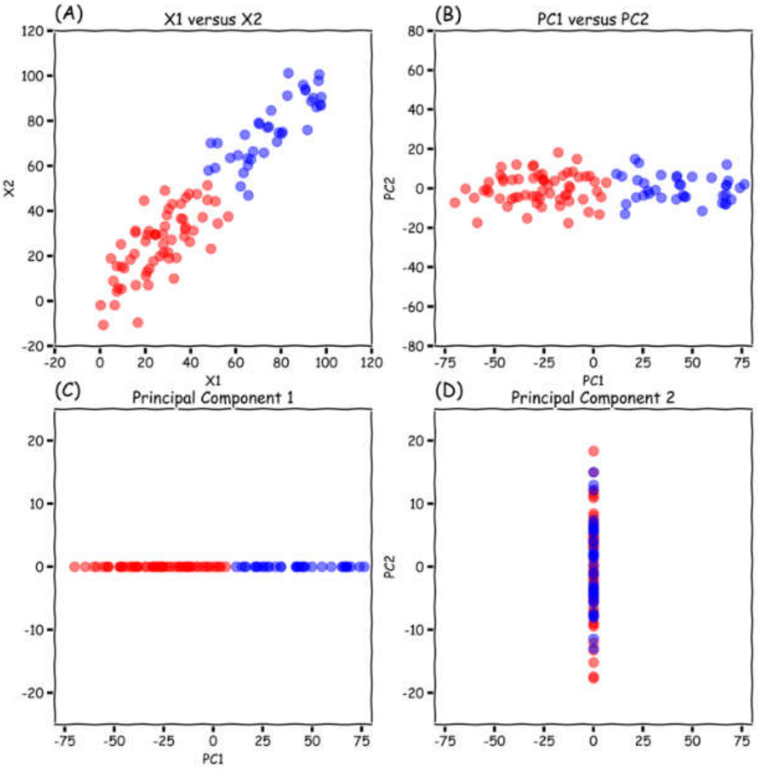
then by the following equation the new data matrix can be obtained with reduced dimension, i.e., *(nXk)* where, k≤ p

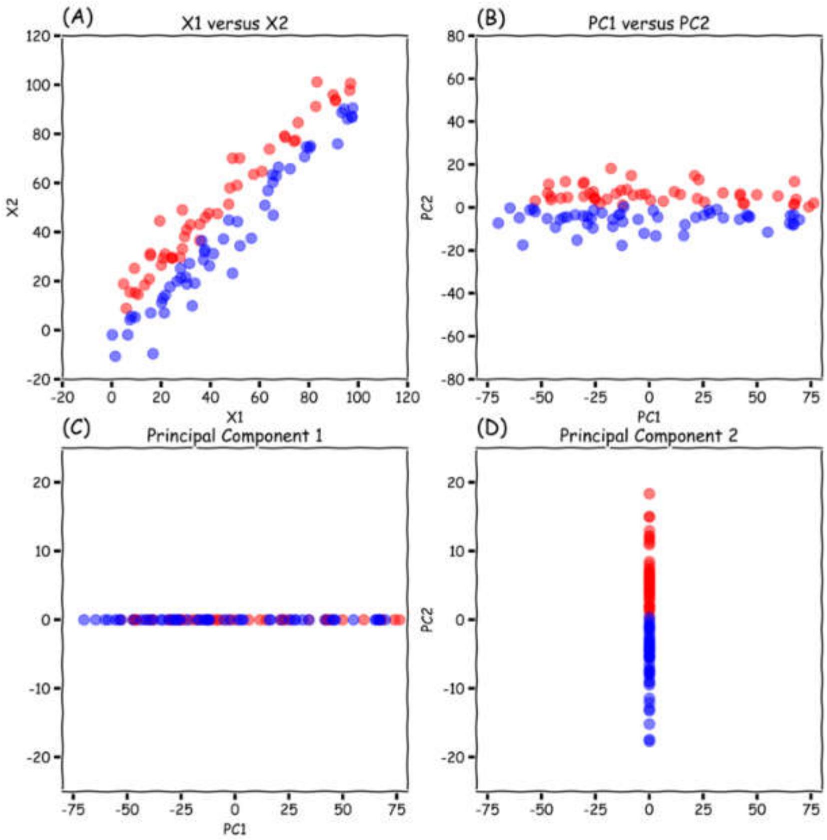
$$[X_{new}]^T{}_{kXn}=[\Lambda]_{kXp}\bullet[X]^T{}_{pXn}$$

Limitations of PCA:

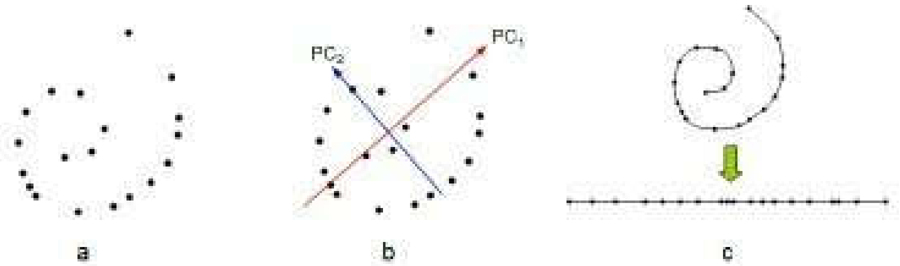
1. If the separation of the classes is more pronounced in the direction of smaller variance


(A)
X1 versus X2
X1
X2
PC1
PC2
(B)
PC1 versus PC2
PC1
PC2
(C)
Principal Component 1
PC1
(D)
Principal Component 2
PC2

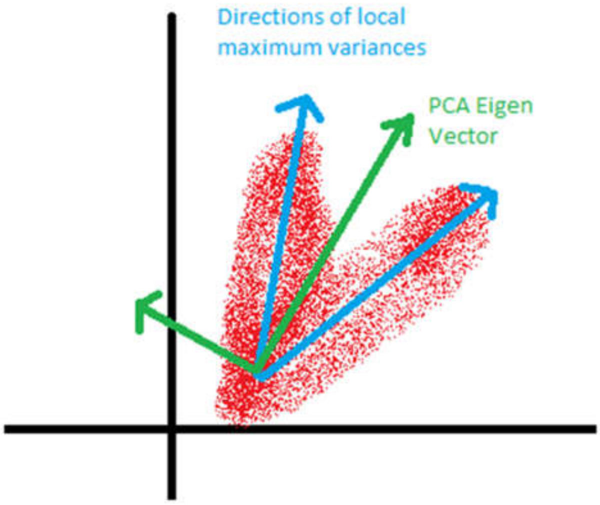

(A)
X1 versus X2
X1
X2
(B)
PC1 versus PC2
PC1
PC2
(C)
Principal Component 1
PC1
(D)
Principal Component 2
PC2



2. If the data is not linearly correlated



3. Orthogonal transformations doesn’t guarantee projections with the highest variance



4. Mean and covariance doesn't describe some distributions

There are many statistics distributions in which mean and covariance doesn't give relevant information of them. In fact, mean and covariance are used (or could be considered important) for Gaussians.

5. Scale variant

   PCA, is a rotation transformation of your dataset, which means that doesn’t affect the scale of your data. That means that if you change the scale of just some of the variables in your data set (e.g., if you normalize the data of some dimensions), you will get different results by applying PCA.