Assignment -2

Name: Nusrat Jannat

ID: 20121176145

Session: 2019-20

Year: 4th year even semester

Topic: Principal Component Analysis

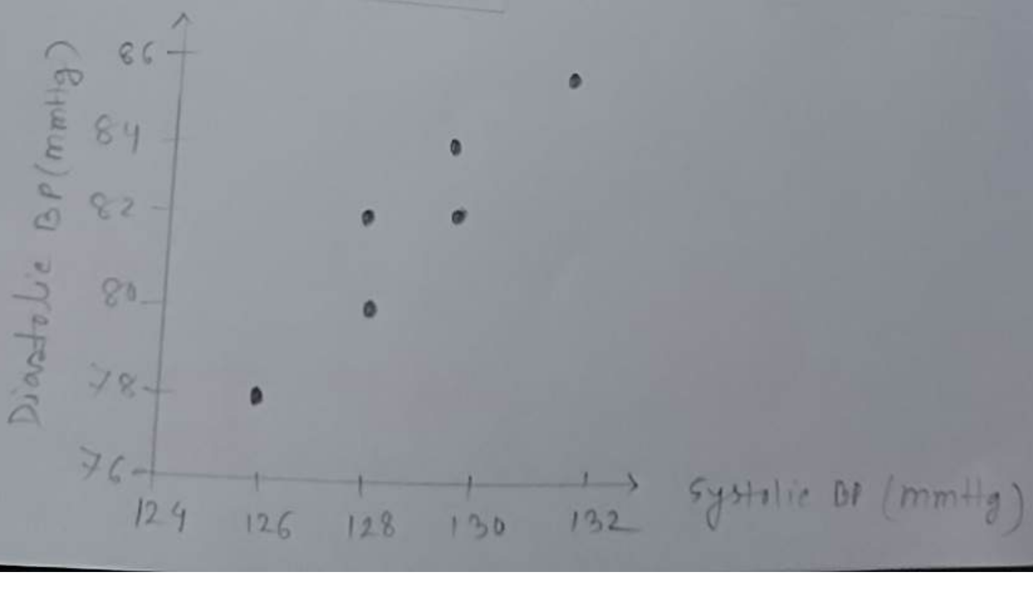
Course: Machine Learning

Principal Component Analysis (PCA): It is a popular unsupervised dimensionality reduction technique in machine learning used to transform high dimensional data into a lower dimensional representation.

To explain how the PCA works, we will use the example data to combine the two blood pressure variables into just one variable based on data from six individuals.

The datasets are as following:

Systolic BP	Diastolic BP
126	78
128	80
128	82
130	82
130	84
132	86

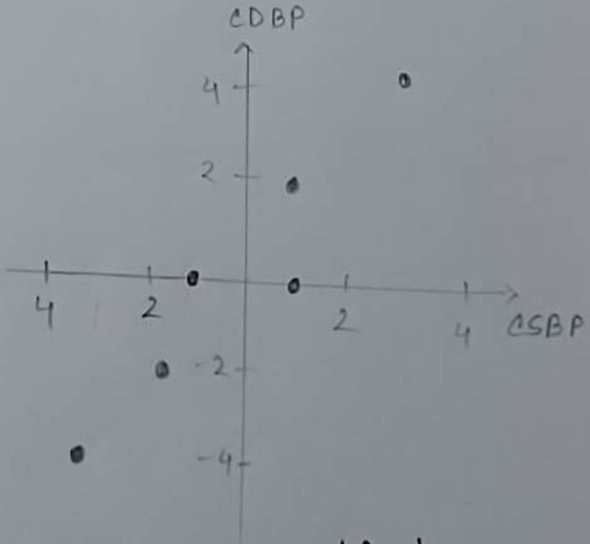


To compute PCA, we need to perform following steps:

1. Center the data.
2. Calculate the covariance matrix.
3. Calculate eigenvalues of the covariance matrix.
4. Calculate eigenvectors of the covariance matrix.
5. Order the eigenvectors.
6. Calculate the principle components.

Step 1: Center the data: In this case, we will only center the data which means that we subtract all the values for each variable by its corresponding mean. The mean by which we subtract systolic blood pressure from each observation is 129 and diastolic BP is 82.

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



When we center the data, it means that we center the data points around the origin. It will help us when we will rotate the data.

step2: Calculate the covariance matrix: We calculate the covariance matrix based on the centered data.

$$\text{var}(CSBP) = \frac{1}{n-1}\sum_{i=1}^{n}(CSBP_i - \overline{CSBP})^2$$

$$= ((-3)^2+(-1)^2+(-1)^2+1^2+1^2+3^2)/(6-1) = 22/5 = 4.4$$

$$\text{var}(CDBP) = \frac{1}{n-1}\sum_{i=1}^{n}(CDBP_i - \overline{CDBP})^2$$

$$= ((-4)^2+(-2)^2+0^2+0^2+2^2+4^2)/(6-1) = 40/5 = 8$$

$$\text{cov}(CSBP, CDBP) = \frac{1}{n-1}\sum(CSBP_i - \overline{CSBP}).(CDBP_i - \overline{CDBP})$$

$$= ((-3).(-4)+(-1).\cancel{x}(-2)+(-1).0+1.0+1.2+3.4)/(6-1)$$

$$= 28/5 = 5.6$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

covarinace matrix,

$$A = \begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

Step3: Calculate the eigenvalues of the covariance matrix

$$\det|A - \lambda I| = 0$$

$$\det\left|\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}\right| - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = 0$$

$$\det\left|\begin{bmatrix} (4.4-\lambda) & 5.6 \\ 5.6 & (8.0-\lambda) \end{bmatrix}\right| = 0$$

Calculating the deterministic of the matrix

$(4.4-\lambda)(8.0-\lambda)-5.6\cdot 5.6=0$

$\Rightarrow 3.84-12.4\lambda+\lambda^2=0$

So $\lambda_1=0.32$, $\lambda_2=12.08$

Step 4: Calculate eigenvectors of the covariance matrix.

while $\lambda_2 = 12.08$

$A.v=\lambda.v$

$$\Rightarrow \begin{bmatrix}4.4 & 5.6\\ 5.6 & 8.0\end{bmatrix}\cdot\begin{bmatrix}x\\ y\end{bmatrix}=12.08\begin{bmatrix}x\\ y\end{bmatrix}$$

So we get

$4.4x+5.6y=12.08x$

$\Rightarrow 5.6y=7.68x$

$\therefore y=1.37x$

$5.6x+8.0y=12.08y$

$\Rightarrow 5.6x=4.08y$

$\therefore y=1.37x$

So eigen vector, $v_2=\cancel{\begin{bmatrix}1\\ 1.3x\end{bmatrix}}\begin{bmatrix}1\\ 1.37\end{bmatrix}$

normalization $v_2=\begin{bmatrix}\frac{1}{\sqrt{1^2+1.37^2}}\\ \frac{1.37}{\sqrt{1^2+1.37^2}}\end{bmatrix}=\begin{bmatrix}0.59\\ 0.81\end{bmatrix}$

Now, while $\lambda_1=0.32$

$A.v=\lambda.v$

$$\Rightarrow \begin{bmatrix}4.4 & 5.6\\ 5.6 & 8.0\end{bmatrix}\cdot\begin{bmatrix}x\\ y\end{bmatrix}=0.32\begin{pmatrix}x\\ y\end{pmatrix}$$

Here eigen vector, $v_1=\begin{bmatrix}-1\\ 0.72\end{bmatrix}$

we get after normalization,

$$v_2 = \begin{bmatrix} \frac{-1}{\sqrt{(-1)^2+(0.72)^2}} \\ \frac{0.72}{\sqrt{(-1)^2+(0.72)^2}} \end{bmatrix} = \begin{bmatrix} -0.81 \\ 0.59 \end{bmatrix}$$

Step 5: Order the eigenvector: We order the eigenvectors based on their corresponding eigen values, where the eigenvector with the largest eigen value becomes our first eigenvector.

So, $v_1 = \begin{bmatrix} 0.59 \\ 0.81 \end{bmatrix}$; $\lambda_2 = 12.08$

$v_2 = \begin{bmatrix} -0.81 \\ 0.59 \end{bmatrix}$; $\lambda_1 = 0.32$

So the matrix $v = \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix}$

Step 6: Calculate the principal components: We will now use this matrix to transform our original centered data so that the two variable are completely uncorrelated.

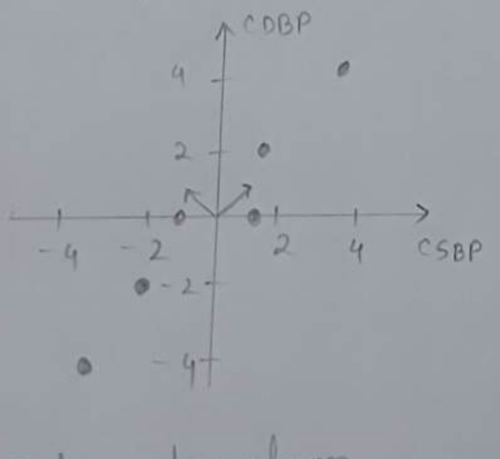
write the centered data as matrix D.

$$D = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix}$$

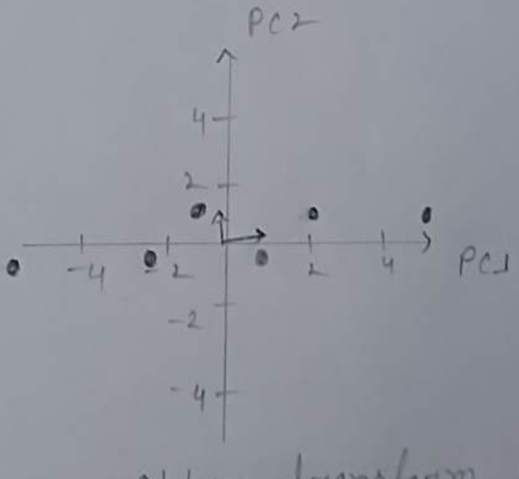
Now we will multiply our data matrix D by matrix v, which includes our eigenvectors as columns to get transformed data.

$$Dv = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{array}{cc} PC1 & PC2 \\ \left[\begin{matrix} -5.0 \\ -2.2 \\ -0.6 \\ 0.6 \\ 2.2 \\ 5.0 \end{matrix}\right. & \left.\begin{matrix} 0.1 \\ -0.4 \\ 0.8 \\ -0.8 \\ 0.4 \\ -0.1 \end{matrix}\right] \end{array}$$

When we go from our original data matrix to the transformed data, this can be seen like we rotate the data clockwise until the two eigenvectors point in the same direction as the x and y-axes of the plot.



Before transform



After transform

Now let's compare the centered data with the transformed data.

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
var = 4.4	var = 8.0

covariance matrix = $\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$

PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
var = 12.08	var = 0.32

covariance matrix = $\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$

Now we can see for

$$\% \text{ var of } PC1 = \frac{12.08}{12.08+0.32} = 97.4\%$$

$$\% \text{ var of } PC2 = \frac{0.32}{12.08+0.32} = 2.6\%$$

As we can see that the first principal component captures 97.4% of the total variance that means that almost all variance is kept in the first principle component. and PC1 is the 1D representation of this 2D dataset.