PCA: Principal Components Analysis

CS277

Principal Components Analysis (PCA)

- An exploratory technique used to reduce the dimensionality of the data set to 2D or 3D
- Can be used to:
  - Reduce number of dimensions in data
  - Find patterns in high-dimensional data
  - Visualize data of high dimensionality
- Example applications:
  - Face recognition
  - Image compression
  - Gene expression analysis

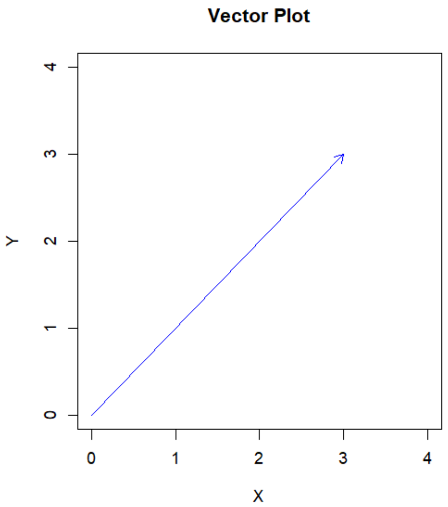
PCA is used to reduce the number of dimensions in the data.

PCA is used to visualize the data in 2D or 3D

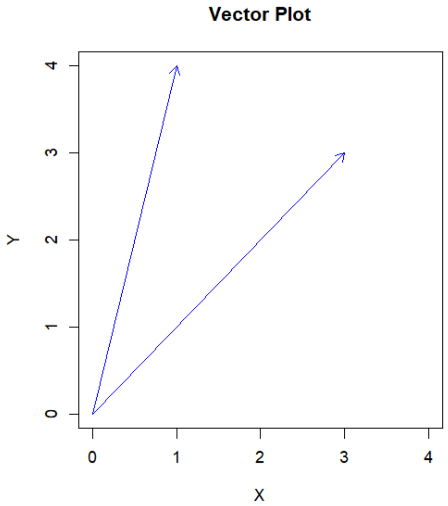
Principal Components Analysis Ideas (PCA)

- Does the data set ‘span’ the whole of d dimensional space?
- For a matrix of *m* samples x *n* genes, create a new covariance matrix of size *n* x *n*.
- Transform some large number of variables into a smaller number of uncorrelated variables called principal components (PCs).
- Developed to capture as much of the variation in data as possible

Consider a vector,
$v = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$



Consider another vector, $v = \begin{bmatrix} 1 \\ 4 \end{bmatrix}$



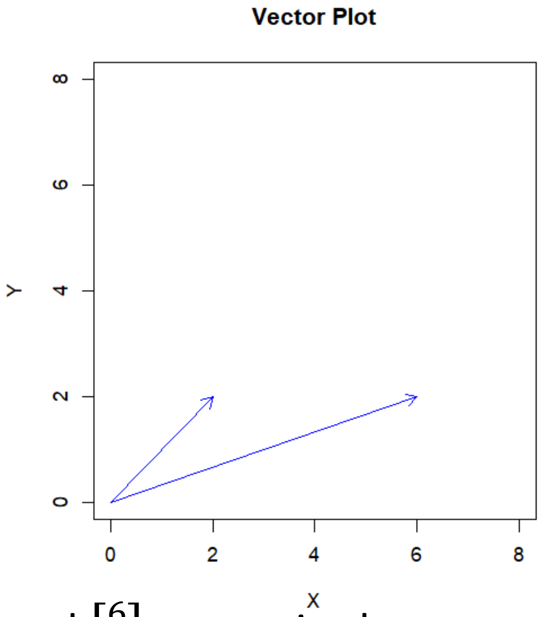
Eigenvector is a non-zero vector that satisfies $Av = \lambda v$ where A is a square matrix and λ is a scalar value

Consider a matrix A as $\begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix}$

$$v = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

Is it an eigenvector of matrix A?

$$Av = \begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 6 \\ 2 \end{bmatrix}$$



$\begin{bmatrix} 2 \\ 2 \end{bmatrix}$ and $\begin{bmatrix} 6 \\ 2 \end{bmatrix}$ are not in the same direction

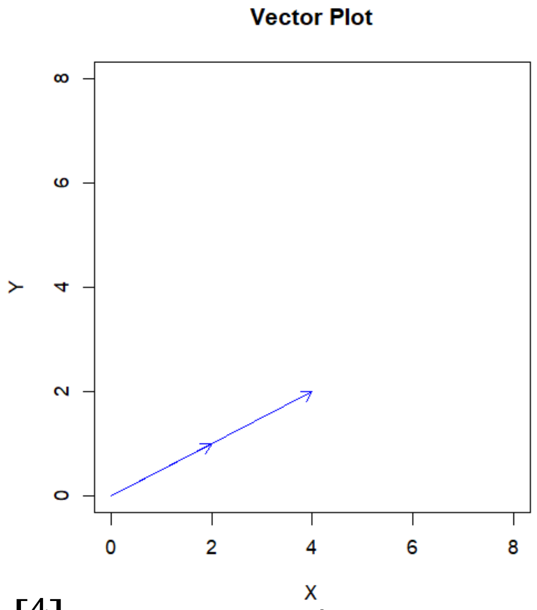
Eigenvector is a non-zero vector that satisfies $Av = \lambda v$ where A is a square matrix and λ is a scalar value

Consider a matrix A as $\begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix}$

$$v = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Is it an eigenvector of matrix A?

$$Av = \begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix} = 2 \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$



$\begin{bmatrix} 2 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 4 \\ 2 \end{bmatrix}$ are in same direction

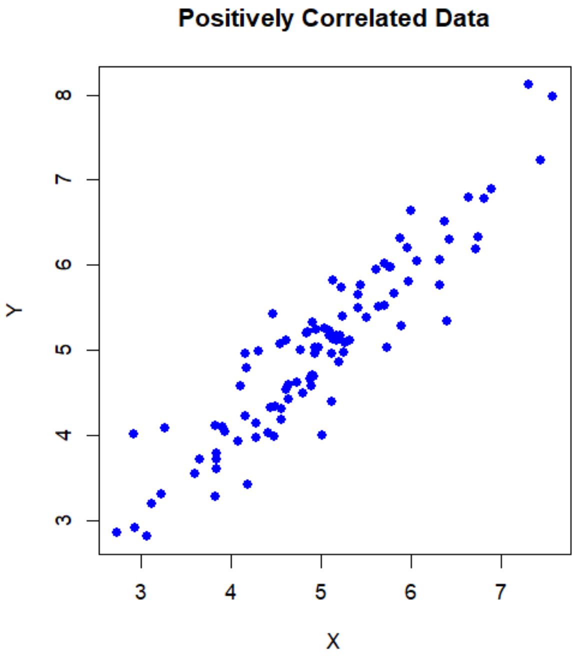


Figure shows two correlated properties (x, y)

Construct a new property $\alpha_1 x_1 + \alpha_2 x_2$

What does this linear transformation signify geometrically?

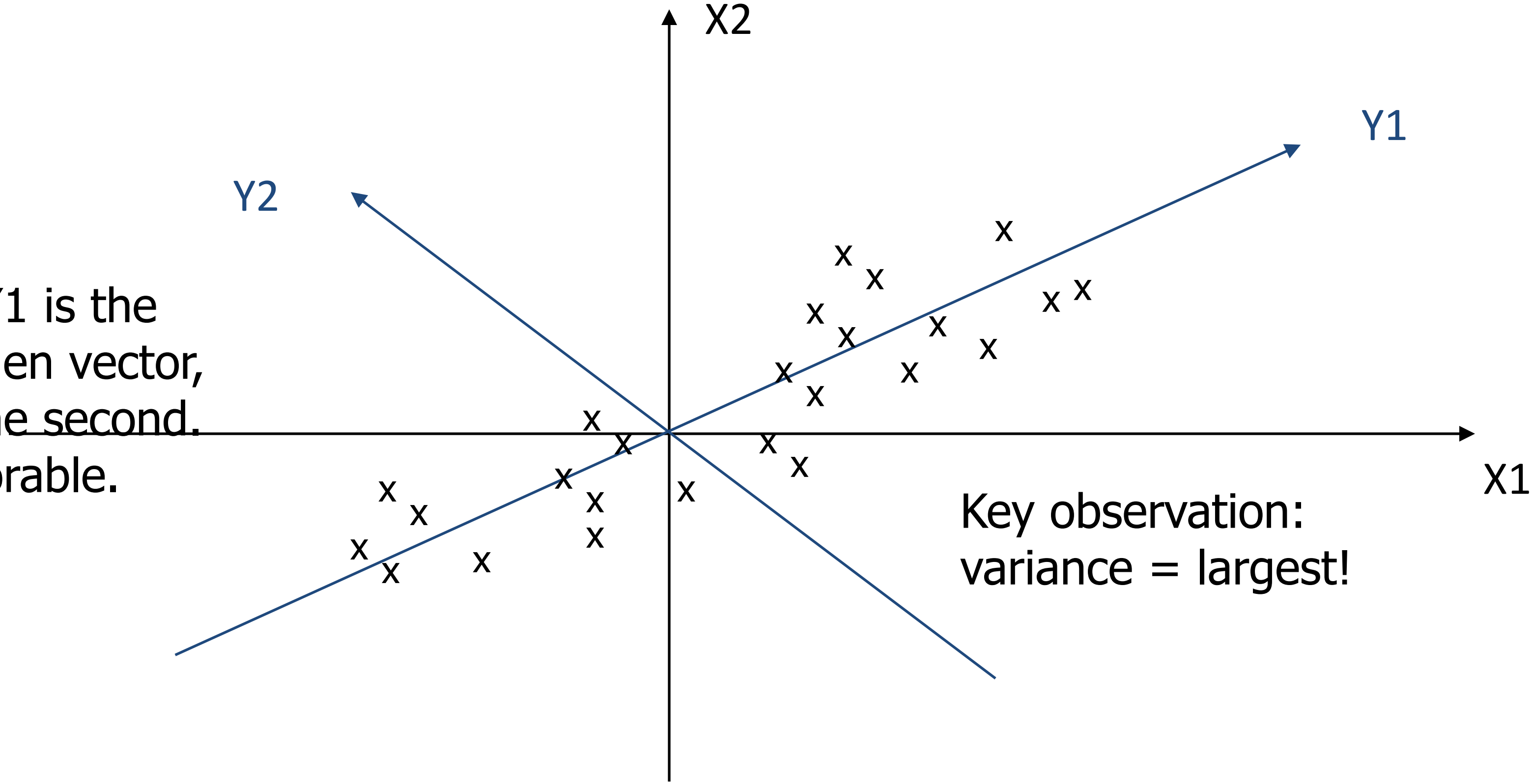
If $[\alpha_1 \ \ \alpha_2]$represents a vector, then

$[\alpha_1 \ \ \alpha_2] \begin{bmatrix} x1 \\ x2 \end{bmatrix}$ is a measure of projection of x onto line α

So what ideally should the vector $[\alpha_1 \ \ \alpha_2]$ be??

In the figure if x varies then y also varies, so both x and y are required to represent the points It would have been ideal if we can transform x and y into x’ and y’ such that with variation in x’, y’ did not vary at all and hence could be dropped from consideration.

Principal Components Analysis (PCA)



Note: Y1 is the first eigen vector, Y2 is the second. Y2 ignorable.

Key observation: variance = largest!

PCA: one attribute

- Question: how much spread is in the data along the axis? (distance to the mean)
- Var=sd^2

$$\mathrm{var}(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)}$$

Temperature
42
40
24
30
15
18
15
30
15
30
35
30
40
30

For two dimensions

Covariance: measures the correlation between X and Y

- Cov(X,Y)=0: independent
- Cov(X,Y)>0: move in same direction
- Cov(X,Y)<0: move in opposite direction

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

X=Temperature	Y=Humidity
40	90
40	90
40	90
30	90
15	70
15	70
15	70
30	90
15	70
30	70
30	70
30	90
40	70
30	90

More than two attributes: covariance matrix

- Contains covariance values between all possible dimensions (=attributes):

$$C^{nxn} = (c_{ij} \mid c_{ij} = \operatorname{cov}(Dim_i, Dim_j))$$

- Example for three attributes (x,y,z):

$$C = \begin{pmatrix} \operatorname{cov}(x,x) & \operatorname{cov}(x,y) & \operatorname{cov}(x,z) \\ \operatorname{cov}(y,x) & \operatorname{cov}(y,y) & \operatorname{cov}(y,z) \\ \operatorname{cov}(z,x) & \operatorname{cov}(z,y) & \operatorname{cov}(z,z) \end{pmatrix}$$

Eigenvalues & eigenvectors

- Vectors **x** having same direction as *A***x** are called *eigenvectors* of *A* (*A* is an n by n matrix).
- In the equation *A***x**=λ**x**, λ is called an *eigenvalue* of *A*.

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} . \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix} = 4. \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

Eigenvalues & eigenvectors

- $A\mathbf{x}=\lambda\mathbf{x} \Leftrightarrow (A-\lambda \mathrm{I})\mathbf{x}=0$
- How to calculate **x** and λ:
  - Calculate $det(A-\lambda I)$, yields a polynomial (degree n)
  - Determine roots to $det(A-\lambda I)=0$, roots are eigenvalues λ
  - Solve $(A- \lambda I)$ $\mathbf{x}=0$ for each λ to obtain eigenvectors **x**

Principal components

- 1. principal component (PC1)
  - The eigenvalue with the largest absolute value will indicate that the data have the largest variance along its eigenvector, the direction along which there is greatest variation
- 2. principal component (PC2)
  - the direction with maximum variation left in data, orthogonal to the PC1.
- In general, only few directions manage to capture most of the variability in the data.

Steps of PCA

1. Get the data
    1. Let $\bar{X}$ be the mean vector (taking the mean of all rows)
2. Adjust the original data by subtracting the mean

    $$X' = X - \bar{X}$$

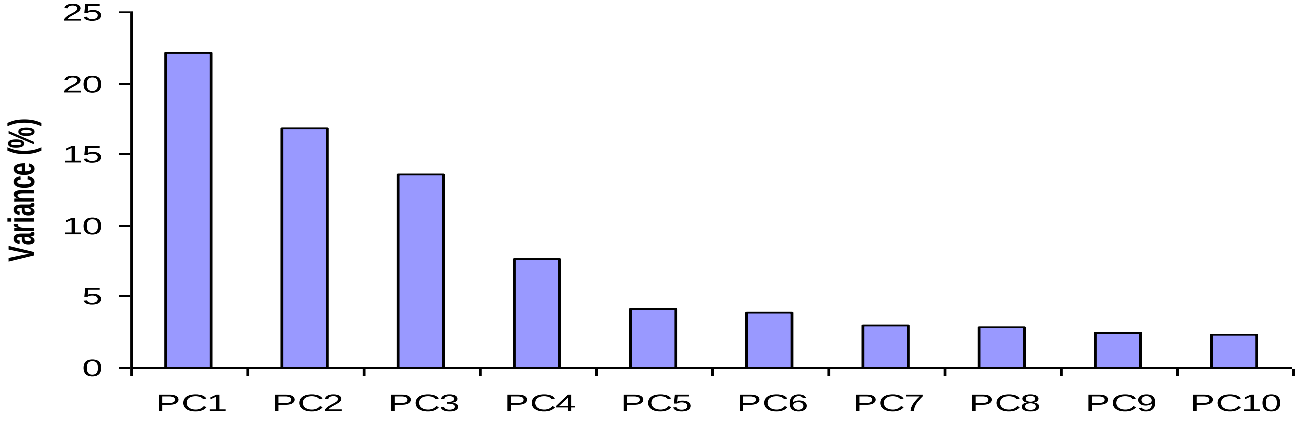
3. Compute the covariance matrix C of adjusted X or centered X
4. Find the eigenvectors and eigenvalues of C.
5. Choose components and form a feature vector
6. Derive the new set of points with respect to the PCs – known as PC Score

Eigenvalues

- Calculate eigenvalues λ and eigenvectors **x** for covariance matrix:
  - Eigenvalues λ_j are used for calculation of [% of total variance] (V_j) for each component j:

$$V_j = 100 \cdot \frac{\lambda_j}{\sum_{x=1}^{n} \lambda_x} \qquad \sum_{x=1}^{n} \lambda_x = n$$

Principal components - Variance



Transformed Data

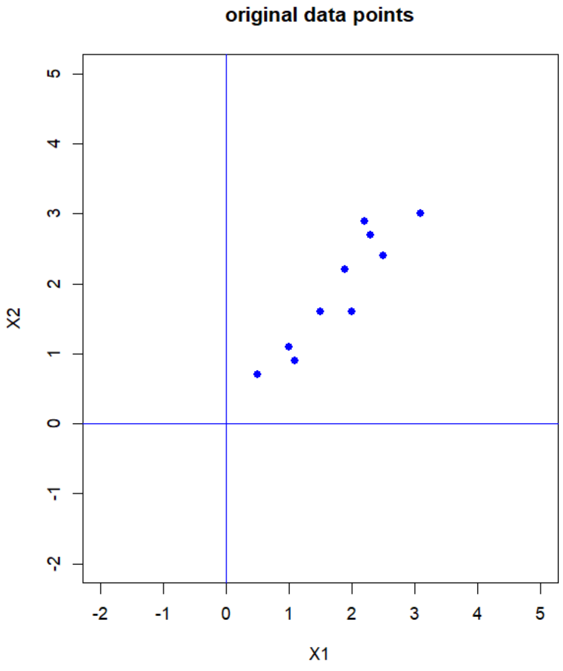
- Eigenvalues λ_j corresponds to variance on each component *j*
- *Thus, sort by* λ_j
- Take the first *p* eigenvectors $\mathbf{e}_{i;}$ where p is the number of top eigenvalues
- These are the directions with the largest variances

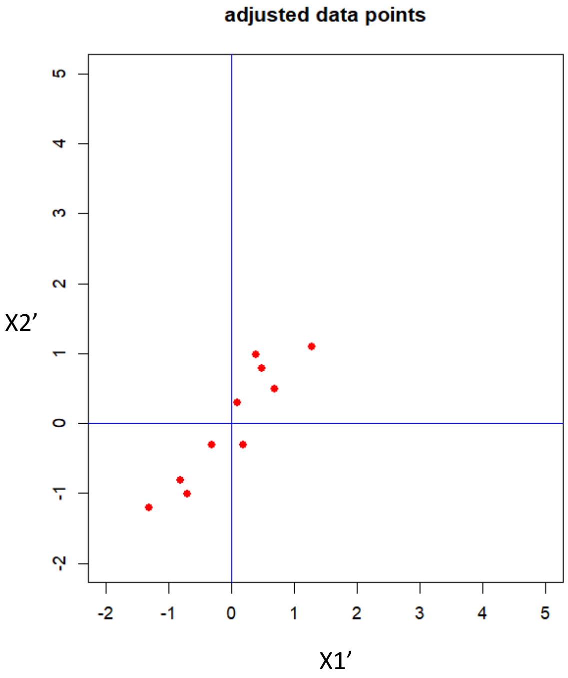
$$\begin{pmatrix} y_{i1} \\ y_{i2} \\ \dots \\ y_{ip} \end{pmatrix} = \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_p \end{pmatrix} \begin{pmatrix} x_{i1} - \overline{x_1} \\ x_{i2} - \overline{x_2} \\ \dots \\ x_{in} - \overline{x_n} \end{pmatrix}$$

An Example

Mean1=1.81
Mean2=1.91

X1	X2	$X1' = X1 - \overline{X1}$	$X2' = X2 - \overline{X2}$
2.5	2.4	0.69	0.49
0.5	0.7	-1.31	-1.21
2.2	2.9	0.39	0.99
1.9	2.2	0.09	0.29
3.1	3.0	1.29	1.09
2.3	2.7	0.49	0.79
2	1.6	0.19	-0.31
1	1.1	-0.81	-0.81
1.5	1.6	-0.31	-0.31
1.1	0.9	-0.71	-1.01


original data points
X2
X1
5
4
3
2
1
0
-1
-2
-2 -1 0 1 2 3 4 5


adjusted data points
X2’
X1’
5
4
3
2
1
0
-1
-2
-2 -1 0 1 2 3 4 5

Covariance Matrix

C=

0.6165	0.6154
0.6154	0.7165

- Eigenvalues:= $\begin{bmatrix} 1.28 \\ 0.05 \end{bmatrix}$
- Eigenvectors=
- $\begin{bmatrix} \boxed{\begin{matrix} 0.6781 \\ 0.7350 \end{matrix}} & \boxed{\begin{matrix} -0.7350 \\ 0.6781 \end{matrix}} \end{bmatrix}$

e1 e2

e1 and e2 are unit eigenvectors corresponding to eigenvalues 1.28 and 0.05

Feature Vector

- Form a feature vector- a matrix of vector
- Constructed by taking the eigenvectors that you want to keep
- Feature Vector = (e1,e2,…,ep)
- For the previous example

Two options

e1 and e2

$$\begin{bmatrix} 0.6781 & -0.7350 \\ 0.7350 & 0.6781 \end{bmatrix}$$

Consider the significant one and drop out the least one

$$\begin{bmatrix} 0.6781 \\ 0.7350 \end{bmatrix}$$

If we only keep one dimension

- We keep the dimension of e1=(0.6781, 0.7350)
- The corresponding PC1 score is obtained using

PC1 = 0.6781.X1’ + 0.7350.X2’

$X1' = X1 - \overline{X1}$	$X2' = X2 - \overline{X2}$	PC1 Score
0.69	0.49	0.828
-1.31	-1.21	-1.778
0.39	0.99	0.992
0.09	0.29	0.274
1.29	1.09	1.676
0.49	0.79	0.913
0.19	-0.31	-0.099
-0.81	-0.81	-1.145
-0.31	-0.31	-0.438
-0.71	-1.01	-1.224

If we keep both dimensions

- We keep the dimension of e1=(0.6781, 0.7350) and e2=(-0.7350,0.6781)
- The corresponding PC1 and PC2 scores are obtained using

PC1 = 0.6781.X1’ + 0.7350.X2’

PC2 = -0.7350.X1’ + 0.6781.X2’

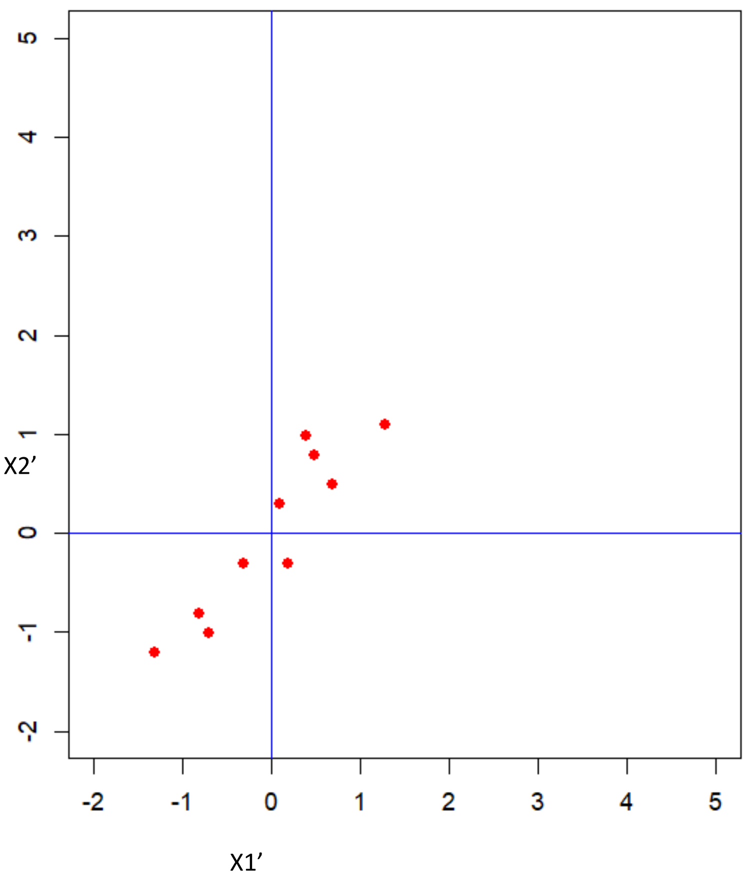
Var(PC1)= 1.28 and Var(PC2)= 0.05

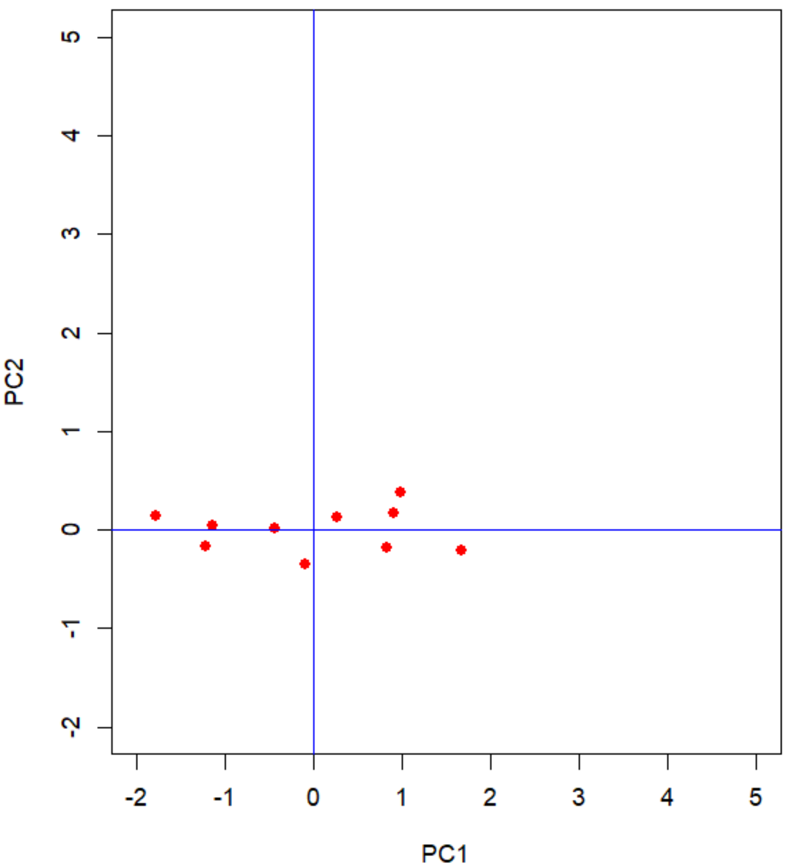
Same as eigenvalues

PC1 captures $=\frac{1.28}{1.28+0.05} = 96\%$ variance

$X1' = X1 - \overline{X1}$	$X2' = X2 - \overline{X2}$	PC1 Score	PC2 Score
0.69	0.49	0.828	-0.173
-1.31	-1.21	-1.778	0.142
0.39	0.99	0.992	0.385
0.09	0.29	0.274	0.130
1.29	1.09	1.676	-0.209
0.49	0.79	0.913	0.176
0.19	-0.31	-0.099	-0.35
-0.81	-0.81	-1.145	0.046
-0.31	-0.31	-0.438	0.018
-0.71	-1.01	-1.224	-0.163

adjusted data points





Principal components

- General about principal components
  - summary variables
  - linear combinations of the original variables
  - uncorrelated with each other
  - capture as much of the original variance as possible

Find PC1 and PC2 score of the following data points

ID	SBP	DBP
1	126	78
2	128	80
3	128	82
4	130	82
5	130	84
6	132	86