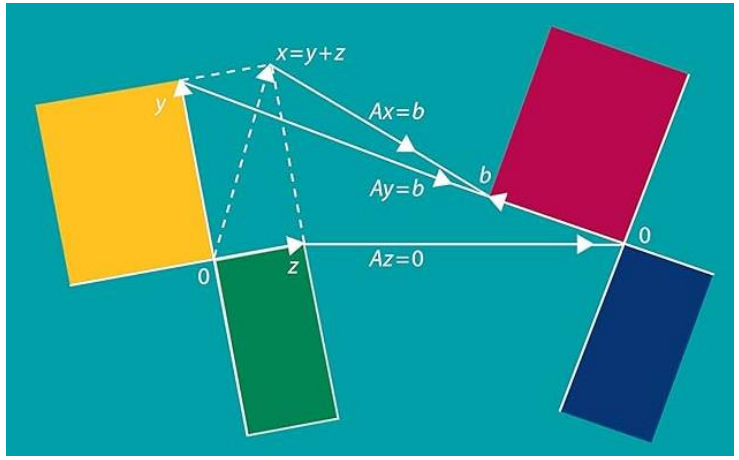
Principle Component Analysis (PCA)

(Linear Algebra)

Randil Pushpananda, PhD

rpn@ucsc.cmb.ac.lk



PCA

- Principal Component Analysis, is a statistical technique used to reduce the dimensionality of a dataset while preserving as much of the original variability as possible.
- PCA transforms the original data into a new set of variables called principal components.
- These principal components are linear combinations of the original variables and are ordered such that the first few retain most of the variation present in the original dataset.

Why PCA?

- It simplifies complex data by reducing the number of variables, making it easier to analyze and visualize.
  - Dimensionality Reduction: PCA reduces the number of variables in a dataset while retaining the most important information.
  - Data Visualization: By reducing data to 2 or 3 dimensions, PCA allows for easier visualization, helping to identify patterns, trends, or clusters in the data.
  - Noise Reduction: PCA can help remove noise by focusing on the components that capture the most variance, effectively filtering out less important information.
  - Feature Extraction: It helps in identifying the most important features in the data

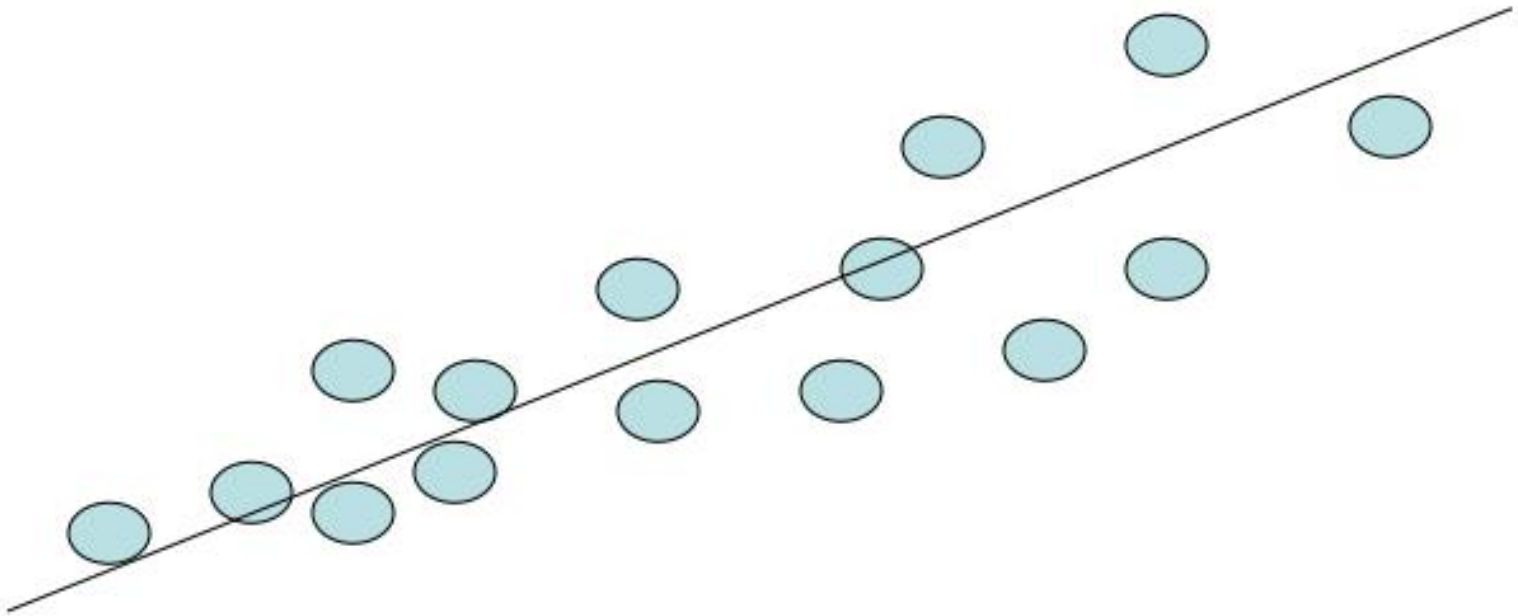
PCA vs SVD

PCA: Involves finding the eigenvectors (principal components) and eigenvalues of the covariance matrix of the data. The eigenvectors correspond to the directions of maximum variance, and the eigenvalues indicate the amount of variance along each direction.

SVD: Decomposes any rectangular matrix A into three matrices: $A = U\Sigma V^T$, where:

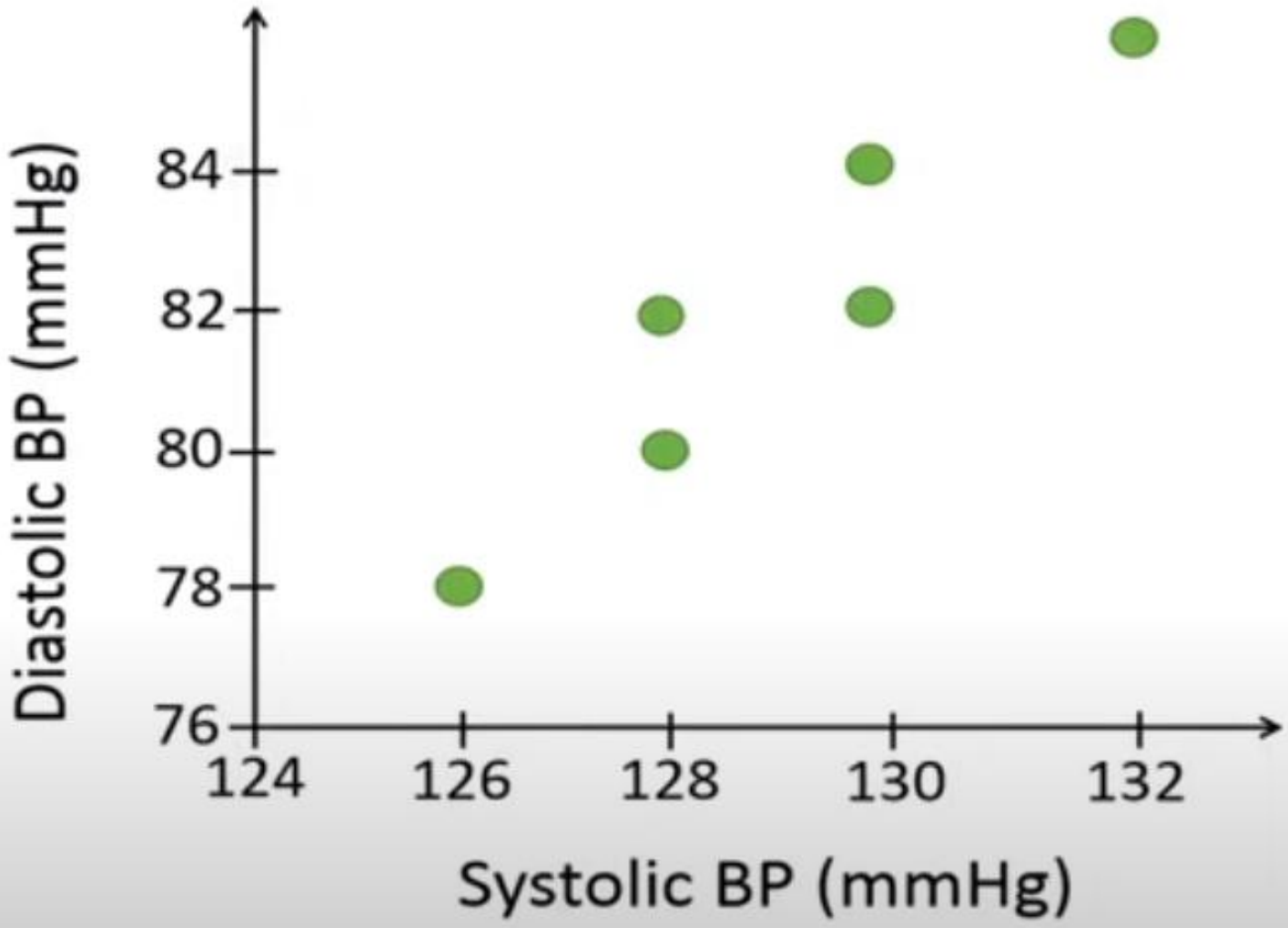
- U contains the left singular vectors (orthonormal columns).
- Σ is a diagonal matrix with singular values (square roots of the eigenvalues).
- V^T contains the right singular vectors (orthonormal rows).

- Given m points in a n dimensional space, for large n, how does one project on to a 1 dimensional space?



- Choose a line that fits the data so the points are spread out well along the line

Example

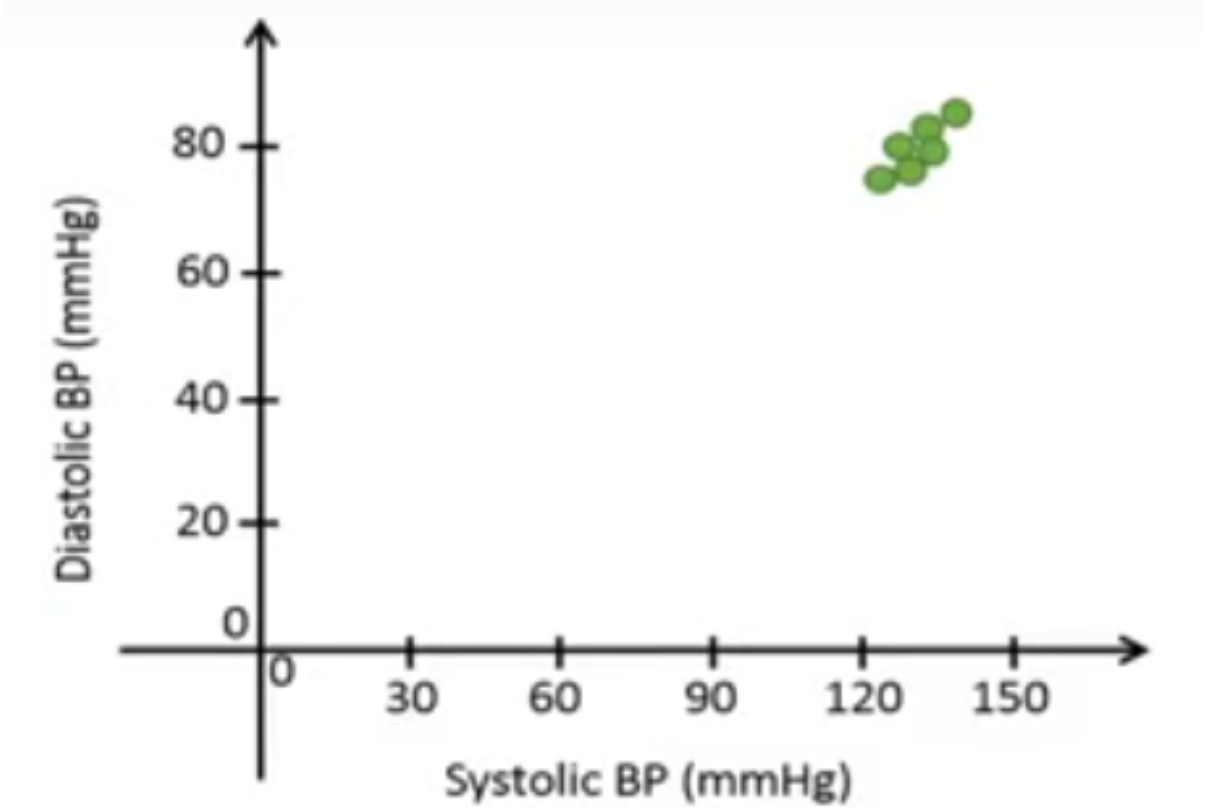


Systolic BP	Diastolic BP
126	78
128	80
128	82
130	82
130	84
132	86

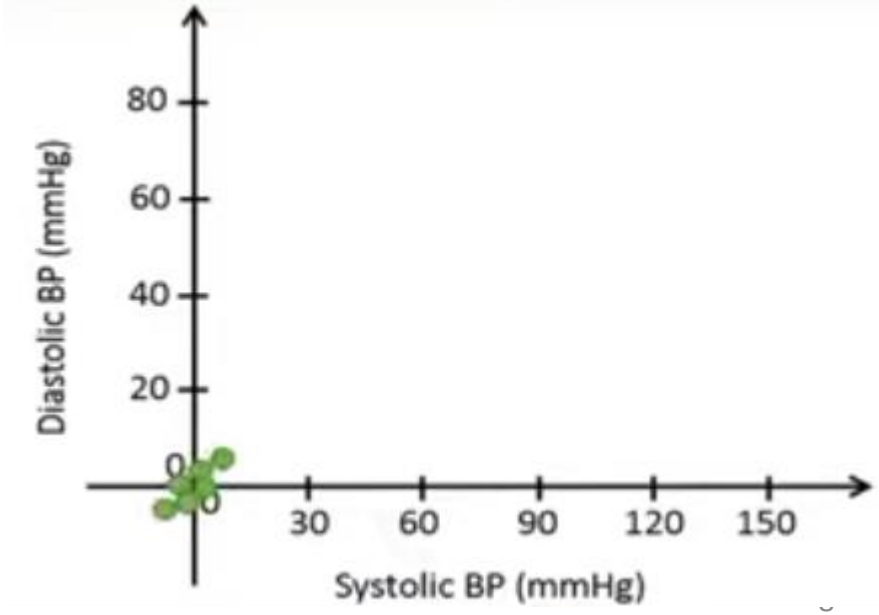
How to Compute

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 principal components

Center Data



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



Calculating Covariance Matrix

	SBP	DBP
SBP		
DBP		

$$cov(x1, x2) = \frac{\sum_{i=1}^{n}(x1_i - \bar{x1})(x2_i - \bar{x2})}{n-1}$$

Where:

- $\bar{x}_1 \, and \, \bar{x}_2$ are the mean values of features $x_1 \, and \, x_2$
- n is the number of data points

The value of covariance can be positive, negative or zeros.

Calculate the eigenvalues of covariance matrix

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

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

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

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

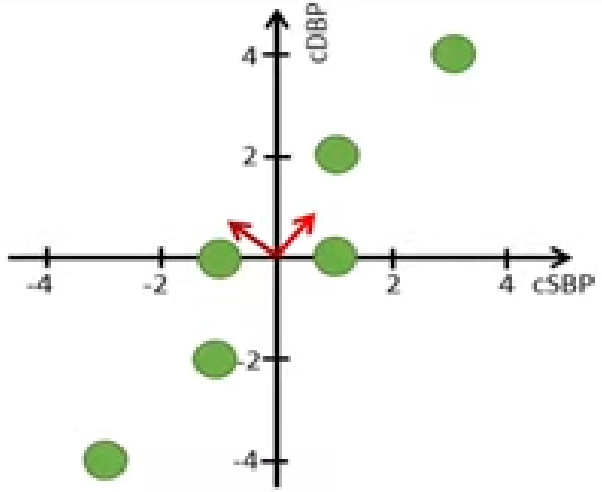
$$\lambda_1=0.32 \qquad \lambda_2=12.08$$

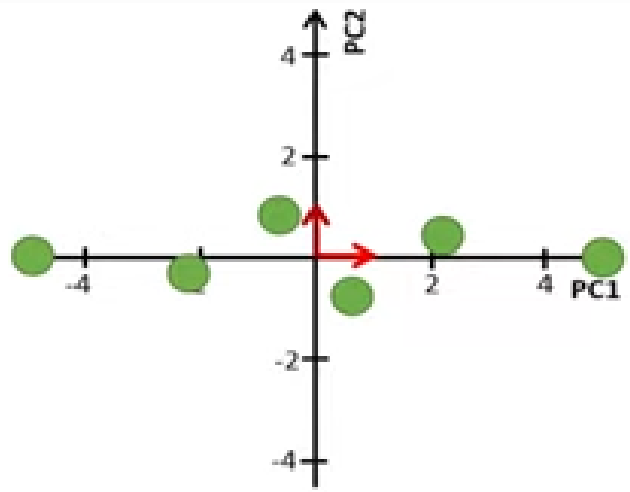
Find Eigenvectors

$$v_2 = \begin{bmatrix} 0.59 \\ 0.81 \end{bmatrix} \quad \lambda_2 = 12.08$$

$$v_1 = \begin{bmatrix} -0.81 \\ 0.59 \end{bmatrix} \quad \lambda_1 = 0.32$$

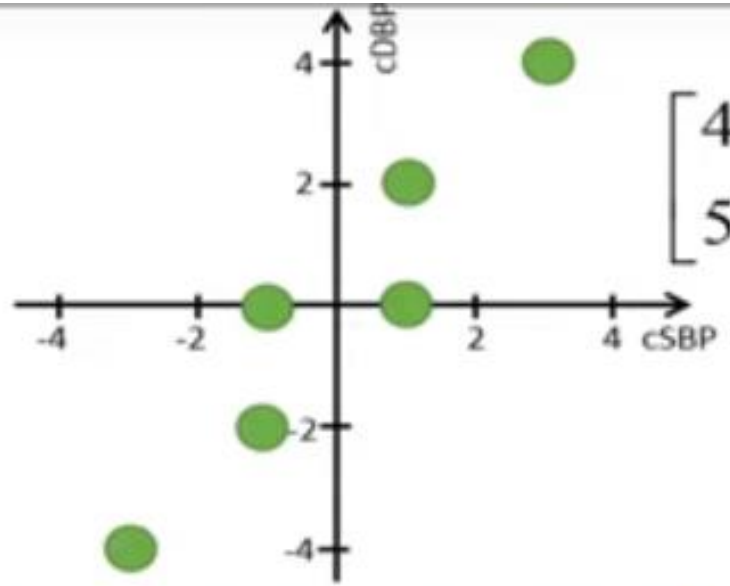
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4





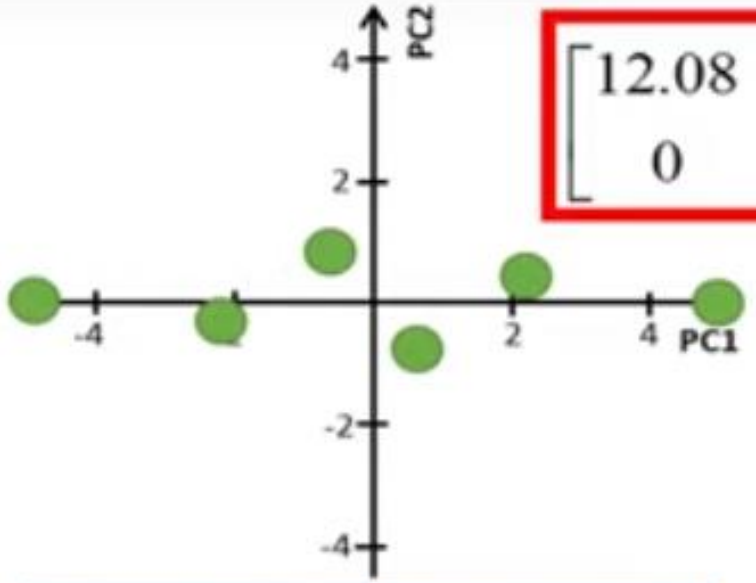
$$V = \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} \qquad D = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix}$$

$$DV = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{matrix} \text{PC1} & \text{PC2} \\ \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix} \end{matrix}$$



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

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



$$\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$$

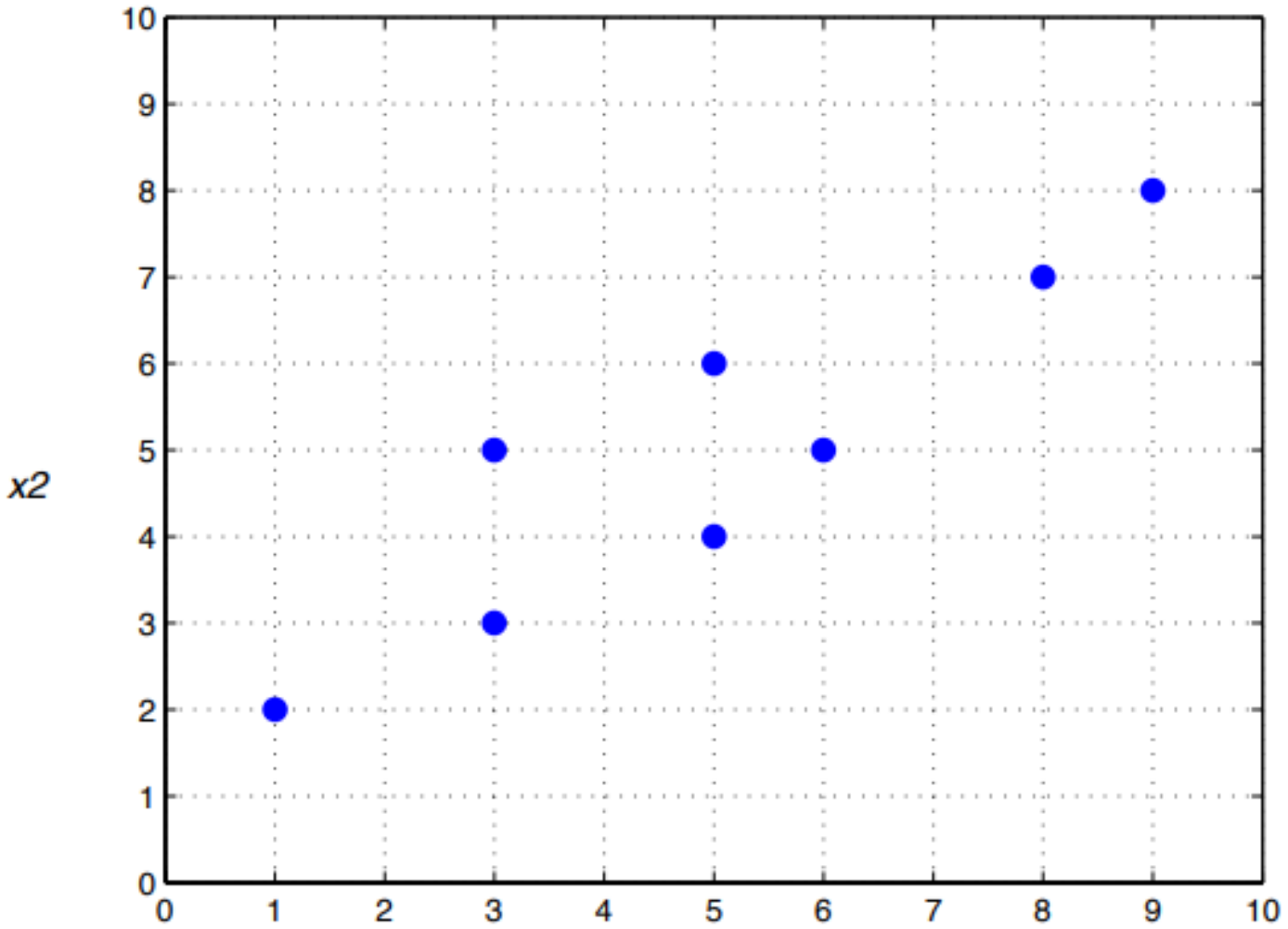
PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

$$PC1 = 0.59 \cdot cSBP + 0.81 \cdot cDBP$$

$$PC2 = -0.81 \cdot cSBP + 0.59 \cdot cDBP$$

PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	[illegible]
0.6	[illegible]
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

- X=(x1,x2)={(1,2),(3,3),(3,5),(5,4),(5,6),(6,5),(8,7),(9,8)}



$$\Sigma = \begin{pmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{pmatrix}$$

$$\lambda = 0.4081 \qquad \lambda = 9.3419$$

$$\vec{v} = \begin{pmatrix} 0.5883 \\ -0.8086 \end{pmatrix} \qquad \vec{v} = \begin{pmatrix} -0.5883 \\ -0.8086 \end{pmatrix}$$