LECTURE 6. PCA

(Principal Component Analysis)

MANU 465

Ahmad Mohammadpanah

PhD, PEng

AIntelligentManufacturing.com

The Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial is a large, randomized trial designed and sponsored by the National Cancer Institute (NCI) in 2006 to determine the chance of getting PLCO cancer. Participants are being followed and additional data will be collected through 2015, for 216 patients, labeled as Cancer and Healthy, with 4000 features (info related to gene, blood, etc.)

	X_1	X_2	X_3	X_4	...	X_{4000}	Result
Person 1	0.063915364	0.025408624	0.025536250	0.012817321		0.036122788	Cancer
Person 2	0.033241734	0.051084790	0.036122788	0.029651841		0.079289645	Cancer
Person 3	0.018484138	0.056304950	0.054195240	0.079289645		0.039348744	Healthy
Person 4	0.0086176926	0.021738490	0.0097349817	0.050676957		0.039736748	Cancer
Person 5	0.035628796	0.027409980	0.027520513	0.039736744		0.039736744	Healthy
Person 6	0.037925478	0.014913797	0.052254751	0.057712860		0.0147349818	Healthy
.	...						...
Person 216	0.035628796	0.027409980	0.027520513	0.039736744		0.039736475	Cancer

$$X_1, X_2, \ldots X_{4000} \longrightarrow$$

97%

- $Z_1 = \ldots$
- $Z_2 = \ldots$
- $Z_3 = \ldots$

PCA (Principal Component Analysis)

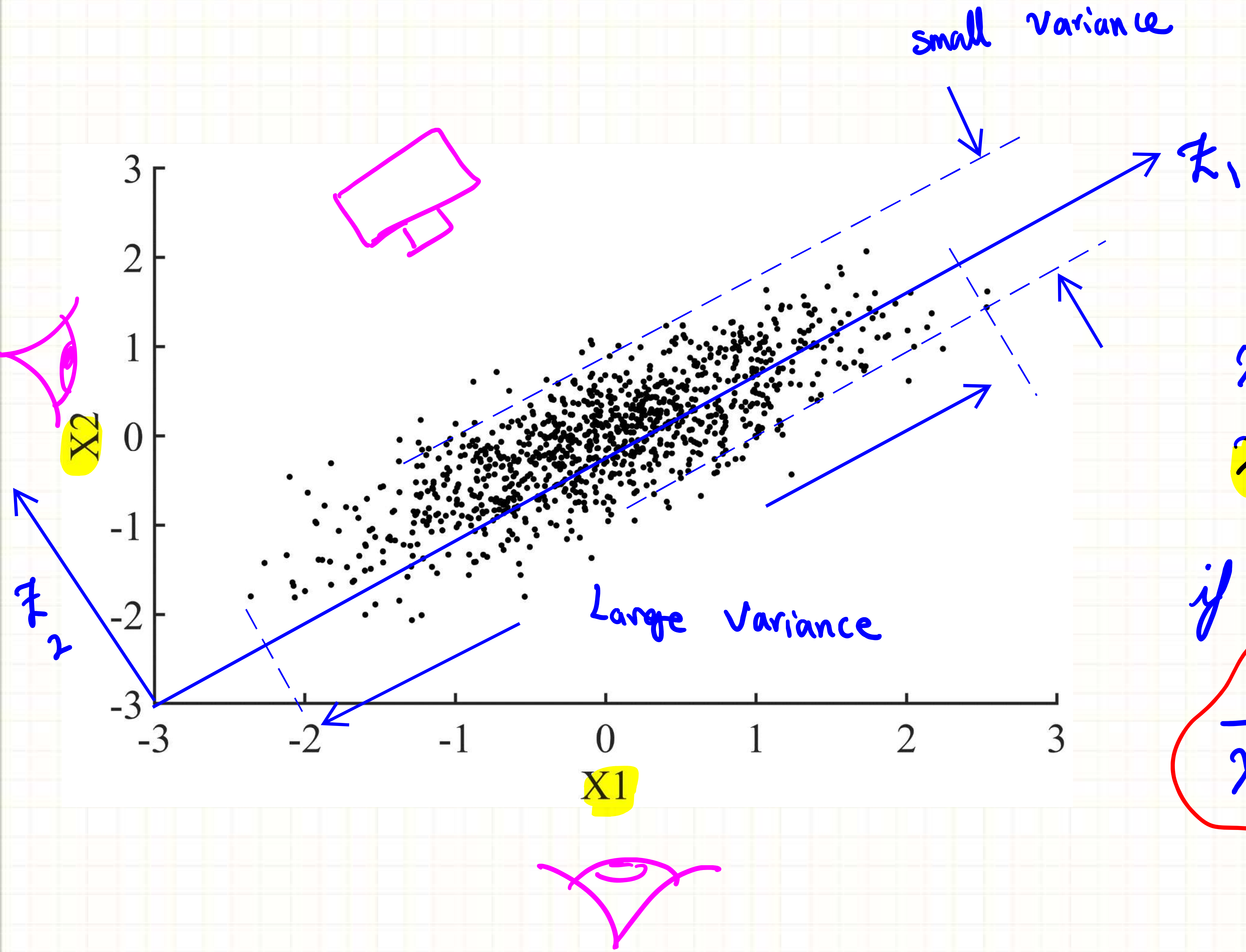
PCA is a powerful tool which:

- Reduces the dimension of the data.
- Takes 4 or more variables and makes a 2D plot.
- Finds the dominant combinations of variables that describes as much of the data as possible.

Objective

In this lecture, you will learn what PCA does, how it does it, and how to apply it to get deeper insight into your data.

What does PCA do?



PCA will find the direction of max variance.

$\lambda_1 = var(Z_1)$

$\lambda_2 = var(Z_2)$

if you only use Z_1, then

$\frac{\lambda_1}{\lambda_1 + \lambda_2} \times 100\%$ of information

preserves.

How does it work?

- Fundamentally, PCA transform the data into directions of maximum variance.
- Mathematically, these directions are the Eigenvectors; and the variances in each direction are the corresponding Eigenvalues of the covariance matrix of data.

$$C = \begin{array}{c|cccc} & x_1 & x_2 & x_3 & \cdots x_n \\ \hline x_1 & Var(x_1) & \cdot & \cdot & \\ x_2 & \cdot & Var(x_2) & & \\ \vdots & & & \ddots & \\ x_n & & & & Var(x_n) \end{array}$$

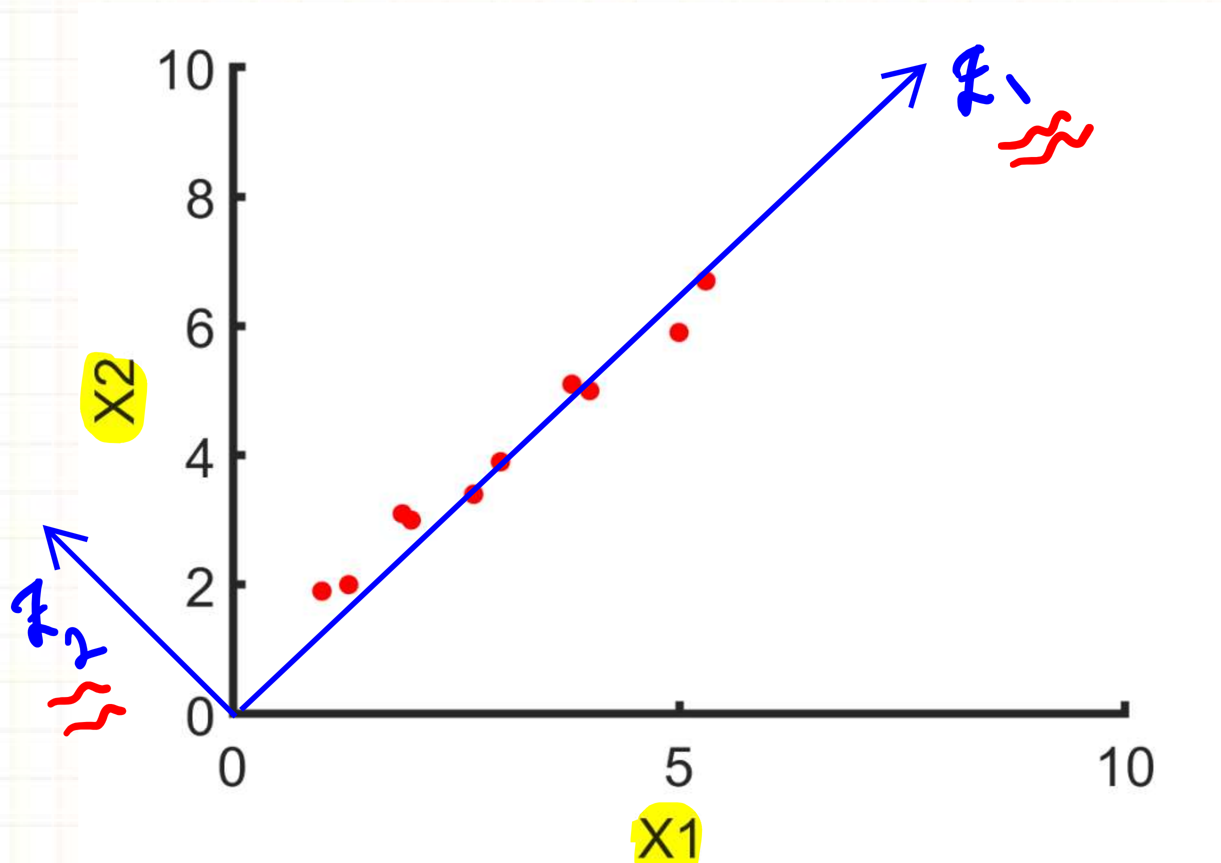
Covar(x_1, x_3)

Covar(x_1, x_2)

Covar(x_2, x_1)

Example) Find the direction of maximum variance for this dataset.

X1	X2
3.8	5.1
2	3
1.9	3.1
5	5.9
1.3	2
3	3.9
5.3	6.7
4	5
1	1.9
2.7	3.4



$Var(x_1) = 2.23$

$Var(x_2) = 2.63$

$Cov(x_1, x_2) = 2.40 = Cov(x_2, x_1)$

$$Var(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}, \quad Covar(x_1, x_2) = \frac{\sum_{i=1}^{n}(x_i^1 - \bar{x}_1)(x_i^2 - \bar{x}_2)}{n-1}$$

$$C = \begin{bmatrix} 2.23 & 2.4 \\ 2.4 & 2.63 \end{bmatrix},$$

finding Eigenvalues & Eigenvectors :

$$|C - \lambda I| = 0 \longrightarrow \boxed{\begin{aligned} \lambda_1 &= 4.84 \\ \lambda_2 &= 0.01 \end{aligned}}$$

, if we only use Z_1 information,

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} \times 100\% = 99\%$$

of information still are available.

$$\begin{cases} [C - \lambda I]\begin{bmatrix} u \\ v \end{bmatrix} = 0 \\ u^2 + v^2 = 1 \end{cases} \longrightarrow$$

for $\lambda = \lambda_1 \longrightarrow V_1 = \begin{bmatrix} 0.67 \\ 0.73 \end{bmatrix}$

for $\lambda = \lambda_2 \longrightarrow V_2 = \begin{bmatrix} -0.73 \\ 0.67 \end{bmatrix}$

eigenvectors

the new variables:

$$Z_1 = [X]\, V_1 = \begin{array}{cc} X_1 & X_2 \\ \begin{bmatrix} 3.8 & 5.1 \\ 2 & 3 \\ 1.9 & 3.1 \\ 5 & 5.9 \\ 1.3 & 2 \\ 3 & 3.9 \\ 5.3 & 6.7 \\ 4 & 5 \\ 1 & 1.9 \\ 2.7 & 3.4 \end{bmatrix} \end{array} \begin{bmatrix} 0.67 \\ 0.73 \end{bmatrix} = \underbrace{\begin{bmatrix} 6.3 \\ 3.5 \\ 3.5 \\ 7.7 \\ 2.3 \\ 4.9 \\ 8.5 \\ 6.3 \\ 2 \\ 4.3 \end{bmatrix}}_{Var(Z_1) = 4.84}, \quad Z_2 = [X]\, V_2 = \underbrace{\begin{bmatrix} 0.65 \\ 0.56 \\ 0.7 \\ 0.3 \\ 0.39 \\ 0.43 \\ 0.63 \\ 0.31 \\ 0.51 \\ 0.44 \end{bmatrix}}_{Var(Z_2) = 0.01}$$

PC1

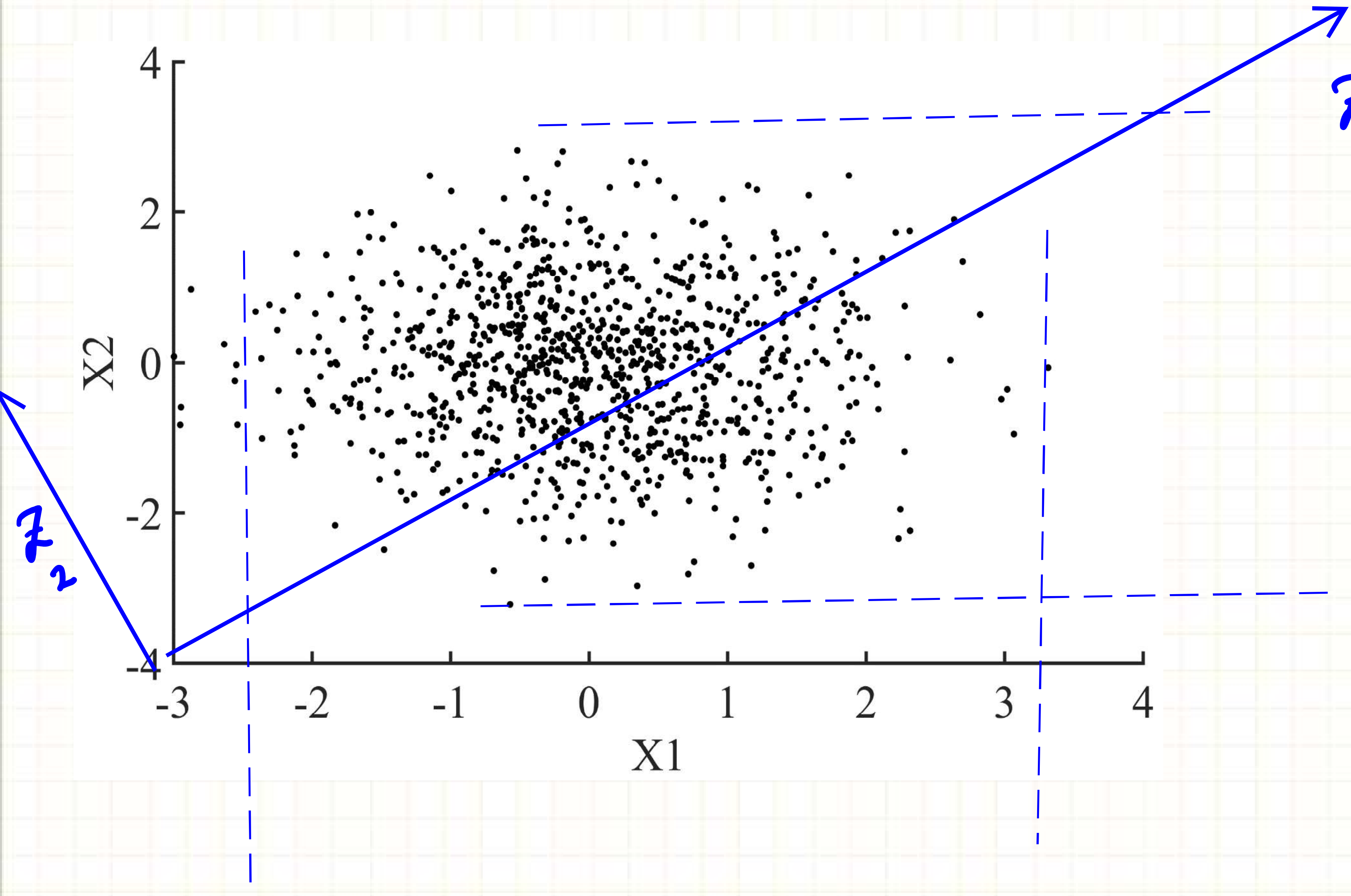
our input matrix

PC2

Variance of the new variables are the same as Eigenvalues.

Does PCA always work?

$$Var(x_1) + Var(x_2) = Var(Z_1) + Var(Z_2)$$



$Z_1 = c_1 x_1 + c_2 x_2$

λ_1

λ_2

$\vdots$

How to apply it?

1. Compute the Covariance Matrix C for X (the input data).

2. Compute the Eigenvectors ($\mathrm{V_i}$) and Eigenvalues (λ_i) of C.

3. Sort the eigenvalues from the maximum to minimum ($\lambda_1 > \lambda_2 > \lambda_3 > \cdots . > \lambda_n$)

4. First Principal Component is PC1=X*V_1 , and PC2=X*V_2 , ...

5. We may just keep the first few PCs and chop of the rest (shrinking the dimension of the data).

6. Contribution of the m PCs we used to describe the data=$\frac{\sum_1^m \lambda_i}{\sum_1^n \lambda_i}$

10 variables $\rightarrow$ $\lambda_1, \lambda_2, \ldots, \lambda_{10}$

only keep

λ_1 & λ_2 $\rightarrow$ $\dfrac{\lambda_1+\lambda_2}{\lambda_1+\lambda_2+\cdots+\lambda_{10}} \times 100\%$ information preserved.

In Python, SKLearn, there is a built-in Function:

```
from sklearn.decomposition import PCA
PrinComp=PCA(n_components=2)
PrinComp.fit(X)
Z=PrinComp.transform(X)
```

Please check the code, Wood_PCA.ipynb on Canvas.

Practice:

1. *Assignment 5.* ✓
2. *Apply this method to a set of images (for example, the Fashion or Digit MNIST).*
3. *Apply this to the Semi-Conductor dataset.* ✓
4. *Attend Tutorial 5.*

Summary

- PCA is a linear transformation which find the direction of maximum variance in the data.
- It reduces the dimension of the data without loosing much information.
- It is a good tool for data visualization.
- In practice, we can simply, use the PCA built-in function in Python to apply PCA to a dataset.