INTRODUCTION TO PRINCIPAL COMPONENT ANALYIS (PCA)

LECTURE 1

Dr Ummugulsum Alyuz

Research Fellow at Centre for Climate Change Research (C3R)
Department of Physics, Astronomy and Mathematics (PAM)
School of Physics, Engineering and Computer Science (SPECS)
University of Hertfordshire

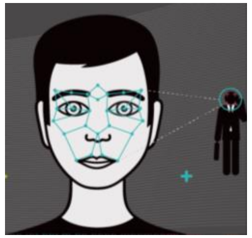
17th October 2024

Learning Outcomes

With this lecture, you will learn:

- What is PCA ?
- When do we need PCA ?
- Basic background of PCA
- How PCA is calculated in Python ?
- How to interpret PCA results ?

Some real-world examples



Face Recognition

Thousands of pixel features → a few key components
For simplifying face data for efficient recognition without losing critical details.

Atmospheric Pollution

Multiple pollutant measurements → a few components
For identifying key pollution patterns and sources (e.g., traffic, industry) for simpler analysis.

Astronomy

Thousands of celestial object measurements (e.g., brightness, distance, spectrum) → a few components
Helping identify key astronomical patterns and phenomena.

Genomics

Thousands of gene expression features → a few components
Helping identify key patterns or variations in biological data.

Financial Markets: Hundreds of stock price features → a few components,
Helps for capturing overall market trends or sector-specific movements for easier analysis

Marketing: Several customer data (e.g., demographics, purchasing behaviour) → a few components
Helping identify key customer segments or purchasing patterns for targeted marketing strategies.

When is PCA useful ?

- Imagine you are an admissions officer at a university.
  - Evaluating students for admission.
- Which student would you offer a spot ?

Dimension 1

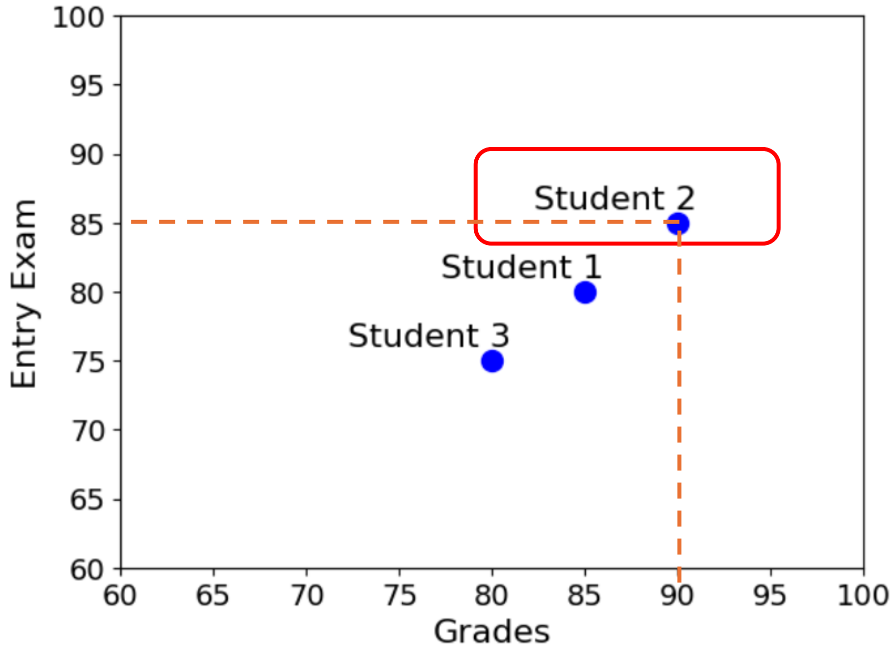
	Grades (0-100)
Student 1	85
Student 2	90
Student 3	80

Variance is a measure of **how spread out** the data points are. It tells us how much the individual values in the dataset differ from the average (mean).

When is PCA useful ?

- Imagine you are an admissions officer at a university.
  - Evaluating students for admission.
  - Which student would you accept ?

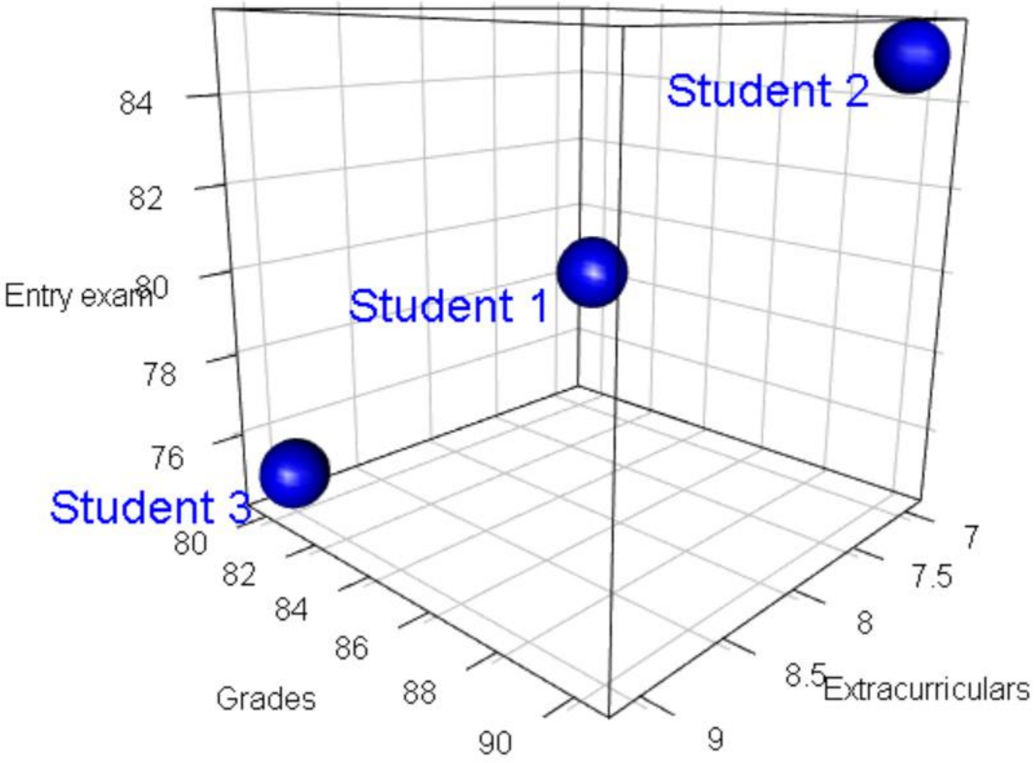
	Dimension 1	Dimension 2
	Grades (0-100)	Entry exam (0-100)
Student 1	85	80
Student 2	90	85
Student 3	80	75



When is PCA useful ?

- Imagine you are an admissions officer at a university.
  - Evaluating students for admission.
  - Which student would you accept ?

	Dimension 1	Dimension 2	Dimension 3
	Grades (0-100)	**Entry exam (0-100)**	**Extracurriculars (0-10)**
Student 1	85	80	8
Student 2	90	85	7
Student 3	80	75	9



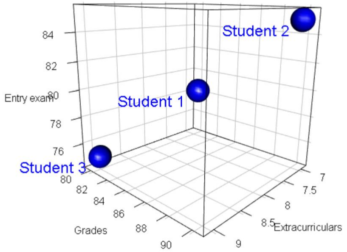
When is PCA useful ?

- Imagine you are an admissions officer at a university.
  - Evaluating students for admission.
  - Which student would you accept ?

	Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5
	Grades (0-100)	**Entry exam (0-100)**	**Extracurriculars (0-10)**	**Recommendation (0-10)**	**Interview (0-10)**
Student 1	85	80	8	9	7
Student 2	90	85	7	8	8
Student 3	80	75	9	10	6

How PCA works ?

Multiple dimensions in the data



To ve covered in the next lecture

PCA

1 –Adjusting the data
so that its mean is zero

2 - Calculate the covariance matrix
to understand variable relationships in the dataset.

3– Calculate eigenvectors
to find the direction of the biggest variance in the data.

4– Calculate eigenvalues
to understand how important each component is.

5– Projecting the data
into top principal components

PRINCIPAL COMPONENTS

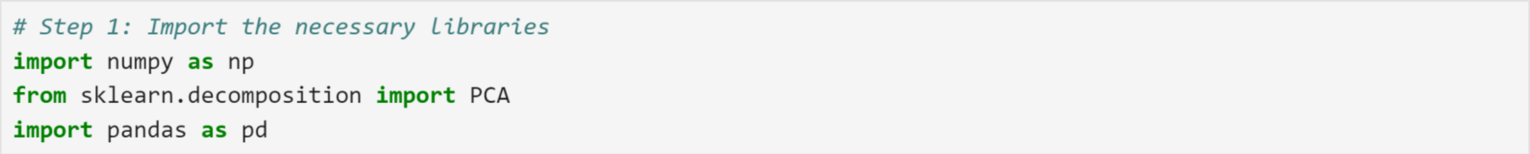
FIRST PRINCIPAL COMPONENT
captures the biggest differences in the data

SECOND PRINCIPAL COMPONENT
captures the next largest variance in the data (but it is in a completely independent direction)

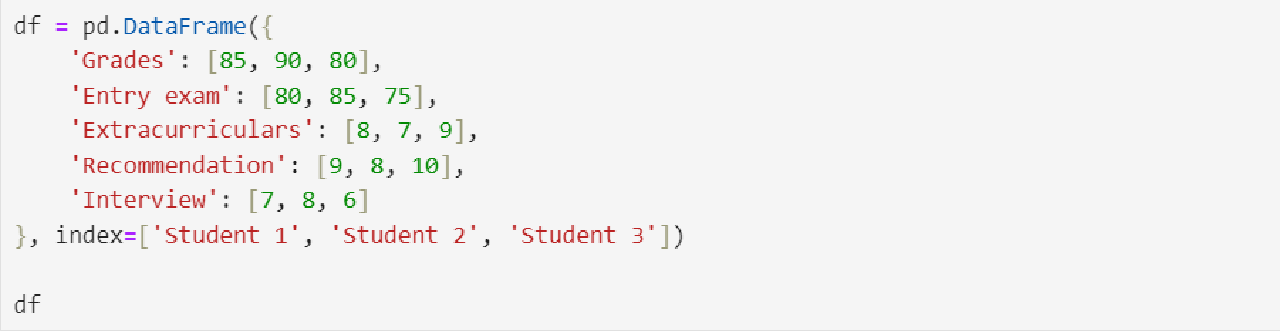
With those new components, PCA:

- **Reduce the number of dimensions** while retaining variability in the original data.
- Makes it **easier to visualise**, analyse, and identify trends.
- Especially useful in machine learning and data analysis by **simplifying data** for algorithms.

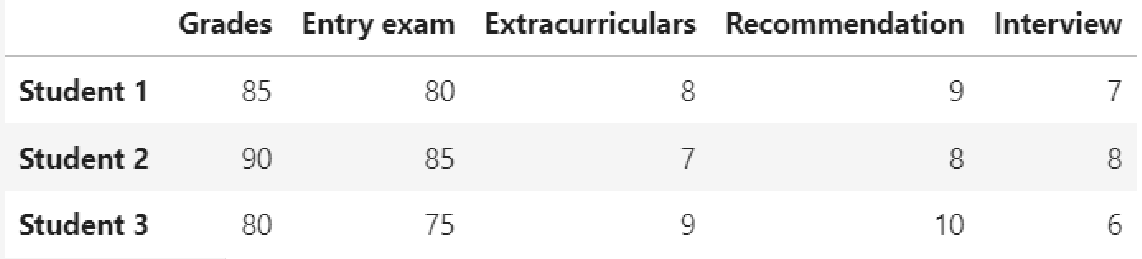
An example Python code for solving the problem with PCA



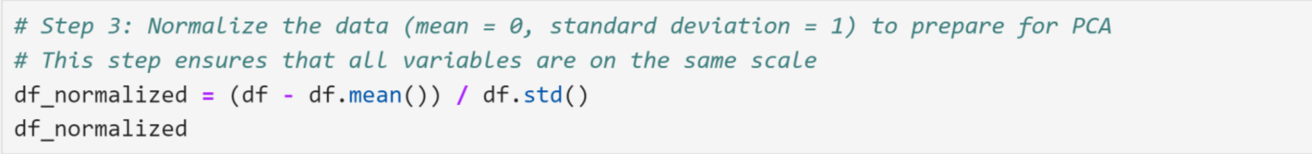
```python
# Step 1: Import the necessary libraries
import numpy as np
from sklearn.decomposition import PCA
import pandas as pd
```



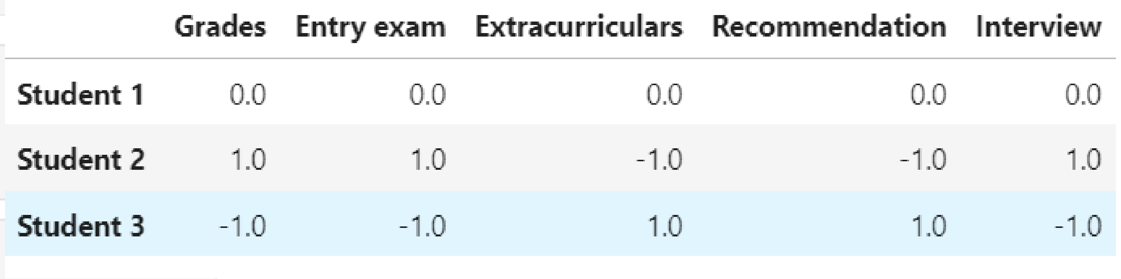
```python
df = pd.DataFrame({
    'Grades': [85, 90, 80],
    'Entry exam': [80, 85, 75],
    'Extracurriculars': [8, 7, 9],
    'Recommendation': [9, 8, 10],
    'Interview': [7, 8, 6]
}, index=['Student 1', 'Student 2', 'Student 3'])

df
```



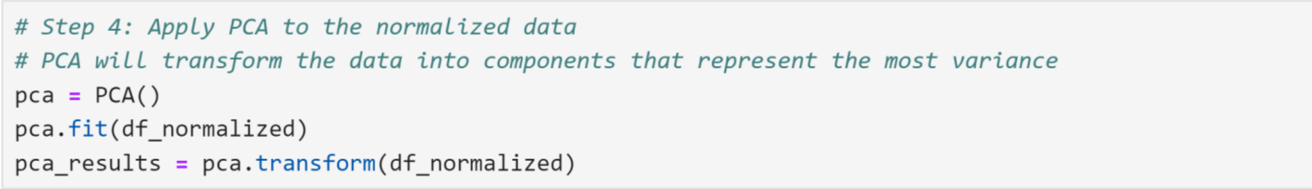
	Grades	Entry exam	Extracurriculars	Recommendation	Interview
Student 1	85	80	8	9	7
Student 2	90	85	7	8	8
Student 3	80	75	9	10	6



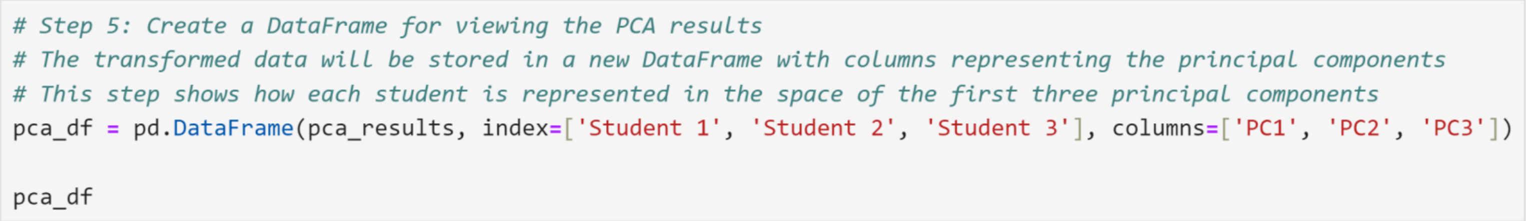
```python
# Step 3: Normalize the data (mean = 0, standard deviation = 1) to prepare for PCA
# This step ensures that all variables are on the same scale
df_normalized = (df - df.mean()) / df.std()
df_normalized
```



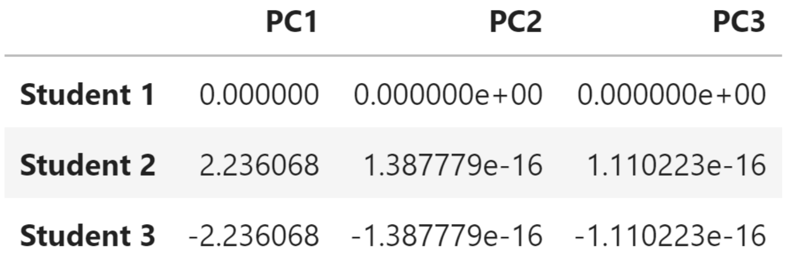
	Grades	Entry exam	Extracurriculars	Recommendation	Interview
Student 1	0.0	0.0	0.0	0.0	0.0
Student 2	1.0	1.0	-1.0	-1.0	1.0
Student 3	-1.0	-1.0	1.0	1.0	-1.0



```python
# Step 4: Apply PCA to the normalized data
# PCA will transform the data into components that represent the most variance
pca = PCA()
pca.fit(df_normalized)
pca_results = pca.transform(df_normalized)
```



```python
# Step 5: Create a DataFrame for viewing the PCA results
# The transformed data will be stored in a new DataFrame with columns representing the principal components
# This step shows how each student is represented in the space of the first three principal components
pca_df = pd.DataFrame(pca_results, index=['Student 1', 'Student 2', 'Student 3'], columns=['PC1', 'PC2', 'PC3'])

pca_df
```



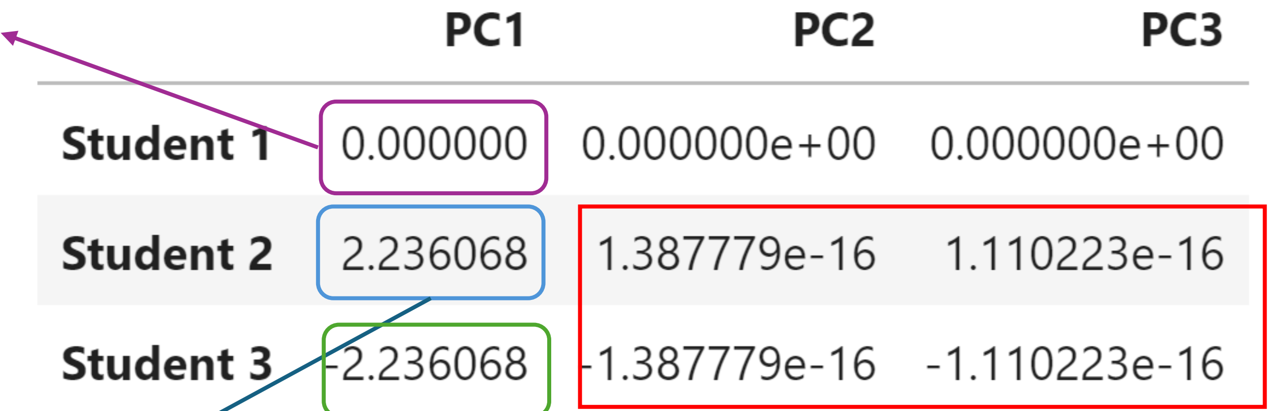
	PC1	PC2	PC3
Student 1	0.000000	0.000000e+00	0.000000e+00
Student 2	2.236068	1.387779e-16	1.110223e-16
Student 3	-2.236068	-1.387779e-16	-1.110223e-16

The code we used in the lecture is available from : https://github.com/umga/Teaching

How we interpret the results ?

- Scores **0** on PC1
- Student 1 is the average case for this particular component, neither excelling nor underperforming compared to the others.

PRINCIPAL COMPONENTS OF THE DATA



	PC1	PC2	PC3
Student 1	0.000000	0.000000e+00	0.000000e+00
Student 2	2.236068	1.387779e-16	1.110223e-16
Student 3	-2.236068	-1.387779e-16	-1.110223e-16

- Positive score
- Perform well across the main factors that contribute the most variance in the data.

- Negative score
- Suggesting they are performing worse than Student 2 in those factors.

- Extremely close to **0**
- **Virtually no additional variance** beyond what PC1 captures.

Homework

- Go to the link : https://www.menti.com/alucv787tywo (or read the barcode)
- The quiz should take around **5-10 minutes** to complete.
- Submit your answers **before the next lecture**.
- We will discuss the quiz results and address any questions during the next lecture.
- The code and the presentation we used in this lecture is available from https://github.com/umga/Teaching

Next Lectures

LECTURE 2 : We will understand the mathematics behind PCA with the same example.

LECTURE 3 : We will solve more complex examples in Python and interpret the results.

Further Reading

1. Chapters 3.5 and 14.5 in Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer. Available online at: https://hastie.su.domains/ElemStatLearn/

2. Jolliffe, I. T. (2002). Principal Component Analysis (2nd ed.). Springer. Part of the Springer Series in Statistics (SSS). https://link.springer.com/book/10.1007/b98835