Welcome to the CoGrammar Principal Component Analysis

The session will start shortly...

Questions? Drop them in the chat. We'll have dedicated moderators answering questions.

CoGrammar

Data Science Session Housekeeping

- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment for all - please engage accordingly. **(Fundamental British Values: Mutual Respect and Tolerance)**
- No question is daft or silly - **ask them!**
- There are **Q&A sessions** midway and at the end of the session, should you wish to ask any follow-up questions. Moderators are going to be answering questions as the session progresses as well.
- If you have any questions outside of this lecture, or that are not answered during this lecture, please do submit these for upcoming Academic Sessions. You can submit these questions here: [Questions]

Data Science Session Housekeeping cont.

- For all **non-academic questions**, please submit a query: www.hyperiondev.com/support

- Report a **safeguarding** incident: www.hyperiondev.com/safeguardreporting

- We would love your **feedback** on lectures: [Feedback on Lectures]()

Skills Bootcamp
8-Week Progression Overview

Fulfil 4 Criteria to Graduation

✅ Criterion 1: Initial Requirements

Timeframe: First 2 Weeks
Guided Learning Hours (GLH): Minimum of 15 hours
Task Completion: First four tasks

Due Date: 24 March 2024

✅ Criterion 2: Mid-Course Progress

60 Guided Learning Hours

Data Science - **13 tasks**
Software Engineering - **13 tasks**
Web Development - **13 tasks**

Due Date: 28 April 2024

CoGrammar

Skills Bootcamp Progression Overview

✅ Criterion 3: Course Progress

Completion: All mandatory tasks, including Build Your Brand and resubmissions by study period end
Interview Invitation: Within 4 weeks post-course
Guided Learning Hours: Minimum of 112 hours by support end date (10.5 hours average, each week)

✅ Criterion 4: Demonstrating Employability

Final Job or Apprenticeship Outcome: Document within 12 weeks post-graduation
Relevance: Progression to employment or related opportunity

CoGrammar

SKILLS FOR LIFE
SKILLS BOOTCAMPS

Department for Education

CoGrammar

Principal Component Analysis

June 2024

Learning Objectives

- ❖ Understand what Principal Component Analysis is and what it aims to achieve.
- ❖ Understand the issues we face with high dimensionality
- ❖ Grasp the Mathematical foundations of PCA
- ❖ Understand how to interpret PCA results

Principal Component Analysis

Introduction

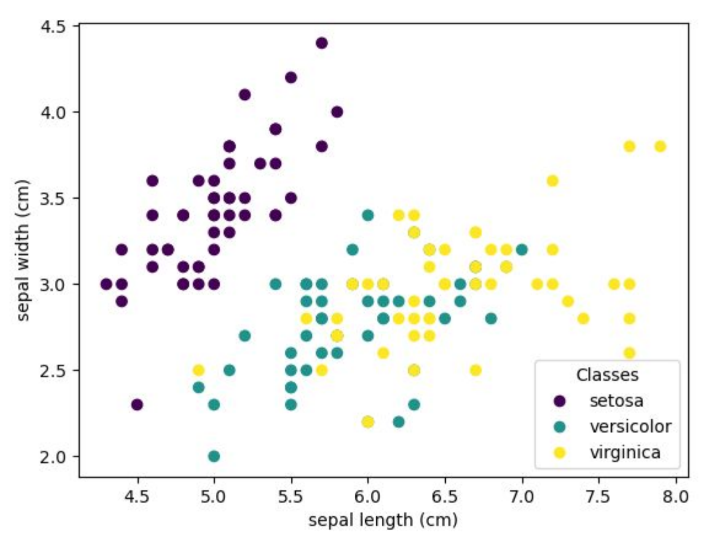
CoGrammar

Principal Component Analysis

- **Principal Component Analysis (PCA)** is a statistical technique used for dimensionality reduction, transforming data into a set of linearly uncorrelated variables called principal components.
- In practice, the number of variables in a machine learning task can be high.
- As the number of variables grows, the data becomes harder to work with. Relationships between variables become harder to see, training slows down, and the chance of overfitting increases. It is, therefore, useful to know a bit about how to reduce the number of variables while still retaining enough information about our dataset.

Dimensionality

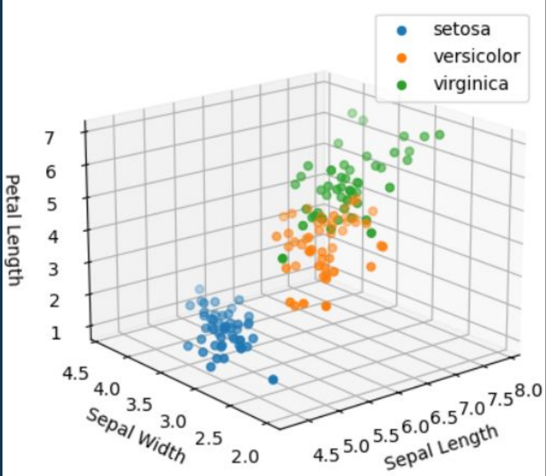
Data points can be represented as a vector, an array of points in a feature space, e.g [4, 3] for the sepal length and sepal width of an instance in the Iris dataset. A subset of the Iris dataset containing only sepal lengths and sepal widths can be visualised in a 2-dimensional space.



CoGrammar

Dimensionality

We can include a third feature, the petal length, and plot in a 3-dimensional space.



There is a fourth feature in the Iris dataset, the petal width. If we wanted to create a visual representation of our observations with this feature included, we would need to plot them in a 4-dimensional space. This is difficult to conceptualise and visualise, and this is one of the problems of high dimensionality.

CoGrammar

Mathematical Foundation

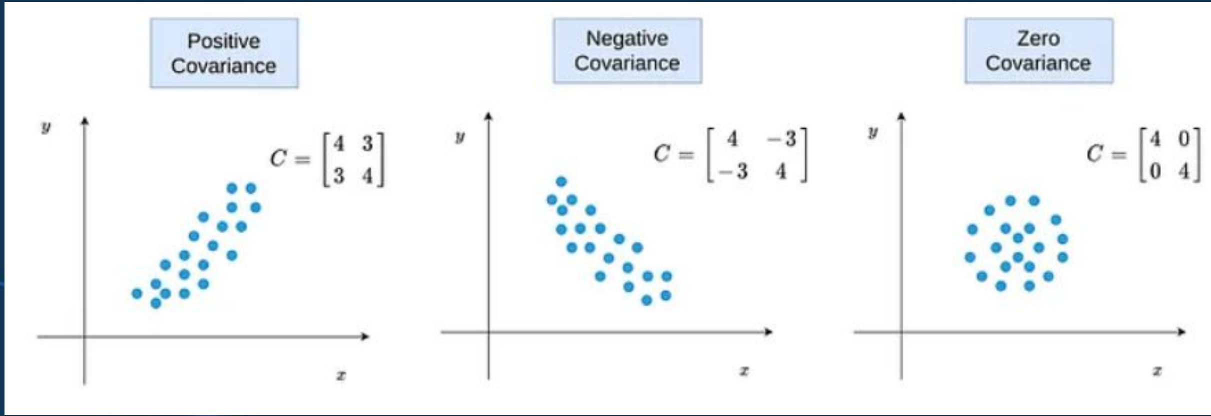
CoGrammar

Mathematical Foundation

❖ The Covariance Matrix measures the linear relationship between variables. It helps us identify the direction of maximum variance in the data. We denote the covariance matrix with an uppercase C.

$$var(x) = \frac{\sum_i^n (x_i - \overline{x})^2}{N - 1}$$

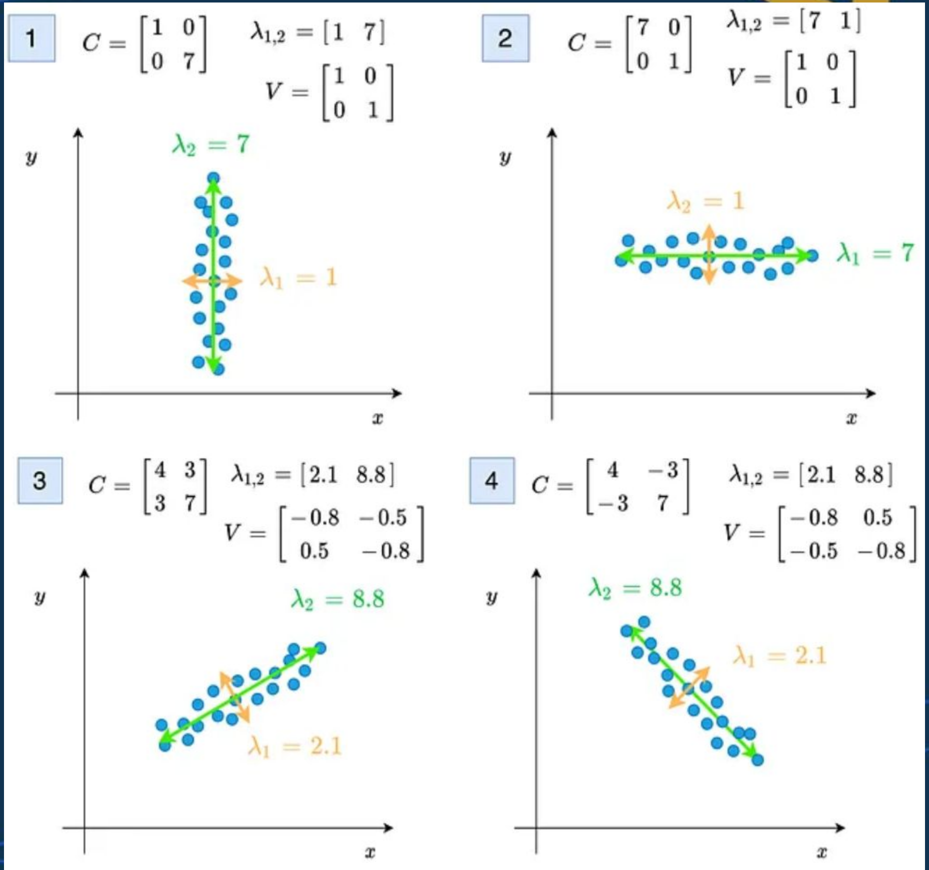
$$\begin{array}{c c} & \begin{array}{cc} x & y \end{array} \\ \begin{array}{c} x \\ y \end{array} & \begin{bmatrix} var(x) & cov(x, y) \\ cov(x, y) & var(y) \end{bmatrix} \end{array} \qquad \begin{array}{c c} & \begin{array}{ccc} x & y & z \end{array} \\ \begin{array}{c} x \\ y \\ z \end{array} & \begin{bmatrix} var(x) & cov(x, y) & cov(x, z) \\ cov(x, y) & var(y) & cov(y, z) \\ cov(x, z) & cov(y, z) & var(z) \end{bmatrix} \end{array}$$



CoGrammar

Mathematical Foundation

- Eigenvalues are scalar values that give the magnitude of variance in the direction of eigenvectors. Eigenvectors indicate directions of principal components.
- In other words, eigenvectors define directions of the new feature space, and eigenvalues determine their magnitude.
- λ = eigenvalues, V = eigenvectors



CoGrammar

Source: https://towardsdatascience.com/5-things-you-should-know-about-covariance-26b12a0516f1

Principal Components

CoGrammar

Principal Components

- Each **Principal Component** is a weighted sum of the original features i.e., PCs are linear combinations of the original variables.
- Components are ordered by the amount of variance they explain, with the first component explaining the most.
- **Formula**: $PC_i = w_{i\,1} X_1 + w_{i\,2} X_2 + \ldots + w_{i\,p} X_p$

where $w_{i\,j}$ are the weights (loadings) of the j-th feature in the i-th principal component. The weights indicate the contribution of each original feature to the principal component.

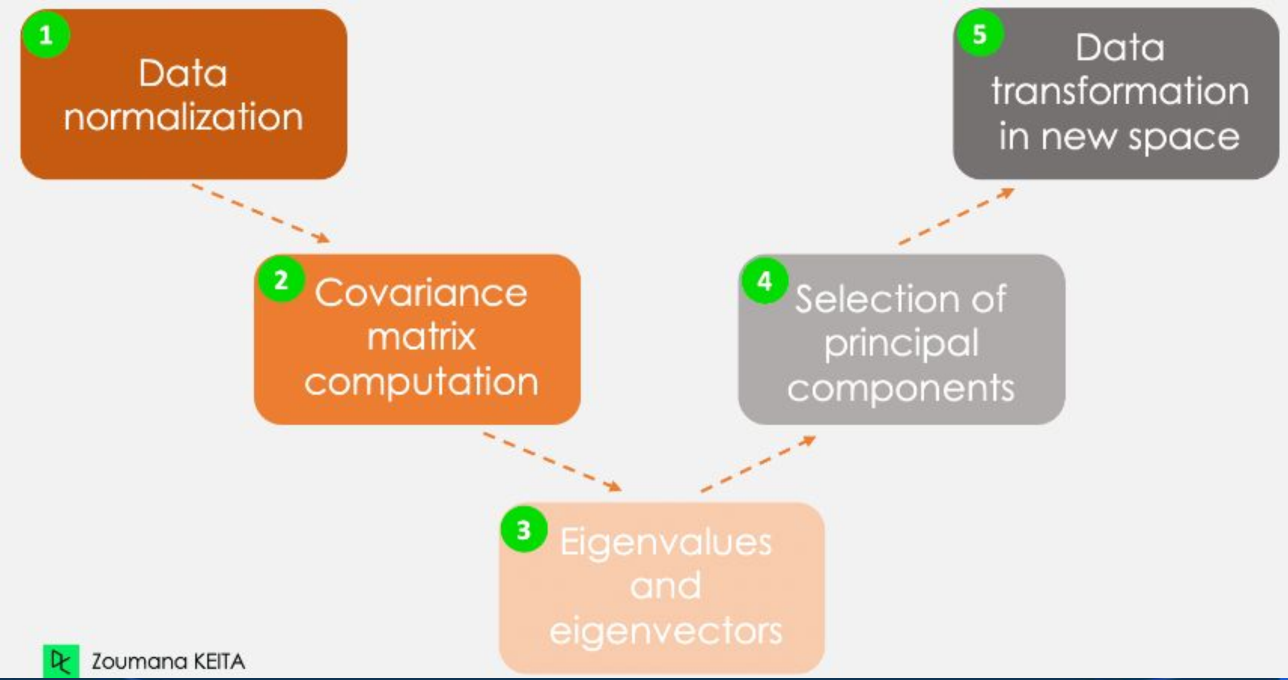
Example: for a dataset with three features X_1, X_2, X_3, the first 2 PCs are:

Principal Component 1: $PC_1 = w_{1\,1} X_1 + w_{1\,2} X_2 + w_{1\,3} X_3$

Principal Component 2: $PC_2 = w_{2\,1} X_1 + w_{2\,2} X_2 + w_{2\,3} X_3$

CoGrammar

Steps to perform PCA



Source: https://www.datacamp.com/tutorial/pca-analysis-r

Interpreting PCA Results

CoGrammar

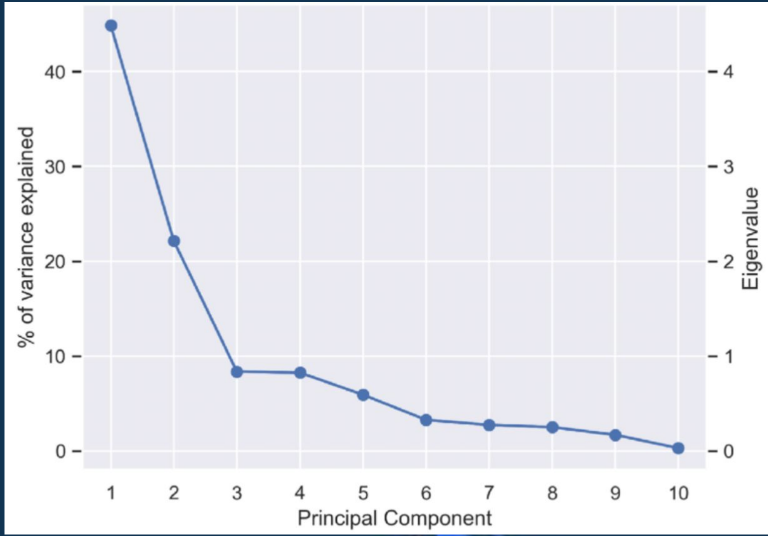
Interpreting PCA Results

- **Variance Explained:** the proportion of the total variance in the data that each principal component (PC) accounts for.
- Think of explained variance as the amount of information about the original data that each PC retains. PCA aims to reduce dimensions while retaining as much information as possible. For example, going from a 10-dimensional space to a 3-dimensional space while preserving the majority of the variance.
- Let's say we have a dataset containing 20 features. We can create up to 20 principal components. If the first 3 PCs explain 92% of the total variance, the remaining 17 PCs explain only 8% of the variance. Retaining just the first 3 PCs is sufficient. By doing so, we:
  - Reduce dimensionality & focus on information that is most important
  - Increase efficiency by simplifying the dataset without significantly losing information

CoGrammar

Scree Plot

- A **Scree plot** is a plot of the eigenvalues or the % of explained variance in descending order. It helps us determine the number of principal components to retain.
- We are interested in the 'elbow' point.
- In the scree plot, the cumulative percentage of variance explained by the first 3 PCs is (approximately) 44% + 22% + 9% = 75%.
- Notice how the proportion of explained variance decreases as we move along the PCs. That is because PCs are ranked, with the first PC explaining the most. This is also why we only choose to work the first few PCs - they explain majority of the variance.



CoGrammar

Biplots

❖ A **Biplot** plots the scores of the first two principal components. It is a scatterplot of observations projected onto a 2-dimensional space defined by the first two PCs.

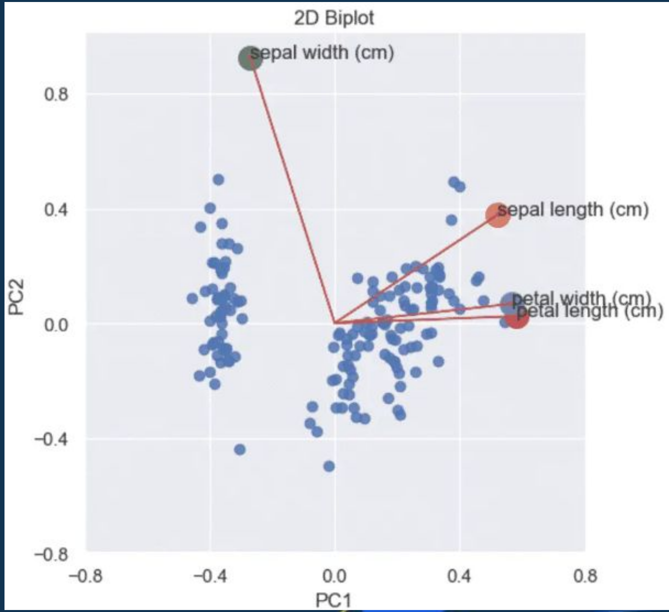
❖ When interpreting biplots, we look for the following:

➢ Clusters of observations - the closer points are to one another, the more similar they are

➢ Distance from origin - the farther points are from the origin, the more extreme their values are on the PCs

➢ Angles between vectors - small angles indicate strong positive correlations between variables, right angles indicate no correlation, and large angles (close to 180°) indicate strong negative correlations.

➢ Projection of points onto vectors - points in the same direction of a vector indicate higher values for that value. Opposite direction indicates lower values.



Source: https://www.jcchouinard.com/python-pca-biplots-machine-learning/

CoGrammar

Questions and Answers

CoGrammar

Thank you for attending

SKILLS FOR LIFE
SKILLS BOOTCAMPS

Department for Education

CoGrammar