Researchers in Schools

Using Mathematics to Analyse "Big Data": finding a needle in the "big data" haystack

Key Stage 4 Programme

Pupil Name

Coursebook Designed by

Contents

Uni Pathways launch event 3
Timetable and Assignment Submission 7
Timetable – Tutorials 7
Timetable – Homework Assignments 7
Assignment Submission – Lateness and Plagiarism 7
Course Rationale 8
Uni Pathways Mark Scheme 2020 9
Baseline Assignment (essay based): Pupil Feedback Report 11
Final Assignment (essay based): Pupil Feedback Report 12
Subject Vocabulary 13
Tutorial 1 – Introduction to Data 15
What is the Purpose of Tutorial 1? 15
By the end of the tutorial, students will be able to: 15
Tutorial 2 – Introduction to Matrices 19
What is the Purpose of Tutorial 2? 19
By the end of the tutorial, students will be able to: 19
Tutorial 3 – Variance and Covariance 27
What is the Purpose of Tutorial 3? 27
By the end of the tutorial, students will be able to: 27
Tutorial 4 – Principal Component Analysis 35
What is the Purpose of Tutorial 4? 35
By the end of the tutorial, students will be able to: 35
Tutorial 5 – Real life Principal Component Analysis 42
What is the Purpose of Tutorial 5? 42
By the end of the tutorial, students will be able to: 42
Tutorial 6 – Feedback tutorial 44
What is the Purpose of Tutorial 6? 44
Final assignment feedback from your Uni Pathways Teacher 44
Tutorial 7 – Final tutorial 46
Appendix 1 – Referencing correctly 48
How do I reference? 49
Notes 50

Uni Pathways launch event

Welcome to Uni Pathways! We are delighted to be able to launch the Uni Pathways programme with you. We hope that you are ready to embark on your Uni Pathways journey and that you enjoy the video.
You will be asked to pause the video at times to complete some work in this workbook, so make sure you have a pen / pencil to hand when you start the video.

By the end of the video, you will have

- Learnt about what studying at university means
- Learnt about some of the skills that you will develop during Uni Pathways
- Heard from pupils who have participated in Uni Pathways or The Scholars Programme (which is the same programme!)
- Heard from current university students talking about what life is like at university

If you are in school your teacher will play the video. If you are at home and logged in to a session with your teacher, your teacher will play the video and show it to you. If you are participating independently your teacher will email you the link to the video or the video file. There are opportunities for you to answer some questions, and you will be told when to pause the video to answer them.

Introduction to Uni Pathways

1. Write down what you think a supra-activity is

_____
_____
_____

2. Independent learning is

_____
_____
_____

3. Write down in your own words what resilience means, and come up with a different example to the one mentioned in the PowerPoint.

_____
_____
_____
_____

4. Why do you think analysis and evaluation skills are useful in your academic career? _____
_____
_____
_____
_____
_____

5. Why do you think analysis and evaluation skills are useful in life in general? _____
_____
_____
_____

6. What are the different types of learning that university students do?

Pupils' experience of Uni Pathways
You will now watch some pupils talk about their experience of learning a PhD topic and producing a final assignment. Some may refer to The Scholars Programme rather than Uni Pathways. As you heard in the introduction PowerPoint, The Scholars Programme is a very similar programme to Uni Pathways. Listen carefully and then respond to the questions below. Be prepared to share some of your responses with your peers.

1. How did the pupils describe talking about writing a long essay for their final assignment?

2. Write down something that one of the pupils mentioned was particularly interesting in their course.

3. What were some challenges that the pupils met?

Virtual campus tours
You will now go on a virtual tour of some universities! Once you have seen some of the university campuses, respond to the questions below.
Note down something that you liked out of any of the campus tours you saw. It may be a particular building, space, city etc.

1. Were there any similarities between the different university campuses? If so, what are the similarities?

2. If you had to pick one of those universities to go and visit in person, which one would it be and why?

3. Note down some of the societies that you could join at different universities

Meet university students!
You are about to virtually meet or hear from some current university students.
If you are virtually meeting them, think about what questions you would like to ask the students, share your questions with a partner and note those questions down in the space below.

If you are not meeting them but watching some videos that they have made, use the space below to note down what else you would like to find out about university. You can then ask your teacher during Uni Pathways tutorials!

Reflection
Congratulations on completing the launch of the Uni Pathways programme. Before you go, take some time to reflect on what you have learnt by answering the following questions:

1. Are there any aspects of university style learning that you would look forward to? If so, what are they?

2. What challenges do you think students face when learning at university? How would you try to overcome these challenges?

3. What skills do you hope to develop during your Uni Pathways course?

4. What part of Uni Pathways seems the most challenging for you?

5. What are you most looking forward to about Uni Pathways?

Timetable and Assignment Submission

Timetable – Tutorials

Tutorial	Date	Time	Location
1			
2			
3			
4			
5			
6 (Feedback)			
7 (Feedback)			

Timetable – Homework Assignments

Homework Assignment	Description	Due Date
Tutorial 1	Baseline assignment	
Tutorial 2	Matrices, vectors and eigenvectors: short report and questions.	
Tutorial 3	Variance and covariance: short report and questions.	
Tutorial 4	Principal component analysis: short report and questions.	
Tutorial 5	Draft final assignment	

Assignment Submission – Lateness and Plagiarism

Lateness	
Submission after midnight on deadline day	10 marks deducted
Plagiarism	
Some plagiarism	10 marks deducted
Moderate plagiarism	20 marks deducted
Extreme plagiarism	Automatic fail

Course Rationale

Here we will look at analysing “big data”, where the amount of data we have is too large to consider using conventional techniques. The modern world is full of “big data”, from the data that Google have on your movements, to what Amazon collect about your shopping, to what Netflix have on your TV preferences. Collecting this data is the easy part, but figuring out what to do with it is tricky. Given your previous TV watching habits, how does Netflix decide what programmes to recommend to you?

In this course we will consider this problem. We will investigate a method that helps to find patterns in high dimensional data (this sounds complicated, but just means that we are measuring lots of things), and also allows us to compress this data. We will introduce the mathematics behind this method, and apply it to some real-life examples.

This area of study is extremely important at the moment, and is often used as an introduction to artificial intelligence and machine learning (specifically, unsupervised machine learning). Knowledge of this field is in high demand by employers.

Uni Pathways Mark Scheme 2020

	Subject Knowledge	Critical Thinking	Written Communication
1st	The work shows a depth of knowledge and understanding of key concepts and scientific methods, through engaging with relevant sources. Knowledge is used to build and support highly effective scientific arguments and explanations.	Analyses key scientific evidence, arguments, and reasoning. Interprets meaning and makes connections. Identifies and critically evaluates key scientific arguments and evidence, deciding on their credibility, strength, and relative significance, drawing convincing conclusions.	The work has a coherent flow and is well structured. The writing style is appropriate; scientific language and key scientific terms are used accurately and effectively to support the arguments and explanations made. There are no, or very few, errors in spelling or grammar. Consistent referencing, appropriate paragraphing and use of correctly labelled tables and graphs matching the style taught in the course.
2:1	The work shows an understanding of key concepts and scientific methods, drawing on relevant sources. Knowledge is used to build and support effective scientific arguments and explanations.	Analyses relevant scientific evidence, arguments, and reasoning. Identifies and critically evaluates relevant scientific arguments and evidence, deciding on their credibility and strength, drawing reasonable conclusions. Shows some understanding of the relative value of evidence and arguments.	The work is well-structured. The writing style is appropriate; scientific language and key terms are used correctly. There are few errors in spelling or grammar. Mostly consistent referencing and use of tables and figures; matching the style taught in the course.
2:2	The work shows an understanding of key concepts and scientific methods, with no major misconceptions. Beginning to apply this knowledge to build and support effective scientific arguments and explanations.	Identifies and uses basic scientific evidence, arguments, and reasoning. Showing some understanding of the quality of scientific arguments and evidence. Not yet showing understanding of the relative value of evidence and arguments.	The work has some structure. The writing style can sometimes be informal; occasionally scientific language and key terms are not used when it would be appropriate to do so. There are some errors in grammar and spelling do not get in the way of communicating the content. Referencing has some consistency; matching the style taught in the course Limited use of tables and graphs.

3rd	Shows a developing understanding of key concepts and scientific methods, with some misconceptions. Does not yet apply this knowledge to build and support scientific arguments and explanations.	Beginning to analyse scientific evidence, arguments, and reasoning. Describes evidence and arguments, while not yet evaluating them.	The grammar, spelling, style, and structure of the work need improving in order to communicate ideas to the reader. Scientific language, key terms and references are not always used correctly. Limited, or no use of tables and graphs.

Baseline Assignment (essay based): Pupil Feedback Report

Name of Pupil	
Name of School	St Gregory's Catholic College
Name of RIS teacher	Mr Smith
Title of Assignment	

How your assignment is graded:

Grade	Marks	What this means
1st	70+	Performing to an excellent standard at A-level
2:1	60-69	Performing to a good standard at A-level
2:2	50-59	Performing to an excellent standard at GCSE
3rd	40-49	Performing to a good standard at GCSE
Working towards a pass	0-39	Performing below a good standard at GCSE
Did not submit	DNS	No assignment received by The Brilliant Club

Lateness	
Any lateness	10 marks deducted
Plagiarism	
Some plagiarism	10 marks deducted
Moderate plagiarism	20 marks deducted
Extreme plagiarism	Automatic fail

Marks

OVERALL MARK / 100		FINAL MARK / 100 (including any deductions)	
DEDUCTED MARKS		FINAL GRADE	

If marks have been deducted (e.g. late submission, plagiarism) the teacher should give an explanation in this section:

Mark Breakdown and Feedback

Subject knowledge	
	mark
Critical thinking	
	mark
Written communication	
	mark

Final Assignment (essay based): Pupil Feedback Report

Name of Pupil	
Name of School	
Name of RIS teacher	
Title of Assignment	

How your assignment is graded:

Grade	Marks	What this means
1st	70+	Performing to an excellent standard at A-level
2:1	60-69	Performing to a good standard at A-level
2:2	50-59	Performing to an excellent standard at GCSE
3rd	40-49	Performing to a good standard at GCSE
Working towards a pass	0-39	Performing below a good standard at GCSE
Did not submit	DNS	No assignment received by The Brilliant Club

Lateness	
Any lateness	10 marks deducted
Plagiarism	
Some plagiarism	10 marks deducted
Moderate plagiarism	20 marks deducted
Extreme plagiarism	Automatic fail

Marks

OVERALL MARK / 100		FINAL MARK / 100 (including any deductions)	
DEDUCTED MARKS		FINAL GRADE	

If marks have been deducted (e.g. late submission, plagiarism) the teacher should give an explanation in this section:

Mark Breakdown and Feedback

Subject knowledge	
	mark
Critical thinking	
	mark
Written communication	
	mark

Subject Vocabulary

Word	Definition	In a sentence
Real number, $\mathbb{R}$	Any value that can represent a continuous distance along a line. We say "x belongs to the set of real numbers" when we see $x \in \mathbb{R}$.	Five is a real number.
Vector, $w \in \mathbb{R}^p$	A collection of p real numbers, where $w = (w_1, w_2, \dots, w_p)$ and each $w_1 \in \mathbb{R}$, $w_2 \in \mathbb{R}$, $\dots, w_p \in \mathbb{R}$. We say "$w$ is a p-dimensional vector" when we see $w \in \mathbb{R}^p$.	$w = (1, 20, -\pi)$ is a 3 dimensional vector.
Orthogonal vectors	Two vectors $v \in \mathbb{R}^p$ and $w \in \mathbb{R}^p$ are orthogonal if they are perpendicular (so that the angle between them is 90 degrees). Here we have $v^\top w = \sum_{i=1}^{p} v_i w_i = 0$, and we can write $v \perp w$.	The vectors $v = (1, 0, -1)$ and $w = (-1, 1, -1)$ are orthogonal, as here $v^\top w = (1 \cdot -1) + (0 \cdot 1) + (-1 \cdot -1) = 0$.
Direction, $w \in \mathbb{R}^p$	The vector $w \in \mathbb{R}^p$ is called a direction if the sum of each element squared is equal to 1. That is, $w^\top w = \sum_{i=1}^{p} w_i^2 = 1$.	The three-dimensional vector $w = (\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}})$ is a direction, as we have $\left(\frac{1}{\sqrt{3}}\right)^2 + \left(\frac{1}{\sqrt{3}}\right)^2 + \left(\frac{1}{\sqrt{3}}\right)^2 = 1$.
Matrix, $X \in \mathbb{R}^{n \times p}$	Matrices are tables of numbers that are put in big brackets. The 'order' or 'size' of the matrix is given by the number of rows and the number of columns. All rows must have the same number of columns.	The 3 by 2 matrix $X \in \mathbb{R}^{3 \times 2}$ has elements, $X = \begin{pmatrix} -100 & \frac{2}{3} \\ \pi & 5 \\ 723843724.2321 & 1 \end{pmatrix}$.
Projection, $z \in \mathbb{R}^n$	The vector $z \in \mathbb{R}^n$ is a projection if $z = X\,w$, where $X \in \mathbb{R}^{n \times p}$ is some matrix and $w \in \mathbb{R}^p$ is a direction.	We project the data X along the direction w to obtain the projection $z = Xw$.
Eigenvector and Eigenvalue	Let $X \in \mathbb{R}^{n \times n}$ be a n by n matrix. Suppose the vector $v \in \mathbb{R}^n$ satisfies $Xv = \lambda v$ for some number $\lambda \in \mathbb{R}$. Then, v is an eigenvector of X, and λ the corresponding eigenvalue.	Let the 2 by 2 matrix be given by, $X = \begin{pmatrix} 1 & 0 \\ 1 & 2 \end{pmatrix}$. Then, $v = (0, 1)$ is an eigenvector of X, with eigenvalue $\lambda = 2$.
Variance and Covariance	Variance is a measure of the spread of some data from the mean. Covariance is a measure of the similarity between two paired sets of data. A positive covariance means that a high value of one set of data means the associated pair from the	We measure the height and arm-span of a sample of students from St Gregory's school. The covariance between the heights and arm-spans is positive, indicating that the taller a student is, the more likely they will have a larger arm-span.

	second data set is also likely to have a high value.	
Population	The population is all of the people, objects or events that you are interested in.	In the investigation to find out what the students think of school dinners, the population is all the students in the school.
Sample	A selection of people, objects or events chosen from a population, such that they are representative of the whole population.	We select a sample of students in the school and ask them about the school dinners. This saves us asking all the students in the school.
Sample point	One event or item from a population.	One student is a sample point from the population of all the students in the school.
Variable	A variable is any characteristics, number, or quantity that can be measured or counted.	One variable of our data is the quality of the school diners, which we measure on a scale from 0 to 10.
High dimensional data	Data where the number of variables is large, such that we are measuring lots of different things. Often, the number of variables is larger than the size of the sample.	The data that Google collect about you is high-dimensional, as they collect data on lots of things you do online.
Dimension reduction	Reducing the number of variables of some data set. This can be done by removing some variables completely, or by combining the variables to create a new variable.	Instead of measuring both the height and the arm-span of our students, we could just measure the height. As the covariance between the two variables is large, we could estimate the associated arm- span from the height. This reduction in the number of variables is known as dimension reduction.
Principal Component Analysis (PCA)	PCA is a method of dimension reduction, that obtains new variables which are a combination of the original variables, chosen such that the variance of the new variables is maximum. This gives variables with a large spread.	We apply PCA to the height and arm-span data, to obtain a combination of the height and arm-span which maximises the spread of the (new, one dimensional) data.

Tutorial 1 – Introduction to Data

What is the Purpose of Tutorial 1?

By the end of the tutorial, students will be able to:

1) Explain what "data" are, and three examples.
2) Discuss two issues that can occur when the amount of data is large.
3) Define key terms: 'variables', 'dimensions', 'population', and 'sample', and know what `high dimensional data' is.

Introduction to Tutorial 1:

The term "data" can refer to any information that can be collected through observation. Data is often referred to as the "new oil" due to its importance in the modern economy. Data can refer to many diverse things, from prices of the chocolate bars in a vending machine, to the value of shares in a collection of companies, to the number of homeless people in Bath.

In the modern world, a huge amount of data is collected on each one of us every day. Google collect information on your movements and your search history; Instagram collect information on what videos you like; and the bus company collect information on when and where you take the bus.

In this tutorial we will discuss some examples of data, and then introduce some key terms that will be used throughout the Uni Pathways course. We will also discuss *high dimensional data*.

Classroom tasks:

1) **In Pairs: Discuss and note down some example of "data" that you can think of.**

2) **Definitions:**

Population:

A population is…

Sample point:

A sample point is…

Sample:

A sample is…

Variable:

A variable is…

3) Examples of populations and sample, and associated variables:

High dimensional data:

We say data is high dimensional if…

Assignment 1 – Baseline Assignment

Write a short summary (300-500 words) on what is meant by "data", describing what we mean by a sample and a population.

You should include:

i) Some examples of data, samples and populations;
ii) A description of "high dimensional" data, giving some real-world examples;
iii) Some reasons why high dimensional data might be difficult to analyse.

Tutorial 2 – Introduction to Matrices

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \dots & n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & a_{32} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \end{array}$$

What is the Purpose of Tutorial 2?

By the end of the tutorial, students will be able to:

1) Write data as a (n x p) matrix, and can describe what the rows and columns represent.
2) Multiply a matrix by a vector.
3) Define eigenvectors and eigenvalues mathematically.

Material covered in tutorial 2:

In this tutorial we will discuss how data can be represented in matrix form.

A **matrix** is a table of numbers enclosed by big brackets. The size of a matrix is given by the $(number\ of\ rows) \times\ (number\ of\ columns)$. Every row of a matrix must contain the same number of elements.

If X is a matrix with n rows and m columns, then we say that X has size $n\ \times m$, and this is notated $X \in \mathbb{R}^{n\ \times m}$. This matrix has elements,

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix}.$$

A **vector** can be thought of as a matrix with only 1 column. The size of a vector is given by the $number\ of\ elements$ that it contains.

If w is a vector with p elements, then we say that w is of size p, and this is notated $w \in \mathbb{R}^p$. This vector has elements,

$$\begin{pmatrix} w_1 \\ \vdots \\ w_p \end{pmatrix}.$$

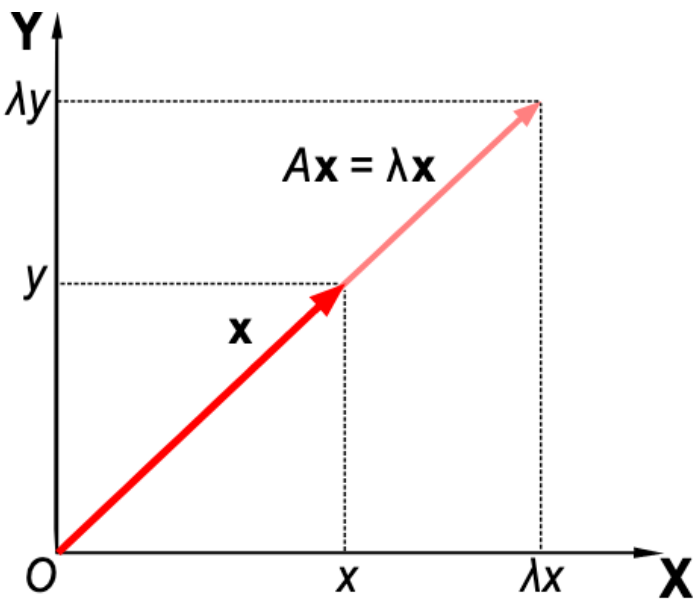
For information on multiplying matrices, see https://ed.ted.com/lessons/how-to-organize-add-and-multiply-matrices-bill-shillito#watch

Before **multiplying a vector by a matrix**, we need to ensure that they are compatible sizes. A vector can be multiplied by a matrix if the size of the vector is equal to the number of columns of the matrix. That is, the vector $w \in \mathbb{R}^p$ can be multiplied by a matrix X if the matrix has size $(n \times p)$ for any number of rows n. In this case the output is a vector $y \in \mathbb{R}^n$ of size n, and has elements,

$$y = X\,w = \begin{pmatrix} x_{11}w_1 + x_{12}w_2 + \cdots + x_{1p}w_p \\ \vdots \\ x_{n1}w_1 + x_{n2}w_2 + \cdots + x_{np}w_p \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{p} x_{1i}w_i \\ \vdots \\ \sum_{i=1}^{p} x_{ni}w_i \end{pmatrix},$$

where here **summation notation** is used to represent the sum of the elements. The dots "⋯" and "⋮" represent missing elements that follow the same pattern as the others.

An **eigenvector** of a matrix is a vector such that when it is multiplied by the matrix, the direction of the vector remains the same. The magnitude of the eigenvector can change when multiplied by the matrix, and this is represented by the associated **eigenvalue**. For example, if we have a matrix $A \in \mathbb{R}^{2\times 2}$ with an eigenvector $x \in \mathbb{R}^2$, then the direction of the product Ax (which is a vector of size 2) is the same as that of the eigenvector. This is shown by the following diagram, with the associated eigenvalue given by $\lambda \in \mathbb{R}$. Mathematically, we have $Ax = \lambda x$.



1) Worked examples: introduction to matrices

a) Matrix 1: The identity matrix in $p = 2$ dimensions is given by,

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Here, we write $X \in \mathbb{R}^{2\times 2}$, as we have two rows and two columns.

b) Matrix 2: The identity matrix $X \in \mathbb{R}^{4\times 4}$ in $p = 4$ dimensions is given by,

$$X =$$

c) Matrix 3: The matrix,

$$X = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}$$

has three rows and two columns, so we write $X \in$

2) Multiplying a matrix and a vector:

A matrix can be multiplied by a vector when…

3) Worked examples: multiplying a matrix by a vector.

a) Example 1: Let,

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ and } v = \begin{pmatrix} 7 \\ -2 \end{pmatrix}.$$

Calculate Xv. Recall that X here is called the identity matrix. Consider why it might be called this?

b) Example 2: Let,

$$X = \begin{pmatrix} 3 & 2 \\ -1 & 1 \end{pmatrix} \text{ and } v = \begin{pmatrix} -2 \\ 2 \end{pmatrix}.$$

Calculate Xv.

c) Example 2: Let,

$$Y = \begin{pmatrix} 1 & -1 & 4 \\ 0 & 3 & -2 \end{pmatrix} \text{ and } z = \begin{pmatrix} 2 \\ -1 \\ 3 \end{pmatrix}.$$

Calculate Yz.

4) Definitions: Let $X \in \mathbb{R}^{p \times p}$ be some matrix. Then,
An eigenvector…

An eigenvalue…

5) Worked examples: Check whether the following vectors are eigenvectors. If so, state the associated eigenvalue.

a) Let $X \in \mathbb{R}^{2 \times 2}$ be given by the identity matrix,

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Is the vector $v = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ an eigenvector of X? If so, find the associated eigenvalue.

b) Let $X \in \mathbb{R}^{3\times 3}$ be given by

$$X = \begin{pmatrix} -2 & -4 & 2 \\ -2 & 1 & 2 \\ 4 & 2 & 5 \end{pmatrix}.$$

Is the vector $v = \begin{pmatrix} 2 \\ -3 \\ -1 \end{pmatrix}$ an eigenvector of X? If so, find the associated eigenvalue.

c) Let $X \in \mathbb{R}^{2\times 2}$ be given by

$$X = \begin{pmatrix} 4 & 1 \\ 2 & 1 \end{pmatrix}.$$

Is the vector $v = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$ an eigenvector of X? If so, find the associated eigenvalue.

Assignment 2 – Matrices, vectors and eigenvectors: short report and questions.

Complete all homework on a separate sheet of paper.

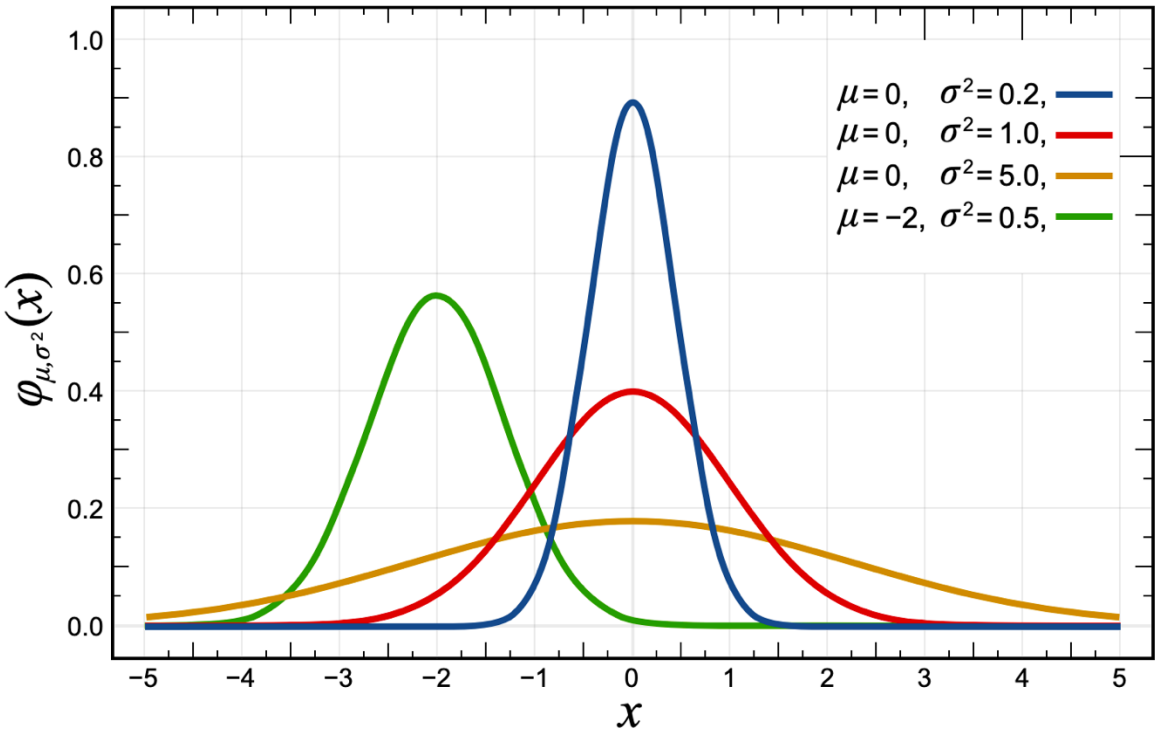
1) Describe, in words and using mathematical symbols, the conditions required to multiply a vector, v, by a matrix, X. Describe what the resulting object $w = Xv$ looks like.

2) When you multiple a vector by a matrix, the matrix transforms the original vector. With this in mind, consider the matrix,
$$X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$
Comment on why this is known as the 'identity matrix'.

3) Let,
$$X = \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix} \text{ and } v = \begin{pmatrix} 7 \\ -2 \end{pmatrix}.$$
Calculate Xv.

4) Let $X \in \mathbb{R}^{2\times 2}$ be given by
$$X = \begin{pmatrix} 1 & 5 \\ 3 & 3 \end{pmatrix}.$$
Is the vector $v = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ an eigenvector of X?

Tutorial 3 – Variance and Covariance



What is the Purpose of Tutorial 3?

By the end of the tutorial, students will be able to:

1) Calculate the sample variance of some data.
2) Understand what a covariance matrix is, and what each element represents.

Material covered in tutorial 3:

In this tutorial we will learn about and discuss the following concepts: **variance, covariance, sample variance** and **sample covariance.**

Before we discuss variance and covariance, we need to understand what a **random variable** is. A random variable is some variable that can take values in a sample space. For example, if we flip a coin, the sample space is either Heads or Tails, and then we can define X to be the random variable which can either equal 'Heads' or 'Tails' each with probability $\frac{1}{2}$. If instead X was the random variable that gives the result of a die throw, then the sample space would be the set of numbers $1, 2, 3, 4, 5, 6$, with each value occurring with probability $\frac{1}{6}$.

Now we can introduce **variance**. Variance is a measure of how far numbers are **spread out from the mean value**. It is often notated by the symbol σ^2, or by $\mathrm{Var}(X)$, where X is a random variable. This is an **exact** number, and is calculated from the properties of the random variable X directly.

For example, with the throw of the die, the variance is calculated using the fact that **we know** that the set of numbers $1, 2, 3, 4, 5, 6$, are each as likely to occur. But, what if we do not know anything about X directly, but instead have just been told the outcome of a number of realisations of X? In this case we would just have a list of numbers, with each number representing the value from a single die throw, e.g. $1, 3, 3, 5, 4, 5, 4, 6, 3, 4, 6, 4, 6, 2$. In this case, we can estimate the variance by calculating the **sample variance**.

The **sample variance** is calculated by finding an average of the **squared differences** from the mean. Suppose we have a sample of size n, given by $x_1, x_2, \ldots, x_n$. Then, the sample variance (given by the letter s^2) can be calculated in three steps:

1. Calculate the mean,

$$\mu = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n};$$

2. Find the squared differences by subtracting the mean from each number and squaring the result, which gives

$$y_1 = (x_1 - \mu)^2; y_2 = (x_2 - \mu)^2; \cdots; y_n = (x_n - \mu)^2;$$

3. Divide the sum of the squared differences by $n - 1$, to give the sample variance,

$$s^2 = \frac{y_1 + y_2 + \cdots + y_n}{n - 1} = \frac{\sum_{i=1}^{n} y_i}{n - 1}.$$

Example: If we have the set of values from fourteen throws of a die, $1, 3, 3, 5, 4, 5, 4, 6, 3, 4, 6, 4, 6, 2$, then the sample variance can be calculated in the following way:

1. Calculate the mean:

$$\mu = \frac{1 + 3 + 3 + 5 + 4 + 5 + 4 + 6 + 3 + 4 + 6 + 4 + 6 + 2}{14} = 4;$$

2. Find the squared differences:

$$\begin{gathered} y_1 = (x_1 - \mu)^2 = (1 - 4)^2 = 9 \\ y_2 = (x_2 - \mu)^2 = (3 - 4)^2 = 1 \\ \vdots \\ y_{14} = (x_{14} - \mu)^2 = (2 - 4)^2 = 4, \end{gathered}$$

which gives the values,

$$\begin{gathered} y_1 = 9, y_2 = 1, y_3 = 1, y_4 = 1, y_5 = 0, y_6 = 1, y_7 = 0, y_8 = 4, y_9 = 1, y_{10} = 0, \\ y_{11} = 4, y_{12} = 0, y_{13} = 4, y_{14} = 4; \end{gathered}$$

3. Calculate the average of the squared differences to give the sample variance,

$$s^2 = \frac{9 + 1 + 1 + 1 + 0 + 1 + 0 + 4 + 1 + 0 + 4 + 0 + 4 + 4}{14 - 1} = \frac{30}{13}.$$

A **covariance matrix** extends the idea of the variance to a random variable that is more than one dimensional. Similarly, the **sample covariance matrix** is an estimate of the true covariance given some data that is more than one dimensional.

Covariance is closely connected to **correlation**: they both measure the extent to which two sets of (ordered) data *move in the same direction*. For example, a persons' height and their arm span are highly (positively) correlated, as we would expect that if one is large then so is the other. They would also have a high covariance. An example of data that would be negatively correlated (and therefore have a negative sample covariance) is altitude and air pressure, as when altitude increases, air pressure decreases.

Examples of data with large positive covariance, near zero covariance and large negative covariance are shown in Figure 1, Figure 2, and Figure 3 respectively.

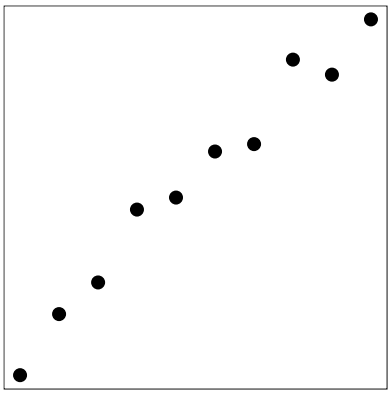


Figure 1: large positive covariance

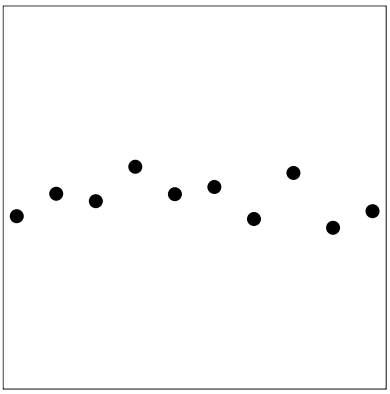


Figure 2: near zero covariance

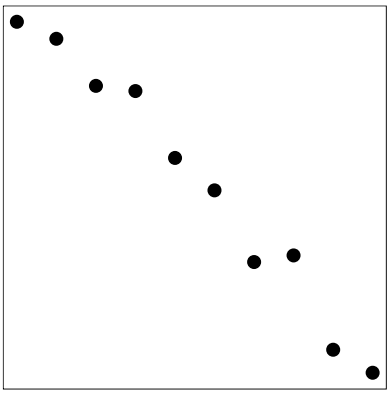


Figure 3: large negative covariance

Let some p-dimensional data be represented by the matrix $X \in \mathbb{R}^{n \times p}$, such that each dimension of the data has n elements. Then, the sample covariance matrix is given by the symbol $C_X \in \mathbb{R}^{p \times p}$. The diagonal of the sample covariance matrix gives the sample variance of each dimension of the data separately. That is, the ith diagonal element gives the sample variance of the ith column of X. The $(i,j)^{th}$ element of the sample covariance matrix C_X gives the covariance between the data from the ith and j^{th} column of X.

1) **Definitions:**

Variance

Sample variance

Steps for calculating sample variance:

2) Worked examples: Calculating sample variance given some data.

a. Suppose we have the heights of eleven students in cm, which are

143cm, 173cm, 120cm, 135cm, 140cm, 141cm, 167cm, 180cm, 173cm, 194cm, 175cm.

Calculate the sample variance of this data:

b. Calculate the sample variance of the following data, presented as a vector $x \in \mathbb{R}^{18}$:

$$x = (-3, 4.1, 700, 2, 53, -32, -10, 34, 1, 2, 3, -10, 9, 3, 42, 4, 1, 654)$$

3) **Covariance: in pairs, discuss what the covariance matrix represents, and the intuition behind it. Write down your conclusions.**

Assignment 3 – Variance and covariance: short report and questions.

Complete all homework on a separate sheet of paper.

1) Describe, in one or two sentences, what the term *variance* means?

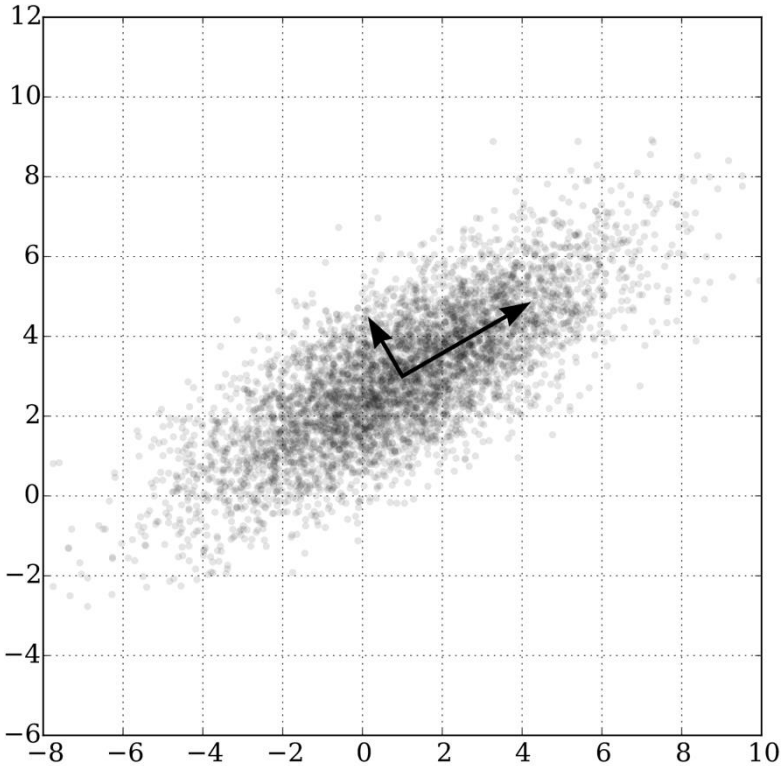
2) Consider the following two sets of data, given by the red and blue points in the plot in Figure 4. Which one would have a larger mean? Which one would have a larger sample variance? Explain your answer.

Figure 4: two sets of data given by blue and red points

3) Suppose we have the weights of nine apples in grams, given by
153g, 131g, 167g, 152g, 193g, 188g, 130g, 187g, 172g.
Calculate the sample variance of this data.

4) If two sets of data have positive covariance, what does this tell us about the relationship between the two?

Tutorial 4 – Principal Component Analysis



What is the Purpose of Tutorial 4?

By the end of the tutorial, students will be able to:

a. Describe the concept behind principal component analysis and can relate this to the line of best fit.
b. Know what principal directions and principal components are.
c. Write down how principal directions are given by the eigenvectors of the covariance matrix.

Material covered in tutorial 4:

In this tutorial we will cover **dimension reduction** using **principal component analysis** (often abbreviated to **PCA**). Principal component analysis is also used to find **hidden patterns** in data.

First, we discuss what **dimension reduction** means. When we talk about dimension reduction, we mean shrinking the number of dimensions of some data. In matrix form, suppose we have data given by the matrix $X \in \mathbb{R}^{n \times p}$, such that we have p dimensions and each dimension has n data points. Then, dimension reduction means that we are looking for a matrix $Y \in \mathbb{R}^{n \times r}$, with number of columns $r < p$, such that Y includes as much information about the original data X as possible. Although we lose information when using dimension reduction, the aim is that the resulting matrix Y still includes most of the important information, or that Y reveals some important patterns that are hidden in X.

Principal component analysis is a dimension reduction method that uses the sample covariance between dimensions to judge what is important or not. We will first introduce it in two dimensions. In this case, principal component analysis works in the exact same way as **line of best fit**.

In the **line of best fit** we find a line that goes through a set of points, minimising the **residual sum of squares.** Suppose we draw a line through a set of points, and measure how far each point is from this line (the *perpendicular* distance, to be precise). Then, the residual sum of squares is the sum of the squares of all these distances. It turns out that minimising the residual sum of squares is exactly the same as maximising the variance of the data projected onto the line. Here, by *projected* we mean that the data is placed onto the line of best fit. We will show what we mean by way of an example.

In Figure 5 below, we have 10 points, and a line of best fit (the red line) has been drawn. The distances between the line and the points are shown by the blue lines. Note that the blue lines are *perpendicular* to the line of best fit. The line of best fit here is the line that minimises the sum of the squares of these distances. It turns out that minimising the residual sum of squares is the same as maximising the spread of the data projected onto the line of best fit (given by the orange arrow).

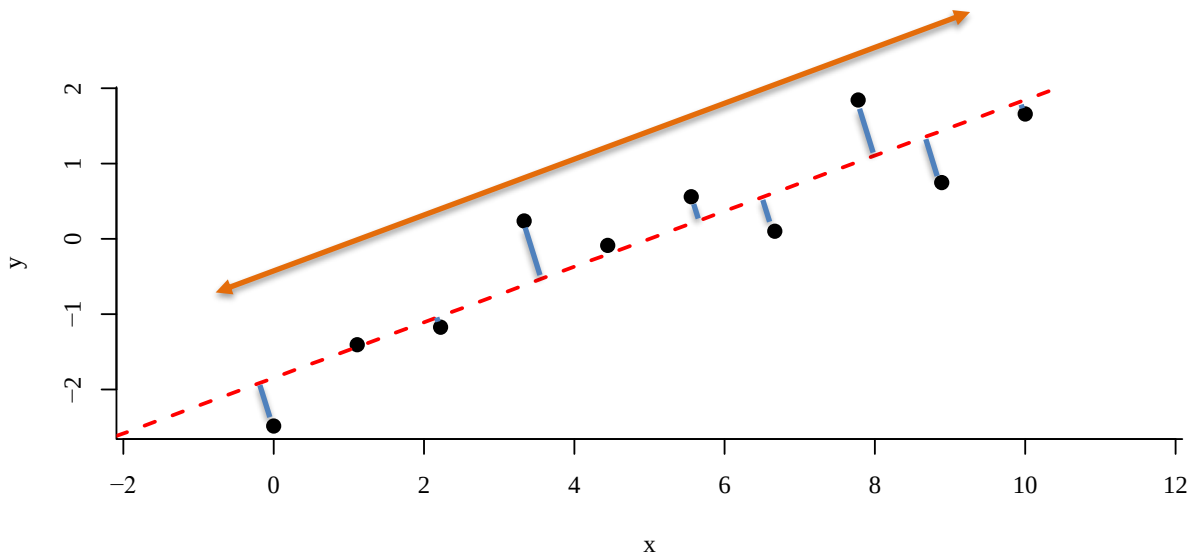


Figure 5: example data (black dots), with line of best fit minimising the residual sum of squares (red dotted line). The blue lines show the distance from the line of best fit to the true points, and the orange line gives the spread of the points along the line of best fit.

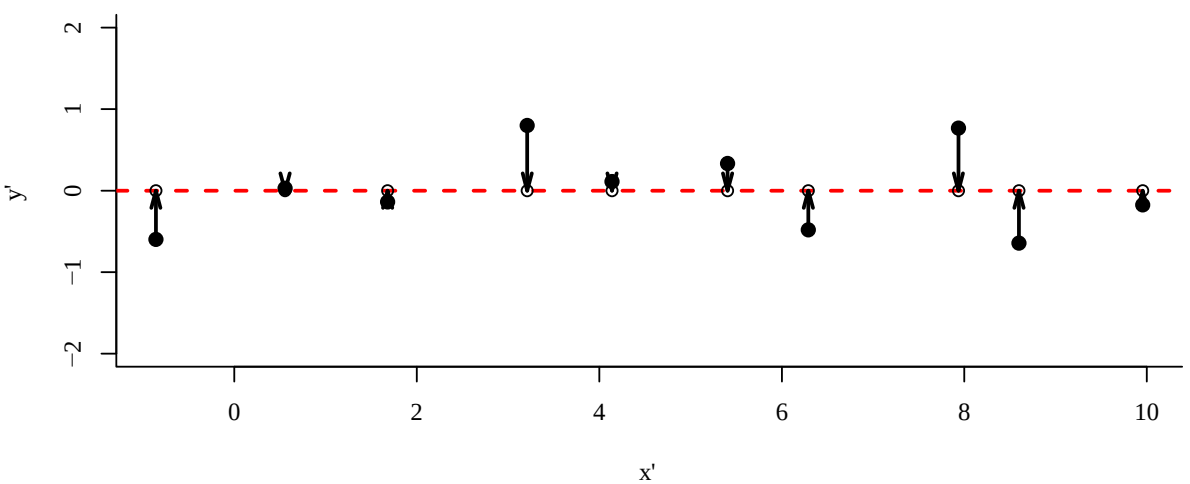


Figure 6: rotated data from Figure 5 (black dots) such that the line of best fit (red dotted line) is horizontal. The data is projected onto the line of best fit (along the arrows) and this is given by the open circles.

Now, suppose we *project* the data onto the line of best fit, such that all the new points lie on the red dotted line. The data projected onto the line of best fit is now in one dimension

(that is, all the new data points *lie on a line*) and therefore we have reduced a two dimensional data set to one dimension. The information we have lost in this process is given by the blue lines. We have lost information about the *spread* of the data perpendicular to the line of best fit. This is shown in Figure 6.

In Figure 7, we show how choosing a different line (given by the green line, which this time is *not of best fit*), results in the spread of the projected data (given by the purple arrow) being smaller than when the line of best fit was chosen.

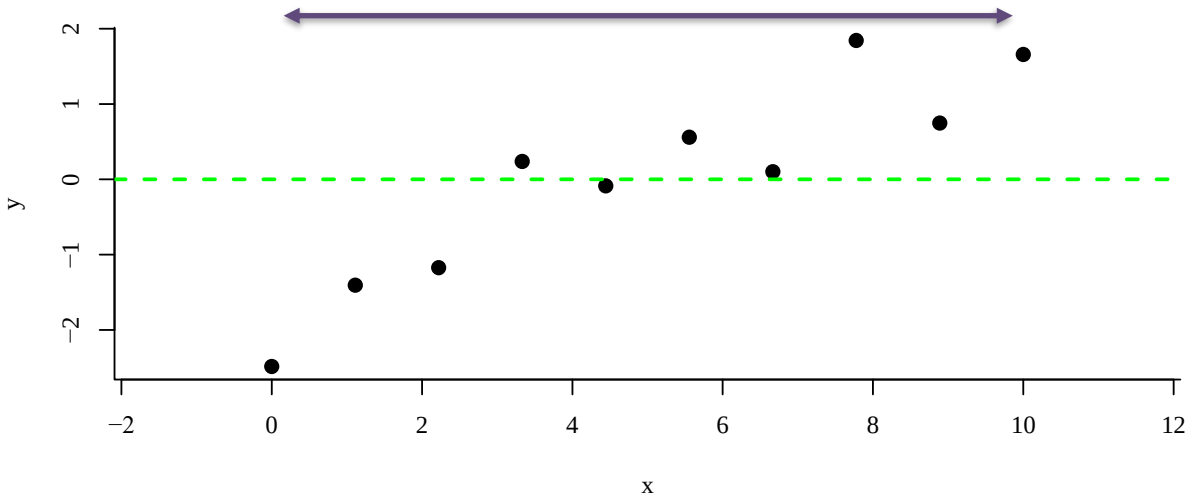


Figure 5: same data as in Figure 5, but here an incorrect line of best fit is chosen (given by the dotted green line). In this case, the spread of the data along this line (given by the purple arrows) is smaller than the spread of the data along the true line of best fit (shown by the orange arrows in Figure).

Principal component analysis expands the line of best fit to data that is higher than two dimensional. When we apply principal component analysis, we look for directions such that the projected data has maximal spread (*i.e.* the sample variance of the projected data is maximised). These directions are call the **principal directions**, and we find these directions one at a time (*sequentially*).

It can be shown that the **principal directions** are given by the **eigenvectors of the sample covariance matrix** of the data (see Tutorial 3 – Variance and Covariance). The first principal direction is the eigenvector that has the largest eigenvalue associated with it. Mathematically, suppose we have data $X \in \mathbb{R}^{n \times p}$, and we have calculated the sample covariance matrix, given by $C_X \in \mathbb{R}^{p \times p}$. Then, each principal direction $w_i \in \mathbb{R}^p$, $i = 1, 2, \dots, p$, satisfies

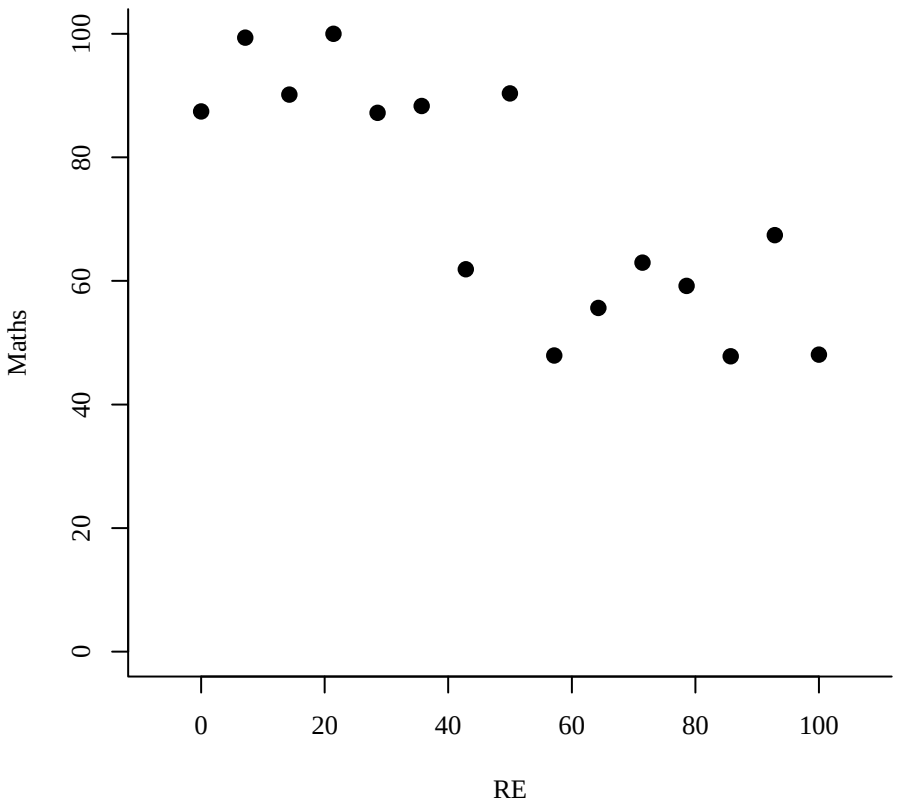
$$C_X w_i = \lambda_i w_i,$$

for some eigenvalue $\lambda_i \in \mathbb{R}$. We order the eigenvectors $w_1, w_2, \dots, w_p$ such that the associated eigenvalues satisfy, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. This means that the direction given by w_1 is called the first principal direction and maximises the sample variance of the data projected onto this direction. It turns out that the sample variance is given by the associated eigenvalue λ_1.

The new data given by $C_X w_1 \in \mathbb{R}^n$ (*i.e.* projected onto the first principal direction) is now one dimensional, so we have shrunk our original p dimensional data to just one. We can choose the number of dimensions to reduce our data to, say r dimensions, by projecting it only onto the first r principal directions. As r increases, the error from reducing the size of our data set decreases, as more of the information from the original data set is kept.

Play around with the plots at https://setosa.io/ev/principal-component-analysis/ to see for yourself how the line of best fit is related to principal component analysis.

1) **Line of best fit:** Consider the following scatter graph of students' maths exams results against their RE exam results.
   - **a.** Describe how the line of best fit is found.
   - **b.** Draw a (estimated) line of best fit.
   - **c.** Given a student scores 50 in their maths test, what would you estimate their RE test result to be?



2) **Describe how principal component analysis works, using the line of best fit in two dimensions as a starting point.**

3) **Describe mathematically how principal directions and principal components are found, using eigenvectors and eigenvalues.**

Assignment 4 – Principal component analysis: short report and questions.

Complete all homework on a separate sheet of paper.

You do not need to write anything for questions 1 and 2 – they are here to help you understand principal component analysis.

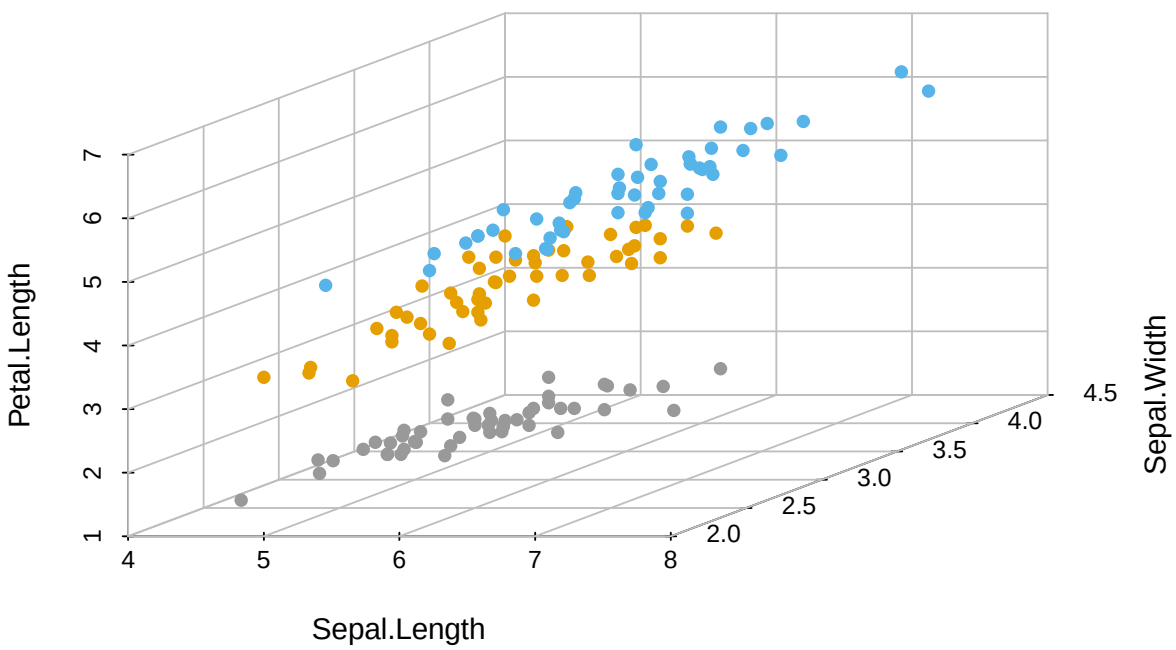
1) Modify the 2D example at https://setosa.io/ev/principal-component-analysis/. Think about how this relates to the example given about in Figure 5 and Figure 6.

2) Modify the 3D example at https://setosa.io/ev/principal-component-analysis/. Rotate the data until you understand what the first principal component shows. Understand how the three clouds have been found in the first principal component. Look at the "Eating in the UK" example. Consider how the 17 dimensions have been reduced to one dimension.

3) Describe what is meant by data being "projected onto a line" when we discussed the line of best fit above. See Figure and Figure above.
   Hint: think about moving the data perpendicular to the line.

4) Principal component analysis finds directions that maximise variance in the data. Why might this be useful for compressing data (/dimension reduction)?
   Hint: think about how the line of best fit was found in Figure 5, by minimising the perpendicular error. This also maximises the spread of the data. When this data was projected onto the line (Figure 6), the information lost was represented by this perpendicular distance.

Tutorial 5 – Real life Principal Component Analysis



What is the Purpose of Tutorial 5?

By the end of the tutorial, students will be able to:

1) Understand what 'R' is and how it can be used for data analysis.
2) Understand how principal component analysis can be applied to some simple data sets, using 'R'.
3) Describe how changing the number of principal components used in dimension reduction has an effect on the quality of a compressed image.

Material covered in tutorial 5:

In this tutorial we will use a programming language called 'R' (https://www.r-project.org/) to see how principal component analysis can be applied to some real-life data. We will also see how principal component analysis can compress images, and consider how changing the number of principal components affects the quality of the compressed imaged.

This tutorial will be split into three parts:

In **part 1,** the programming language 'R' will be introduced. This is a popular computer language (similar to Python) and free software that is used by statisticians and for data analysis.

In **part 2**, we will use 'R' to apply principal component analysis to some data. The 'R' software comes with some pre-loaded data, including the *iris* dataset. We will perform principal component analysis on this dataset, alongside others.

Part 3 involves using principal component analysis to compress images. Recall from Tutorial 4 – Principal Component Analysis that we can chose how many principal components to keep (and this affects the size of the resulting matrix), with fewer principal components meaning that more information is lost in the dimension reduction procedure. We will show this information loss by applying principal component analysis to an image, and visually comparing the different compressed images obtained when a differing number of principal components are kept.

Assignment 5 – Final Assignment

Complete all homework on a separate sheet of paper.

In your report you will include a short discussion of what high dimensional data is, and some of the issues inherent with analysing large data sets. In addition, you will describe how the dimension reduction technique, principal component analysis, can be used to find patterns in the data, including its benefits and drawbacks. A mathematical description of principal component analysis is required.

Produce a report (maximum 2000 words) as described above. Your report should include the following sections:

1. Abstract: A summary of the main points discussed on this course.

2. Review: An in-depth discussion of the main points. In this, include:
   i. An introduction to dimension reduction techniques and principal component analysis, including a comparison of the line of best fit to principal component analysis in two dimensions.
   ii. A short mathematical description of principal component analysis. This should be written in full sentences, including mathematical notation.
   iii. An explanation of the result that you might expect to get when applying principal component analysis to some data (which will be given to you). Note here that you do not need to actually apply principal component analysis, but just comment on the expected patterns that might be found. Annotated plots describing this result by showing the first two principal directions should be included.

3. Conclusion: A summary of the strengths and weaknesses of principal component analysis.

Your homework assignments and your baseline assignments should help you to form this report.

Success criteria
- Your report is coherent (each section should lead into the next).
- Your spelling, punctuation, grammar and the use of subject vocabulary is correct.
- The mathematical description of principal component analysis is correct, and written in full sentences.

Understand	Be able to
What high dimensional data is, and how it is represented in matrix form.	Use correct mathematical language.
The reasoning behind principal component analysis.	Explain what properties principal component analysis tries to obtain from the new variables.
The mathematics behind principal component analysis.	Assess how principal component analysis will work on simple toy examples.
	Assess the advantages and dis- advantages of principal component analysis in a clear manner.

Tutorial 6 – Feedback tutorial

What is the Purpose of Tutorial 6?

- To receive feedback on your final assignment
- To respond to the feedback from your Uni Pathways teacher
- To write targets for improvement on your final assignment

Final assignment feedback from your Uni Pathways Teacher

(Remember to look at the mark scheme to help you understand what you have done well so far, and how you can do even better in your final assignment)

Here are three things that my Uni Pathways Teacher thought I did well in my draft assignment
•   •   •

Here are three things that my Uni Pathways Teacher thinks that I could do to get a higher mark in my final assignment
•   •

•

Tasks from my Uni Pathways Teacher to do during the feedback tutorial to help me improve
•   •   •
My response:

Actions I will take to improve my final assignment after this tutorial…
•   •   •

Hand in date for my final assignment:	

Tutorial 7 – Final tutorial

What is the Purpose of Tutorial 7?

- To receive feedback and a grade on your final assignment.
- To reflect on the programme including what you enjoyed and what was challenging.
- To ask any questions you may have about university.

Final assignment feedback from my Uni Pathways Teacher

Final mark:	University style grade:
Feedback: Here are three things that my Uni Pathways teacher thought I did well in my final assignment • • •	

Here are three things that my Uni Pathways teacher thinks I should remember for when I am doing this kind of study in the future
• • •

University

What questions do you still have about University after taking part in Uni Pathways?
•   •   •

Reflecting on Uni Pathways

What did you most enjoy about Uni Pathways?
•   •   •

What did you find challenging about the programme?	How did you overcome these challenges?
• • •	• • •

Appendix 1 – Referencing correctly

When you get to university, you will need to include references in the assignments that you write, so we would like you to start getting into the habit of referencing in your Brilliant Club assignment. This is really important, because it will help you to avoid plagiarism. Plagiarism is when you take someone else's work or ideas and pass them off as your own. Whether plagiarism is deliberate or accidental, the consequences can be severe. In order to avoid losing marks in your final assignment, or even failing, you must be careful to reference your sources correctly.

What is a reference?

A reference is just a note in your assignment which says if you have referred to or been influenced by another source such as book, website or article. For example, if you use the internet to research a particular subject, and you want to include a specific piece of information from this website, you will need to reference it.

Why should I reference?

Referencing is important in your work for the following reasons:

- It gives credit to the authors of any sources you have referred to or been influenced by.
- It supports the arguments you make in your assignments.
- It demonstrates the variety of sources you have used.
- It helps to prevent you losing marks, or failing, due to plagiarism.

When should you use a reference?

You should use a reference when you:

- Quote directly from another source.
- Summarise or rephrase another piece of work.
- Include a specific statistic or fact from a source.

How do I reference?

There are a number of different ways of referencing, and these often vary depending on what subject you are studying. The most important to thing is to be consistent. This means that you need to stick to the same system throughout your whole assignment. Here is a basic system of referencing that you can use, which consists of the following two parts:

1. **A marker in your assignment**: After you have used a reference in your assignment (you have read something and included it in your work as a quote, or re-written it your own words) you should mark this is in your text with a number, e.g. [1]. The next time you use a reference you should use the next number, e.g. [2].
2. **Bibliography**: This is just a list of the references you have used in your assignment. In the bibliography, you list your references by the numbers you have used, and include as much information as you have about the reference. The list below gives what should be included for different sources.

    a. **Websites** – Author (if possible), title of the web page, website address, [date you accessed it, in square brackets].

    E.g. Dan Snow, 'How did so many soldiers survive the trenches?', http://www.bbc.co.uk/guides/z3kgjxs#zg2dtfr [11 July 2014].

    b. **Books** – Author, date published, title of book (in italics), pages where the information came from.

    E.g. S. Dubner and S. Levitt, (2006) *Freakonomics*, 7-9.

    c. **Articles** – Author, 'title of the article' (with quotation marks), *where the article comes from* (newspaper, journal etc.), date of the article.

    E.g. Maev Kennedy, 'The lights to go out across the UK to mark First World War's centenary', Guardian, 10 July 2014.

Notes

UNI PATHWAYS

UNI PATHWAYS

thebrilliantclub.org