Mathematics for Data Science

Revision Notes

Henry Baker

Fall 2023

Hertie School

Contents

Part I

Probability Theory

Chapter 1

Probability Theory

Mid Terms Fall 2023 – Henry Baker

M4DS Mid-term Revision
Session 1: Probability Theory

De Morgan's Law:

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

Notation:
Let A_n be the event that the nth flip is Heads

- $A_1 = \{(1, s_2, \ldots, s_{10}) : s_j \in \{0, 1\} \text{ for } 2 \leq j \leq 10\}$: **First flip = H**
- $B = \bigcup_{n=1}^{10} A_n = A_1 \cup A_2 \cup \cdots \cup A_{10}$: **Event that at least one flip was Heads**
- $C = \bigcap_{n=1}^{10} A_n = A_1 \cap A_2 \cap \cdots \cap A_{10}$: **Event that all flips were Heads**
- $D = \bigcup_{n=1}^{9} (A_n \cap A_{n+1})$: **Event that there were at least two consecutive Heads**

Naive Definition of Prob:

$$P_{\text{naive}}(A) = \frac{|A|}{|S|} = \frac{\text{number of outcomes favorable to } A}{\text{total number of outcomes in } S}$$

In general, $P_{\text{naive}}(A^c) = 1 - - - P_{\text{naive}}(A)$

Multiplication Rule:

- Think of it in terms of trees (see slide 26)
- NB: doesn't matter which order the events are structured in (counter intuitive).
- All sampling derives from multiplication rule

Combinations = when order does not matter: think of multiplication rule
permutations = when order/position matters: $n!$

With Repl
Without Re

NB: order matters, sampling w/ replacement also written as $n \cdot (n-1) \cdot (n-2) \cdot \ldots (n-k+1)$
Sampling with replacement: n^k
Sampling w/o replacement: $n * (n-1) * \ldots (n - - - k+1)$
Birthday Problem — counting complement
What's the probability of no matching birthdays?
This amounts to sampling the days of the year without replacement:

$$\begin{aligned} P(\text{no birthday match}) &= \frac{\text{number of ways to not repeat birthdays}}{\text{number of total possibilities}} \\ &= \frac{365 \times 364 \times \cdots \times (365 - - - k+1)}{365^k} \\ P(\text{birthday match}) &= 1 - - - \frac{365 \times 364 \times \cdots \times (365 - - - k+1)}{365^k} \end{aligned}$$

Leibniz' mistake: when order matters, make sure to label items:
2 dice –> 2x ways to get 11, 1x way to get 12
Binomial Coefficient:

$$\binom{n}{k}$$

= number of subsets of size k for set n.
Order does not matter.

$$\begin{aligned} \binom{n}{k} &= \frac{n(n-1)\ldots(n-k+1)}{k!} \\ &= \frac{n!}{(n-k)!k!} \end{aligned}$$

For $k > n$, we have $\binom{n}{k} = 0$.
PRACTICE QS: SLIDE 35
Non-Naive Probabilities: Probability Functions

- a probability space consisting of a *sample space* S and a *probability function* P
- which takes an event $A \subseteq S$ as input and returns $P(A)$, a real number between 0 and 1, as output.
- The function P must satisfy the following two axioms:
  - $P(\emptyset) = 0, P(S) = 1$.
  - If $A_1, A_2, \ldots$ are disjoint events, then

$$P\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} P(A_j)$$

.

Properties of Prob (cheat sheet)

- $P(A^c) = 1 - - - P(A)$
- If $A \subseteq B$, then $P(A) \leq P(B)$.
- $P(A \cup B) = P(A) + P(B) - - - P(A \cap B)$.

Chapter 2

Conditional Probability and Random Variables

Mid Terms Fall 2023 – Henry Baker

M4DS Mid Terms Revision: Session 2
Conditional Probability + Random Variables

2.1 Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

Where

- P(A) is prior
- P(B) is posterior

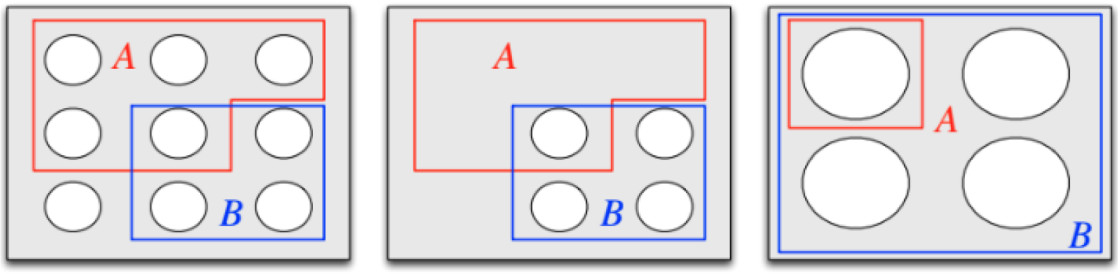


Figure 2.1: For P(A|B): we know B occurred, so get rid of outcomes in B^c; re-normalise by dividing by B

Playing card example:

- $A =$ 1st card is a heart.
- $B =$ 2nd card is red.
- There are 52 cards.
- What is $P(B|A)$?

Using the formula:

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

we can calculate:

1. $P(A) = \frac{13}{52}$
2. $P(B) = \frac{25}{51}$ (without replacement)
3. $P(B \cap A) = \frac{13 \times 25}{52 \times 51} = \frac{25}{204}$
4. $P(B|A) = \frac{\frac{25}{204}}{\frac{13}{52}} = \frac{25}{51}$

Conditional Probability Notes:

- $P(A|B) \neq P(B|A)$
- chronology unimportant: even if A occurs before B — conditional probability is about what info one event gives about the other, not whether one caused the other
- P6 THERE'S AN EXAMPLE OF THIS
- P7 PRACTICE QS

2.2 Bayes Rule (w/LOTP combo)

Bayes Rule

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

LOTP

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Expanded:

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \ldots + P(B|A_n)P(A_n)$$

p10 for lake pic
p11 FOR QS

Combining Bayes with LOTP:

$$P(A|B) = \frac{P(B|A) \times P(A)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

Bayes' Rule w/ Extra Conditioning:

$$P(A|B,E) = \frac{P(B|A,E) \times P(A|E)}{P(B|E)}$$

LOTP w/ Extra Conditioning:

$$P(B|E) = \sum_{i=1}^{n} P(B|A_i,E)P(A_i|E)$$

Both combined:

$$P(A|B,E) = \frac{P(B|A,E) \times P(A|E)}{\sum_{i=1}^{n} P(B|A_i,E)P(A_i|E)}$$

Examples on p 17

2.3 Independence of Events

$$P(A \cap B) = P(A) \cdot P(B)$$

Equivalent to

$$P(A|B) = P(A)$$

- Independence is symmetric
- different from disjoint (A gives you no info re: B), whereas disjoint means $P(A \cap B) = 0$. So, knowing A occured actually tells you that B did not occur. *(Disjoint events can only be independent if P(A) = 0 or P(B) = 0).*
- If A and B are independent, then A and B^c are independent, A^c and B are independent, and A^c and B^c are independent

Independence of 3 events:
Needs to be more than pairwise independence (conditions 1 — 3)

$$P(A \cap B) = P(A)P(B) \tag{2.1}$$
$$P(A \cap C) = P(A)P(C) \tag{2.2}$$
$$P(B \cap C) = P(B)P(C) \tag{2.3}$$
$$P(A \cap B \cap C) = P(A)P(B)P(C) \tag{2.4}$$

Example p21

Conditional Independence

$$P(A \cap B|E) = P(A|E)P(B|E)$$

- **Conditional independence given E does not imply conditional independence given E^c** (Example 2.5.9)
- **Conditional independence does not imply independence** (Example 2.5.10)
- **Independence does not imply conditional independence** (Example 2.5.11) Further intuition: see Example 2.5.12.23

For all: consider phone call example:

- Friend A calling; friend B calling are independent
- But, given event E: that one friend called, now they are conditionally dependent.
- But E_c: that a friend did not call, now means they are independent again.

SKIPPED MONTY HALL + OTHER P27 — 29

2.4 Random Variables

- **r.v.** is a function from the sample space S to the real number line $\mathbb{R}$; assigns a numerical value $X(s)$ to each possible outcome s of the experiment.
- **Support of X** is defined as all the values x such that $P(X = x) > 0$.
- **PMF** of X is the function $pX(x) = P(X = x)$. This is positive if x is in the support of X, and 0 otherwise.

Steps to construct PMF:

- Enumerate all possible outcomes of the r.v. *Example: For the coin, the possible outcomes are: HH, HT, TH, TT. So, X can take values: 0, 1, 2.*
- Calculate probabilities for each outcome. *Example: P(X=0) = P(TT) = 1/4, P(X=1) = P(HT or TH) = 2/4 = 1/2, P(X=2) = P(HH) = 1/4.*

Coin Toss Example:

- Let X be the number of heads. This assigns the values: $X(HH) = 2; X(HT) = 1; X(TH) = 1; X(TT) = 0$
- Let Y be the number of tails. This assigns the values: $Y(HH) = 0; Y(HT) = 1; Y(TH) = 1; Y(TT) = 2$. Note that $Y(s) = 2 - - - X(s)$ for all s.
- Let I be an indicator for whether the first toss is heads. Then I assigns 1 to HH and HT
- *and 0 to TH and TT.*

PMF:

- $pX(0) = P(X = 0) = \frac{1}{4}$
- $pX(1) = P(X = 1) = 1/2$
- $pX(2) = P(X = 2) = 1/4$
  and pX (x) = 0 for all other values of x.

PMFs must (1) be non negative, and (2) sum to 1.

2.5 Common Distributions & their PMFs

Bernoulli Distribution (i.e. $1\times$ trial)

- parameter p = prob of success
- written as $X \sim Bern(P)$
- $P(X = 1) = p$ and $P(X = 0) = 1 - - - p$

PMF:

$$P(X = k) = \begin{cases} 1 - - - p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases}$$

Binomial Distribution

- parameters n and p
- n independent Bernoulli trials

PMF:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

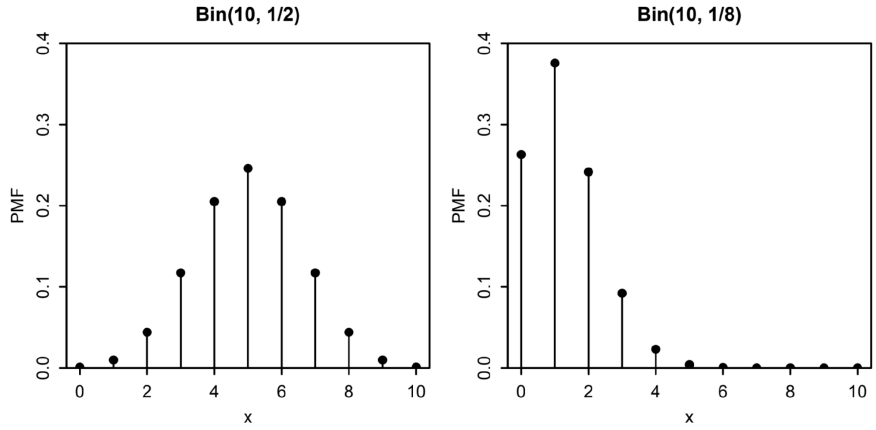


Figure 2.2: Binomial PDFs

Discrete Uniform Distributions

- parameter C
- $X \sim \text{DUnif}(C)$

PMF:

$$P(X = x) = \frac{1}{|C|}$$

$$P(X \in A) = \frac{|A|}{|C|}$$

Where $A \subseteq C$.

2.6 Common Distributions & their CDFs

CDF of X, is the function $F_X(x) = P(X \leq x)$.

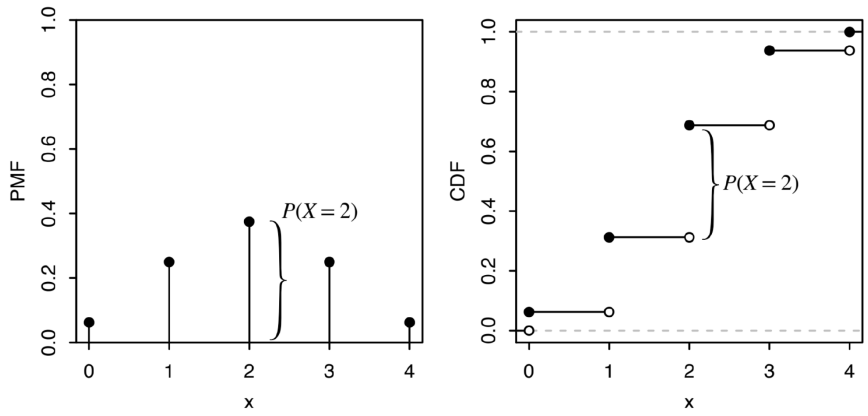


Figure 2.3: PMF, CDF of $X \sim Bin(4; 1 = 2)$

Chapter 3

Joint Random Variables

Mid Terms Fall 2023 – Henry Baker

M4DS Mid Terms Revision: Session 3
Conditional Probability & Random Variables

3.1 (Joint) Random Variables and their Distributions

(I skipped p2: random walk — I didn't understand the final PMF bit)
A function of 2 r.v.s = joint distribution?
Given an experiment with sample space S, if X and Y are r.v.s that map $s \in S$ to $X(s)$ and $Y(s)$ respectively, then $g(X, Y)$ is the r.v. that maps s to g(X(s), Y (s)).
Independence of r.v.s for joint r.v.s (???)
Continuous r.v.s

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$$

Discrete r.v.s

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

Conditional Independence

$$P(X \leq x, Y \leq y|Z = z) = P(X \leq x|Z = z)P(Y \leq y|Z = z)$$

NB:

- **Independence does not imply conditional independence**.
  - Let X be an indicator for whether my friend Bob calls me next Friday and Y be an indicator for whether my friend Alice calls next Friday, and suppose X and Y are independent
  - Let Z be an indicator for exactly one of my friends calling me next Friday
  - Then, X and Y are (perfectly) dependent given Z.
- **Conditional independence does not imply independence**. . . A lot of applied causal inference is built on finding conditional independence where there is not independence, e.g. by applying statistical controls. E.g., no selection into treatment given some conditions.

3.2 Expectation

Expectation, Expected Value, Mean
= weighted avg of possible values X can take:

$$E(X) = \sum_x x \cdot \underbrace{P(X = x)}_{\text{PMF at } x}$$

Eg Let X be the result of rolling a 6-sided die:

$$E(X) = (1 + 2 + 3 + 4 + 5 + 6) \cdot \frac{1}{6} = 3.5$$

NB X never equals its mean here.
Eg let X be the result of 2x coin flip (heads)

$$E(X) = 0 \times \frac{1}{4} + 1 \times \frac{1}{2} + 2 \times \frac{1}{4} = 1$$

In Bernoulli: $X \sim Bern(p)$:

$$E(X) = 1p + 0(1 - p) = p$$

This is the value (1, 0) multiplied by its probability of occurring.
Linearity of Expectation:

1. $E(cX) = cE(x)$
2. $E(X + Y) = E(X) + E(Y)$

Even if X and Y are not independent
BUT this only works when functions are linear.... p.12

Returning to coin flip example:
1) $E(cX) = cE(x)$
For a constant $c = 3$, let $Y = 3X$. Thus, the values of Y are 0, 3, and 6.

$$E(3X) = 0 \times \frac{1}{4} + 3 \times \frac{1}{2} + 6 \times \frac{1}{4} = 3$$

According to the property:

$$E(3X) = 3E(X) = 3 \times 1 = 3$$

2) $E(X + Y) = E(X) + E(Y)$
Property: The expected value of the sum of two random variables is the sum of their expected values.
Example: Let X be the number of heads when flipping a fair coin twice and Z represents the outcome when rolling a fair die.

$$E(X) = 1$$
$$E(Z) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5$$

For $W = X + Z$:

$$E(X + Z) = E(X) + E(Z) = 1 + 3.5 = 4.5$$

3.3 Variance

$$Var(X) = E((X - - - EX)^2)$$
$$S.d = \sqrt{var}$$

This means: it is the actual value (x), minus mean(μ), squared, then multiplied by probability of that value, then all summed.
Example: What is the variance of X, if X is the result of one roll of a fair six-sided die?
E(X) = 3.5.

$$\frac{1}{6}\left[(1 - - - 3.5)^2 + (2 - - - 3.5)^2 + \cdots + (6 - - - 3.5)^2\right] \approx 2.9$$

More useful:

$$Var(X) = E(X^2) - - - (EX)^2$$

Example: six-sided die. Let X be the result of one roll.

$$E(X) = \frac{1}{6}\sum_{i=1}^{6} i = \frac{1+2+3+4+5+6}{6} = 3.5 \tag{3.1}$$

$$E(X^2) = \frac{1}{6}\sum_{i=1}^{6} i^2 = \frac{1^2+2^2+3^2+4^2+5^2+6^2}{6} = \frac{91}{6} \tag{3.2}$$

$$\text{Var}(X) = E(X^2) - - - (E(X))^2 \tag{3.3}$$

$$\text{Var}(X) = \frac{91}{6} - - - (3.5)^2 = \frac{91}{6} - - - 12.25 = 2.92 \tag{3.4}$$

Variance facts:

- $Var(c) = 0$ for any constant c
- $Var(X + c) = Var(X)$ for any constant c
- $Var(cX) = c^2 Var(X)$ for any constant c <— **NB**
- $Var(X + Y) = Var(X) + Var(Y)$ only if X and Y are independent.
  Caution: unlike expectation, variance is not linear
  - $Var(cX) \neq cVar(X)$
  - $Var(X + Y) \neq Var(X) + Var(Y)$ (in general)

3.4 Joint Distributions — how 2 r.v.s interact

3.4.1 PMF

Still has to sum to 1 over all values of X and Y.

$$p_{X,Y}(x, y) = P(X = x, Y = y)$$

3.4.2 CDF

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y)$$

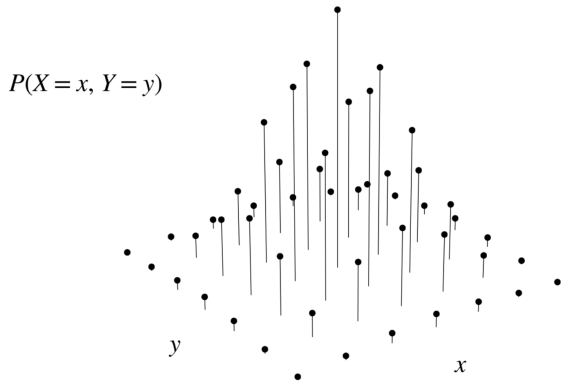


Figure 3.1: Joint PMF

3.4.3 Marginal PMF of X

Sum over all y:

$$P(X = x) = \sum_{y} P(X = x, Y = y)$$

$\rightarrow$ the PMF of X

We've marginalised out Y — by summing over all values of Y.

Heuristic: no longer interested in Y; it's sort of flattened the square base into a line.

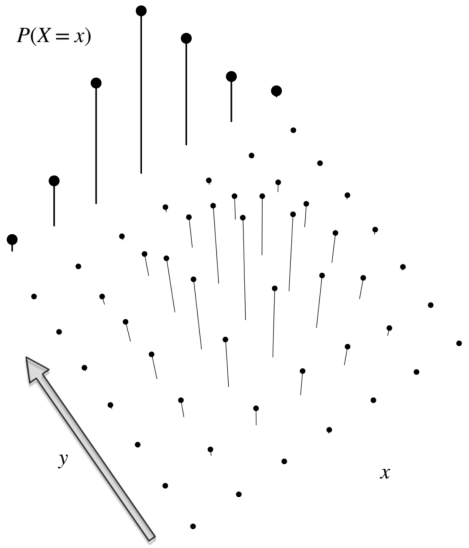


Figure 3.2: Marginal PMF of X

3.4.4 Conditional PMF of Y

Joint divided by marginal.

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

= the joint PMF over the PMF of 2nd variable at a certain value of (ie the marginal)

You need to renormalise based on PMF of $X = x$ (ie a certain value of x)

Heuristic: taking a cross section of the joint based on a value of X.

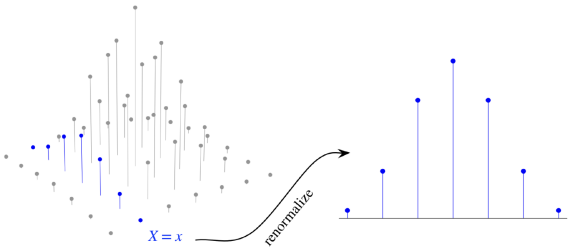


Figure 3.3: Conditional PMF of Y

Joint PMF > Marginal PMF of X > Conditional PMF of Y

Example 1:
randomly sample from population:

- Let X be an indicator for the presence of a gene
- Y be an indicator for developing a certain disease at some point in his life.

			$Y =$
$X = 1$			$\frac{5}{100}$
$X = 0$			$\frac{3}{100}$

Table 3.1: Contingency table for X and Y

What are joint PMF of X, Y?, marginal prob of X, Y? conditional prob of Y given X = 1?

- joint PMF: given by table itself
- marginal: we work out onto the margins

	$Y = 1$	$Y = 0$	Marginal X
$X = 1$	$\frac{5}{100}$	$\frac{20}{100}$	$\frac{25}{100}$
$X = 0$	$\frac{3}{100}$	$\frac{72}{100}$	$\frac{75}{100}$
Marginal Y	$\frac{8}{100}$	$\frac{92}{100}$	1

Table 3.2: Contingency table for X and Y with Marginal Distributions

- conditional distribution of Y|X = 1:
  $P(Y=1|X=1) = \frac{P(X=x,Y=y)}{P(X=x)} = \frac{\frac{5}{100}}{\frac{25}{100}} = 0.2$
  So conditional distribution of Y given $X = 1$ is Bernoulli(0.2)

NB we can work out if they are independent r.v.s.

$$\text{If independent: } P(X=x, Y=y) = P(X=x)P(Y=y)$$

Finding even one pair, such that $P(X=x, Y=y) \neq P(X=x)P(Y=y)$ is sufficient to rule out independence.

To determine if the variables X and Y are independent, we need to check if:

$$P(X=x, Y=y) = P(X=x) \times P(Y=y)$$

for all values of x and y.
Checking marginals against individual values in the joint PMF.
Using the values from the table:

1. For $X=1$ and $Y=1$:
$$\frac{5}{100} \stackrel{?}{=} \frac{25}{100} \times \frac{8}{100}$$
2. For $X=1$ and $Y=0$:
$$\frac{20}{100} \stackrel{?}{=} \frac{25}{100} \times \frac{92}{100}$$
3. For $X=0$ and $Y=1$:
$$\frac{3}{100} \stackrel{?}{=} \frac{75}{100} \times \frac{8}{100}$$
4. For $X=0$ and $Y=0$:
$$\frac{72}{100} \stackrel{?}{=} \frac{75}{100} \times \frac{92}{100}$$

Given that none of the joint probabilities match the product of the marginal probabilities, we can conclude that X and Y are **not independent**.

FROM LAB NEED TO ADD THE MEAN AND VARIANCE OF THE BINOMIAL / THE MULTINOMIAL / POISSON

ChatGPT's Joint > Marginal > Conditional example:

Joint Distribution to Marginal and Conditional PMFs

Consider two discrete random variables X and Y with the following joint distribution:

$P(X,Y)$	$Y=1$
$X=1$	0.1
$X=2$	0.05
$X=3$	0.05

Marginal PMFs

Marginal PMF of X:

$$P(X=1) = \sum_y P(X=1, Y=y) = 0.1 + 0.2 + 0.1 = 0.4$$
$$P(X=2) = \sum_y P(X=2, Y=y) = 0.05 + 0.25 + 0.1 = 0.4$$
$$P(X=3) = \sum_y P(X=3, Y=y) = 0.05 + 0.1 + 0.05 = 0.2$$

Marginal PMF of Y:

$$P(Y=1) = \sum_x P(X=x, Y=1) = 0.1 + 0.05 + 0.05 = 0.2$$
$$P(Y=2) = \sum_x P(X=x, Y=2) = 0.2 + 0.25 + 0.1 = 0.55$$
$$P(Y=3) = \sum_x P(X=x, Y=3) = 0.1 + 0.1 + 0.05 = 0.25$$

Conditional PMFs

Conditional PMF of X given $Y=1$:

$$P(X=1|Y=1) = \frac{0.1}{0.2} = 0.5$$
$$P(X=2|Y=1) = \frac{0.05}{0.2} = 0.25$$
$$P(X=3|Y=1) = \frac{0.05}{0.2} = 0.25$$

Conditional PMF of X given $Y=2$:

$$P(X=1|Y=2) = \frac{0.2}{0.55} \approx 0.3636$$
$$P(X=2|Y=2) = \frac{0.25}{0.55} \approx 0.4545$$
$$P(X=3|Y=2) = \frac{0.1}{0.55} \approx 0.1818$$

Conditional PMF of X given $Y=3$:

$$P(X=1|Y=3) = \frac{0.1}{0.25} = 0.4$$
$$P(X=2|Y=3) = \frac{0.1}{0.25} = 0.4$$
$$P(X=3|Y=3) = \frac{0.05}{0.25} = 0.2$$

Conditional PMF of Y given $X=1$:

$$P(Y=1|X=1) = \frac{0.1}{0.4} = 0.25$$
$$P(Y=2|X=1) = \frac{0.2}{0.4} = 0.5$$
$$P(Y=3|X=1) = \frac{0.1}{0.4} = 0.25$$

Conditional PMF of Y given $X = 2$:

$$P(Y = 1|X = 2) = \frac{0.05}{0.4} = 0.125$$
$$P(Y = 2|X = 2) = \frac{0.25}{0.4} = 0.625$$
$$P(Y = 3|X = 2) = \frac{0.1}{0.4} = 0.25$$

Conditional PMF of Y given $X = 3$:

$$P(Y = 1|X = 3) = \frac{0.05}{0.2} = 0.25$$
$$P(Y = 2|X = 3) = \frac{0.1}{0.2} = 0.5$$
$$P(Y = 3|X = 3) = \frac{0.05}{0.2} = 0.25$$

Part II

Calculus

Chapter 4

Calculus I

Mid Terms Fall 2023 – Henry Baker

M4DS Mid Terms Revision: Session 4
Calculus I

4.1 Differentiation Rules

- **Rule 0: Constants**: derivative = 0 –> can take out of the any differentiation
- **Rule 1: Powers**: $\frac{d}{dx}x^n = nx^{n-1}$
- **Rule 2: Sum/Differences**:

$$\frac{d}{dx}(\mathrm{f}(x) \pm g(x)) = \frac{d}{dx}\mathrm{f}(x) \pm \frac{d}{dx}g(x)$$

- **Rule 3: Constant Multiples**

$$\frac{d}{dx}[kf(x)] = k\frac{d}{dx}f(x)$$

- **Rule 4: Products**

$$\frac{d}{dx}[g(x)f(x)] = g'(x) \cdot f(x) + g(x) \cdot f'(x)$$

- **Rule 5: Quotients**

$$\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) = \frac{g(x) \cdot f'(x) - - - f(x) \cdot g'(x)}{g(x)^2}$$

- **Rule 6: Chain**
  If y is a function of u, and u is a function of x, then:

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$

  Steps

  1. Identify inner + outer

2. differentiate both
3. multiply derivatives together

Example:

$$\frac{d}{dx}f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$$

1) strip out the constant:

$$f(x) = \frac{1}{\sqrt{2\pi}} \cdot \frac{d}{dx}e^{-\frac{1}{2}x^2}$$

2) break it into 2:

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx} \tag{4.1}$$
$$u = -\frac{1}{2}x^2 \tag{4.2}$$
$$y = e^u \tag{4.3}$$
$$\frac{dy}{du} = e^u \tag{4.4}$$
$$\frac{dy}{dx} = -x \tag{4.5}$$
$$\frac{dy}{dx} = e^u \cdot (-x) \tag{4.6}$$
$$\frac{dy}{dx} = -xe^{-\frac{1}{2}x^2} \tag{4.7}$$
$$\tag{4.8}$$

- **Rule 7: Natural Exponential**

$$y(x) = e^x \rightarrow \frac{dy}{dx} = e^x$$

- **Rule 8: Natural Logarithms**

$$y = \ln(x) \rightarrow \frac{dy}{dx} = \frac{1}{x}$$

- **Exponential Functions**
  Power rule is for when x is a **constant**.
  Exponential functions take x as the exponent itself.

$$\frac{d}{dx}a^x = \ln(a) \times a^x$$

$$\frac{d}{dx}a^{bx} = b \times \ln(a) \times a^{bx}$$

Example: a^x
10^x becomes $ln(10)$ $\times 10^x$:

$$f(x) = \frac{10^x}{\ln(10)}$$
$$f'(x) = \ln(10) \times 10^x \times \frac{1}{\ln(10)} = 10^x$$

Example: a^{bx}

$$f(x) = 2^{4x} + 4x^2$$
$$f'(x) = 4\ln(2) \times 2^{4x} + 8x$$

However, when differentiating a^u where u is a function of x not just a simple constant multiplier like in the previous rule), we have to use the chain rule:

$$\frac{d}{dx}a^u = u' \times \ln(a) \times a^u$$

4.2 Power Series

$$\sum_{k=0}^{\infty} c_k x^k = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + \dots$$

Where x is coefficient, c_k are some constants

$$\sum_{k=0}^{\infty} c_k (x - - - a)^k = c_0 + c_1(x - - - a) + c_2(x - - - a)^2 + \dots$$

Where a is some constant → "centered at $x = a$." Can diverge / converge — for the following:
if $x = 0.5$ → converges
if $x = 2$ → diverges

$$\sum_{k=0}^{\infty} x^k = 1 + x + x^2 + x^3 + x^4 \dots$$

Intuition: tricky functions can be represented as infinite polynomials (when they converge to a approximate a function). Infinite polynomials are easy to differentiate, integrate, manipulate.
We can even truncate inf polynomial for some k for a good enough approx
DIDN'T UNDERSTAND SOME OF THIS — P31, 32, 34

4.3 Maximum Likelihood Estimation (MLE)

- We have some data.
- We assume the events in the data are i.i.d.
- Following an {insert distribution}, with {insert parameters}, but some unknown success probability p.
- Aim: find value of p that maximises the probability that this particular dataset is observed.
- We can't observe the true p, but every possible value that p could take corresponds to some joint probability of observing the data that we see → find value of p that makes data most likely.

4.3.1 Theoretical Overview

Given a set of independent and identically distributed (i.i.d) data points $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, the likelihood function, denoted by $\mathcal{L}(\theta; \boldsymbol{X})$, represents the joint probability of observing the given data as a function of the parameter(s) θ. The aim is to find the value of θ that maximizes this likelihood function.

4.3.2 Steps for Maximum Likelihood Estimation

1. **Specify the Statistical Model:** Define the probability distribution that models your data, parameterized by θ.
2. **Construct the Likelihood Function:** Write down the likelihood function $\mathcal{L}(\theta; \boldsymbol{X})$. For i.i.d data, this is often the product of the individual probabilities or probability densities.
3. **Take the Logarithm:** To simplify calculations, take the natural logarithm of the likelihood function. This is called the log-likelihood, and is denoted as $\ell(\theta) = \log \mathcal{L}(\theta; \boldsymbol{X})$. Maximizing the log-likelihood is equivalent to maximizing the likelihood because the natural logarithm is a monotonically increasing function.
4. **Differentiate and Set to Zero:** Find the derivative of the log-likelihood with respect to the parameter(s) θ. Set the derivative (or gradient, for multiple parameters) to zero to find the critical points.
5. **Solve for θ:** Solve the equation from the previous step to get the MLE of θ, often denoted as $\hat{\theta}_{MLE}$.
6. **Check for Maximum:** Ensure that the critical point obtained is a maximum by checking the second derivative or the Hessian matrix (for multiple parameters).

4.3.3 Worked Example: Direct Approach (Without Logging)

We will use goalie example:

Express likelihood function

Specify the model Let X_i (number of saves in game i) be i.i.d., with distribution

$$X_i \sim Binomial\binom{5}{\theta}$$

Here θ represents unknown value of p.
Given a statistical model with parameter(s) θ, and some observed data X, the likelihood function $L(\theta; X)$ quantifies how likely the observed data is for different values of θ:

$$L(\theta; X) = f(X; \theta)$$

Where $f(X; \theta)$ is the PMF/PDF of the oberved data, given the parameters θ.

$$L(x_1, x_2, \ldots, x_n; \theta) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n; \theta)$$

Write the PMF of X_i

1. Express the **generic PMF** given the distribution
2. Express the **probabilities for each X_i**
3. Express the **joint probabilities** of all events X_i.

> Our dataset:
> 1st game: 1 save / 2nd game: 3 saves / 3rd game 2: saves / 4th game: 2 saves)

Generic PMF Binomial PMF for 5 trials is given by:

$$P(X_i = k) = \binom{5}{k} p^k (1 - - - p)^{5-k}$$

Probabilities of each X_i

$$P(X_1 = 1) = \binom{5}{1} \theta^1 (1 - - - \theta)^{5-1}$$
$$P(X_2 = 3) = \binom{5}{3} \theta^3 (1 - - - \theta)^{5-3}$$
$$P(X_3 = 2) = P(X_4 = 2) = \binom{5}{2} \theta^2 (1 - - - \theta)^{5-2}$$

Joint probabilities of all X_i Since events are i.i.d, the joint probability of observing them all is the product of individual probabilities.

$$\begin{aligned} P(X_1 = 1, X_2 = 3, X_3 = 2, X_4 = 2) &= \binom{5}{1} \theta^1 (1 - - - \theta)^{5-1} \times \binom{5}{3} \theta^3 (1 - - - \theta)^{5-3} \times \left(\binom{5}{2} \theta^2 (1 - - - \theta)^{5-2} \right) \\ &= \binom{5}{1} \times \binom{5}{3} \times \binom{5}{2} \times \binom{5}{2} \theta^8 (1 - - - \theta)^{12} \\ &= 5000 \theta^8 (1 - - - \theta)^{12} \end{aligned}$$

Maximise θ

Maximising $5000 \cdot \theta^8 (1 - \theta)^{12}$ tells us exactly which value of θ makes the joint probability of observing our data the largest.

$$P(X_1 = 1, X_2 = 3, X_3 = 2, X_4 = 2) = L(1, 3, 2, 2; \theta) = 5000 \cdot \theta^8 (1 - \theta)^{12}$$

> Maximisation:
>
> 1. take first derivative with respect to θ,
> 2. set equal to 0.
> 3. solve for θ

$$\begin{aligned}\frac{d}{d\theta}L(1,3,2,2;\theta) &= \frac{d}{d\theta}(5000\cdot\theta^8(1-\theta)^{12})\\ &= 5000((8\theta^7)(1---\theta)^{12}+(\theta^8)(12(1---\theta)^{11})(-1))\\ &= 5000((8\theta^7)(1---\theta)^{12}---(\theta^8)(12(1---\theta)^{11})\end{aligned}$$

Set that equal to 0.

When we do so, θ becomes $\hat{\theta}$, the solution to the maximization problem.

$$\begin{aligned}0 &= 5000((8\hat{\theta}^7)(1---\hat{\theta})^{12}---(\hat{\theta}^8)(12(1---\hat{\theta})^{11}))\\ 0 &= 8\hat{\theta}^7(1---\hat{\theta})^{12}---\hat{\theta}^8(12(1---\hat{\theta})^{11})\\ \frac{12\hat{\theta}^8}{8\hat{\theta}^7} &= 1---\hat{\theta}\\ \frac{12\hat{\theta}}{8} &= 1---\hat{\theta}\\ 20\hat{\theta} &= 8\\ \hat{\theta} &= \frac{8}{20}\end{aligned}$$

4.3.4 Log-Likelihood Approach

Better behaved function; easier to maximise.

Log rules:

$$\begin{aligned}\log(ab) &= \log(a)+\log(b)\\ \log\left(\frac{a}{b}\right) &= \log(a)---\log(b)\\ \log\left(\prod_{j=1}^{n}x_j\right) &= \sum_{j=1}^{n}\log(x_j)\\ \log(a^b) &= b\log a\end{aligned}$$

Take the log

Generic likelihood given n events with m trials

$$L(x_1,\ldots,x_n;\theta)=\prod_{j=1}^{n}\binom{m}{x_j}\cdot\theta^{x_j}\cdot(1---\theta)^{m-x_j}$$

Log-likelihood

$$\ell(x_1,\ldots,x_n;\theta)=\sum_{j=1}^{n}\left[\log\binom{m}{x_j}+x_j\log\theta+(m---x_j)\log(1---\theta)\right]$$

Derive the log Term 1: constant drops out
Term 2: $log\theta$ becomes $\frac{1}{\theta}\rightarrow\frac{x_j}{\theta}$
Term 3: For $log(1-\theta)$ we apply the chain rule:

- inner: -1
- outer: $\frac{1}{1-\theta}$

- inner × outer: $-\frac{1}{1-\theta}$

$$\frac{d\ell}{d\theta} = \sum_{j=1}^{n}\left(\frac{x_j}{\theta}\right) - - - \sum_{j=1}^{n}\left(\frac{m - - - x_j}{1 - - - \theta}\right)$$

Set derivative to 0, solve for θ THIS BELOW IS NOT RIGHT, LOOK AT SLIDES

$$0 = \sum_{j=1}^{n} x_j(\frac{1}{\hat{\theta}} - - - \sum_{j=1}^{n}(m - x_j)(\frac{1}{1-\hat{\theta}}$$

$$\sum_{j=1}^{n}\frac{5 - - - x_j}{1 - - - \hat{\theta}} = \sum_{j=1}^{n}\frac{x_j}{\hat{\theta}}$$

$$\hat{\theta}\sum_{j=1}^{n}(5 - - - x_j) = (1 - - - \hat{\theta})\sum_{j=1}^{n} x_j$$

$$5n\hat{\theta} - - - \hat{\theta}\sum_{j=1}^{n} x_j = \sum_{j=1}^{n} x_j - - - \hat{\theta}\sum_{j=1}^{n} x_j$$

$$5n\hat{\theta} = \sum_{j=1}^{n} x_j$$

$$\hat{\theta} = \frac{\sum_{j=1}^{n} x_j}{5n}$$

$$0 = \sum_{j=1}^{n} x_j\left(\frac{1}{\hat{\theta}}\right) - \sum_{j=1}^{n}(5 - x_j)\left(\frac{1}{1-\hat{\theta}}\right)$$

$$\sum_{j=1}^{n}(5 - x_j)\left(\frac{1}{1-\hat{\theta}}\right) = \sum_{j=1}^{n} x_j\left(\frac{1}{\hat{\theta}}\right)$$

$$\hat{\theta}\sum_{j=1}^{n}(5 - x_j) = (1-\hat{\theta})\sum_{j=1}^{n} x_j$$

$$5n\hat{\theta} - \hat{\theta}\sum_{j=1}^{n} x_j = \sum_{j=1}^{n} x_j - \hat{\theta}\sum_{j=1}^{n} x_j$$

$$5n\hat{\theta} = \sum_{j=1}^{n} x_j$$

$$\hat{\theta} = \frac{\sum_{j=1}^{n} x_j}{5n}$$

Figure 4.1: NEED TO INTEGRATE

4.3.5 Practice Exercises

1. Given a sample from a Poisson distribution with parameter λ, find the MLE of λ.
2. Given a sample from a binomial distribution with parameters n and p, and known n, find the MLE of p.
3. Given a sample from a uniform distribution on the interval $[0, \theta]$, find the MLE of θ.

Note: Solutions to the practice exercises can be found in [reference textbook or solution manual].

4.3.6 Summary: What is MLE?

Maximum Likelihood Estimation (MLE) is a statistical method used for estimating the parameters of a statistical model. It is a fundamental concept in probability theory and statistics, widely used in various fields, including machine learning and optimization. The core idea of MLE is to find the parameter values that maximize the likelihood function, which measures how well the model with those parameters explains the observed data.
Here's a more detailed breakdown of the MLE concept:
Statistical Model: A statistical model is a mathematical representation of a real-world process, described by a set of parameters. For example, in a normal distribution, the mean and variance are its parameters.
Likelihood Function: The likelihood function is a function of the model parameters given the observed data. It represents the probability of observing the given data under different parameter values of the model. The likelihood is different from probability as it treats the observed data as fixed and the parameters as variables.
Maximizing the Likelihood: MLE seeks to find the parameter values that maximize the likelihood function. These values are called the maximum likelihood estimates. The idea is that the best model parameters are those under which the observed data is most probable.
Log-Likelihood: In practice, it's common to work with the logarithm of the likelihood function, known as the log-likelihood. This transformation simplifies the calculations, especially for products, as it turns them into sums. The maximization principle remains the same since logarithm is a monotonic function.
Finding the Maximum: To find the maximum of the likelihood or log-likelihood, we often use calculus, setting the derivative with respect to the parameters to zero and solving for the parameters. In complex models, this process may require numerical methods.
Applications: MLE is used in various contexts like regression analysis, time series analysis, and machine learning models. It provides a way of fitting a model to data and is foundational in the field of inferential statistics.
For example, in the case of a normal distribution, the MLE for the mean and variance can be calculated using the observed data, and these estimates will be the sample mean and sample variance, respectively.

Chapter 5

Power Series

Mid Terms Fall 2023 – Henry Baker

M4DS Mid Terms Revision: Session 5
Power Series

5.1 Taylor Series Approximation

We have an ugly function: difficult to differentiate, integrate, manipulate, maximise.
Assume: it has a power series representation at some point $x = a$, ie:

$$p(x) = f(x) \text{ at } x = a$$

Recall, polynomial function p(x) can be written as:

$$p(x) = \sum_{j=0}^{\infty} c_j (x - - - a)^j$$

Expanded:

$$p(x) = c_0 + c_1 x + c_2 x^2 + \dots$$

So, setting it to $x = 0$:
if we want $p(x)$ to represent $f(x)$ at $x = 0$, what conditions would we like to hold?

$$p(0) = f(0) \tag{5.1}$$
$$p'(0) = f'(0) \tag{5.2}$$
$$p''(0) = f''(0) \tag{5.3}$$
$$p'''(0) = f'''(0) \tag{5.4}$$
$$\dots \tag{5.5}$$

We can solve for the values of c in the Talor series that make these conditions hold: *the basic intuition is that as we continue to differentate the power series n times, the x in the nth term drop out leaving an associated constant; when we set $x = 0$ all the other terms drop out, this gives us the constant value for the nth term in the power series*

5.1.1 First Condition $p(0) = f(0)$

$$p(x) = c_0 + c_1 x + c_2 x^2 + \dots$$

Setting to 0, $p(0) \to$ all x terms drop out $\to = \dots c_0$. the first Constant term
So, first condition:

$$p(0) = f(0) \tag{5.6}$$
$$p(0) = c_0 \tag{5.7}$$
$$c_0 = f(0) \tag{5.8}$$
$$\tag{5.9}$$

5.1.2 2nd Condition $p'(0) = f'(0)$

$$p'(x) = c_1 + 2c_2 x + 3c_3 x^2 + 4c_4 x^3 \dots$$

Setting to 0: $p'(0) \to$ all x terms drop out $\to = \dots c_1$. the first Constant term
So, second condition:

$$p'(0) = f'(0) \tag{5.10}$$
$$p'(0) = c_1 \tag{5.11}$$
$$c_1 = f'(0) \tag{5.12}$$
$$\tag{5.13}$$

5.1.3 Third Condition: $p''(0) = f''(0)$

$$p''(x) = 2c_2 + 6c_3 x + 12c_4 x^2 \dots$$

Setting to 0: $p''(0) \to$ all x terms drop out $\to = \dots 2c_2$. the first Constant term
So, third condition:

$$p''(0) = f''(0) \tag{5.14}$$
$$p''(0) = 2c_2 \tag{5.15}$$
$$2c_2 = f''(0) \tag{5.16}$$
$$c_2 = \frac{1}{2} f''(0) \tag{5.17}$$

5.1.4 Fourth Condition: $p'''(0) = f'''(0)$

$$p'''(x) = 6c_3 + 24c_4 x + \dots$$

Setting to 0: $p'''(0) \to$ all x terms drop out $\to = \dots 6c_3$. the first Constant term
So, fourth condition:

$$p'''(0) = f'''(0) \tag{5.18}$$
$$p'''(0) = 6c_3 \tag{5.19}$$
$$6c_3 = f'''(0) \tag{5.20}$$
$$c_c = \frac{1}{6} f'''(0) \tag{5.21}$$

5.1.5 Putting it Together

We have following conditions:

1. $c_0 = f(0)$
2. $c_1 = f'(0)$
3. $c_2 = \frac{1}{2}f''(0)$
4. $c_3 = \frac{1}{6}f'''(0)$

to represent $f(x)$ at $x = 0$, our polynomial:

$$p(x) = \underbrace{f(0)}_{c_0} + \underbrace{f'(0)}_{c_1}x + \frac{1}{2}\underbrace{f''(0)}_{c_2}x^2 + \frac{1}{6}\underbrace{f'''(0)}_{c_3}x^3 + \dots$$

If you kept going, a pattern emerges:

$$p(x) = \underbrace{f(0)}_{c_0} + \underbrace{f'(0)}_{c_1}x + \frac{1}{2!}\underbrace{f''(0)}_{c_2}x^2 + \frac{1}{3!}\underbrace{f'''(0)}_{c_3}x^3 + \dots$$

Generalising to the form $x = 0$, to $x = a$

$$p(x) = \underbrace{f(a)}_{c_0} + \underbrace{f'(a)}_{c_1}(x-a) + \frac{1}{2!}\underbrace{f''(a)}_{c_2}(x-a)^2 + \frac{1}{3!}\underbrace{f'''(a)}_{c_3}(x-a)^3 + \dots$$

This is the Taylor series for the function $f(x)$ at a. In the special case when $a = 0$, we also call this the Maclaurin series.
From this, we can write the nth *partial sum* of the Maclaurin series as:

$$\begin{aligned}
p_0(x) &= f(0) \\
p_1(x) &= f(0) + f'(0)x \\
p_2(x) &= f(0) + f'(0)x + \frac{1}{2}f''(0)x^2 \\
p_3(x) &= f(0) + f'(0)x + \frac{1}{2}f''(0)x^2 + \frac{1}{6}f'''(0)x^3
\end{aligned}$$

Each partial sum is an approximation to the function $f(x)$, which gets better and better as you add more terms.

5.1.6 Example:

Find the Taylor polynomials p_0, p_1, p_2, and p_3 for the function $f(x) = \log(x)$ at $x = 1$.
We will require the derivatives of $f(x)$, and we should also evaluate them at $x = 1$:

$$\begin{aligned}
f(x) &= \log(x) && \to \log(1) = 0 \\
f'(x) &= \frac{1}{x} && \to f'(1) = 1 \\
f''(x) &= -\frac{1}{x^2} && \to f''(1) = -1 \\
f'''(x) &= \frac{2}{x^3} && \to f'''(1) = 2
\end{aligned}$$

$$p(x) = \underbrace{f(a)}_{c_0} + \underbrace{f'(a)}_{c_1}(x-a) + \frac{1}{2!}\underbrace{f''(a)}_{c_2}(x-a)^2 + \frac{1}{3!}\underbrace{f'''(a)}_{c_3}(x-a)^3 + \dots$$

$$\begin{aligned}
p_0(x) &= f(1) = 0 \\
p_1(x) &= f(1) + f'(1)(x---1) \\
&= 0 + 1(x---1) \\
&= x---1 \\
p_2(x) &= x---1 + \frac{1}{2}f''(1)(x---1)^2 \\
&= x---1---\frac{1}{2}(x---1)^2 \\
p_3(x) &= x---1---\frac{1}{2}(x---1)^2 + \frac{1}{3!}(x---1)^3 \\
&=
\end{aligned}$$

5.2 Integration

Chapter 6

Continuous Random Variables I

Finals Fall 2023 – Henry Baker

M4DS Finals Revision: Session 6
Calculus Meets Probability / Continuous Random Variables I

6.1 Continuous r.v.s: relationship between PDF-CDF

6.1.1 an r.v. has a continuous distribution if its CDF is differentiable

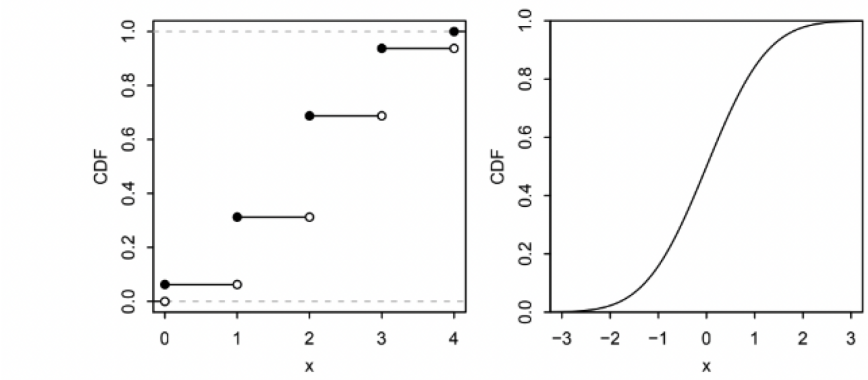


Figure 6.1: Enter Caption

Left: CDF of a discrete r.v.
Right: CDF of a continuous r.v.

6.1.2 PDF of X is the Derivative of the CDF

Remember: PDF is a function that describes the relative likelihood for this random variable to take on a given value.
For a continuous r.v. X with CDF F, the PDF of X is the derivative of f of the CDF, given

$$f(x) = F'(x)$$

The support of X is the set of x where $f(x) > 0$
The height is NOT a probability (it is mostly meaningless; simply standardised so that integrates

(ie the AUC) is equal to 1. Instead, the area (integral) is the probability.

NB: with CDFs: you can always just take the integral from minus infinite to infinite, this is the same as just giving the integral for the common support of x

An important way in which continuous r.v.s differ from discrete r.v.s is that for continuous r.v.s, $P(X = x) = 0$ for all x.

- Why? Continuous r.v.s can take on infinite values
- So, we do NOT interpret PDF of X as $P(X = x)$
- Instead, PDF gives density function from which probabilities can be obtained for intervals of values. The value of the PDF at any given point can be interpreted as a density, not a probability.
- However, CDF remains interpretable as $P(X \leq x)$

6.1.3 Probability of a Continuous Random Variable

Whereas for discrete variables' PDFs we were interested in probabilities associated with values X could take...
... for continuouse variable PDFs we are interested in the probability of X falling in some interval (a, b) (or $[a, b]$ or $(a, b]$, or $[a, b)$:

$$P(a < X \leq b) = F(b) - - - F(a) = \int_a^b f(x)\,dx$$

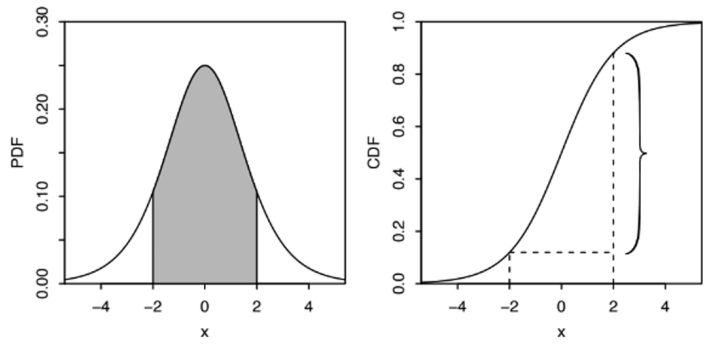


FIGURE 5.2
Logistic PDF and CDF. The probability $P(-2 < X < 2)$ is indicated by the shaded area under the PDF and the height of the curly brace on the CDF.

Figure 6.2: Enter Caption

The probability that a continuous random variable falls within a particular interval is given by the area under the PDF curve over that interval. This is why the total area under the PDF curve across all possible values of the variable is always equal to 1.
In areas where the PDF curve is higher, the continuous random variable is ***more likely*** *to fall within these values.*

6.1.4 Valid PDFs

A valid PDF must satisfy 2 conditions:

1. Non-negative: $f(x) \geq 0$
2. integrates to 1: $\int_{-\infty}^{\infty} f(x)\,dx = 1$

6.2 Expectation of a Continuous r.v

Mean / expected value:

$$E(X) = \int_{-\infty}^{\infty} x f(x)\,dx$$

Broken down:

- x = possible value of r.v. X
- $f(x)$ = PDF of X — ie the relative likelihood for this continuous r.v. to occur at each value of x
- $\int_{-\infty}^{\infty}$ = integral over all possible values of X. We use integral because for continuous variables, probabilities are integrals of the PDF.
- $xf(x)dx$ = weighted value of the r.v. ($f(x$ acts as weight indicating how much each value of x contributes to the avg.)

... so expected value E(X) interpreted as weighted avg, where each possible value of X weighted according to its prob density.
... gives a single number that represents "center of mass" or "balance point" of the distribution X.
for continuous PDFs, think of y-dimension as density → with the mean: trying to balance this object: the mean is at the balancing point of the mass, which is the integral of the density across all common support (ie minus infinite to infinity)

In simpler terms: "To find the expected value of a continuous random variable, take every possible value that variable can have, multiply each by the probability of that value occurring, and then sum all these products together."

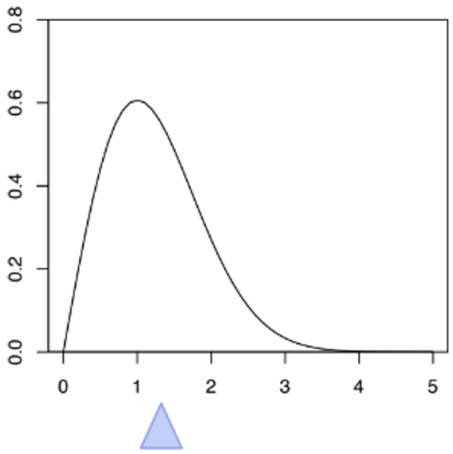


FIGURE 5.5
The expected value of a continuous r.v. is the balancing point of the PDF.

Figure 6.3: Enter Caption

6.3 E.g. 1: Uniform Distribution, Continuous

6.3.1 PDF

A continuous r.v. X is said to have uniform distribution on the interval (a,b) if its PDF is:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

NB: Uniform distribution's PDF function doesn't actually have x parameter: is the same whatever x is.
Proof it is a valid PMF: it sums (integrates) to 1 : rectangle (b-a)*(1/b-a) = 1

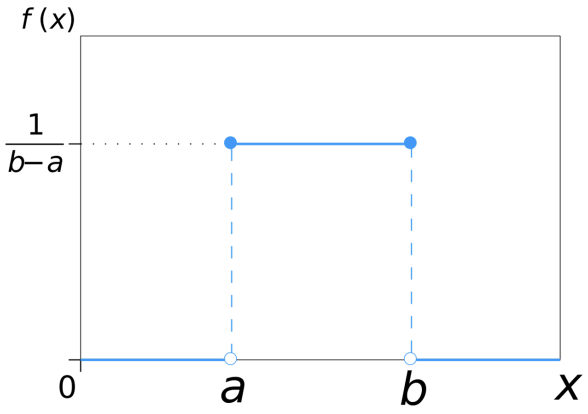


Figure 6.4: Enter Caption

6.3.2 CDF

$$F(x) = \begin{cases} 0 & \text{if } x \leq a \\ \frac{x---a}{b---a} & \text{if } a < x < b \\ 1 & \text{if } x \geq b \end{cases}$$

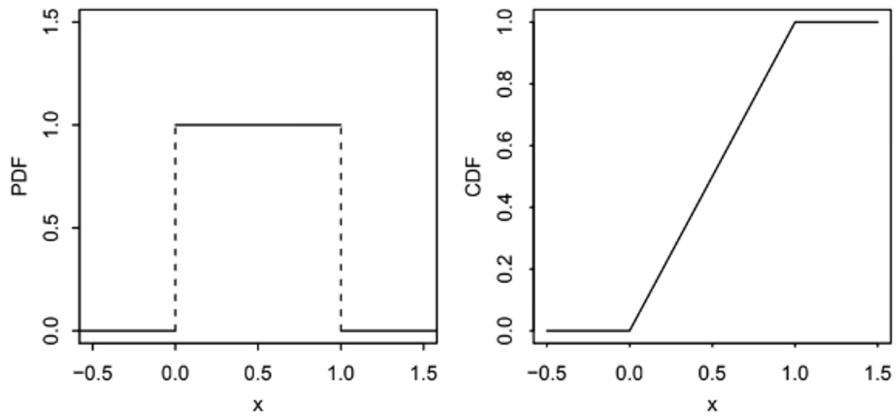


Figure 6.5: Enter Caption

6.3.3 Mean

Mean / expected value:

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx$$

Plug in the function for the distribution (ie the PDF).
Plug in $f(x)$ for the Uniform
(and focus on the support of X because for all non (a, b) values $X = 0$, so they do not contreibute to the mean).

$$\begin{aligned}\int_a^b x \frac{1}{b---a}\, dx &= \frac{x^2}{2(b---a)} + c\Big|_a^b \\ &= \frac{b^2---a^2}{2(b---a)} \\ &= \frac{(b+a)(b---a)}{2(b---a)} \\ &= \frac{a+b}{2}\end{aligned}$$

NB: The integral of x with respect to x is $\frac{x^2}{2}$, so when we include the constant, it becomes $\frac{x^2}{2(b-a)}$.
DO I NEED TO KNOW HOW TO DO THIS????

6.4 E.g.2: Normal Distribution

(NB: is always continuous)

6.4.1 PDF

$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x---\mu}{\sigma}\right)^2}$$

NB: 2x params: μ (mean) and σ^2 (variance).
Standard Normal is just the Normal with $\mu = 0$ and $\sigma^2 = 1$:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

Technically, the Normal's support is -infinite to infinite.
NB as it is continuous, the height is NOT a probability (it is mostly meaningless; simply standardised so that integrates (ie the AUC) is equal to 1. Instead, the area (integral) is the probability.

6.4.2 CDF

By convention, CDF of X written $\phi(x)$ — the functional form of CDF is ugly and unimportant.

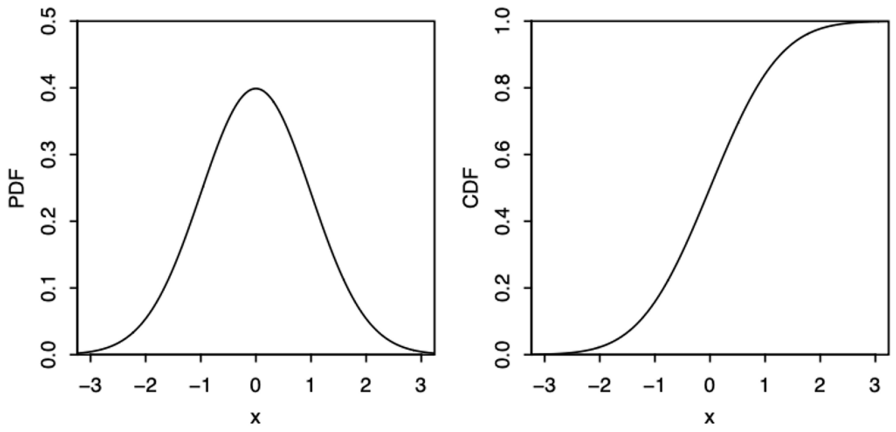


Figure 6.6: Enter Caption

6.4.3 Parameters

Mean μ = 'location' parameter (ie where the distribution is centred).
Variance σ^2 is called the scale parameter (ie determines shape)

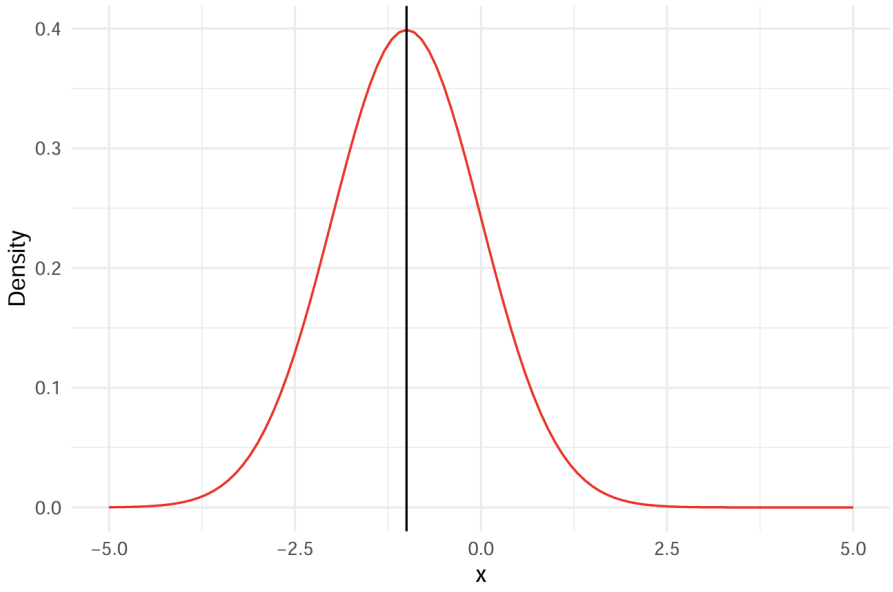
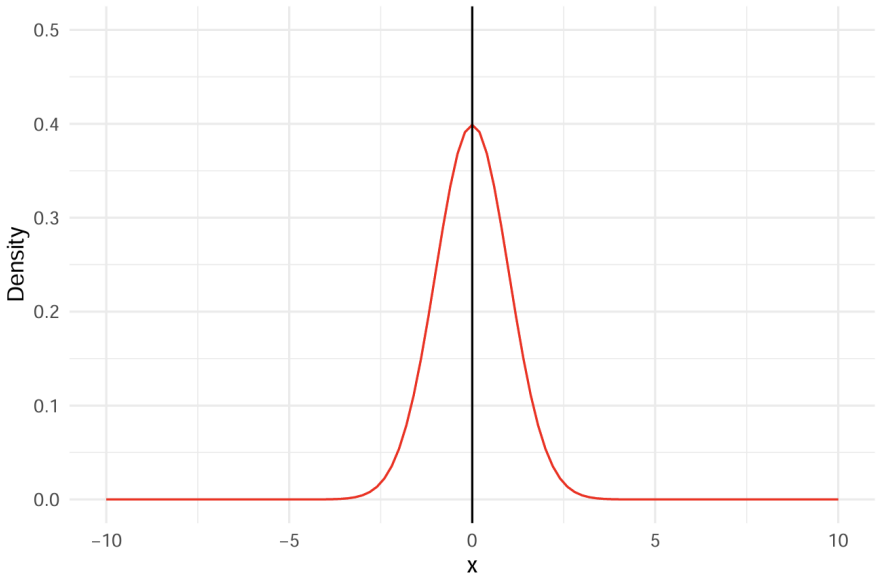
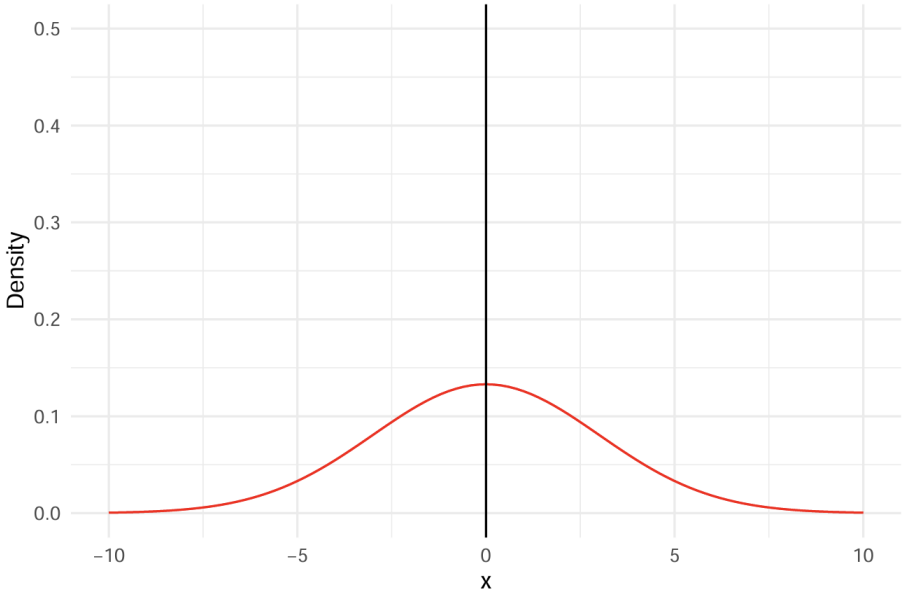


Figure 6.7: Enter Caption



Normal PDF with mean 0, variance 1^2

Figure 6.8: Enter Caption



Normal PDF with mean 0, variance 3^2

Figure 6.9: Enter Caption

6.4.4 Standardization

Any Normal-distributed r.v. X with mean μ and variance σ^2 $\rightarrow$ transformable into Standard Normal:
If

$$Z \sim \mathcal{N}(0,1), \text{ then } X = \mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$$

→ can transform any Normal-distributed r.v. into a Standard Normal:

$$Z = \frac{X - - - \mu}{\sigma}$$

1. demean
2. divide by σ

This is just what z-scores are. Allows for regression interpretation: 1 unit of X associated with a β change in Y.
NB: r.v. itself does NOT have to be Normal distribution — can be ANY distribution; so long as it is i.i.d → you can take the mean and that will be normally distributed.

6.4.5 Features of the Normal

1. Central Limit Theorem: *For i.i.d r.v.s, sampling distribution of the standardized sampled mean tends towards the Standard Normal distribution, even if the original variables themselves are not normally distributed.*
   - underlying r.v. distribution doesn;t have to be Normal distributed.
   - if i.i.d → its sample mean will be normally distributed.
   - this insight allows us to do statistical inference: the sample means are Normal distributed around the mean, so we standardise this.
2. Symmetry of the PDF: $\phi(z) = \phi(-z)$
3. Symmetry of the tails: $\phi(z) = \phi(-z)$

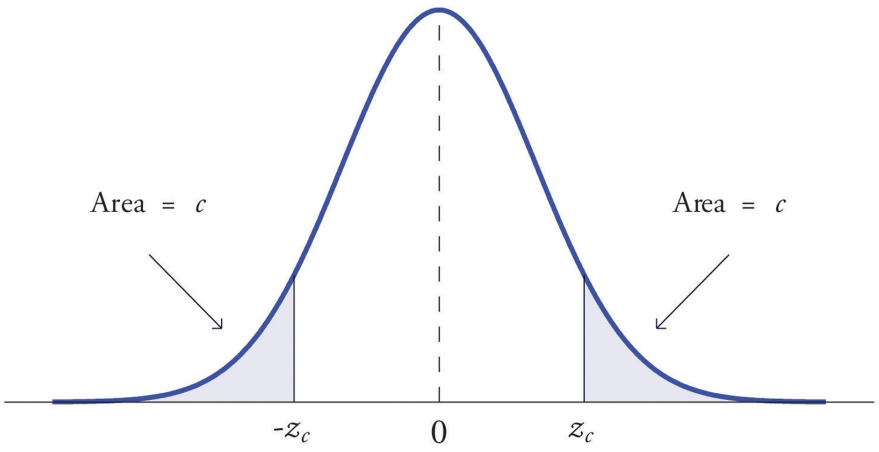


Figure 6.10: Enter Caption

6.4.6 Benchmarks of the Normal

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then:

$$P(|X - - - \mu| < \sigma) \approx 0.68$$
$$P(|X - - - \mu| < 2\sigma) \approx 0.95$$
$$P(|X - - - \mu| < 3\sigma) \approx 0.997$$

6.5 E.g. 3: Exponential

6.5.1 PDF

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0$$

NB: 1x parameter: λ: $X \sim \text{Expo}(\lambda)$

6.5.2 CDF

$$F(x) = 1 - - - e^{-\lambda x}, \quad x > 0$$

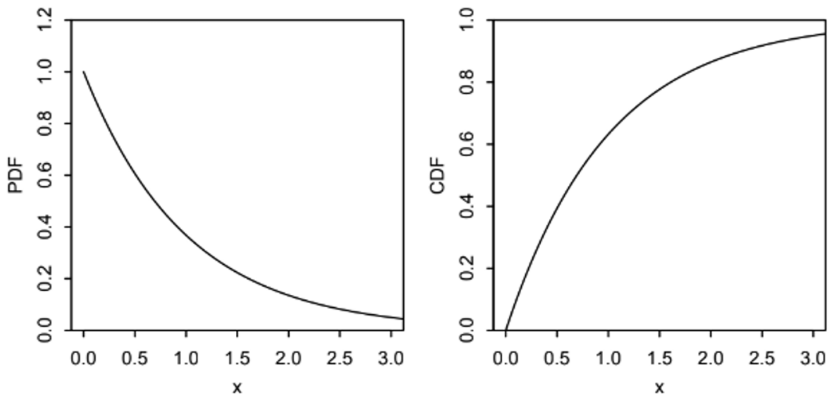


FIGURE 5.12
Expo(1) PDF and CDF.

Figure 6.11: Enter Caption

6.5.3 Modelling purpose

= the continuous analog of the Geometric

- Geometric r.v. counts number of failures until first success in a sequence of Bernoulli trials.
- Exponential r.v. represents continuous time you have to wait before arrival of first success.
- Parameter λ = rate of success per some unit of time.
- NB: connection to the Poisson (part of the Exponential family); the discrete number of events per some unit of time.

6.5.4 Features: memorylessness

$$P(X \geq s + t | X \geq s) = P(X \geq t)$$

- time spent waiting already has no effect on time you will spend waiting for the event
- ie. after waiting s minutes, probability you'll have to wait another t mins is same as prob of having to wait t mins with no prev waiting time under your belt
- if human lifetimes were Exponential: life expectancy age 80 same as if newborn baby
- exponential model is particularly useful in situations where a process decreases or increases at a rate proportional to its current value.

- radioactive decay / compound interest in finance / population growth-decline / cooling-heating in thermodynamics / spread of infectious diseases
- more importantly, Exponential is a building block for more flexible distributions that do account for the passage of time (eg the Weibull)

6.6 Continuous variables applied to Probability

6.6.1 Joint Distribution of Continuous r.v.s

CDF

If X and Y have a continuous joint distribution, we require that the **joint CDF**...

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y)$$

... be differentiable with respect to both x and y.
The partial derivative with respect to x and y is the **joint PDF**
This is quite similar to single r.v.s, where the derivative of the CDF is the PDF, but here since we have two variables we need to differentiate it with respect to both (i.e partial derivatives).

PDF

Joint PDF gives the likelihood of X and Y both taking on specific values x and y

The **joint PDF** is obtained by taking the partial derivatives of the **joint CDF**.

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y)$$

This means you first take the partial derivative of $F_{X,Y}(x,y)$ with respect to x, and then take the partial derivative of the result with respect to y.
E.g. $F(x,y) = \frac{1}{2}x^2y^3$

CDF: $F(x,y) = \frac{1}{2}x^2y^3 \rightarrow$ PDF: $f(x,y) = 3xy^2$

To find the joint PDF, you would take the partial derivatives of this CDF function with respect to x and then y, in turn:

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y}\left(\frac{1}{2}x^2y^3\right)$$

results in a function that represents the joint PDF, which describes how the probability density is distributed across the different values x and Y.

NB this is not the same as just taking the partial derivatives in parrallel, here we are taking them sequentially
Partial Derivative with respect to x ($\frac{\partial F}{\partial x}$):
When differentiating with respect to x, y is treated as a constant. So,

$$\frac{\partial}{\partial x}\left(\frac{1}{2}x^2y^3\right) = y^3\left(\frac{\partial}{\partial x}\left(\frac{1}{2}x^2\right)\right).$$

Differentiate $\frac{1}{2}x^2$ with respect to x, which gives x. Therefore,

$$\frac{\partial F}{\partial x} = xy^3.$$

Partial Derivative with respect to y ($\frac{\partial F}{\partial y}$):
When differentiating with respect to y, x is treated as a constant. So,

$$\frac{\partial}{\partial y}\left(\frac{1}{2}x^2y^3\right) = x^2\left(\frac{\partial}{\partial y}(y^3)\right).$$

Differentiate y^3 with respect to y, which gives $3y^2$. Therefore,

$$\frac{\partial F}{\partial y} = \frac{3}{2}x^2y^2.$$

So, the partial derivatives of $F(x,y) = \frac{1}{2}x^2y^3$ are $\frac{\partial F}{\partial x} = xy^3$ and $\frac{\partial F}{\partial y} = \frac{3}{2}x^2y^2$.

... that is NOT the same as the mixed partial derivative you end up with when getting Joint PDF from Joint CDF of continuous variables

First, differentiate $f(x,y) = \frac{1}{2}x^2y^3$ with respect to x:

$$\frac{\partial f}{\partial x} = xy^3.$$

Then, differentiate the result with respect to y:

$$\frac{\partial^2 f}{\partial x \partial y} = x \cdot 3y^2 = 3xy^2.$$

Thus, the mixed partial derivative $\frac{\partial^2}{\partial x \partial y}$ of $F_{X,Y}(x,y) = \frac{1}{2}x^2y^3$ is $3xy^2$.

NB some of the F vs f notation not right

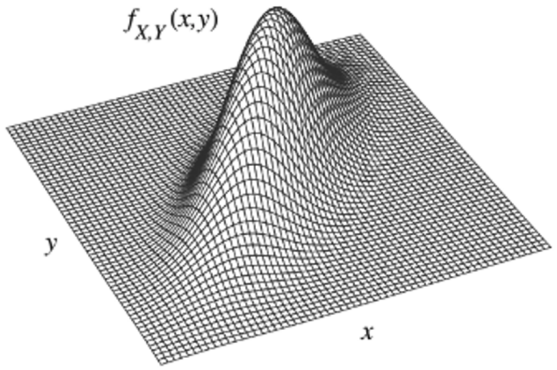


FIGURE 7.4
Joint PDF of continuous r.v.s X and Y.

Figure 6.12: Enter Caption

6.6.2 Marginal Distribution of Continuous r.v.s

To get marginal distribution of X, integrate over all values of Y from the Joint PDF

This process essentially sums out the effect of Y, leaving the distribution of X alone. We have collapsed the 2D distribution into a 1D distribution

...rather than summing (as we did for discrete r.v.s); in effect integration is a form of summing if you think about it in terms of the area under the curve.

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$$

Marginal = when we strip out one of the variables (we are no longer interested in, by summing/integrating all instances)

The marginal PDF $f_X(x)$ tells you about the probability distribution of X irrespective of Y. It is now just the PDF of X
... it also allows us to renormalise from the Joint PDF *(ie the Joint / Marginal)* to get to the Conditional Joint PDF:

6.6.3 Conditional PDF

Conditional PDF of Y given $X = x$ is:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

As with discreet r.v.s, the conditional is the joint over the marginal.
NB: How can we condition on $X = x$ for a continuous r.v. when we learned that the event has prob 0? Technically we're conditioning on the event that X falls in some small interval containing x, say $(x - - - \varepsilon, x + \varepsilon)$, and then taking the limit as ε goes to 0.

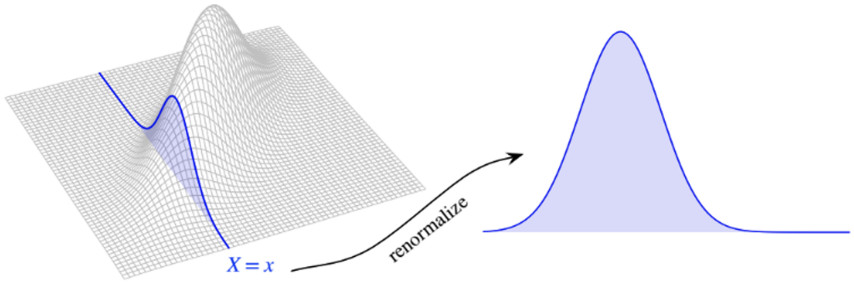


FIGURE 7.5
Conditional PDF of Y given $X = x$. The conditional PDF $f_{Y|X}(y|x)$ is obtained by renormalizing the slice of the joint PDF at the fixed value x.

Figure 6.13: Enter Caption

6.6.4 Bayes Rule and LOTP for continuous r.v.s

Bayes rule:

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)f_Y(y)}{f_X(x)}$$

LOTP:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y)f_Y(y)\,dy$$

We simply replaced sum from the discrete world with an integral.

6.6.5 Can combine discreet and continuous r.v.s

	Y **discrete**	Y **continuous**
X **discrete**	$P(Y=y\|X=x) = \frac{P(X=x\|Y=y)P(Y=y)}{P(X=x)}$	$f_Y(y\|X=x) = \frac{P(X=x\|Y=y)f_Y(y)}{P(X=x)}$
X **continuous**	$P(Y=y\|X=x) = \frac{f_X(x\|Y=y)P(Y=y)}{f_X(x)}$	$f_{Y\|X}(y\|x) = \frac{f_{X\|Y}(x\|y)f_Y(y)}{f_X(x)}$

Figure 6.14: Enter Caption

Chapter 7

Continuous Random Variables II

Finals Fall 2023 – Henry Baker

M4DS Finals Revision: Session 7
Calculus Meets Probability / Continuous Random Variables II

7.1 Covariance

- Covariance between two r.v.s X and Y is a measure of the amount they *vary together*.
- If the variables tend to show similar behavior (i.e., both increase or decrease together), the covariance is positive. If one variable tends to increase when the other decreases, the covariance is negative. If the variables do not show any consistent relationship, the covariance is close to zero.

7.1.1 Covariance Definition

- Covariance is just how two variables move together.
- Covariance is the "expectation of the product, minus the product of the expectations"
- Independent r.v.s: no pattern of how they move together $\rightarrow$ covariance is 0

$$\text{Cov}(X, Y) = \mathbb{E}((X - - - \mathbb{E}X)(Y - - - \mathbb{E}Y))$$

Break down:

- $\mathbb{E}(X)$ and $\mathbb{E}(Y)$ = expected values / means of the r.v.s.
- $(X - - - \mathbb{E}X)$ and $(Y - - - \mathbb{E}Y)$ = deviation from the mean (how far each individual observation of the variables is from their average values.)
- $(X - - - \mathbb{E}X)(Y - - - \mathbb{E}Y)$ = product of deviations (for each pairs of observations of X and Y. IS THIS BASIS OF LEAST SQUARED ERRORS????
- $\mathbb{E}(X - - - \mathbb{E}X)(Y - - - \mathbb{E}Y)$ = expected value (or average) of these products over all pairs of observations. **Quantifies the average product of deviations, thus giving a measure of how much X and Y co-vary**.

By linearity of expectation:

Covariance is the "expectation of the product, minus the product of the expectations"

$$\text{Cov}(X,Y) = \mathbb{E}[XY] - - - \mathbb{E}[X]\mathbb{E}[Y]$$

*** NB: definition of independence from wk 3:

- Continuous r.v.s: $P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$
- Discrete r.v.s: $P(X = x, Y = y) = P(X = x)P(Y = y)$

This is related? Here rather than probabilities, we are dealing with expectations. . . :

$$\begin{aligned} \text{Independence} &: \text{Covariance} = 0 \\ \text{Covariance} = 0 &: \mathbb{E}[XY] - - - \mathbb{E}[X]\mathbb{E}[Y] = 0 \\ &: \mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \end{aligned}$$

7.1.2 Some Covariance rules

1. $\text{Cov}(X,X) = \text{Var}(X)$
2. $\text{Cov}(X,Y) = \text{Cov}(Y,X)$
3. $\text{Cov}(X,c) = 0$ for any constant c
4. $\text{Cov}(aX,Y) = a \cdot \text{Cov}(X,Y)$ for any constant a
5. $\text{Cov}(X+Y,Z) = \text{Cov}(X,Z) + \text{Cov}(Y,Z)$
6. $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2 \cdot \text{Cov}(X,Y)$
7. **additional:** $E(EX)$: expectation of a constant is just a constant .

7.2 Correlation

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}}$$

This scaling puts correlation between -1:1 (for ease of interpretation)

7.2.1 Correlation Imposes Linearity

- Independence $\rightarrow$ uncorrelated
- Uncorrelated $\rightarrow \times \rightarrow$ Independence
- Non-linearity will mess you up. E.g if you're trying to fit a curve through the parabola: you'd get a straight line
- if you have any non-linearity $\rightarrow$ throw some other things (other than correlation) before you conclude that they are independent

If X and Y are independent $\rightarrow$ uncorrelated

Proof:

$$\begin{aligned} \text{Cov definition: } Cov(X,Y) &= E(XY) - - - E(X)E(Y) \\ \text{for independent i.v.s: } E(XY) &= E(X)E(Y) \\ \text{thus Cov} &= 0 \end{aligned}$$

BUT
If X and Y are uncorrelated $\rightarrow$ not necessarily independent.

Consider:

$$\begin{aligned} X &\sim \mathcal{N}(0,1) \\ Y &= X^2 \end{aligned}$$

These are perfectly dependent (Y is a function of X), but uncorrelated:

$$\begin{aligned} Cov(X,Y) &= E(XY) - - - E(X)E(Y) \\ &= E(X \cdot X^2) - - - E(X)E(Y) \\ &= X^3 - - - E(X)E(Y) \\ &= 0 - - - 0 \cdot E(Y) \\ &= 0 \end{aligned}$$

the important point is that $\mathbf{E(X) = 0}$ since it is Standard Normal distribution.

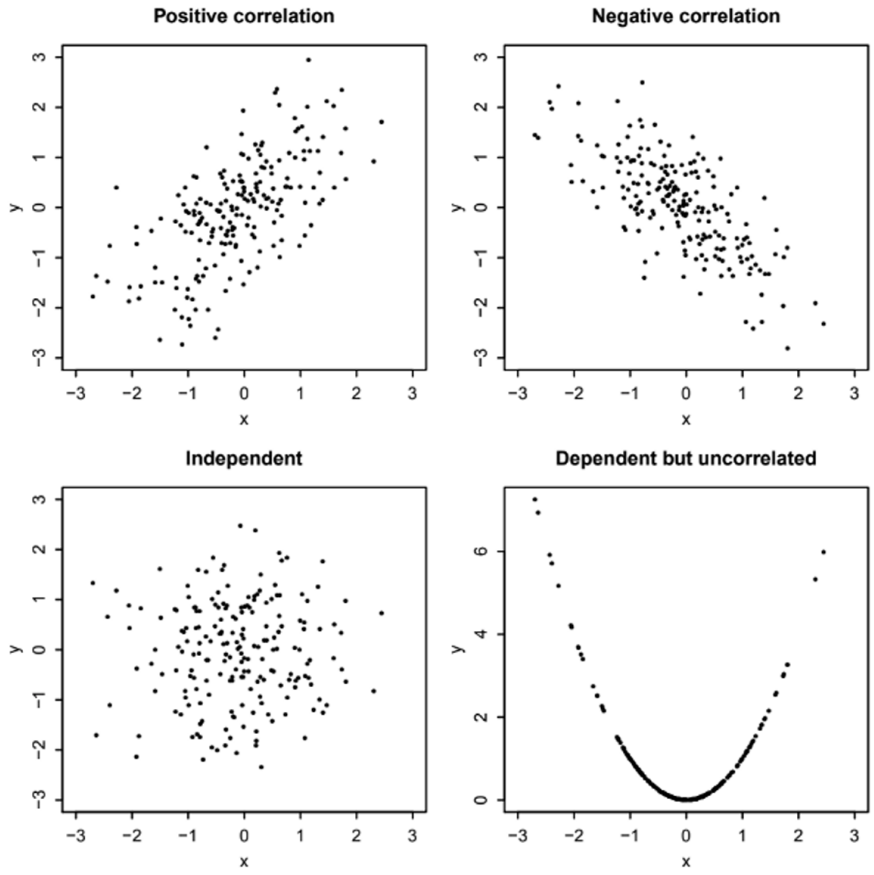


Figure 7.1: Enter Caption

These plots display how correlation imposes linearity (unlike dependence)

7.2.2 Further Explanation: Proof of Cov > Corr

Two random variables X and Y are said to be **independent** if the occurrence of an event related to X does not affect the probability of an event related to Y, and vice versa.

Mathematically, independence means that for any events A and B:

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

Uncorrelated random variables are those for which there is no linear relationship between them. The covariance between uncorrelated variables is zero.

Covariance, denoted as $\text{Cov}(X, Y)$, measures the joint variability of two random variables. It's defined as

$$\text{Cov}(X, Y) = E[(X - - - E[X])(Y - - - E[Y])]$$

which simplifies to

$$E(XY) - - - E(X)E(Y)$$

For independent r.v.s, the expectation of their product is the product of their expectations.

$$E(XY) = E(X)E(Y)$$

Substituting this into the covariance formula:

$$\begin{aligned} Cov(X,Y) &= E(XY) - - - E(X)E(Y) \\ &= E(X)E(Y) - - - E(X)E(Y) \\ &= 0 \end{aligned}$$

Since covariance is zero, X and Y are uncorrelated.

Implication of Zero Covariance: Zero covariance implies that there is no linear relationship between X and Y. In other words, knowing the value of X gives no information about the value of Y, and vice versa, which is consistent with the definition of independence.

The **correlation coefficient**, often denoted as ρ, is a normalized measure of the strength and direction of the linear relationship between two variables.

It's defined as

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

where σ_X and σ_Y are the standard deviations of X and Y, respectively.

If $\text{Cov}(X,Y) = 0$, then $\rho_{X,Y} = 0$, indicating no linear correlation.

In summary, if two random variables are independent, they do not affect each other, and therefore, their covariance (and thus their correlation) is zero. This means independent variables are always uncorrelated. However, the reverse is not always true; uncorrelated variables are not necessarily independent.

Correlation is just a measure of covariance and linearity, whereas dependence measures something different.

7.2.3 Example proving independence

Suppose X and Y are r.v.s where X can take values with 1 or 2 with equal probability, and Y can take values 3 or 4 with equal probablity. Assume X and Y are independent.

$$\begin{aligned} E(X) &= \frac{1}{2}(1) + \frac{1}{2}(2) = 1.5 \\ E(Y) &= \frac{1}{2}(3) + \frac{1}{2}(4) = 3.5 \end{aligned}$$

$E(XY)$ can be calculated by considering all combinations of X and Y
We have four combinations: (1, 3), (1, 4), (2, 3), and (2, 4). Each combination occurs with a probability of 1/4, since the probabilities of X and Y are each 1/2.

$$\begin{aligned} E(XY) &= \frac{1}{4}(1 \times 3) + \frac{1}{4}(1 \times 4) + \frac{1}{4}(2 \times 3) + \frac{1}{4}(2 \times 4) = 5.25 \\ \text{Cov}(X,Y) &= E(XY) - - - E(X)E(Y) = 5.25 - - - (1.5 \times 3.5) = 0 \end{aligned}$$

Since the covariance is 0, it suggests that X and Y are uncorrelated

7.3 Law of Large Numbers

Assume i.i.d rvs X_1, X_2, $X_3 \cdots$ with mean μ and variance σ^2.
We take sample of size n and define the sample mean as:

$$\bar{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Assume the daily temperatures for five consecutive days are represented by random variables X_1, X_2, X_3, X_4, and X_5. These could be, for instance:
$X_1 = 3°\text{C}$ (Temperature on Day 1)
$X_2 = 5°\text{C}$ (Temperature on Day 2)
$X_3 = 7°\text{C}$ (Temperature on Day 3)
$X_4 = 2°\text{C}$ (Temperature on Day 4)
$X_5 = 4°\text{C}$ (Temperature on Day 5)
The average temperature over these five days, denoted as $\bar{X}_5$, is calculated as:

$$\bar{X}_5 = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5} = \frac{3 + 5 + 7 + 2 + 4}{5} = 4.2°\text{C}$$

This sample mean is itself an r.v.
What are its expectation and variance?
Expectation

$$\begin{aligned} E(\bar{X}_n) &= E\left(\frac{1}{n}(X_1 + \ldots + X_n)\right) \\ &= \frac{1}{n}(E(X_1) + \ldots + E(X_n)) \\ &= \frac{1}{n}(\mu_1 + \ldots + \mu_n) \\ &= \frac{1}{n}(n\mu) \\ &= \mu \end{aligned}$$

Variance

$$\begin{aligned} \text{Var}(\bar{X}_n) &= \frac{1}{n^2}\text{Var}(X_1 + \ldots + X_n) \\ &= \frac{\sigma^2}{n} \end{aligned}$$

Law of Large Numbers: as n grows large, the sample mean $\bar{X}$ converges to the true mean μ.

- Sample mean (if i.i.d): $\bar{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$
- sample mean is itself r.v.:
  - Expectation = μ (i.e. the population mean — the sample mean and the population mean converge as n grows)
  - Variance = $(\frac{\sigma}{n})^2 = \frac{\sigma^2}{n}$
  - Standard Deviation = $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$

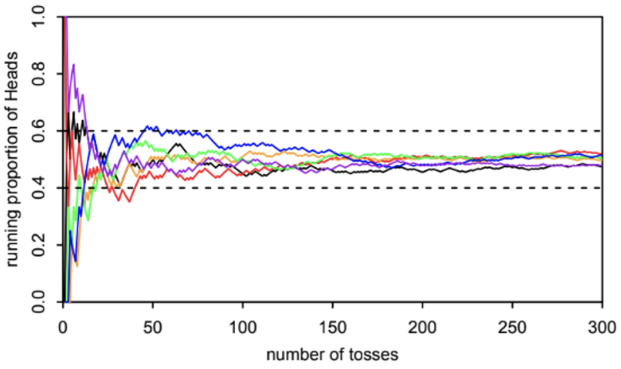


FIGURE 10.2
Running proportion of Heads in 6 sequences of fair coin tosses. Dashed lines at 0.6 and 0.4 are plotted for reference. As the number of tosses increases, the proportion of Heads approaches 1/2.

Figure 7.2: Enter Caption

Law of Large Numbers: as n grows large, the sample mean $\bar{X}$ converges to the true mean μ.
NB: LLN does not contradict a coin toss (or other r.v. being memoryless: convergence takes place through swamping: past tosses are "swamped" by the infinitely many tosses yet to come.

7.4 Central Limit Theorem

CLT = that the standardised sample mean (standardised $\bar{X}$) converges in distribution to the standard Normal as $n \to \infty$
i.e. given a sufficiently large sample size, the sampling distribution of the sample means will be approximately normally distributed, regardless of the shape of the population distribution

7.4.1 Calculating the Standardised Sample Mean

Calculating the Standardised Sample Mean:

1. subtract expectation μ
2. dividing by standard deviation $\frac{\sigma}{\sqrt{n}}$

NB: the variance of the sampling distribution as defined above was $\frac{\sigma^2}{n}$ so the standard deviation is in turn the square root of that: $\sqrt{\frac{\sigma^2}{n}} \to \frac{\sigma}{\sqrt{n}}$

7.4.2 Convergence to Standard Normal

LLN says this quantity (the Standardised Sample Mean) converges in distribution to the Standard Normal as $\to \infty$:

$$\frac{\bar{X}_n - - - \mu}{\frac{\sigma}{\sqrt{n}}} = \sqrt{n}\left(\frac{\bar{X}_n - - - \mu}{\sigma}\right)$$

$$\sqrt{n}\left(\frac{\bar{X}_n - - - \mu}{\sigma}\right) \xrightarrow{d} \mathcal{N}(0,1) \tag{7.1}$$

This holds regardless of distribution of Xs

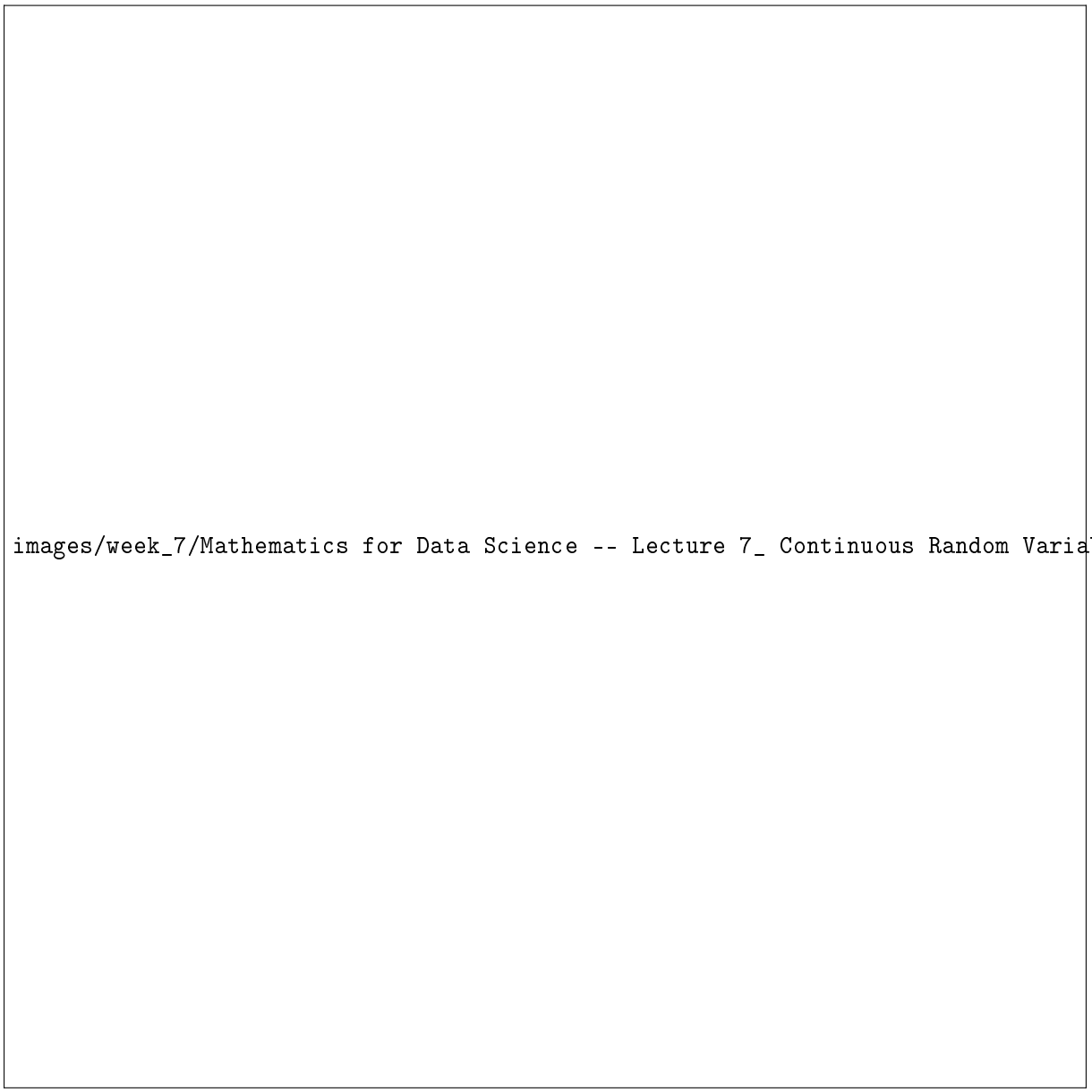


Figure 7.3:

This shows:

- for sample size $n = 1$ it is just the PDF of the underlying distribution:
- practical e.g: if we were sampling number of passengers in cars which we take to be geometrically distributed, when sample size $n = 1$, each sample mean is just the value of the underlying r.v. so it acts as just taking observations of the underlying r.v.
- when $n > 1$ we are now starting to get sample means which will move towards the centre of the distribution (which is the true population mean μ according to Law of Large Numbers).
- the distribution of these sample means approach Normal distribution as we increase n
- the variance of this sample mean distribution is the squared population standard deviation divided by n: Variance $= \frac{\sigma^2}{n}$. This formula shows that the variance of the sample means decreases as the sample size n increases. In other words, larger samples lead to a narrower spread in the sampling distribution of the mean.
- NB: $n = 30$ is the magic number: here everything starts to approach Normal.

7.4.3 Further Notes on CLT

NB: you can un-standardise the sample mean to restate the CLT as:

$$\bar{X}_n \xrightarrow{d} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{as} \quad n \to \infty$$

NB: CLT works just as well for sum rather than the mean:

$$\sum_{i=1}^{n} X_i \xrightarrow{d} \mathcal{N}(n\mu, n\sigma^2) \quad \text{as} \quad n \to \infty$$

While CLT works for any distribution with finite mean and variance, underlying distirbution matters for how large n has to be before the Normal approximation starts to look accurate.
As shortcut: can write CLT in approximate form:

$$\bar{X}_n \approx \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

7.4.4 Example: Normal Approximation to the Binomial

Recalling that the Binomial (n, p) is the sum of n Bernoullis with probability p, we can even use the Normal distribution to approximate the Binomial.

$$\sum_{i=1}^{n} X_i \approx \mathcal{N}(n\mu, n\sigma^2)$$

and recalling that the mean of a Bernoulli is p and its variance is $p(1 - - - p)$, we can use the CLT to say:

$$\sum_{i=1}^{n} X_i \approx \mathcal{N}(np, np(1 - - - p))$$

7.5 Lab: EM Algorithm

Mid Terms Fall 2023 – Henry Baker

M4DS Mid Terms Revision: Lab 7
EM Algorithm

High Level explanation of EM Algo

1. Initialise parameters
   - μ_1, μ_2 = means
   - σ_1, σ_2 = standard deviations
   - $\pi_1.\pi_2$ = mixing properties (the initial prob of being in one distribution)
2. E-step: Expectation — compute responsibilities of each data point (= calculate the γ of each data point: the prob that each data point belongs to each component given current parameter values)
3. M-step: Maximisation — update the parameters based on the responsibilities (= MLE of each parameter given the γ values for each data point (the parameters defined as some function involving sums of gamma-data point, so will give a single value).
4. Evaluate the new log-likelihood with new (i) parameter, (ii) responsibilities.
5. Check for convergence

7.6 Context; Set up

- Used to find Maximum Likelihood parameters of a statistical model with **latent variables**, which are hidden or unobserved characteristics of our data (i.e. where straightforward MLE cannot be applied).
- Going to model Old Faithful eruptions as a Gaussian Mixture (a mixture of 2 Normal distribution, each woth theeir own mean and variance).
- Eruption can either be short type or long type, correspondinbg to a different (unobservced) geological process.
- Define Z_1 (first eruption) as r.v.:

$$Z_1 = \begin{cases} 0 & \text{if eruption is long type} \\ 1 & \text{if eruption is short type} \end{cases}$$

  Z_2 (second eruption):

$$Z_2 = \begin{cases} 1 & \text{if eruption is long type} \\ 0 & \text{if eruption is short type} \end{cases}$$

- Define

$$\pi_1 = \text{probability eruption is short type}$$

$$\pi_2 = \text{probability eruption is long type}$$

  Assume $\pi_1 + pi_2 = 1$ (i.e. there are no other types.

- Define: X = r.v. representing duration of eruption.

Task is to produce an MLE estimate for the type probabilities, and the means and variances of the two eruption types

7.7 Attempting MLE

Using **Law Of Total Probability** to write PDF of X:

$$\begin{aligned} Pr(X = x) &= Pr(Z_1 = 1)Pr(X = 1|Z_1 = 1) + Pr(Z_2 = 1)Pr(X = x|Z_2 = 1) \\ &= \pi_1\mathcal{N}(x|\mu_1, \sigma_1) + \pi_2\mathcal{N}(x|\mu_2, \sigma_2) \end{aligned}$$

With this PDF, we can write down the **likelihood** of the dataset (where n is number of observations in the dataset.

$$L(x_1, \ldots, x_n; \mu_1, \mu_2, \sigma_1, \sigma_2, \pi_1, \pi_2) = \prod_{i=1}^{n}(\pi_1\mathcal{N}(x_i|\mu_1, \sigma_1) + \pi_2\mathcal{N}(x_i|\mu_2, \sigma_2))$$

As a **Log-Likelihood function**

$$\ell(x_1, \ldots, x_n; \mu_1, \mu_2, \sigma_1, \sigma_2, \pi_1, \pi_2) = \sum_{i=1}^{n}\log(\pi_1\mathcal{N}(x_i|\mu_1, \sigma_1) + \pi_2\mathcal{N}(x_i|\mu_2, \sigma_2))$$

Here, each observation x_i has a probability of being from either the first normal distribution ($\mathcal{N}(x_i|\mu_1, \sigma_1)$ with weight π_1, or from the second normal distribution $\mathcal{N}(x_i|\mu_2, \sigma_2)$ with weight π_2.

The weights pi_1 and pi_1 are essentially the mixing proportions (NB, as above, they should sum to 1).

For a single observation x_i, the likelihood is the sum of the probabilities of x_i coming from each distribution.

When we have multiple independent observations, the total likelhood is the product of the individual likelihoods.

7.7.1 MLE for mu_1

Take first derivative of the log likelihood with respect to μ_1.

By the Chain Rule:

$$\frac{d}{dx}\left(f(g(x))\right) = f'(g(x)) \cdot g'(x)$$

Let $f(y) = \log(y)$, and $g(x_i) = \pi_1\mathcal{N}(x_i|\mu_1, \sigma_1) + \pi_2\mathcal{N}(x_i|\mu_2, \sigma_2)$.

The derivative of $f(y)$ with respect to y is:

$$f'(y) = \frac{1}{y}.$$

The derivative of $g(x_i)$ with respect to μ_1 is:

$$\frac{\partial}{\partial \mu_1} g(x_i) = \frac{\partial}{\partial \mu_1} \left(\pi_1 \mathcal{N}(x_i|\mu_1, \sigma_1)\right) + \frac{\partial}{\partial \mu_1} \left(\pi_2 \mathcal{N}(x_i|\mu_2, \sigma_2)\right).$$

Since $\pi_2 \mathcal{N}(x_i|\mu_2, \sigma_2)$ does not depend on μ_1, its derivative is 0. Thus,

$$\frac{\partial}{\partial \mu_1} g(x_i) = \frac{\partial}{\partial \mu_1} \left(\pi_1 \mathcal{N}(x_i|\mu_1, \sigma_1)\right).$$

Then:

$$\begin{aligned}
\text{Then, } \frac{\partial \ell}{\partial \mu_1} &= \sum_{i=1}^{n} \frac{\partial}{\partial \mu_1} \log\left(g(x_i)\right) \\
&= \sum_{i=1}^{n} \frac{1}{g(x_i)} \cdot \frac{\partial}{\partial \mu_1} g(x_i) \text{ by the chain rule}
\end{aligned}$$

Since only the first term of $g(x_i)$ depends on μ_1, its derivative is:

$$\begin{aligned}
\frac{\partial}{\partial \mu_1} g(x_i) &= \frac{\partial}{\partial \mu_1} \left(\pi_1 \mathcal{N}(x_i|\mu_1, \sigma_1)\right) \\
\text{Thus, } \frac{\partial \ell}{\partial \mu_1} &= \sum_{i=1}^{n} \frac{1}{\pi_1 \mathcal{N}(x_i|\mu_1, \sigma_1) + \pi_2 \mathcal{N}(x_i|\mu_2, \sigma_2)} \cdot \frac{\partial}{\partial \mu_1} \left(\pi_1 \mathcal{N}(x_i|\mu_1, \sigma_1)\right)
\end{aligned}$$

$$\frac{\partial \ell}{\partial \mu_1} = \sum_{i=1}^{n} \frac{1}{\pi_1 \mathcal{N}(x_i|\mu_1, \sigma_1) + \pi_2 \mathcal{N}(x_i|\mu_2, \sigma_2)} \underbrace{\frac{\partial}{\partial \mu_1} \pi_1 \mathcal{N}(x_i|\mu_1, \sigma_1)}_{\text{let's take this piece}}$$

I SKIPPED THE REST HERE ...
set it to 1, and solve for μ_1

7.8 Bayes rule to the Rescue

7.8.1 Gammas

At this point, lets take a moment to look at the expression we're trying to solve:

$$0 = \sum_{i=1}^{n} \underbrace{\frac{\pi_1 \mathcal{N}(x_i|\mu_1, \sigma_1)}{\pi_1 \mathcal{N}(x_i|\mu_1, \sigma_1) + \pi_2 \mathcal{N}(x_i|\mu_2, \sigma_2)}}_{\textbf{what is this? We'll call this } \gamma} \cdot \left(\frac{x_i - - - \mu_1}{\sigma_1^2}\right)$$

This is the probability that data point i belongs to the first eruption time, given its eruption duration.
We write down this probability and apply Bayes rule:

$$\begin{aligned}
\Pr(z_{1i} = 1|x_i) &= \frac{f(x_i|z_{1i} = 1) \Pr(z_{1i} = 1)}{f(x_i)} \\
&= \frac{\pi_1 \mathcal{N}(x_i|\mu_1, \sigma_1)}{\pi_1 \mathcal{N}(x_i|\mu_1, \sigma_1) + \pi_2 \mathcal{N}(x_i|\mu_2, \sigma_2)}
\end{aligned}$$

We will call this γ_{1i}; similarly, let $\gamma_2 i = Pr(z_{2i} = 1|x_i)$

So the gammas are the probability that a certain data point belongs to their associated distribution (given the values observed).

7.8.2 Back to the maximisation problem

(bracketing the γ for a moment:

Solving for mu_1, we get:

$$\begin{aligned}
0 &= \sum_{i=1}^{n} \gamma_{1i} \frac{x_i - - - \hat{\mu}_1}{\sigma_1^2} \\
0 &= \sum_{i=1}^{n} \gamma_{1i}(x_i - - - \hat{\mu}_1) \\
0 &= \sum_{i=1}^{n} \gamma_{1i} x_i - - - \sum_{i=1}^{n} \gamma_{1i}\hat{\mu}_1 \\
\sum_{i=1}^{n} \gamma_{1i}\hat{\mu}_1 &= \sum_{i=1}^{n} \gamma_{1i} x_i \\
\hat{\mu}_1 &= \frac{\sum_{i=1}^{n} \gamma_{1i} x_i}{\sum_{i=1}^{n} \gamma_{1i}}
\end{aligned}$$

And for $\hat{\mu}_2$ you would get the same thing:

$$\hat{\mu}_2 = \frac{\sum_{i=1}^{n} \gamma_{2i} x_i}{\sum_{i=1}^{n} \gamma_{2i}}$$

In similar fashion, we can find the MLE for $\hat{\sigma}_1, \hat{\sigma}_2, \hat{\pi}_1,$ and $\hat{\pi}_2$ as:

7.9 The EM Algorithm

1. **Initiatize your parameters** means μ_1, μ_2, standard deviations σ_1, σ_2 mixing properties π_1, π_2 to some starting points. (i.e. literally just assign them some numbers based from educated guesses based on your data. E.g. you might start with equal mixing properties ($\pi_1 = \pi_2 = 0.5$)

   Evaluate the log likelihood under these values.

2. **E-Step — Expectation: (re)Compute the *responsibilities* for each data point**: Evaluate $\gamma_{1i} = Pr(z_{1i} = 1|x_i)$ under the current parameter values. Also evaluate $\gamma_{2i} = Pr(z_{2i} = 1|x_i)$. You do this for each x_i These are the ***responsibilities***. These are the probability that each data point belongs to each component of the mixture model.

$$\gamma_{1i} = \frac{\pi_1 \mathcal{N}(x_i|\mu_1, \sigma_1)}{\pi_1 \mathcal{N}(x_i|\mu_1, \sigma_1) + \pi_2 \mathcal{N}(x_i|\mu_2, \sigma_2)}$$
$$\gamma_{2i} = \frac{\pi_2 \mathcal{N}(x_i|\mu_2, \sigma_2)}{\pi_1 \mathcal{N}(x_i|\mu_1, \sigma_1) + \pi_2 \mathcal{N}(x_i|\mu_2, \sigma_2)}$$

3. **M-step — Maximisation: update parameters based on *responsibilities***: Using the γ_{1i} and $\gamma_2 i$, update the parameters using the current responsibilities to maximise the expected log-likelihood.

$$\hat{\mu}_1 = \frac{\sum_{i=1}^n \gamma_{1i} x_i}{\sum_{i=1}^n \gamma_{1i}}$$
$$\hat{\mu}_2 = \frac{\sum_{i=1}^n \gamma_{2i} x_i}{\sum_{i=1}^n \gamma_{2i}}$$
$$\hat{\sigma}_1 = \sqrt{\frac{\sum_{i=1}^n \gamma_{1i}(x_i - - - \hat{\mu}_1)^2}{\sum_{i=1}^n \gamma_{1i}}}$$
$$\hat{\sigma}_2 = \sqrt{\frac{\sum_{i=1}^n \gamma_{2i}(x_i - - - \hat{\mu}_2)^2}{\sum_{i=1}^n \gamma_{2i}}}$$
$$\hat{\pi}_1 = \frac{1}{n} \sum_{i=1}^n \gamma_{1i}$$
$$\hat{\pi}_2 = \frac{1}{n} \sum_{i=1}^n \gamma_{2i}$$

   These updates are done to improve the fit of the model to the data, based on the current responsibilities.

4. **Evaluate the new log likelihood with your new paramaters and responsibilities**.

$$\text{Log-Likelihood}(\mu_1, \mu_2, \sigma_1, \sigma_2, \pi_1, \pi_2) = \sum_{i=1}^n \log\left(\pi_1 \mathcal{N}(x_i|\mu_1, \sigma_1) + \pi_2 \mathcal{N}(x_i|\mu_2, \sigma_2)\right)$$

5. Check for **convergence** — i.e when the log likelihood has not moved far from the log likelihood in the previous iteration.

Part III

Linear Algebra

Chapter 8

Linear Algebra I

Mid Terms Fall 2023 – Henry Baker

M4DS Mid Terms Revision: Session 8
Linear Algebra

8.1 Data Structures

8.1.1 Basics

1. Scalar: $x = 1$
2. Vector: x = $\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$
3. Matrix: A = $\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$

   - Matrix dimensions are m
     - m = rows (/observations)
     - n = columns (/variables / features)
   - A generic element of this matrix is a_{ij}
     (same as indexing in R)
   - This feels counter-intuitive: the index is [row, column]

4. Tensor: an array of matrices, with elements a_{ijk}

8.1.2 Compact Notation

1. Scalar: $x \in \mathbb{R}^{\not\perp}$
2. Vector: $\mathbf{x} \in \mathbb{R}^{\times}$
3. Matrix: $\mathbf{X} \in \mathbb{R}^{m \times n}$

8.1.3 Transpose

$$\mathrm{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$
$$\mathrm{A}^T = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ a_{13} & a_{23} \end{bmatrix}$$

Same with vectors; vector can be treated as matrix with 1 column; a scalar is matrix with 1 row, 1 columns).

> Heuristic: all the row values become columns values, and all the column values become row values... so whereas the usual [r,c] indexing feels un-intuitive, it transforms to a more comfortable, [c,r] format.
>
> Visually: get the first column, string it out into a row, let the rest of the row fall into the associated column.

8.2 Basic Transformations

8.2.1 Adding Matrices

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} + \begin{bmatrix} 1 & 2 & 1 \\ 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 4 & 4 \\ 5 & 6 & 6 \end{bmatrix}$$

To add two matrices, they must have the same dimensions.

8.2.2 Multiplying a Matrix by a Scalar

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$
$$cA = \begin{bmatrix} c \times a_{11} & c \times a_{12} & c \times a_{13} \\ c \times a_{21} & c \times a_{22} & c \times a_{23} \end{bmatrix}$$

8.3 Multiplying Vectors

Just a simplified form of matrix multiplication.

$$\mathbf{a}^T\mathbf{b} = [a_1, a_2, a_3] \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = a_1b_1 + a_2b_2 + a_3b_3$$

Heuristic Algorithm:

- tip the RHS vector over to its left
- multiply each of the overlapping elements (such that each of the indexes are aligned with their equivalent)
- add the products together
- becomes a single scalar

8.4 Multiplying Matrices

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{bmatrix}$$

- To multiply matrices, they must be conformable:
  - mn and np
  - the inner values: n
  - LHS columns = RHS rows
- gives an output of mp — the outer values
- i.e. the dot product dimensions given by the otuer values

NB a matrix's dimensions are also given rows x columns Thus, in matrix multiplication, order matters

Small Matrix Multiplication cheat sheets:
2x2:

$$AB = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \times \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{pmatrix}$$

3x3

$$AB = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \times \begin{pmatrix} j & k & l \\ m & n & o \\ p & q & r \end{pmatrix} = \begin{pmatrix} aj+bm+cp & ak+bn+cq & al+bo+cr \\ dj+em+fp & dk+en+fq & dl+eo+fr \\ gj+hm+ip & gk+hn+iq & gl+ho+ir \end{pmatrix} \tag{8.1}$$

2x3; 3x2

$$AB = \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix} \times \begin{pmatrix} g & h \\ i & j \\ k & l \end{pmatrix} = \begin{pmatrix} ag+bi+ck & ah+bj+cl \\ dg+ei+fk & dh+ej+fl \end{pmatrix}$$

3x2; 2x2

$$ABc = \begin{pmatrix} a & b \\ c & d \\ e & f \end{pmatrix} \times \begin{pmatrix} g & h \\ i & j \end{pmatrix} = \begin{pmatrix} ag+bi & ah+bj \\ cg+di & ch+dj \\ eg+fi & eh+fj \end{pmatrix}$$

General Heuristic:

- to be conformable: inner terms the same (n: RHS columns, LHS rows)
- the new matrix's dimensions will be outer terms ($m \times p$)
- for each element a_{ij}: sum of the products of the elements of the corresponding row of A and the corresponding column of B.

  SHORT CUT:

  - **identify dimensions of resultant matrix**
  - **identify specific elements within this matrix: → identify their indexes (row, column), write out as a matrix of generic placeholder: (e.g. a_35; a_71; etc)**
  - **for row index value: take elements of equivalent row from matrix A**
  - **for column index value: take elements of equivalent column from matrix B (i.e. the vector)**
  - **write them as sum of multiples**
    - **write out the row values of Matrix A, spaced out with "(" before and "×" after each**
    - **write out the column values of Matrix B in the remaining spaces, with ") +" after each.**
- so for item a_12:
  - all the elements of row 1 from A
  - all the elements of column 2 from B
  - multiplied by each other
  - summed
- for item a_21:
  - all the elements of row 2 of A
  - all the elements of column 1 of B
  - multiplied by eac hother
  - summed
- eg for $a_{2,1}$: this is the 2nd row of the 1st column; sum of the products of all of the LHS matrix's 2nd row, RHS matrix's first column

8.5 Mutliplying Matrices with Vectors

1. **Check conformable**: ensure number of columns = length of vector.
   If matrix $m \times n$, and vector is $n \times 1$ (remember, vectors are vertical): resulting vector will be $m \times 1$
   i.e. will be a vector
2. **set up multiplication**: multiply elements of the row of the matrix, with corresponding element of the matrix then summing.

As before:

- identify dimensions of resultant matrix (in this case a vector)
- identify specific elements within this → identify their indexes (row, column)
- for row index value: take elements of equivalent row from matrix A
- for column index value: take elements of equivalent column from matrix B (i.e. the vector)
- write them as sum of multiples

Easier short cut:

- visually: pick up the vector, tip it over, place it above the matrix
- multiply each of the elements of matrix's columns by the associated vector element above
- sum each row

$$\begin{aligned}
\text{Matrix A (2x3):} \quad & \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \\
\text{Vector x (3x1):} \quad & \begin{bmatrix} 7 \\ 8 \\ 9 \end{bmatrix} \\
\text{Resulting Vector y (2x1):} \quad & \begin{bmatrix} (1 \cdot 7) + (2 \cdot 8) + (3 \cdot 9) \\ (4 \cdot 7) + (5 \cdot 8) + (6 \cdot 9) \end{bmatrix} = \begin{bmatrix} 50 \\ 122 \end{bmatrix}
\end{aligned}$$

8.6 Transpose Facts

$$\begin{aligned}
(A^T)^T &= A \\
(A + B)^T &= A^T + B^T \\
(AB)^T &= B^T A^T \\
a^T b &= b^T a
\end{aligned}$$

NB: line 3 and 4: the order reverses!!

8.7 Systems of Equation

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\
&\vdots \\
a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_n
\end{aligned}$$

In matrix notation:

$$Ax = b \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

Where:
A = matrix of data points $(m \times n)$
x = vector of coefficients we are trying to estimate $(n \times 1)$
b = vector of response variables $(m \times 1)$

8.8 Identity Matrix

The Identity Matrix I_n is a matrix of size $n \times n$ with 1's across the main diagonal and 0's everywhere else.

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Identity Matrix useful for:

- takes a vector x_n to itself : $I_n x_n = x_n$
- takes a matrix $A_{m \times n}$ to itself: $I_m A_{m \times n} = A_{m \times n}$ and $A_{m \times n} I_n =$A$_{m \times n}$
- it defines the inverse of a matrix: $A^{-1}A = I_n$

8.9 The Inverse of a Matrix

$$A^{-1}A = I_n$$

- The inverse does not always exist **only square matrices may have an inverse** (and sometimes they still don't)
- We have several different algorithms to find it when it does exist
- However, it's mainly a theoretical tool and doesn't often need to be computed directly

8.10 Vector Norm

Norm of a vector is a measure of its magnitude or distance from 0

$$|x|_1 = \sum_i |x_i|$$
$$|x|_2 = \sqrt{\sum_i x_i^2}$$
$$|x|_\infty = \max_i |x_i|$$

- L1 Norm (Manhattan / Taxicab) = take the absolute values of the elements of the vector before summing them.
- L2 Norm (Euclidean / most common) = square root of the sum of the squares of the vector's elements (for 2D: this is the hypotenuse)
- L3 Norm (less common) = cube root of the sum of the cubes of the vector's elements

8.11 Lab: Regression Using Matrix Algebra

Mid Terms Fall 2023 – Henry Baker

M4DS Mid Terms Revision: Session 8
Matrix Algebra and Calculus applied to Linear Regression

8.12 Linear Regression

Task: Modelling home values in Boston

8.12.1 Set up

We require that our model be linear in the parameters, which means that for any observation / person i, we can express the response y_i (the median home value) as a *linear combination* of the variables, plus some error term.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i \tag{8.2}$$

NB: the $x_{i1} \rightarrow$ the i indexes observation, the number value indexes the variable

Our goal is the find the best values of the beta coefficients where 'best' has specific meaning: the coefficients that minimise the sum of squared errors over the dataset.

8.12.2 Objective Function for Least Squares

The objective function for least squares regression is:

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} \varepsilon_i^2$$

$$= \min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} (y_i - - - \beta_0 - - - \beta_1 x_{i1} - - - \beta_2 x_{i2} - - - \ldots - - - \beta_p x_{ip})^2$$

8.12.3 System of Equations = inefficient approach

We could set this up as a maximisation problem using a system of equations to solve, where the **partial derivatives of the least squares objective function with respect to each coefficient are set to zero** for optimization. These conditions are given by:

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (y_i - - - \beta_0 - - - \beta_1 x_{i1} - - - \beta_2 x_{i2} - - - \ldots - - - \beta_p x_{ip})^2 = 0 \tag{8.3}$$

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} (y_i - - - \beta_0 - - - \beta_1 x_{i1} - - - \beta_2 x_{i2} - - - \ldots - - - \beta_p x_{ip})^2 = 0 \tag{8.4}$$

$$\frac{\partial}{\partial \beta_2} \sum_{i=1}^{n} (y_i - - - \beta_0 - - - \beta_1 x_{i1} - - - \beta_2 x_{i2} - - - \ldots - - - \beta_p x_{ip})^2 = 0 \tag{8.5}$$

...

$$\frac{\partial}{\partial \beta_p} \sum_{i=1}^{n} (y_i - - - \beta_0 - - - \beta_1 x_{i1} - - - \beta_2 x_{i2} - - - \ldots - - - \beta_p x_{ip})^2 = 0 \tag{8.6}$$

NOTE TO SELF — THIS IS WHY REGRESSION OUTPUT INTERPRETATION IS "HOLDING EVERYTHING ELSE CONSTANT, THE EFFECT OF A 1 UNIT CHANGE IN X ON Y IS...". BETA COEFFICIENTS ARE JUT THE PARTIAL DERIVATIVE OF A MULTIVARIATE FUNCTION?"

Or, we could solve same problem using linear algebra and calculus — much more efficient!

8.12.4 Matrix approach = efficient

Representing the linear regression model in Matrix form:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{8.7}$$

Let's collapse this into:

$$Y = X\beta + \varepsilon$$

Where:

Y $(n \times 1)$ is called the response vector
X $(n \times (p+1))$ is called the design matrix
β $((p+1) \times 1)$ is our vector of coefficients
ε $(n \times 1)$ is our vector of error terms

Think about how this works:

- the design matrix — coefficient product produces a vector
- each element of the response vector represents the summed coefficient-matrix × element products (i.e. for all the variables), plus an error term.
- for each observation you thus have an outcome value = coefficient-matrix-element-product + error term.

Representing the squared error minimisation problem (the cost function) in Matrix form:

$$\varepsilon^T \varepsilon = Y^T Y - - - Y^T X\beta - - - \beta^T X^T Y + \beta^T X^T X\beta$$

This expanded **cost function** (expressed in marix notation) is what we will be minimising (by taking first derivative, set to zero)

Proof:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \ldots & \varepsilon_n \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \varepsilon^T \varepsilon \tag{8.8}$$

Some matrix manipulation of the cost function:

$$\begin{aligned}\varepsilon^T\varepsilon &= (Y - - - X\beta)^T(Y - - - X\beta) \\ &= Y^TY - - - Y^TX\beta - - - \beta^TX^TY + \beta^TX^TX\beta\end{aligned}$$

We still take the first derivative, set it equal to 0, to solve for error-miniming coefficients — but it's efficient, as we'll do it with the whole vector of coefficients.

Solving minimisation problem (set objective function to 0; differentiate; solve

(1) Setting to 0 → (2) taking first derivative with respect to β → (3) simplifying:

$$-2X^TY + 2X^TX\beta = 0$$

Proof:

This equation below takes the first derivative of the expanded squared error term from above, with respect to the coefficient **vector (NB!!)** β, and sets it equal to 0:

$$\frac{\partial}{\partial\beta}(Y^TY - - - Y^TX\beta - - - \beta^TX^TY + \beta^TX^TX\beta) = 0$$

$$\frac{\partial}{\partial\beta}Y^TY - - - \frac{\partial}{\partial\beta}Y^TX\beta - - - \frac{\partial}{\partial\beta}\beta^TX^TY + \frac{\partial}{\partial\beta}\beta^TX^TX\beta = 0$$

Setting this derivative equal to zero allows us to solve for the values of $\beta (a vector!) that minimize the sum of square$

Using the Matrix Cookbook, this reduces to:

$$-2X^TY + 2X^TX\beta = 0$$

Solving for beta:

$$\beta = (X^TX)^{-1}X^TY$$

8.13 Penalised Regression

Penalized regression is a method used in statistical modeling to prevent overfitting, improve prediction accuracy, and sometimes enable model interpretation. The idea is to introduce a penalty term to the regression model that constrains the coefficients, making the model simpler and less prone to overfitting. The two most common types of penalized regression are Lasso Regression (which uses the L1 norm) and Ridge Regression (which uses the L2 norm).

8.13.1 L1 Norm (Lasso Regression)

The L1 norm, used in Lasso Regression, is the sum of the absolute values of the coefficients. In mathematical terms, if you have coefficients $\beta_1, \beta_2, \ldots, \beta_n$, the L1 norm is:

$$\text{L1 norm} = \sum_{i=1}^{n} |\beta_i|$$

In Lasso Regression, the penalty term added to the regression model is proportional to the L1 norm. The objective function in Lasso Regression becomes:

$$\text{Minimize} \left(\text{Residual Sum of Squares} + \lambda \sum_{i=1}^{n} |\beta_i| \right)$$

where λ is a tuning parameter that determines the strength of the penalty. A higher value of λ results in more regularization.

The L1 penalty has the effect of forcing some of the coefficient estimates to be exactly zero when the tuning parameter is sufficiently large. This leads to sparsity, which means that the model automatically selects variables by setting some coefficients to zero, effectively performing variable selection. This is useful in both improving model interpretability and dealing with high-dimensional data.

8.13.2 L2 Norm (Ridge Regression)

The L2 norm, used in Ridge Regression, is the sum of the squares of the coefficients. For coefficients $\beta_1, \beta_2, \ldots, \beta_n$, the L2 norm is:

$$\text{L2 norm} = \sum_{i=1}^{n} \beta_i^2$$

In Ridge Regression, the penalty term added to the regression model is proportional to the L2 norm. The objective function in Ridge Regression is:

$$\text{Minimize} \left(\text{Residual Sum of Squares} + \lambda \sum_{i=1}^{n} \beta_i^2 \right)$$

As with Lasso, λ is a tuning parameter that controls the strength of the penalty. Unlike Lasso, Ridge Regression does not produce sparse models; it doesn't set coefficients to zero but instead shrinks them towards zero. This is particularly useful when dealing with multicollinearity or when you have more predictors than observations.

8.13.3 Summary

Both L1 and L2 regularization techniques are used to prevent overfitting, but they do so in different ways:

- L1 (Lasso) can produce simpler and more interpretable models because it can reduce some coefficients to zero, effectively selecting more relevant features.
- L2 (Ridge) is better suited for dealing with multicollinearity and does not exclude variables from the model but rather reduces their impact.

Choosing between Lasso and Ridge (or combining them, as in Elastic Net Regression) depends on the specific dataset, the underlying problem, and the need for model interpretability.

8.14 Bringing together the above expanded cost function with the L2 Norm constraint, to give the Ridge Regression objective function we try to minimise)

Above we worked out the cost function (i.e. the residual sum of squares / the squared error term) in matrix notation, which we were trying to minimise (by taking first derivative and setting to

0)

Here we are combining that objective function with the L2 Norm constraint. This gives us a new objective function, typical of ridge regression.

To solve the given minimization problem, we need to find the value of β that minimizes the objective function. The objective function can be written as:

$$Y^TY - - - Y^TX\beta - - - \beta^TX^TY + \beta^TX^TX\beta + \lambda\sum_{i=1}^{n}\beta_i^2$$

This is a typical objective function in ridge regression, a method used in linear regression to introduce regularization. The regularization term, $\lambda\sum_{i=1}^{n}\beta_i^2$, helps to prevent overfitting by penalizing large coefficients.

To minimize this function with respect to β, we take its derivative and set it to zero. Let's start by taking the derivative with respect to β:

$$\frac{\partial}{\partial\beta}\left(Y^TY - - - Y^TX\beta - - - \beta^TX^TY + \beta^TX^TX\beta + \lambda\sum_{i=1}^{n}\beta_i^2\right) = 0$$

Solving this, we get:

$$-2X^TY + 2X^TX\beta + 2\lambda I\beta = 0$$

where I is the identity matrix. Simplifying, we have:

$$X^TX\beta + \lambda I\beta = X^TY$$

This leads to the normal equation for ridge regression:

$$(X^TX + \lambda I)\beta = X^TY$$

To find β, we solve this equation:

$$\beta = (X^TX + \lambda I)^{-1}X^TY$$

Chapter 9

Linear Algebra II

Mid Terms Fall 2023 – Henry Baker

M4DS Mid Terms Revision: Session 9
Linear Algebra II

9.1 Linear Dependence

Linear dependence means redundant information: at least one of the vectors can be expressed as a linear combination of the others.

A set of vectors is linearly independent if and only if there is **no** redundancy in the information they provide.

If even one row (or column) of a matrix is linearly dependent on others, then the entire set of rows (or columns) is considered linearly dependent. For a matrix to be completely linearly independent, every single row (or column) must be linearly independent of the others.

For square matrices, another way to think about this is in terms of the determinant. A square matrix is linearly independent (and hence invertible) if and only if its determinant is non-zero. If even one row or column is linearly dependent, the determinant of the matrix will be zero, indicating linear dependence.

TL;DR: Linear Dependence: there exists a set of scalar coefficients, not all zero, such that a weighted sum of the columns (or rows) equals a zero vector

$$c_1v_1 + c_2v_2 + \cdots + c_nv_n = 0$$

- Let the sets $S = v_1, v_2, \cdots, v_n$ be a set of vectors.
- Members of the set s are **linearly dependent if at least one of the members of the set can be written as a linear combination of the other members.**
- i.e. if no scalar multiple of one vector can be added to a scalar multiple of the other to produce the zero vector, unless both scalars are zero.

- In simpler terms, one vector cannot be expressed as a multiple of the other.
- formally, Linear Dependence: if and only if

$$c_1v_1 + c_2v_2 + \cdots + c_nv_n = 0$$

where at least one of the c_is is non-zero (non-trivial solution)

- intuitively:
  - for Linear Dependence, although it looks like we are constraining oursleves by setting the condition "= 0", we are in fact not. We can just throw in some given/arbitrary scalar values for c that make this true, i.e. we can simply just rescale some of the vectors such that they cancel out other (and thus reach 0). This is because they are ultimately acting on the same dimensional plane; you can rescale them to express the movement of the other(s). **Thus, by saying that some rescaled combination of vectors come to 0, we are saying that they are acting on / moving across / spanning the same dimensions; this is thus redundant information, and thus the columns are linearly dependent.**
  - Whereas, for Linear Independence, we are in fact **more constrained: there is NO way to just scale vectors such that the above equation is true, in that they don't operate on the same dimensions, so can't just be rescaled to counter the movements/effects of another(s)** (and thus reach 0). Thus by NOT coming to 0, we are specifying that each column here must provide new information re: movements through dimensional space. Each has to be unique, in that you cannot rescale another vector to express that same information (if you could, you could rescale that vector in the opposite direction to the initial vector to reach 0.) Thus by saying it cannot = 0, we are saying it has to be unique; thus linear independence.
  - Eg — two Linearly Independent vectors: $\begin{bmatrix}4\\0\end{bmatrix}$ and $\begin{bmatrix}0\\3\end{bmatrix}$ no way to express a function that gets us from 4, 0 to 0, 3; **the only unique solution to get them to equal each other is by setting both C_1 and C_2 to 0** — which is the trivial example which doesn't satisfy linear dependence.
  - Linearly Independent vectors represent arrows pointing in different dimensions; they are different pieces of information.

The set $\left\{\begin{bmatrix}2\\3\end{bmatrix}, \begin{bmatrix}4\\6\end{bmatrix}\right\}$ is linearly dependent.

The set $\left\{\begin{bmatrix}0\\3\end{bmatrix}, \begin{bmatrix}4\\0\end{bmatrix}\right\}$ is linearly independent.

The set $\left\{\begin{bmatrix}0\\3\end{bmatrix}, \begin{bmatrix}4\\0\end{bmatrix}, \begin{bmatrix}2\\2\end{bmatrix}\right\}$ is linearly dependent.

9.1.1 Examples of proving linear (in)dependence

1) set up the equation $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 \cdot + c_n\mathbf{v}_n = \mathbf{0}$
2) solve the system of equations: take one c_nv_n over to the other side so you can write one of the c_nv_n as a function of the others, then plug back in, etc.

E.g.1 Proving Linear Dependence: finding non-trivial solution

Vectors: $\mathbf{v}_1 = \begin{bmatrix}1\\2\\3\end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix}2\\4\\6\end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix}-1\\-2\\-3\end{bmatrix}$

Step 1: Set up the equation

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 = \mathbf{0}$$

This leads to the system:

$$\begin{aligned} 1c_1 + 2c_2 - - - 1c_3 &= 0 \\ 2c_1 + 4c_2 - - - 2c_3 &= 0 \\ 3c_1 + 6c_2 - - - 3c_3 &= 0 \end{aligned}$$

Step 2: Solve the system:

- **Finding a Non-Trivial Solution:**
  - e.g: let's assign $c_2 = 1$ and $c_3 = 0$...
  - ...then from the first equation: $c_1 + 2(-1) + -1(0) = 0$...
  - ...gives $c_1 = 2$.
  - Thus $c_1 = 2; c_2 = -1; c_3 = 0$ is a valid solution
  - when dependent, a non-trivial solution always exists (ie other than $c_1 = c_2 = c_3 = 0$)
- **Observation:**
  - each row is a multiple of the first row: not independent.
  - example of a homogeneous system, where all the equations are multiples of each other

We were able to find the non-zero coefficients c_1, c_2, c_3 that satisfy the linear combination $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 = \mathbf{0}$.
We found a non-trivial solution where not all c_i are 0: so they are linearly dependent.

E.g.2 Proving Linear Independence: only a trivial solution

$$\textbf{Vectors:} \quad \mathbf{v}_1 = \begin{bmatrix}1\\0\\0\end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix}0\\1\\0\end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix}0\\0\\1\end{bmatrix}$$

Step 1: Set up the equation

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 = \mathbf{0}$$

This leads to the system:

$$\begin{aligned} 1c_1 + 0c_2 + 0c_3 &= 0 \\ 0c_1 + 1c_2 + 0c_3 &= 0 \\ 0c_1 + 0c_2 + 1c_3 &= 0 \end{aligned}$$

Simplify:

$$\begin{aligned} c_1 &= 0 \\ c_2 &= 0 \\ c_3 &= 0 \end{aligned}$$

Step 2: Solve the system:
Since the only solution to this system is the trivial solution $c_1 = c_2 = c_3 = 0$, the vectors are linearly independent.
Observation:
Can see that there is no way to multiply one of the vectors to map it onto another = independent.

E.g.3 Proving Linear Dependence (Mixed Case)

$$\textbf{Vectors:} \quad \mathbf{v}_1 = \begin{bmatrix}1\\3\\4\end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix}2\\6\\8\end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix}0\\1\\2\end{bmatrix}$$

Step 1: Set up the equation

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 = \mathbf{0}$$

This leads to the system:

$$\begin{aligned} c_1 + 2c_2 &= 0 \\ 3c_1 + 6c_2 + c_3 &= 0 \end{aligned}$$

Step 2: Solve the system:
Finding a Non-Trivial Solution:

- Solving the first equation for c_1 gives $c_1 = -2c_2$.

- Substitute c_1 into the second equation: $3(-2c_2) + 6c_2 + c_3 = 0$ simplifies to $c_3 = 0$.
- Substitute c_1 and c_3 into the third equation: it holds true for any value of c_2.

We found a non-trivial solution where not all c_i are 0: so the vectors are linearly dependent.
Observation:
Can see that v_1 and v_2 are linearly dependent.

Determining Linear Dependence or Independence

Given Vectors:
$\mathrm{v}_1 = \begin{bmatrix}0\\3\end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix}4\\0\end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix}2\\2\end{bmatrix}$
Step 1: Set up the equation for linear combination:

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 = \mathbf{0}$$

This leads to the system of linear equations:

$$0c_1 + 4c_2 + 2c_3 = 0$$
$$3c_1 + 0c_2 + 2c_3 = 0$$

Step 2: Solve the system:

- From the first equation, we get: $2c_2 + c_3 = 0 \rightarrow c_3 = -2c_2$.
- Substituting $c_3 = -2c_2$ into the second equation gives $3c_1 - - - 4c_2 = 0 \rightarrow 3c_1 = 4c_2$.

Conclusion:
The system does not provide a clear non-trivial solution for c_1, c_2, c_3, suggesting that the vectors are linearly independent under typical circumstances.

9.2 The Span

9.2.1 in terms of Vector Space / Basis vectors

We define the span of a the set of vectors S as **all linear combination of the vectors in the set**.

- the span of v_1 and v_2 incl all vectors that can be formed by taking $a \cdot v_1 + b \cdot v_2$, where both vectors are linearly independent (independence not a hard requirement, but otherwise, we are just including redundant info if you think about it geometrically)
- the span of 2 linearly independent vectors $v_1 \in \mathbb{R}^2$ and $v_2 \in \mathbb{R}^2$ is $\mathbb{R}^2$: this means that any vector in the 2D space can be expressed as a linear combination of v_1 or v_2
- this implies that v_1 and v_2 form a basis for R^2 since they can generate the entire 2D space through their linear combos (i.e. they provide a complete representation of the 2D space)

9.2.2 in terms of the Spanning Set

Going the other way, if V is a vector space and S a set of vectors in V, then we say that S **is a spanning set for V if all the dimensions of V can be represented by linear combinations of S**.

- a Spanning Set S of a vector space V is a set of vectors, such that every vector in V can be expressed as a linear combo of the vectors in S.
- if S spans V, it means that S contains enough linearly independent vectors to cover the entire vector space V
- generally a spanning set S must contain as least as many elements as the linearly independent vectors from V... but a Spanning set not a minimal thing: as long as it contains n linearly independent vectors then it can also include additional vectors, but they will be redundant info.

There are exactly n linearly independent vectors in $\mathbb{R}^n$ (no more, no less!)

- ... 2 in $\mathbb{R}^2$
- ... 3 in $\mathbb{R}^3$
- .... n in $\mathbb{R}^n$
- these vectors = the basis for the space
- Geometrically: visualise $\mathbb{R}^2$ vector space: every movement can be made up of linear combination is movements along y and along x — these are the fundamental building blocks which are **orthogonal (linearly independent)** — everything else is linearly dependent.
- linearly dependent vectors present redundant information, in that they are **colinear** (geometric interpretation in R2) or **co-planar** (geometric interpretation of R2+) to the basis of the space,
- i.e. one of the vectors of the set can be represented as a combination of others **(it doesn't give us any new dimensions)**
- A set of vectors is said to be linearly independent if no vector in the set can be written as a linear combination of the others: none of the vectors in the set is redundant in terms of the information it provides about the vector space.

The **dimension of a vector space V is defined as the maximum number of linearly independent vectors in V**. It represents the number of coordinates needed to specify any vector in the space uniquely.

- For a set S to span a vector space V, it must contain at least as many vectors as the dimension of V.
- However, having exactly as many vectors in S as the dimension of V does not automatically guarantee that S spans V. Those vectors also need to be linearly independent.

A Spanning Set S must contain at last as many elements as the linearly independent vectors from V.

There are exactly n linearly independent vectors in R^n

Dimension of a vector space V is defined as the maximum number of linearly independent vectors in V. It represents the number of coordinates needed to specify any vector in the space uniquely

9.3 The Determinant

The determinant is a scalar value that can be computed from the elements of a square matrix. Often denoted as $det(\mathbf{A})$ or $|\mathbf{A}|$
= a way to multiply entries of the matrix to produce a single number.

9.3.1 Uses

- **Linear Transformations** — the absolute value of the determinant of a matrix representing a linear transformation reflects how the transformation changes the area (in 2D) or volume (in 3D) of shapes. A negative determinant indicates that the transformation also involves a reflection.
- **Eigenvalues and Characteristic Polynomial**: The characteristic polynomial of a matrix, used to find its eigenvalues, is derived from its determinant. (I think this is related to the above, as a determinant tells you nature of the transformation, and eigenvalue-eigenvector pairings do the same??)
- **to determine if a matrix has an inverse** (is non-singular). A square matrix is singular (non-invertible) if and only if its determinant is zero. (see Savov p. 169-171)
- **to check for linear independence** — related to checking if has inverse: as a matrix has an inverse if and only if its columns are linearly independent (see Savov p. 167)
- **to compute areas and volumes** — the absolute value of the determinant of a matrix formed by vectors representing the sides of a parallelogram (in 2D) or a parallelepiped (in 3D) gives the area (in 2D) or volume (in 3D) of that geometric shape. (see Savov p. 161)
- **solving systems of equations** (see Cramer's Rule, Savov p. 166)

9.3.2 Calculation

For a 1x1 Matrix (scalar):

$$A = [a]$$

the determinant is simply the value of that single element:

$$\det(A) = a$$

For a 2x2 Matrix:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

$$\det(A) = ad - - - bc$$

For a 3x3 Matrix:

$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

$$\det(A) = a(ei - - - fh) - - - b(di - - - fg) + c(dh - - - eg)$$

The matrices bigger than 3x3 ,the determinant calculation becomes complex and typically calculated using methods such as expansion by minors or Laplace expansion.

go through https://www.khanacademy.org/math/linear-algebra/vectors-and-spaces/linear-independence/v/linear-algebra-introduction-to-linear-independence from 7 mins onwards

- general intutition product of top left x bottom right, minus product of top right x bottom left...
- ... but this only works properly for 2x2 matrix, after that it gets complicated...
- ... for 3x3: take each element in term as a constant, multiply the remaining minor, alternately subtract then add each element.

General formula:

$$det(A) = \sum_{j=1}^{n} (-1)^{1+j} a_{1j} M_{1j}$$

where M_ij is called the minor associated with a_i: the determinant of the submatrix generated by removing row i and column j from the matrix A.

In the determinant formula, the $(-1)^{1+j}$ just reverses the sign starting with +, then -, then +, etc. This means we go up to the power of $+j$ where j is the number of columns in the matrix, so 1+[odd value column] is to raise it to an even power.
2x2 Matrix:

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - - - a_{12}a_{21}$$

3x3 Matrix

$$\det(A) = a \cdot \det \begin{vmatrix} e & h \\ f & i \end{vmatrix} - - - b \cdot \det \begin{vmatrix} d & g \\ f & i \end{vmatrix} + c \cdot \det \begin{vmatrix} d & g \\ e & h \end{vmatrix}$$

$$\rightarrow \det(A) = a(ei - - - fh) - - - b(di - - - fg) + c(dh - - - eg)$$

Resources for determinants: Savov, p. 158-161

9.4 Matrix Inverse

$$A^{-1}A = I$$

Not all matrices invertible: only exists if and only if:

$$det(A) \neq 0$$

Inverse defined:

$$A^{-1} = \frac{1}{\det(A)} \text{adj}(A)$$

Where adj(A) is the adjugate matrix

The inverse of a 2x2 matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is given by

$$A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

where $\det(A) = ad - - - bc$ is the determinant of A. The matrix A is invertible if and only if $\det(A) \neq 0$.

Given the matrix $A = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}$.
The determinant of A is calculated as

$$\det(A) = 3 \times 3 - - - 0 \times 0 = 9$$

The formula for the inverse of a 2x2 matrix is

$$A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Applying this to matrix A:

$$A^{-1} = \frac{1}{9} \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{3} \end{pmatrix}$$

Thus, the inverse of matrix A is $A^{-1} = \begin{pmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{3} \end{pmatrix}$.

OR AN EASIER WAY is just to eyeball it for easy matrices:

$$\begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{3} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

So, what do you have to multiply the values, to get to 1, 0, 0, 1.

9.4.1 Adjugate Matrix

= the tranpose of the matrix of cofactors:

$$adj(A) = C^T$$

... see Savov p169-171 for matrix of cofactors

9.5 To know

1. **a matrix has an inverse, if and only if it has a non-zero determiant**
2. **a matrix has an inverse, if and only if its columns are linearly independent**
3. **thus, a matrix has a non-zero determinant if and only if its columns are linearly independent.**

This is actually intuitive: if you think of linear dependence as about vectors/columns providing information about the vector space, and the determinant tells you about the nature of the linear transformation.... if you have redundant (dependent) columns then this means the matrix is not a clean representation of a linear transformation, qhich is they it has implications for its determinant.

So the direction of causality is:
Linear Independence → Non-zero Determinant → has Inverse

9.5.1 Conditions under which a matrix is invertible:

1. **matrix is square**
2. **columns are linearly independent (full rank)**
3. **non-zero determinant**

Given the matrix $A = \begin{pmatrix} 1 & 0 \\ 2 & 0 \end{pmatrix}$, the determinant of A is calculated as

$$\det(A) = 1 \times 0 - - - 0 \times 2 = 0$$

Since the determinant of A is zero, the columns are not linearly independent, the matrix A is not invertible.

Proof by contradiction:

Suppose a linearly dependent Matrix *is invertible*...

Let A be a square matrix with dependent columns: $a_1, a_2 \cdots a_n$

$$A = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}_i$$

By definition of Linear Independence:

$$c_1 a_1 + c_2 a_2 + \ldots + c_n a_n = 0$$

and $c_i \neq 0$ for some i

We can write this equivalently as:

$$A\mathbf{c} = \mathbf{0}, \text{where } \mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}$$

This equation means that c is a vector that, when multiplied by A, results in the zero vector. This is the definition of a vector being in the null space of A.

If A was invertible, then we could multiply both sides by its inverse:

$$\begin{aligned} A^{-1}A\mathbf{c} &= A^{-1}\mathbf{0} \\ I\mathbf{c} &= \mathbf{0} \\ \mathbf{c} &= \mathbf{0} \end{aligned}$$

BUT, by the definition of linear dependence, c cannot be the zero vector.
So we have a contradiction.
So A must be Linearly independent.

9.6 Eigenvalues and Eigenvectors

Eigenvalue-vector pairing effectively reduces matrix to a single scalar (λ), along a particular path (provided by the vector).

Eigenvectors represent directions in the data space, and eigenvalues indicate the magnitude of variance along these directions.

The matrix describes a transformation / a journey / a movement, and that lambda value is the most direct route to replicate that journey (but it is specific / associated with that vector)

Definition of an eigenvector-eigenvalue pairing:

$$Av = \lambda v$$

Where v is an eigenvector, and λ is an eigenvalue (a scalar)

I.e.

- if you multiply the matrix by a specific vector (eigenvector) you get the same result as multiplying that same vector by a single number (eigenvalue).
- = a specific vector-scalar pairing that reduces the matrix to a single number (for a given vector)
- Think of the eigenvector as a projected movement in a specific direction, and the eigenvalue as the extent / magnitude / degree of that movement.
- we are reducing the matrix to a single scalar value (its essence), which it can be reduced to when transforming a specific vector.
- Alternatively, think of A as some transformation of v: it projects v from one place to another. **If A is the journey, then the eigenvalue λ is the most direct route; the essence of the transformation** (the eigen = the self)

For example:

$$\underbrace{\begin{bmatrix}1 & 2\\ 4 & 3\end{bmatrix}}_{A}\underbrace{\begin{bmatrix}1\\ 2\end{bmatrix}}_{v} = \underbrace{5}_{\lambda}\underbrace{\begin{bmatrix}1\\ 2\end{bmatrix}}_{v}$$

In this case, $v = \begin{bmatrix}1\\ 2\end{bmatrix}$ is an eigenvector and $\lambda = 5$ is an eigenvalue for $A = \begin{bmatrix}1 & 2\\ 4 & 3\end{bmatrix}$

NB Chat GPT gave the below automatically when I fed it the above, which is interesting, it's presenting the data matrix A, a given vector, the eigen value, and the eigen vectors
For example:

$$\underbrace{\begin{bmatrix}1 & 2\\ 4 & 3\end{bmatrix}}_{A}\underbrace{\begin{bmatrix}v_1\\ v_2\end{bmatrix}}_{v} = 5\underbrace{\begin{bmatrix}\lambda_1\\ \lambda_2\end{bmatrix}}_{\lambda}$$

9.6.1 Conditions for existence of (non-trivial) Eigenvector/values

Not every matrix can be decomposed into eigenvalue-vector pairs.The conditions for the existence of non-trivial eigenvectors and eigenvalues for a square matrix A are:

1. The matrix A must be a **square** matrix. The existence of eigenvalues and eigenvectors is a fundamental property of square matrices. Everything comes from there??
2. There must exist a **scalar λ such that** $\det(\lambda\mathbf{I} - - - \mathbf{A}) = 0$, where I is the identity matrix of the same dimension as A. This is derived from the **characteristic equation** $\det(\lambda I - - - A) = 0$.
3. For each eigenvalue λ, there must exist a non-zero solution $\mathbf{v}$ to the equation $(A - - - \lambda I)\mathbf{v} = 0$.
4. Eigenvectors must be non-zero. The zero vector is not considered an eigenvector.
5. The algebraic multiplicity of an eigenvalue must be at least as great as its geometric multiplicity, which is the number of linearly independent eigenvectors corresponding to that eigenvalue.

Proof for finding eigenvalues

Start with how we defined the eigenvector and eigenvalue:

$$Av = \lambda v$$

Simply rearrange:

$$0 = \lambda v - - - Av$$

Now, let's replace λ with λI_n (this is the same):

$$0 = \lambda I_n v - - - Av$$

Which allows us to write:

$$\begin{aligned} 0 &= (\lambda I_n - - - A)v \\ 0 &= \underbrace{(\lambda_n I - - - A)}_{\text{call this } \mathbf{B}\mathbf{v}} \\ 0 &= \mathbf{B}\mathbf{v} \end{aligned}$$

This is trivially satisfied by $v = 0$.
We're looking for some non-trivial solution; i.e. some non-zero v

But more importantly, this is *precisely* the definition of **linear dependence for the columns of B**, where our current vector v is the vector of constants $c_1, c_2, \ldots, c_n$.

If a matrix has linearly dependent columns, then its determinant is zero; if we want a nontrivial solution to the above (IS THIS THE DEFINITION / EXISTENCE OF EIGEN VECTORS?), we require:

$$det(\lambda I_n - - - A) = 0$$

$$det(\lambda I_n - - - A) = 0$$

Same as

$$det(A - - - \lambda I_n) = 0$$

This is the eigenvalue equation.
It is a generic form of the characteristic equation of a matrix (IS IT?).
Eigenvalues are the values of $\lambda thatsatisfytheequation.$

where:

- A = *square* $n \times n$ *matrix; whose eigenvalues we are interested in finding.*
- λ = *(scalar) eigenvalue of matrix A; goal is to finv values of* λ *that satisfy this equation.*
- I_n = *identify matrix of same size as A*
- $\rightarrow \lambda I_n - - - A)$ = ***matrix formed by subtracting the matrix*** A ***from the scalar*** λ ***multiplied by the identity matrix*** I_n. *The result is a matrix where the diagonal elements are* $\lambda - - - a_{ij}$ *where* a_{ij} *are diagonal elements of A. WHAT ABOUT NON-DIAGONAL ELEMENTS*
- *the determinant of that matrix*
- *set that determinant to 0* $\rightarrow$ *becomes the characteristic equation. Solve to find eigenvalues of A.*

Essentially: looking for values of λ that make the determinant of $(\lambda I_n - - - A)$ equal to zero.

Explanation

The determinant of a square matrix is a scalar value that provides important information about the matrix, including whether its columns (or rows) are linearly independent. When a matrix has linearly dependent columns, its determinant is zero.

- **Linear Dependence:** If the columns (or rows) of a matrix are linearly dependent, it means that at least one column (or row) can be expressed as a linear combination of the others. In other words, there exists a set of scalar coefficients, not all zero, such that a weighted sum of the columns (or rows) equals a zero vector.

$$c_1 v_1 + c_2 v_2 + \cdots + c_n v_n = 0$$

- **Determinant and Linear Transformations:** The determinant can be interpreted as a scaling factor of the linear transformation represented by the matrix. If a matrix has linearly dependent columns, it means that the space it maps to is "flattened" in at least one dimension (since one column is a combination of others). This "flattening" means the matrix compresses the space into a lower dimension, resulting in a volume of zero in the original space.
- **Geometric Interpretation:** Geometrically, the determinant of a matrix can be thought of as the volume of the parallelepiped spanned by its columns (in the case of a 3x3 matrix)

or rows. If columns are linearly dependent, this parallelepiped collapses into a lower-dimensional space (like a line or a plane), which has zero volume in the context of the original space.

- **Properties of the Determinant Function:** The determinant function has a property that if two columns (or rows) of a matrix are identical, or one column is a scalar multiple of another, the determinant is zero. This is a specific case of linear dependence. In a more general sense, any linear combination leading to a redundant or dependent column (or row) results in a zero determinant.

9.6.2 Finding Eigenvalues/vectors

Start with condition:

$$det(\lambda I_n - - - A) = 0$$

Applying this to our matrix:

$$\det\left(\lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - - - \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}\right) = 0$$

Computing this yields the **characteristic polynomial**:

$$\lambda^2 - - - 4\lambda - - - 5 = 0$$

Giving us $\lambda = 5, \lambda = -1$.

These are the eigenvalues.

9.6.3 Eigenspace

From Eigenvalues $\rightarrow$ *Eigenvectors*.
There will be more than one eigenvector corresponding to a given eigenvalue λ.

The **eigenspace of λ, E_λ, is all vectors v that satsify the condition** $Av = \lambda v$

So, we find all vectors v such that

$$\begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 5 \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

Which is the span of

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

See below for guidance:

Step 1: Matrix Multiplication

First, perform the matrix multiplication on the left side of the equation.

$$\begin{bmatrix}1 & 2\\4 & 3\end{bmatrix}\begin{bmatrix}v_1\\v_2\end{bmatrix} = \begin{bmatrix}1\cdot v_1 + 2\cdot v_2\\4\cdot v_1 + 3\cdot v_2\end{bmatrix}$$
$$= \begin{bmatrix}v_1 + 2v_2\\4v_1 + 3v_2\end{bmatrix}$$

Step 2: Scalar Multiplication

Next, perform the scalar multiplication on the right side of the equation.

$$5\begin{bmatrix}v_1\\v_2\end{bmatrix} = \begin{bmatrix}5v_1\\5v_2\end{bmatrix}$$

Step 3: Setting Up the Equation

Now, equate the results of the left and right side multiplications.

$$\begin{bmatrix}v_1 + 2v_2\\4v_1 + 3v_2\end{bmatrix} = \begin{bmatrix}5v_1\\5v_2\end{bmatrix}$$

Step 4: Solving the System of Equations

This results in a system of linear equations:

$$v_1 + 2v_2 = 5v_1$$
$$4v_1 + 3v_2 = 5v_2$$

These equations can be solved to find the values of v_1 and v_2.

9.6.4 Putting it all together

$$\underbrace{\begin{bmatrix}1 & 2\\4 & 3\end{bmatrix}}_{A}\underbrace{\begin{bmatrix}v_1\\v_2\end{bmatrix}}_{\mathbf{v}} = \underbrace{5}_{\lambda}\underbrace{\begin{bmatrix}v_1\\v_2\end{bmatrix}}_{\mathbf{v}}$$

We have found one eigenvector-eigenvalue combination for the matrix A (there was another for $\lambda = -1$).
see Khan academi links p23

9.7 Eigen decomposition

9.7.1 Definition

We can use this eigenvalue-vector pairings to find the **eigendecomposition** of A

$$A = Q\Lambda Q^{-1}$$

Where:

- A is a square matrix.
- Λ is a diagonal matrix with eigenvalues of A on the main diagonal (and 0 everywhere else). *(An $n \times n$ square will have n eigenvalues, so Λ will also be $n \times n$)*
- Q is a matrix where the columns are eigenvectors of A, correspeonding to the eigenvalues in Λ

9.7.2 The conditions for eigendecomposition

1. The matrix A must be **square** ($n \times n$.
2. The matrix must **have n linearly independent eigenvectors**, where n is the size of the matrix.
3. The matrix A is diagonalizable if it is similar to a diagonal matrix Λ, i.e., there exists an invertible matrix Q such that $Q^{-1}AQ = \Lambda$. This is possible if and only if A has n independent eigenvectors. (this comes from having n independent eigenvectors condition above.)

9.7.3 Calculation

Taking our matrix A from before, we write:

$$A = \underbrace{\begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix}}_{Q} \underbrace{\begin{bmatrix} 5 & 0 \\ 0 & -1 \end{bmatrix}}_{\Lambda} \underbrace{\begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix}^{-1}}_{Q^{-1}}$$

- recalling our solutions $\lambda_1 = 5$ and $\lambda_2 = -1$,
- and the corresponding eigenvectors $v_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ (which we found above)
- and $v_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

9.7.4 Why?

Reducing a matrix to its essence is efficient for computing further transformations

9.8 Singular Value Decomposition of a Matrix

9.8.1 SVD Definition

Eigendecomposition only works for square matrices.
SVD allows us to get a matrix of any dimension to its essence.

$$A = U\Sigma V^T$$

where $A \in \mathbb{R}^{m \times n}$

- U is a matrix containing left singular vectors of A (contains the m eigenvectors of the square matrix AA^T) $(n \times n)$
- V^T is a matrix containing the right singular vectors of A (contains the n eigenvectors of the square matrix A^TA $(m \times n)$
- Σ is a matrix where the diagonal entries are the singular values of A (square roots of the eigenvalues of A^TA (or AA^T) on the diagonal. $(n \times n)$

Dimensions:

Matrix A: Suppose A is an $m \times n$ matrix.
Matrix U: U is an $m \times m$ orthogonal matrix. The columns of U are the left singular vectors of A.
Matrix Σ: Σ is an $m \times n$ diagonal matrix. The non-zero elements of Σ (which are on the diagonal) are the singular values of A. If $m > n$, the additional rows will be filled with zeros. If $n > m$, there will be additional zero columns.
Matrix V^T: V^T is the transpose of an $n \times n$ orthogonal matrix V. The columns of V (or the rows of V^T) are the right singular vectors of A.
So, in the SVD $A = U\Sigma V^T$:

- U aligns with the row dimension of A.
- Σ matches the overall dimensions of A, but is diagonal.
- V^T aligns with the column dimension of A.

Think of AA^T or A^TA as "squaring" the matrix in two ways: 1) it is like A^2 in some ways, 2) it crucially transforms a non-square matrix into a square, so it can be decomposed.
This is why we use singular values of A in *Sigma*'s diagonal: singular values are square roots of the eigenvalues of the "squared A" matrix (AA^T or A^TA)

9.8.2 SVD Is Profoundly Informative About the Structure and Dimensionality of Your Data

- SVD like an an x-ray of a matrix:
  - U **and** V^T **represent how much it rotates an object** when multiplied by it, as they contain the singular vectors (like eigenvectors, but for non-square matrices)
  - Σ **represents associated scaling**, as it contains the associated singular values across its diagonal. (NB remember, these are the square roots of the A's eigenvalues)
- Large singular values correspond to large "strength" of the transformation that A makes along the subspace created by the associated left and right singular vectors.
  - NB: a 0 = redundant
  - the **number of non-zero singular values gives the rank of the matrix, or the number of linearly independent columns**
  - if **many singular values are zero (or very small), it suggests redundancy** in the data: data can be represented in a lower dimensional space without much loss of info

9.9 Lab: PCA as Eigendecomposition

Mid Terms Fall 2023 – Henry Baker

M4DS Finals Revision: Session 9
Linear Algebra II
Lab: Principle Component Analysis

9.10 Set Up: PCA applied to image compression

- PCA = dimension reduction technique. Efficiently summarizes and describes large datasets.
- PCA has large variety of applications in genomics, facial recognition, computervision, finance and econ, climate science, social science, etc.
- PCA foundations:
  - maximisation variance
  - eigendecomposition

9.11 PCA as a Variance-Maximisation Problem

9.11.1 As variance in a matrix

- Take a data matrix X $(n \times p)$ — centered around the mean (i.e. demeaned).
- Challenge: **how to convey as much info about X as possible in only 1 column** of data (one n lengthed vector)? What vector to choose?
  - suppose X contains one column that was pretty similar across observations (e.g. everyone's a Hertie student)
  - and another that was q different (e.g. country of origin).
  - **we would want to keep the column with *higher variance***.
- Now, rather than keeping one column from X, **we can take a column $\mathbf{z}_1$ that is a *linear combination* of all the columns of x.**
  - A linear combination is an expression constructed from a set of terms, by multiplying each term by a constant and adding the results. **It's a way of combining a set of vectors by scaling and adding them together**.
  - Each column/vector is multipled by a corresponding coefficient ϕ, after which the modified columns are added together, to form a new column/vector.

  Here, this means that:

  $$\mathrm{z}_1 = \phi_{11}x_1 + \phi_{21}x_2 + \ldots + \phi_{p1}x_p$$

  where:

  * $x_1, x_2, \ldots, x_p$ are the columns of the matrix X. **They are vectors**
  * $\phi_{11}, \phi_{21}, \ldots, \phi_{p1}$ are the coefficients (the scales by which we multiply each column to be added). **They are scalar values**
  * The product $\phi_{11}x_1, \phi_{21}x_2, \ldots$ scales each column of X by its respective coefficient. **They are vectors**
  * The sum of these products, $\phi_{11}x_1 + \phi_{21}x_2 + \ldots + \phi_{p1}x_p$, creates the new vector z_1. **It is a new vector - *adding vectors together is performed element-wise***

  - **$\mathbf{z}_1$ is formed by 1) multiplying each column in X by its corresponding coefficient ϕ, then 2) each scaled column is added together**.

- Another way of visualising this is :

$$\begin{aligned} \mathrm{z}_{11} &= \phi_{11}x_{11} + \phi_{21}x_{12} + \ldots + \phi_{p1}x_{1p} \\ \mathrm{z}_{21} &= \phi_{11}x_{21} + \phi_{21}x_{22} + \ldots + \phi_{p1}x_{2p} \\ &\vdots \\ \mathrm{z}_{n1} &= \phi_{11}x_{n1} + \phi_{n1}x_{n2} + \ldots + \phi_{n1}x_{np} \end{aligned}$$

  Where
  - each ϕ is a constant.
  - z_i is just a vector containing all the z_{ik} values. In this sense **it stores data from all of the individual x_k elements of the original matrix X**

- This generates a new vector z_1 because each x_i is a vector, and when you multiply a vector by a scalar ϕ the result is a vector.

$$\mathbf{z}_1 = \begin{bmatrix} z_{11} \\ z_{21} \\ \vdots \\ z_{n1} \end{bmatrix}$$

  Here z_1 is a column vector **where each element is a linear combination of the columns of the matrix X based on the PCA loadings (ϕ)**.

- Where each element of z_1: z_{i1} (where $i = 1, 2 \cdots n$ corresponds to the rows of the data), as:

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \ldots + \phi_{p1}x_{ip}$$

9.11.2 As a maximisation problem

- Which linear combination would you choose? The **linear combination (ie the values of $\phi_{11}, \phi_{21}, \cdots, \phi_{p1}$) that maximises the variance of your z_1 vector**.
  - by choosing different coefficients of ϕ, you can emphasize certain features over others
  - so for PCA: we are trying to find/choose coefficients ϕ that results in greatest variance across z_1 vector.
  - it in effect takes a bit of each column [develop this explanation].
  - or, if you purpose was feature engineering, you can uncover hidden structures in the data, such that z_1 reveals patterns or relationships that aren't obvious in the original dataset.

That is, you **solve the maximisation problem:**

$$\begin{aligned} \max_{\phi_{11},\phi_{21},\ldots,\phi_{p1}} \mathrm{Var}(\mathrm{z}_1) &= \max_{\phi_{11},\phi_{21},\ldots,\phi_{p1}} \frac{1}{n}\sum_{i=1}^{n}(z_{i1} - - - \bar{\mathrm{z}}_1)^2 \\ &= \max_{\phi_{11},\phi_{21},\ldots,\phi_{p1}} \frac{1}{n}\sum_{i=1}^{n} z_{i1}^2 \end{aligned}$$

Where:

- first line is the definition of variance: $\bar{\mathrm{z}}_1$ is the column mean of z_1.

- second line follows from the fact that we demeaned our columns such that $\bar{z}_1 = 0$

So variance of z_1 (the PCA vector) can be calculated simply as the avg of the squared elements of z_{i1} (the elements of the PCA vector, which are themselves linear combination of the columns of the matrix X based on the PCA loadings)

9.11.3 Resulting components

Principal components are new, uncorrelated variables that are linear combinations of the original variables. They are aligned with the directions of maximum variance in the dataset. The maximal-variance z_1 is called the **first principle component**.

We call $\phi_1 = \begin{bmatrix} \phi_{11} \\ \phi_{21} \\ \vdots \\ \phi_{p1} \end{bmatrix}$ its loadings.

We **constrain the loadings so that their sum of squares is equal to one**, since allowing these elements to be arbitrarily large could result in an arbitrarily large variance:

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1$$

For constrained optimisation, see next session 10. But instead, here we will solve the same maximisation problem using eigendecomposition.

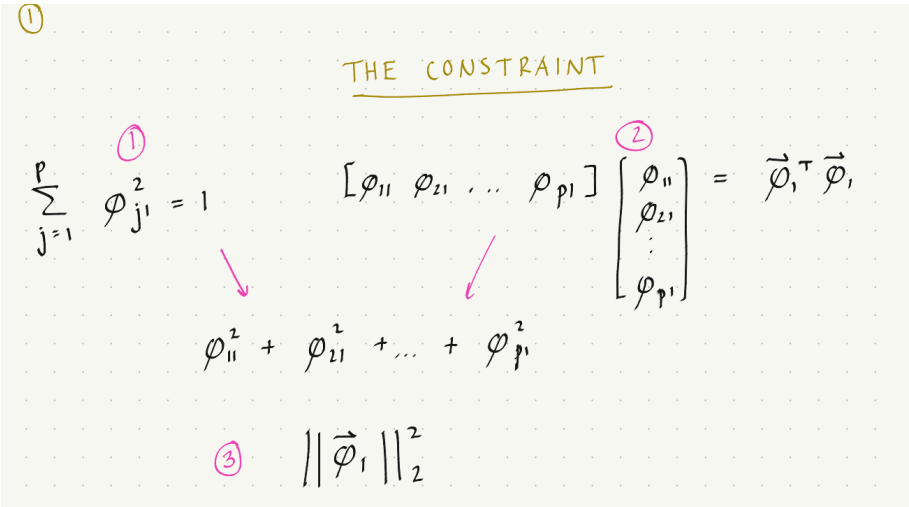


Figure 9.2: Enter Caption

9.12 PCA as Eigendecomposition of the Variance-Covariance Matrix

NB: this equivalently solves the above maximisation problem. Unlike regression application from Wk 8 which used matrix structure and multiplication properties to effectively reengineer the same calculation, this fundamentally leverages an alternate set of properties of eigendecompositions to fundamentally approach the problem from a different angle. Namely that eigendecomposition gives you all the linearly independent vectors within/UNDERLYING? a matrix (IS THIS RIGHT? DO YOU GET P EIGEN VECTORS / LINEARLY

INDEPENDENT VECTORS IF YOU HAVE A PXP MATRIX, AS EACH VECTOR HAS TO BE LINEARLY INDEPENDENT, OTHERWISE DETERMINANT COLLAPSES TO 0 AND YOU CANT DO EIGEN DECOMPOSITION??? OR IS IT JUST THAT EIGEN DECOMPOSITION EXTRACTS THE UNDERLYING LINEARLY INDEPENDENT VECTORS / TRANSFORMATIONS / MOVEMENTS OF ANY GIVEN DATA SET, WHICH JUST HAPPENS TO BE THE NATURE OF WHAT A PRINCIPLE COMPONENT IS?
Chat GPT:

- Eigenvectors associated with different eigenvalues are linearly independent. This is a fundamental property of eigenvectors. If a $p \times p$ has p distinct eigenvalues, it will have p linearly independent eigenvectors.
- However, if some eigenvalues are repeated (i.e., not all eigenvalues are distinct), the matrix may have fewer than p linearly independent eigenvectors. The matrix is then said to be "defective," and not every defective matrix has a complete basis of eigenvectors.
- Eigendecomposition of a matrix involves finding eigenvalues and their corresponding eigenvectors. If a $p \times p$ matrix has p distinct eigenvalues, it will have p linearly independent eigenvectors.
- In PCA, eigendecomposition is used to identify principal components, which are orthogonal and linearly independent directions in the dataset that maximize variance.
- The determinant of a matrix being zero indicates that the matrix does not have an inverse and is not full rank. A zero determinant is related to the presence of an eigenvalue of zero, but a zero eigenvalue does not necessarily mean the absence of a full set of linearly independent eigenvectors.
- PCA leverages eigendecomposition to extract linearly independent directions (principal components) that capture the most variance in the data, making it powerful for dimensionality reduction and feature extraction.

9.12.1 Derive a Variance-Covariance Matrix of X

Variance-covariance matrix of centered matrix X, is given by:

$$\begin{aligned} S &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - - - \bar{\mathbf{x}})(\mathbf{x}_i - - - \bar{\mathbf{x}})^T \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T \end{aligned}$$

Where:

- x_i is a p-length vector of the data to row i,
- $\bar{x}$ is a p-length vector of the means of the data taken over $i = 1$ to n (which all equal to 0 for centered data, hence second line)

Think of each column vector and a transpose vector being multiplied together to collapse into a single scalar value (the covariance between the two columns). This is a single element of the new matrix. So it contains a lot of info! Think of it as a particular form of collapsing of information...

Basically, for each row-observation and extracts the squared differences (ie the variation) for each column-variable, by averaging those differences (between n observations, across the variable) to give the variation across each vector expressed in a new matrix (since a vector and a transpose vector multiply out to a matrix, the output is a matrix).

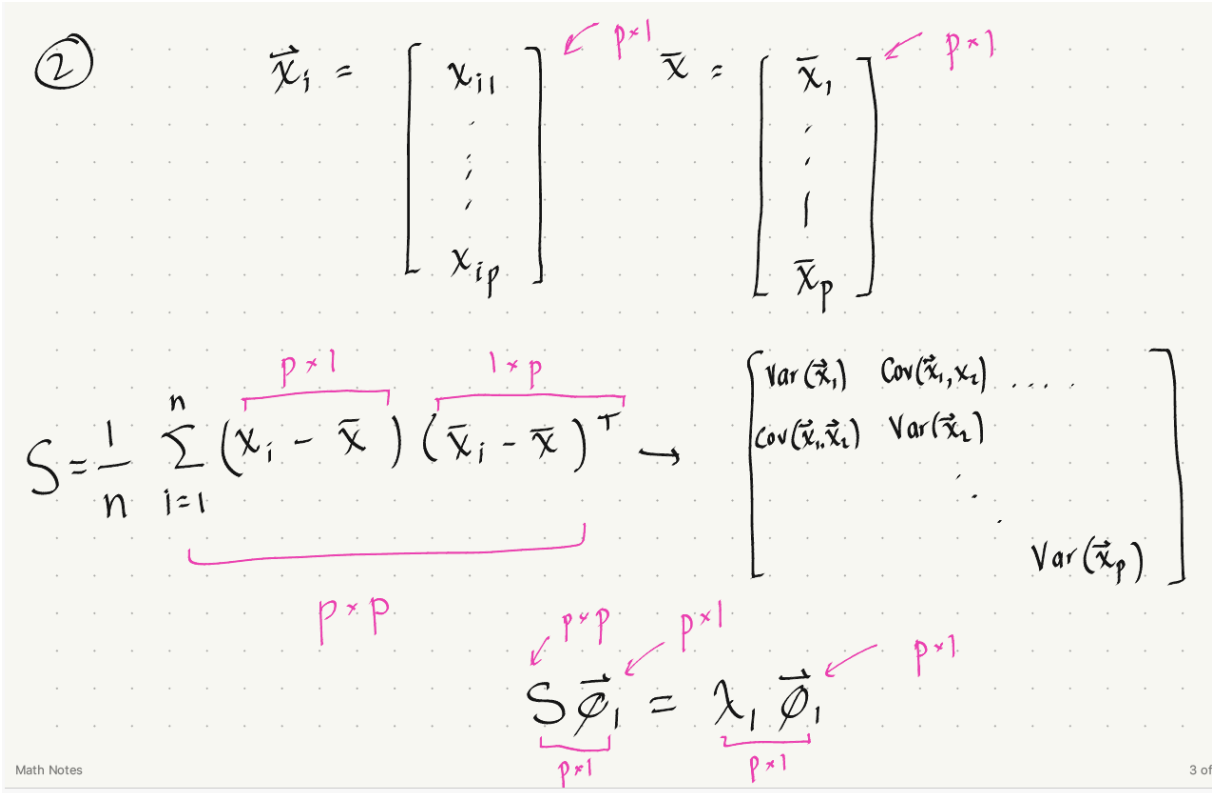


Figure 9.3: Enter Caption

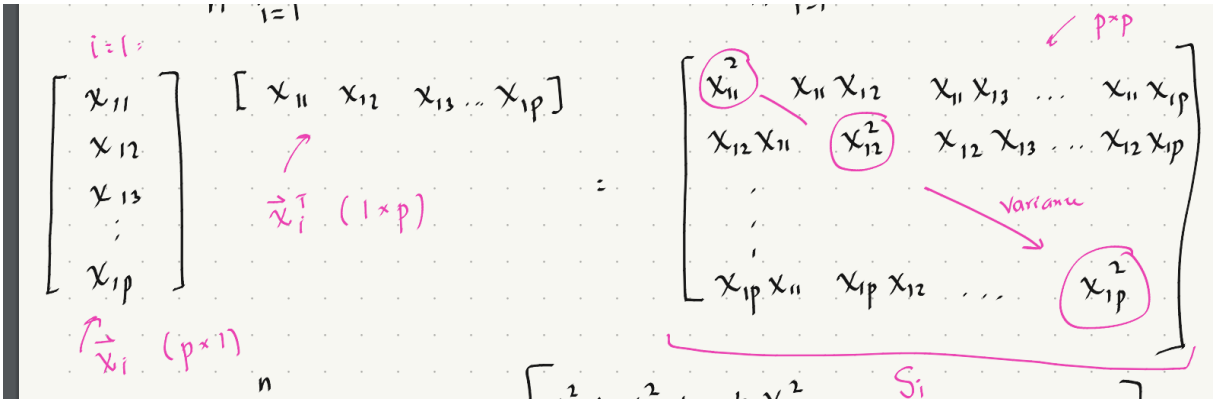


Figure 9.4: Enter Caption

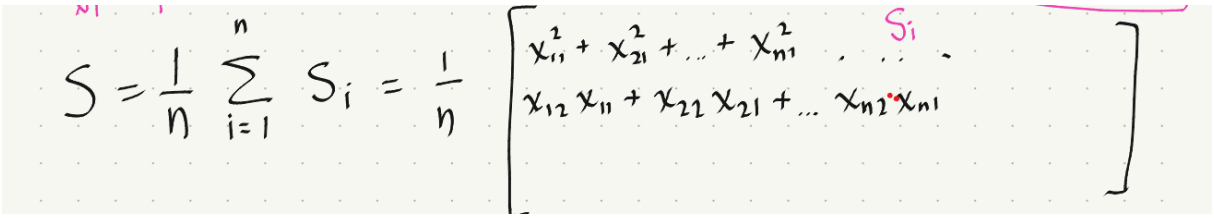


Figure 9.5: WHAT IS THIS????

The **resultant matrix represents the covariance between pairs of variables** in a dataset.

- It expresses the variance of each column (along its diagonal)
- and the off-diagonal elements represent the covariance between two variables/columns
- formally: each element s_ij of this matrix indicates the covariance between the i-th and j-th variable.
- for a dataset with p variables, this is a $p \times p$ matrix

,

9.12.2 Eigendecomposition of Variance-Covariance Matrix

We perform eigendecomposition on this $p \times p$ matrix:

$$S\phi_1 = \lambda_1\phi_1$$

Where

- λ_1 = the largest eigenvalue
- ϕ_1 = the corresponding eigenvector — it gives the loadings of the first principal component.

You would then solve: HOW? YOU WOULD SET THE DETERMINANT TO 0 TO GIVE THE CHARACTERISTIC POLYNOMIAL??? –> GIVES YOU THE EIGENVALUES –> FROM THERE PLUG BACK IN TO GET ASSOCIATED EIGENVECTORS???
As before, the first principal component is given by:

$$\begin{aligned} z_1 &= \phi_{11}x_1 + \phi_{21}x_2 + \ldots + \phi_{p1}x_p \\ &= X\phi_1 \end{aligned}$$

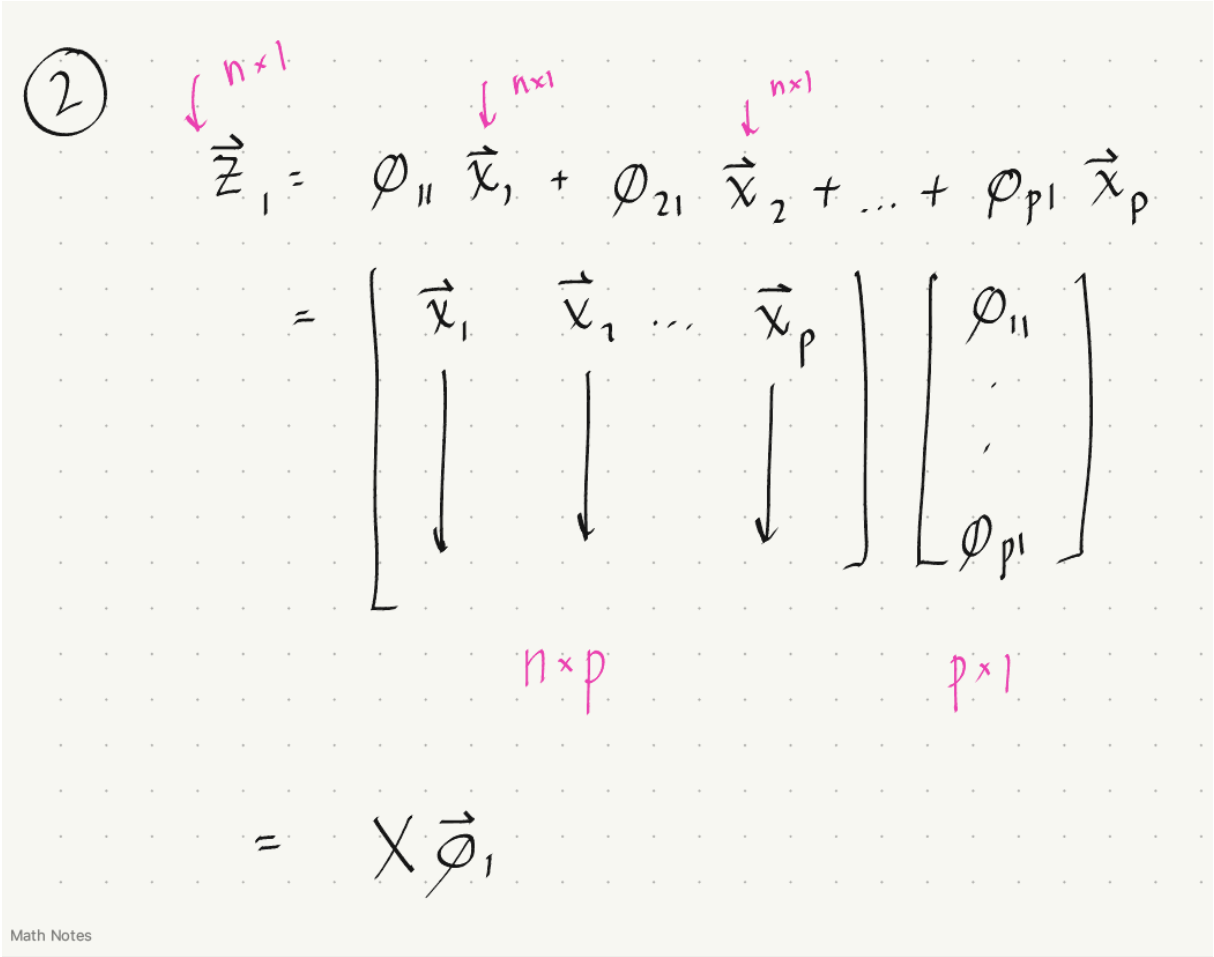


Figure 9.6: Enter Caption

Again, remember that each x_i is a vector of n length representing a variable-column from X
So adding all these vectors element-wise gives us z_1 as a vector.
ABOVE EACH X WAS A P LENGTH VECTOR, BUT HERE THEY ARE N LENGTH???z
Similarly, multiplying a matrix X by a vector ϕ_1 will produce a vector z_1.

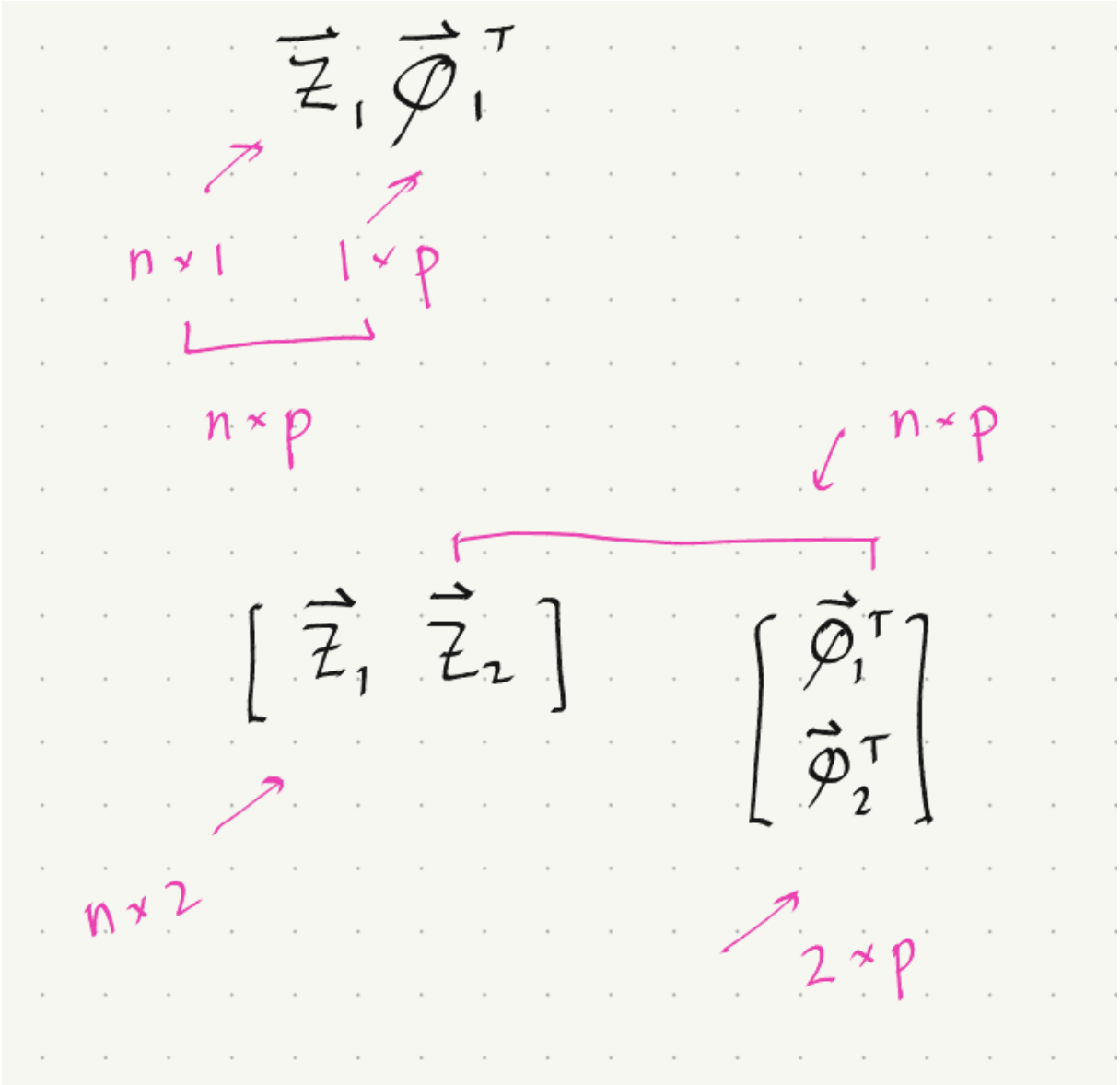


Figure 9.7: Enter Caption

WHAT DOES THIS REPRESENT — IT SEEMS TO BE THAT THE PRINCIPAL COMPONENT VECTOR MULTIPLIED BY THE ASSOCIATED LOADINGS VECTOR PRODUCES AN NXP MATRIX... BUT I THOUGHT THE PRINCIPAL COMPONENT VECTOR WAS A SUM OF THE PRODUCTS OF THE LOADINGS AND VARIABLES... SO THE LOADINGS INFO IS EMBEDDED IN THE PRINCIPLE COMPONENT, WHY WOULD WE WANT TO MUTLIPLY THEM TOGHETER?
IS THIS THE WAY TO 'EXPAND OUT THE PRINCIPLE COMPONENTS' IE TO MOVE BACK FROM PRINCIPLE COMPONENT VECTORS TO THE ORIGINAL DATA (OR AT LEAST REPRESENT THE HIGHEST VARIANCE VERSIONS OF IT) ?
THIS WOULD MAKE SENSE: THE IF YOU HAVE THE 1) LOADINGS, 2) PRINCIPLE COMPONENT –> YOU SHOULD BE ABLE TO INFER BACK TO THE ORIGINAL DATA

9.12.3 Beyond the First Principal Component

- Next, we should choose the linear combination of the columns of X that has maximal variance out of all the linear combinations that are *uncorrelated* with the first principal

component Z_1.

- This ensures that your next column captures the most of the remaining variance in the data without duplicating information you already have
- this is the same as taking the eigenvector corresponding to the second-largest eigenvalue of the variance-covariance matrix.

$$S\phi_2 = \lambda_2\phi_2 \rightarrow z_2 = X\phi_2$$
$$S\phi_3 = \lambda_3\phi_3 \rightarrow z_3 = X\phi_3$$
$$\vdots$$

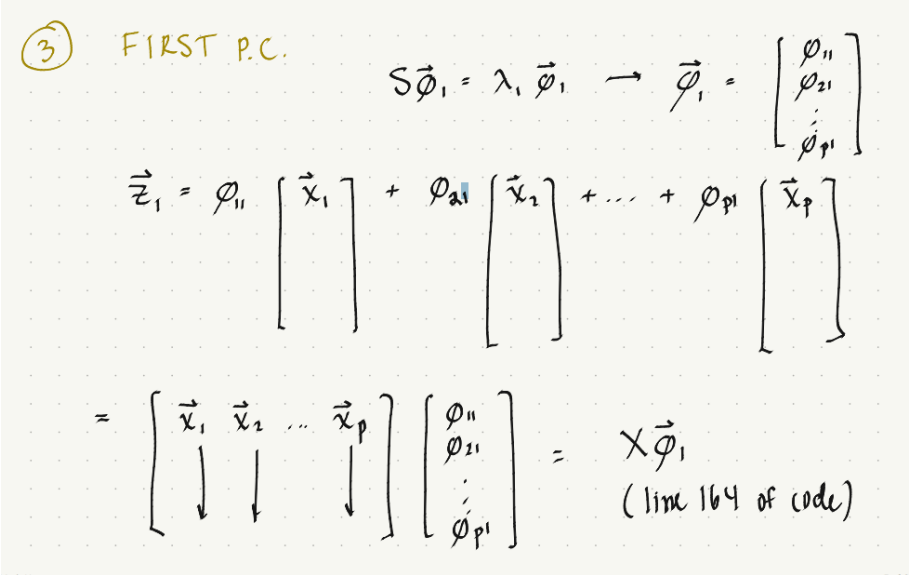


Figure 9.8: Enter Caption

9.12.4 Summary steps

Taking the first principle component = data matrix multiplied by eigenvector corresponding to largest eigenvalue of the variance-covariance matrix of X

1. ?????

9.13 Interpretation

First principal component is a projection of our $n \times p$ matrix onto one dimention (i.e. into one n-length vector.
This is the most efficient way to store our data matrix in one vector: we have max'd variation captured.

9.14 How this works

Eigenvectors represent directions in the data space, and eigenvalues indicate the magnitude of variance along these directions.

Here ϕ is a vector that, when multiplied by S, changes only by a scalar factor λ.
Here eigendecomposition identifies the principal components of the data (ie new, uncorrelated variables, that are linear combinations of the original variables). They are alignes with the directions of maximum variance in the dataset.
The eigenvector associated with the largest eigenvalue points in the direction of greatest variance; i.e. z_1.
The second largest eigenvector is orthogonal to the first, and points in the direction of the second greatest variance. etc.

9.15 Real World applications

By performing eigendecomposition, you can understand the underlying structure of the data, reduce noise, and simplify the dataset while retaining most of the important information.
In many real-world applications, this leads to better performance in machine learning models, as it reduces overfitting and computational cost.

9.16 lab 10: More PCA

Reconstituting data

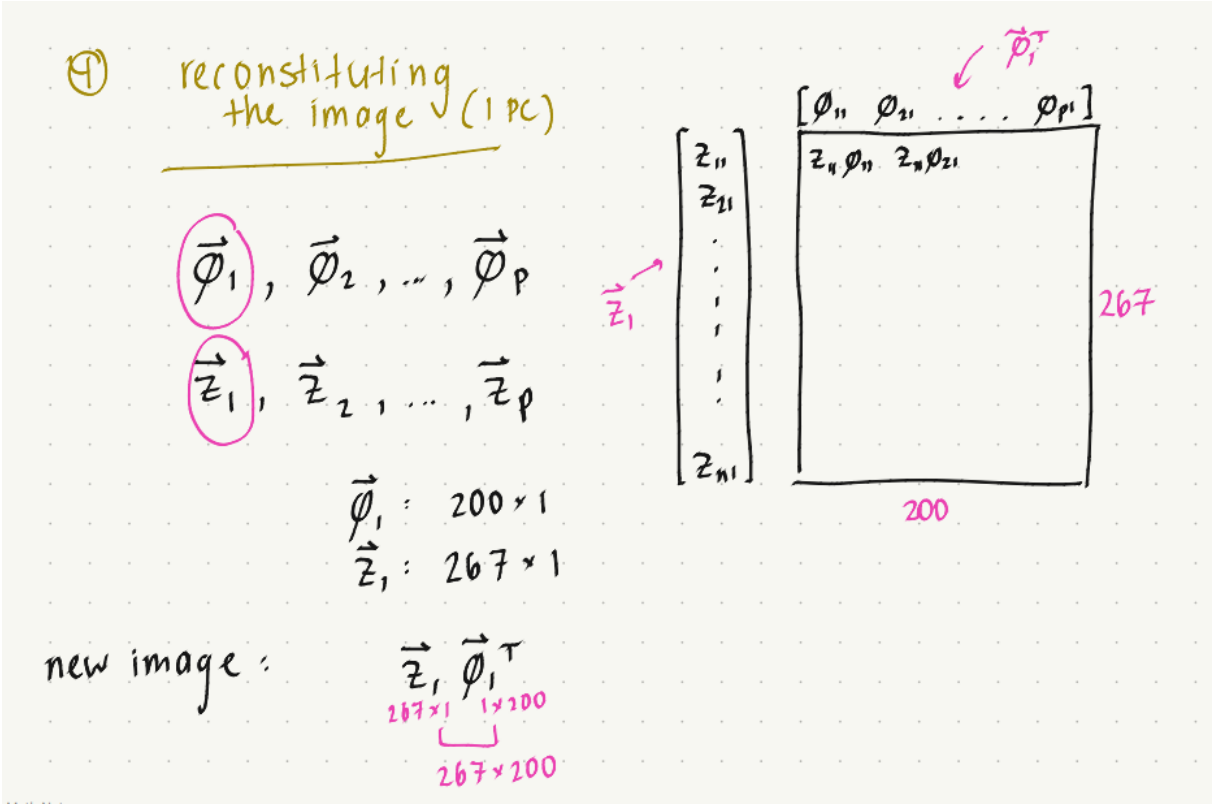


Figure 9.9: Enter Caption

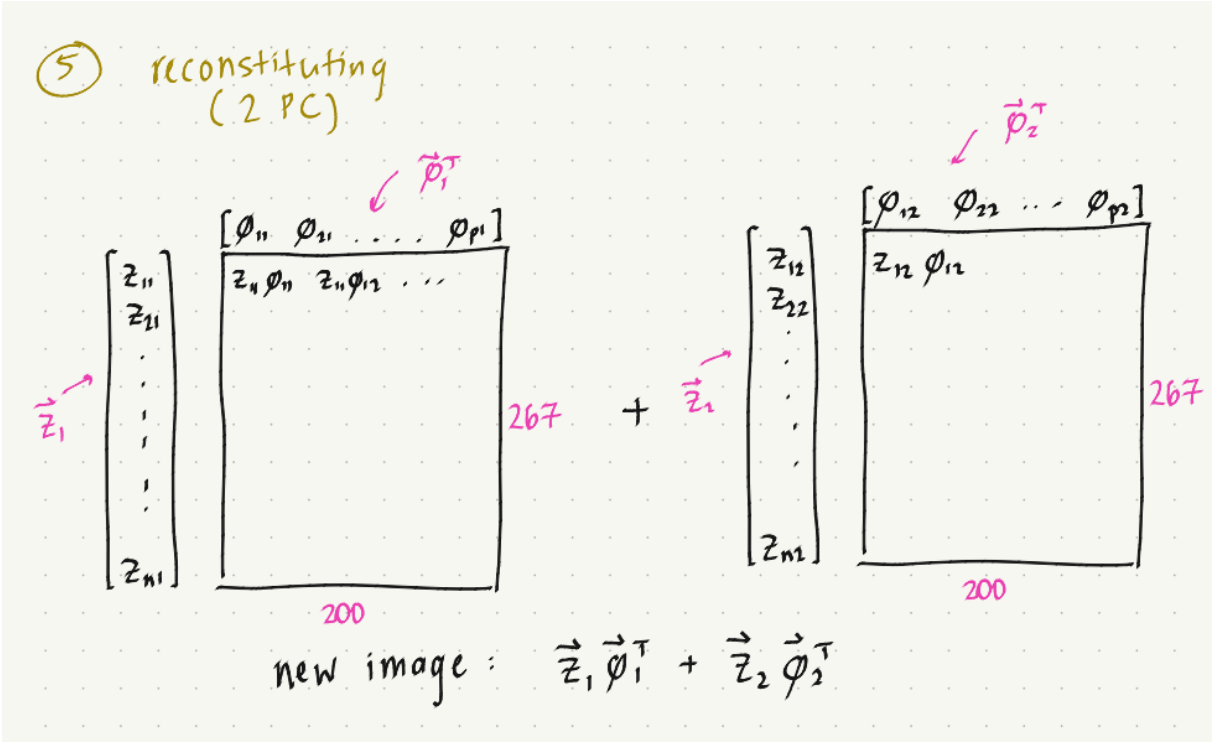


Figure 9.10: Enter Caption

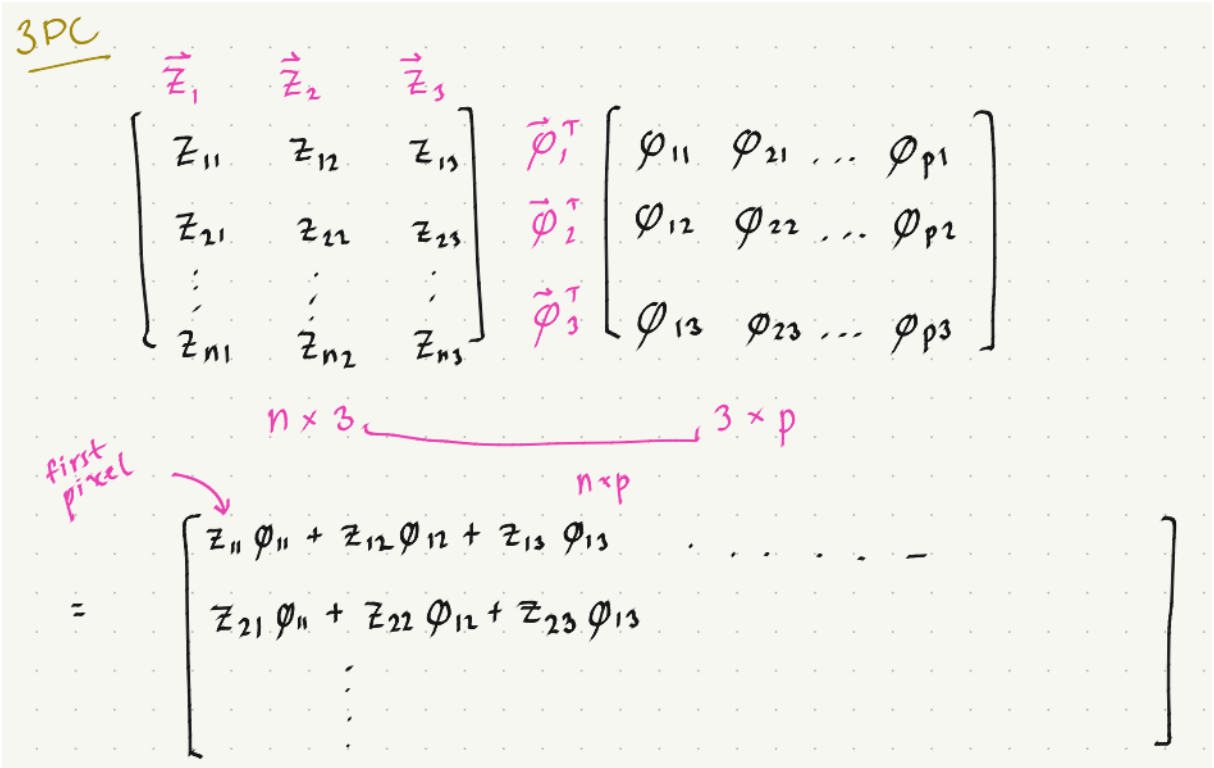


Figure 9.11: Enter Caption

Part IV

Optimisation

Chapter 10

Optimisation

Finals Fall 2023 – Henry Baker

M4DS Finals Revision: Session 10
Optimization

10.1 Second Derivative Test — determines if we've found a min / max

- $\frac{\partial^2 f}{\partial x^2} > 0 \Rightarrow$ the rate of change of the rate of change is positive at that point $\Rightarrow$ minimum.
- $\frac{\partial^2 f}{\partial x^2} < 0 \Rightarrow$ the rate of change of the rate of change is negative at that point $\Rightarrow$ maximum.

10.2 Constrained Optimization (in 2-variable setting)

Suppose we want our solution to be subject to some constraint.
E.g. 1) minimize:

$$\min_{x,y} x + y$$

and 2) we want the solution to live on the outside of a circle, given by

$$x^2 + y^2 = 1.$$

Then, we would:

- Set up our constraint as a function $g(x) = 0$:

$$x^2 + y^2 - - - 1 = 0$$

- Form the Lagrangian:

$$= x + y + \lambda(x^2 + y^2 - - - 1)$$

- Take the first derivatives $\frac{\partial L}{\partial x}$, $\frac{\partial L}{\partial y}$, and $\frac{\partial L}{\partial \lambda}$ and set them equal to 0, forming our first-order conditions.
- Solve this system of equations for x and y.

10.3 Min or Max (for all multivariate calculus?)

To know if min or max: Hessian Matrix

$$H = \begin{bmatrix} \frac{\partial^2 L}{\partial x^2} & \frac{\partial^2 L}{\partial x \partial y} & \frac{\partial^2 L}{\partial x \partial \lambda} \\ \frac{\partial^2 L}{\partial y \partial x} & \frac{\partial^2 L}{\partial y^2} & \frac{\partial^2 L}{\partial y \partial \lambda} \\ \frac{\partial^2 L}{\partial \lambda \partial x} & \frac{\partial^2 L}{\partial \lambda \partial y} & \frac{\partial^2 L}{\partial \lambda^2} \end{bmatrix} \tag{10.1}$$

- All eigenvalues of H are positive $\rightarrow$ critical point is a local min.
- If all eigenvalues of H are negative $\rightarrow$ critical point is a local max.
- If a mix $\rightarrow$ saddle point.

Additional Context

- Hessian matrix of a function $d : R^n \Rightarrow R$ is a square matrix of second-order partial derivatives of the function:

$$H(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_3 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} & \frac{\partial^2 f}{\partial x_3 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \frac{\partial^2 f}{\partial x_2 \partial x_n} & \frac{\partial^2 f}{\partial x_3 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

- helps us understand curvature of multivariable functions: to study local extrema of functions of several variables.
- the sign of the determinant of the Hessian at a critical point of a function can indicate whether is a local max/min/saddle point:
  - if all eigenvalues of H are positive, the critical point is a minimum
  - if all eigenvalues of H are negative, the critical point is a local maximum.
  - if a mix, you have a saddle point
- for a bivariate function, the Hessian a 2x2 matrix; for a function of 3 variables it s a 3x3 matrix, etc.

10.4 Matrix Optimization: the Gradient

The gradient is a vector of first-order partial derivatives of a function.
For a scalar function $f(\mathbf{x})$ where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, the gradient is given by:

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right) = \left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_n} \right) f$$

The gradient points in the direction of the steepest ascent of the function at a given point. In other words, it indicates the direction in which the function increases most rapidly.
The magnitude of the gradient gives the rate of increase of the function in that direction.

10.4.1 Eg f takes some input in $\mathbb{R}^m$ (a vector) $\rightarrow$ outputs some real value $\in \mathbb{R}$ (a scalar)

The gradient of $f$at a specific point $\mathbf{x}$ (NB this point is itself a vector) is a vector that described how the scalar output of f changes with respect to each component of the input vector at that particular point.
The gradient of f is a vector that helps you understand the sensitivity of the output of f to changes in the input vector. Each component of the gradient vector corresponds to a partial derivative of f with respect to a specific component of the input vector.
For instance:

$$f(\mathbf{z}) = \mathbf{z}^T\mathbf{z}$$

for some m-length vector $\mathbf{z}$.
We can write the gradient of $f(\mathbf{z})$ as the **vector of first derivatives** of $\mathbf{z}$ with respect to $\mathbf{z}$:

$$\nabla_{\mathbf{z}} f(\mathbf{z}) = \begin{bmatrix} \frac{\partial f}{\partial z_1} \\ \frac{\partial f}{\partial z_2} \\ \vdots \\ \frac{\partial f}{\partial z_m} \end{bmatrix}$$

In this particular case, the gradient of $f(\mathbf{z}) = \mathbf{z}^T\mathbf{z}$ for some m-length vector $\mathbf{z}$: $\nabla_{\mathbf{z}} f(\mathbf{z}) = 2\mathbf{z}$.
NB: 1) quite similar to how standard calculus works $x^2 \rightarrow 2x$
2) Then the gradient is still a vector, here x is a vector, so $2x$ is just a vector multiplied by a scalar (element-wise)

10.4.2 E.g. 2 f takes some input in $\mathbb{R}^{m\times n}$ (a matrix) $\rightarrow$ outputs some real value $\in \mathbb{R}$ (a scalar

Then, the gradient will be a matrix $\in \mathbb{R}^{m\times n}$
Gradients of matrices: another matrix.

$$\nabla A f(A) = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} & \cdots & \frac{\partial f(A)}{\partial A_{1n}} \\ \frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots & \frac{\partial f(A)}{\partial A_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{m1}} & \frac{\partial f(A)}{\partial A_{m2}} & \cdots & \frac{\partial f(A)}{\partial A_{mn}} \end{bmatrix}$$

where, generally, $(\nabla A f(A))_{ij} = \frac{\partial f(A)}{\partial A_{ij}}$.

this generic formula is just saying that the gradient of a matrix is the change of the output with respect to the change of input element.

10.5 Constrained Optimization Using the Gradient

Solve:

$$\nabla L(x, y, \ldots, \lambda) = 0$$

see the Final Review on this

10.6 Optimization as Eigen decomposition

1. **Form the Lagrangian** (from the objective function and the constraints: $L(x, \lambda)$ where x is the vector variable λ represents the Lagrange multipliers associated with the constraints.
2. **Take the gradient** of the Lagrangian with respect to the vector variable x.
3. **Set to 0**. This results in a system of equations to solve for the optimal values of x
4. **Solve**. When you solve the system of equations derived from the gradient of the Lagrangian, you end up with an equation of the form $Ax = \lambda x$, which is the eigenvalue-vector equation. The solutions to the eigenvalue-eigenvector equation $Ax = \lambda x$ are the eigenvectors of matrix A, and the corresponding λ values are the eigenvalues. These eigenvectors often represent the critical points or optimal solutions of the constrained optimization problem.

Why? Eigenvectors indicate the directions along which the objective function is most sensitive to changes. The eigenvalues provide information about how much the objective function changes in those directions.

Consider the following, **constrained optimization** problem:

$$\max_{x \in \mathbb{R}^n} \quad x^T A x \quad \text{subject to} \quad \|x\|^2 = 1$$

for a symmetric matrix $A \in S^n$.

- The objective function above is one where x is an n-dimensional vector and A is an $n \times n$ symmetric matrix.
- The constraint specifies the Euclidean norm of x must be equal to 1. I.e x must lie on the unit sphere in n-dimensional space.

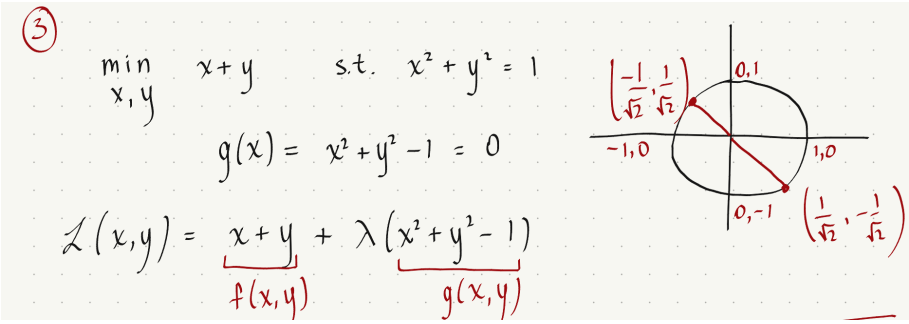


Figure 10.1: Enter Caption

Form the **Lagrangian** (an objective function that includes the equality constraints):

$$L(x, \lambda) = x^T A x - - - \lambda(x^T x)$$

where λ is called the Lagrange multiplier associated with the equality constraint.

For x^* to be a optimal point, **gradient of the Lagrangian has to be 0** at x^*

$$\nabla_x L(x, \lambda) = \nabla_x (x^T Ax - - - \lambda x^T x) = 0$$
$$= 2A^T x - - - 2\lambda x = 0$$

NB
the gradient of the Lagrangian is the Hessian. This is because the Lagrangian it set up with respect to the vector variable x; so the Lagrangian is a function that maps $R^n \to R$, and so ... IS THIS RIGHT? OR SHOULD IT BE A VECTOR AS THE GRADIENT OF A VECTOR IS A VECTOR, ALSO HESSIAN IS 2ND ORDER DERIVATIVES, THIS IS FIRST ORDER? NB this is just the linear equation $Ax = \lambda x$. **This is the eigen-value-vector equation**/ This shows that the only points which can possibly maximize (or minimize) $x^T Ax$ assuming $x^T x = 1$ (i.e. the constraints) are the eigenvectors of A. When we solve this equation we are essentially finding the vectors the eigenvalues-vector pairs.

So these optimal points of the Lagrange Multipler are the eigenvectors of A, and the corresponding eigenvalues λ are the max/min values of $x^T Ax$

10.7 Manual Constrained Optimisation using Lagrange multiplier

1. **Set up the Objective Function:**
   - Identify your objective function, $f(x, y)$, which you want to maximize or minimize.
2. **Set up the Constraint:**
   - Identify the constraint $g(x, y) = 0$. This is the condition that must be satisfied by the variables x and y.
3. **Formulate the Lagrangian:**
   - Introduce a Lagrange multiplier λ and set up the Lagrangian function: $L(x, y, \lambda) = f(x, y) - - - \lambda \cdot g(x, y)$.
4. **Find the First Derivatives and Set Them to Zero:**
   - Compute the partial derivatives of the Lagrangian with respect to each variable (x, y, and λ):
   $$\frac{\partial L}{\partial x} = 0, \quad \frac{\partial L}{\partial y} = 0, \quad \frac{\partial L}{\partial \lambda} = 0$$
   - This step involves taking derivatives and then equating them to zero.
5. **Solve the System of Equations:**
   - From the previous step, you will have a system of equations. Solve these equations simultaneously to find the values of x, y, and λ.
   - This might involve solving for one variable and then substituting it into another equation, or using methods like substitution or elimination.
6. **Determine Maxima/Minima Using the Hessian Matrix:**

- To determine if the solutions are maxima, minima, or saddle points, compute the second-order partial derivatives to form the Hessian matrix. The Hessian matrix H is given by:

$$H = \begin{pmatrix} \frac{\partial^2 L}{\partial x^2} & \frac{\partial^2 L}{\partial x \partial y} \\ \frac{\partial^2 L}{\partial y \partial x} & \frac{\partial^2 L}{\partial y^2} \end{pmatrix}$$

- Evaluate the Hessian matrix at the critical points found in step 5.
- The nature of the critical points is determined by the eigenvalues of the Hessian:
  - If all eigenvalues are positive, it's a local minimum.
  - If all are negative, it's a local maximum.
  - If they have different signs, it's a saddle point.

10.7.1 Examples

10.7.2 Example from Final Review

In the context of portfolio optimization, we consider two assets, Asset A and Asset B, with returns r_A and r_B respectively. We aim to choose weights w_A and w_B for these assets to maximize the expected return of the portfolio. The total investment is constrained to 100% (i.e., $w_A + w_B = 1$). Additionally, the client is risk-averse, leading to a penalty on the variance of the portfolio's returns, denoted by γ.

1) Set up the Lagrangian for your Optimisation Problem

Objective Function The objective function for the portfolio is to maximize the expected return, adjusted for risk aversion. It is given by:

$$\text{Maximize} \quad w_A r_A + w_B r_B + \gamma \sigma^2, \tag{10.2}$$

where σ^2 represents the variance of the portfolio's returns, and $\gamma \geq 0$ is the penalty coefficient for risk aversion.
Constraint The constraint for the portfolio is that the sum of the weights of Assets A and B should equal 1. Mathematically, this is expressed as:

$$w_A + w_B = 1. \tag{10.3}$$

Lagrangian Formulation To incorporate the constraint into the optimization problem, we use a Lagrange multiplier, denoted as λ. The Lagrangian $\mathcal{L}$ is formulated as:

$$\mathcal{L}(w_A, w_B, \lambda) = w_A r_A + w_B r_B - - - \gamma \sigma^2 + \lambda(w_A + w_B - - - 1). \tag{10.4}$$

In this formulation:

- $w_A r_A + w_B r_B$ represents the expected return from the portfolio.
- $-\gamma \sigma^2$ represents the penalty for portfolio variance, reflecting risk aversion.
- $+\lambda(w_A + w_B - - - 1)$ ensures that the weights of the investments sum to 1.

10.7.3 2) suppose you have the following variance-covariance matrix of the returns of your 2 assets, reduce your Lagrangian as much as possible

$$\Sigma = \begin{bmatrix} rA & rB & \\ rA & 1 & 0.5 \\ rB & 0.5 & 2 \end{bmatrix}$$

This relies on formula: $Var(X+Y) = Var(X) + Var(Y) + 2 \cdot Cov(X+Y)$:

Using fact that Var(cX) = $c^2 Var(X)$:

= $w_A^2 Var(r_A) = w_B^2 Var(r_B) + 2(w_A w_B Cov(r_A, r_B)$

Plugging it back in:

L(w_A, w_B, λ) = $w_A r_A + w_B r_B - - - \gamma[w_A^2 sigma_A^2 + w_B^2 sigma_B^2 + 2 w_A w_B \sigma_A B] + \lambda(w_A + w_B - - - 1)$

= $w_A r_A + w_B r_B - - - \gamma[w_A^2 + 2w_B^2 + w_A w_B] + \lambda(w_A + w_B - - - 1)$

Compute the gradient of L, just maximise with respect to w_A, w_B, λ; hold γ as user-provided parameter
Partial Derivative with respect to w_A:

$$\frac{\partial L}{\partial w_A} = r_A - - - 2\gamma w_A - - - \gamma w_B + \lambda$$

Partial Derivative with respect to w_B:

$$\frac{\partial L}{\partial w_B} = r_B - - - 4\gamma w_B - - - \gamma w_A + \lambda$$

Partial Derivative with respect to λ:

$$\frac{\partial L}{\partial \lambda} = w_A + w_B - - - 1$$

BELOW AREN"T RIGHT!

Example 1: Maximizing $f(x, y) = xy$

Objective Function: $f(x, y) = xy$
Constraint: $g(x, y) = x + y - - - 10 = 0$
Lagrangian: $L(x, y, \lambda) = xy - - - \lambda(x + y - - - 10)$
First Derivatives:

$$\frac{\partial L}{\partial x} = y - - - \lambda = 0$$
$$\frac{\partial L}{\partial y} = x - - - \lambda = 0$$
$$\frac{\partial L}{\partial \lambda} = -(x + y - - - 10) = 0$$

Solving the system of equations:

$$\begin{aligned} \lambda = x = y \\ x + y = 10 \\ 2x = 10 \quad \Rightarrow \quad x = 5 \\ y = 10 - - - x = 5 \end{aligned}$$

Solution: $x = 5, y = 5$

Example 2: Minimizing $f(x, y) = x^2 + 3y^2$

Objective Function: $f(x, y) = x^2 + 3y^2$
Constraint: $g(x, y) = x + 2y - - - 5 = 0$
Lagrangian: $L(x, y, \lambda) = x^2 + 3y^2 - - - \lambda(x + 2y - - - 5)$
First Derivatives:

$$\begin{aligned} \frac{\partial L}{\partial x} &= 2x - - - \lambda = 0 \\ \frac{\partial L}{\partial y} &= 6y - - - 2\lambda = 0 \\ \frac{\partial L}{\partial \lambda} &= -(x + 2y - - - 5) = 0 \end{aligned}$$

Solving the system of equations:

$$\begin{aligned} 2x &= \lambda \\ 6y &= 2\lambda \\ x + 2y &= 5 \\ \text{From the first two equations: } x &= \frac{\lambda}{2}, \quad y = \frac{\lambda}{3} \\ \text{Substituting into the constraint: } \frac{\lambda}{2} + 2 \cdot \frac{\lambda}{3} &= 5 \\ \lambda &= \frac{15}{2} \\ x = \frac{\lambda}{2} = \frac{15}{4}, &\quad y = \frac{\lambda}{3} = \frac{5}{2} \end{aligned}$$

Example 2: Minimizing $f(x, y) = x^2 + 3y^2$

Objective Function: $f(x, y) = x^2 + 3y^2$
Constraint: $g(x, y) = x + 2y - - - 5 = 0$
Lagrangian: $L(x, y, \lambda) = x^2 + 3y^2 - - - \lambda(x + 2y - - - 5)$
First Derivatives:

$$\begin{aligned} \frac{\partial L}{\partial x} &= 2x - - - \lambda = 0 \\ \frac{\partial L}{\partial y} &= 6y - - - 2\lambda = 0 \\ \frac{\partial L}{\partial \lambda} &= -(x + 2y - - - 5) = 0 \end{aligned}$$

Solving the system of equations:

$$\begin{aligned} 2x &= \lambda \\ 3y &= \lambda \\ x + 2y &= 5 \end{aligned}$$

Substituting these expressions for λ into equation 1 and 3: $x = \frac{\lambda}{2}, \quad y = \frac{\lambda}{3}$

Substituting into the constraint: $\frac{\lambda}{2} + 2 \cdot \frac{\lambda}{3} = 5$

$$\lambda = \frac{15}{2}$$

$$x = \frac{\lambda}{2} = \frac{15}{4}, \quad y = \frac{\lambda}{3} = \frac{5}{2}$$

Example 3: Maximizing $f(x, y) = x^2 + y^2$

Objective Function: $f(x, y) = x^2 + y^2$
Constraint: $g(x, y) = x^2 + y^2 - - - 4 = 0$
Lagrangian: $L(x, y, \lambda) = x^2 + y^2 - - - \lambda(x^2 + y^2 - - - 4)$
First Derivatives:

$$\frac{\partial L}{\partial x} = 2x(1 - - - \lambda) = 0$$
$$\frac{\partial L}{\partial y} = 2y(1 - - - \lambda) = 0$$
$$\frac{\partial L}{\partial \lambda} = -(x^2 + y^2 - - - 4) = 0$$

Solving the system of equations:

$$\begin{aligned} 2x(1 - - - \lambda) = 0 \quad &\Rightarrow \quad x = 0 \text{ or } \lambda = 1 \\ 2y(1 - - - \lambda) = 0 \quad &\Rightarrow \quad y = 0 \text{ or } \lambda = 1 \\ x^2 + y^2 &= 4 \end{aligned}$$

For $\lambda = 1, x^2 + y^2 = 4$ gives us the circle of radius 2.

Part V

Appendices

Appendix A

Common Distributions

Mid Terms Fall 2023 – Henry Baker

M4DS Mid Terms Revision: Distributions

Bernoulli trials > Binomials > Multinomial
Binomial = n choo

Distribution Name: Bernoulli

- **Definition and Parameters:**
  - Definition: Represents an experiment with exactly two possible outcomes, often referred to as "success" and "failure". It is the simplest discrete distribution.
  - Parameters:
    - p: Probability of success.
- **Probability Mass Function (PMF):**

$$P(X = k) = \begin{cases} p & \text{if } k = 1 \\ 1 - p & \text{if } k = 0 \end{cases}$$

- **Cumulative Distribution Function (CDF):**

$$P(X \leq k) = \begin{cases} 0 & \text{for } k < 0 \\ 1 - p & \text{for } 0 \leq k < 1 \\ 1 & \text{for } k \geq 1 \end{cases}$$

- **Expected Value and Variance:**

$$E(X) = p$$
$$\text{Var}(X) = p(1 - p)$$

- **Worked-out Example:**

  - Problem Statement: A coin is biased with a probability $p = 0.7$ of landing heads. What is the probability it lands tails?
  - Solution: Using the Bernoulli PMF, $P(X = 0) = 1 - 0.7 = 0.3$.

- **Relation to Other Distributions:**
  - A Bernoulli distribution with $n = 1$ trial is a special case of the binomial distribution.
- **Use Cases:**
  - Modeling the outcome of a single trial in any scenario with two possible outcomes, such as a coin toss, a yes/no question, or a pass/fail test.
- **Miscellaneous Notes:**
  - It is named after Jacob Bernoulli, a Swiss mathematician.
  - The Bernoulli distribution can be thought of as a single trial of a binomial experiment.

Distribution Name: Binomial

- **Definition and Parameters:**
  - Definition: Represents the number of successes in n independent Bernoulli trials, each with the same probability p of success.
  - Parameters:
    * n: Number of trials.
    * p: Probability of success on a single trial.
- **Probability Mass Function (PMF):**

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- **Cumulative Distribution Function (CDF):**

$$P(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}$$

- **Expected Value and Variance:**

$$E(X) = np$$
$$\text{Var}(X) = np(1-p)$$

- **Worked-out Example:**
  - Problem Statement: A fair coin is tossed 5 times. What is the probability of getting exactly 3 heads?
  - Solution: Using the Binomial PMF, $P(X = 3) = \binom{5}{3} \times 0.5^3 \times 0.5^2 = 0.3125$.
- **Relation to Other Distributions:**
  - A Binomial distribution with $n = 1$ trial reduces to a Bernoulli distribution.
  - As n becomes large and p is small, the Binomial distribution approaches the Poisson distribution.

- **Use Cases:**
  - Modeling the number of successes in a fixed number of independent trials, such as the number of heads in a certain number of coin tosses, or the number of defective items in a batch.
- **Miscellaneous Notes:**
  - It is one of the most widely used probability distributions in statistics.
  - The binomial distribution assumes that each trial is independent, and the probability of success remains constant across trials.

Distribution Name: Multinomial

- **Definition and Parameters:**
  - Definition: An extension of the binomial distribution to experiments with more than two possible outcomes. It represents the outcomes of n independent trials, each of which can result in one of k possible categories.
  - Parameters:
    - n: Total number of trials.
    - k number of possible outcomes / categories
    - $p_1, p_2, \ldots, p_k$: The probabilities of the k outcomes, where $\sum_{i=1}^{k} p_i = 1$.
- **Probability Mass Function (PMF):**

$$P(X_1 = n_1, X_2 = n_2, \ldots, X_k = n_k) = \frac{n!}{n_1! n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$$

  subject to the constraints $n_1 + n_2 + \ldots + n_k = n$ and $0 \leq n_i \leq n$ for all i.
- **Expected Value and Variance:**

$$E(X_i) = np_i$$
$$\text{Var}(X_i) = np_i(1 - p_i)$$

- **Worked-out Example:**
  - Problem Statement: A die is rolled 10 times. What is the probability of getting exactly 2 ones, 3 twos, and 5 threes?
  - Solution: Using the Multinomial PMF, the probability is calculated as

$$\frac{10!}{2!3!5!} \times \left(\frac{1}{6}\right)^2 \times \left(\frac{1}{6}\right)^3 \times \left(\frac{1}{6}\right)^5$$

- **Relation to Other Distributions:**
  - If $k = 2$, the Multinomial distribution reduces to the Binomial distribution.
- **Use Cases:**
  - Modeling the outcomes of experiments with multiple categories, such as the number of times each face of a die appears in a series of rolls.
  - Used in linguistics to model the occurrence of different words or phrases in a text.
- **Miscellaneous Notes:**
  - The Multinomial distribution generalizes the Binomial distribution, allowing for more than two possible outcomes.

Distribution Name: Geometric

- **Definition and Parameters:**
  - Definition: Describes the number of Bernoulli trials needed for a single success. It represents the probability that the first success will occur on the kth trial.
  - Parameters:
    - p: Probability of success on a single trial.
- **Probability Mass Function (PMF):**

$$P(X = k) = (1 - p)^{k-1}p$$

- **Cumulative Distribution Function (CDF):**

$$P(X \leq k) = 1 - - - (1 - p)^k$$

- **Expected Value and Variance:**

$$E(X) = \frac{1}{p}$$
$$\text{Var}(X) = \frac{1 - p}{p^2}$$

- **Worked-out Example:**
  - Problem Statement: A fair coin is tossed repeatedly. What is the probability that the first head appears on the 3rd toss?
  - Solution: Using the Geometric PMF, $P(X = 3) = (1 - 0.5)^{3-1} \times 0.5 = 0.125$.
- **Relation to Other Distributions:**
  - A special case of the Negative Binomial distribution when the number of successes required is 1.
- **Use Cases:**
  - Modeling the number of trials required before a success in scenarios with two possible outcomes, such as the number of coin tosses before seeing the first head.
- **Miscellaneous Notes:**
  - The Geometric distribution gives the probability distribution of the number of Bernoulli trials needed for one success, which might not necessarily be the first success.

Distribution Name: Negative Binomial

- **Definition and Parameters:**
  - Definition: Describes the number of Bernoulli trials needed for a specified number (r) of successes. It represents the probability that the r-th success will occur on the kth trial.
  - Parameters:
    - r: Number of successes.

    - p: Probability of success on a single trial.
    - k: number of trials

- **Probability Mass Function (PMF):**

$$P(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$$

- **Expected Value and Variance:**

$$E(X) = \frac{r}{p}$$
$$\text{Var}(X) = \frac{r(1-p)}{p^2}$$

- **Worked-out Example:**
  - Problem Statement: A biased coin with a probability $p = 0.4$ of landing heads is tossed repeatedly. What is the probability that the 5th head appears on the 8th toss?
  - Solution: Using the Negative Binomial PMF, $P(X = 8) = \binom{8-1}{5-1} \times 0.4^5 \times (1-0.4)^{8-5}$.

- **Relation to Other Distributions:**
  - If $r = 1$, the Negative Binomial distribution reduces to the Geometric distribution.

- **Use Cases:**
  - Modeling the number of trials required to observe a fixed number of successes, such as the number of patients needed to observe a fixed number of recoveries in a medical study.

- **Miscellaneous Notes:**
  - The Negative Binomial distribution can be thought of as an extension of the Geometric distribution to more than one success.
  - It provides the distribution of the number of successes before a specified number of failures occur.

Distribution Name: Hypergeometric

- **Definition and Parameters:**
  - Definition: Describes the probability of obtaining a specific number of successes when drawing samples without replacement from a finite population containing a fixed number of successes.
  - Parameters:
    - N: Total number of items in the population.
    - K: Number of success items in the population.
    - n: Number of items sampled.
    - k: Number of successes in sample.

- **Probability Mass Function (PMF):**

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

- **Expected Value and Variance:**

$$E(X) = n\frac{K}{N}$$
$$\text{Var}(X) = n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1}$$

- **Worked-out Example:**
  - Problem Statement: From a deck of 52 cards, 5 are drawn. What is the probability that exactly 3 of them are spades?
  - Solution: Using the Hypergeometric PMF, $P(X=3) = \frac{\binom{13}{3}\binom{39}{2}}{\binom{52}{5}}$.

- **Relation to Other Distributions:**
  - If items are drawn with replacement (meaning the drawn items are returned to the population before drawing again), the Hypergeometric distribution becomes the Binomial distribution.

- **Use Cases:**
  - Used in situations where sampling is done without replacement, like drawing cards from a deck, or selecting a committee from a larger group.
  - Common in biology and genetics studies where a subset of a population is examined.

- **Miscellaneous Notes:**
  - The Hypergeometric distribution provides exact probabilities whereas the Binomial distribution provides approximations when sampling without replacement.

Distribution Name: Poisson

- **Definition and Parameters:**
  - Definition: Describes the probability of a given number of events happening in a fixed interval of time or space. The events are assumed to be rare and independent.
  - Parameters:
    - λ: Average rate (mean number) of events in the given interval.

- **Probability Mass Function (PMF):**

$$P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- **Expected Value and Variance:**

$$E(X) = \lambda$$
$$\text{Var}(X) = \lambda$$

- **Worked-out Example:**
  - Problem Statement: A call center receives an average of 5 calls per hour. What is the probability that exactly 7 calls are received in a given hour?
  - Solution: Using the Poisson PMF with $\lambda = 5$, $P(X=7) = \frac{5^7 e^{-5}}{7!}$.

- **Relation to Other Distributions:**
  - The Poisson distribution can be derived as a limit of the Binomial distribution when the number of trials n is large, the probability of success p is small, and np is kept constant.
- **Use Cases:**
  - Modeling the number of events in fixed intervals of time or space, such as the number of phone calls to a call center in an hour, or the number of decay events per unit time from a radioactive source.
- **Miscellaneous Notes:**
  - Named after the French mathematician Siméon Denis Poisson.
  - It is particularly known for modeling rare events.
  - The distribution is defined for all non-negative integers but is most commonly used for small values of k.

wh

A.1 Hypergeometric

double tagged / W vs B balls sampled / stags released....

Appendix B

Cheat Sheet I

Mid Terms Fall 2023 – Henry Baker

M4DS Mid Terms Revision:
Cheat Sheet I

B.1 Session 1: Probability Theory

De Morgan's Law for Unions & Complements:

$$(A \cup B)^c = A^c \cap B^c$$

$$A \cap B)^c = A^c \cup B^c$$

Multiplication rule: think in terms of trees –> n^k
NB chronological order doesn't actually matter here — counter intuitive

Combinations = when order does not matter
permutations = when order/position matters $n!$

---**With Repl**
---**Without Re**

*NB: order matters, sampling w/o replacement also written as $n \cdot (n-1) \cdot (n-2) \cdot \ldots (n-k+1)$

Birthday Problem — counting complement
What's the probability of no matching birthdays?
This amounts to sampling the days of the year without replacement:

$$\begin{aligned} P(\text{no birthday match}) &= \frac{\text{number of ways to not repeat birthdays}}{\text{number of total possibilities}} \\ &= \frac{365 \times 364 \times \cdots \times (365 - - - k + 1)}{365^k} \\ P(\text{birthday match}) &= 1 - - - \frac{365 \times 364 \times \cdots \times (365 - - - k + 1)}{365^k} \end{aligned}$$

Factorial Overcounting:
when arranging n distinct items: $n!$ ways to do so...
..BUT if k items are identical –> divide by $k!$

- when assigning to multiple groups: we are overcounting by the number of groups factorial –> divide by groups factorial
- STATISTICS : overcounts Ss, Ts, Is, there are 3 Ss –> divide 3!3!2!
- same with the multinational: you divide by the repeats factorial: eg number of ways to sort 10 ppl into a group: $\frac{0!}{3!3!4!}$
- problem of non-repeat sampling (ie bday problem)
  - numerator = successes = order matters, w/o replacement $(n \times n-1 \times n-2 \times \ldots (n-k+1)$
  - Denominator = tota = order matters, w/ replacement (n^k)

  When working with multiple events (eg 10 heads); often easier ro say what is the probability of that NEVER happening: ie 1 single event...
  ... if something seems tedious: check its complement

Any probability function P must satisfy the following two axioms:

- $P(\emptyset) = 0, P(S) = 1.$
- If $A_1, A_2, \ldots$ are disjoint events, then

$$P\left(\bigcup_{j=1}^{\infty} A_j\right) \sum_{j=1}^{\infty} P(A_j)$$

.

Properties of Probabilities

- $P(A^c) = 1 - - - P(A)$
- If $A \subseteq B$, then $P(A) \leq P(B)$.
- $P(A \cup B) = P(A) + P(B) - - - P(A \cap B)$.

B.2 Session 2: Conditional Probability & Random Variables

B.2.1 Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

Intuitively: Venn diagram overlap, renormalised for $((B)$

Bayes Rule

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

LOTP

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Putting them together:
With Bayes: when applying LOTP to bottom: ensure you multiply by prob of the conditional.
- Eg: $P(T_c|D) \times P(D) + P(T_c|D_c) \times P(D_c)$
- Eg: $P(ObservedData|Coin_1) \times P(Coin_1) + P(OD|Coin_2) \times P(Coin_2) + P(OD|Coin_3) \times P(Coin_3)$

Bayes' Rule w/ Extra Conditioning:

$$P(A|B, E) = \frac{P(B|A, E) \times P(A|E)}{P(B|E)}$$

LOTP w/ Extra Conditioning:

$$P(B|E) = \sum_{i=1}^{n} P(B|A_i, E)P(A_i|E)$$

Independence of Events

$$P(A \cap B) = P(A) \cdot P(B)$$

> NB:
>
> - in a Venn diag, if the overlap is equal to the product of P(A) and P(B).
> - in a Contingency table, this means cells equal to marginals.

Equivalent to

$$P(A|B) = P(A)$$

- Independence is symmetric
- If A and B are independent:
  - A and B^c are independent,
  - A^c and B are independent,
  - A^c and B^c are independent

This doesn't carry through for conditional independence.

Independence of 3 events:
Needs to be more than pairwise independence (conditions 1 — 3)

$$P(A \cap B) = P(A)P(B) \tag{B.1}$$
$$P(A \cap C) = P(A)P(C) \tag{B.2}$$
$$P(B \cap C) = P(B)P(C) \tag{B.3}$$
$$P(A \cap B \cap C) = P(A)P(B)P(C) \tag{B.4}$$

Conditional Indepdence

- Conditional independence given E does not imply conditional independence given E^c
- Conditional independence does not imply independence
- Independence does not imply conditional independence

B.2.2 Random Variables

- **r.v.** is a function from the sample space S to the real number line $\mathbb{R}$; assigns a numerical value $X(s)$ to each possible outcome s of the experiment.
- **Support of X** is defined as all the values x such that $P(X = x) > 0$.
- **PMF** of X is the function $pX(x) = P(X = x)$. This is positive if x is in the support of X, and 0 otherwise.

Building PMF:

1. Immediately write $P(X = k) = \dots$
2. Enumerate all possible outcomes $(X = 1, X = 2, X = 3 \dots)$.
   consider what support of X could be
3. calculate probabilities for each outcome. (*Example: P(X=0) = P(TT) = 1/4)*
   is there a funcitonal form you can generalise to?
   NB: it it is NOT a binary outcome, might have to permute

PMF:

- $P(X = 0) = \frac{1}{4}$
- $P(X = 1) = 1/2$
- $P(X = 2) = 1/4$
- and pX (x) = 0 for all other values of x.

PMFs must (1) be non negative, and (2) sum to 1.

Bernoulli:

$$P(X = k) = \begin{cases} 1 - - - p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases}$$

Binomial:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Discrete Uniform:

$$P(X \in A) = \frac{|A|}{|C|}$$

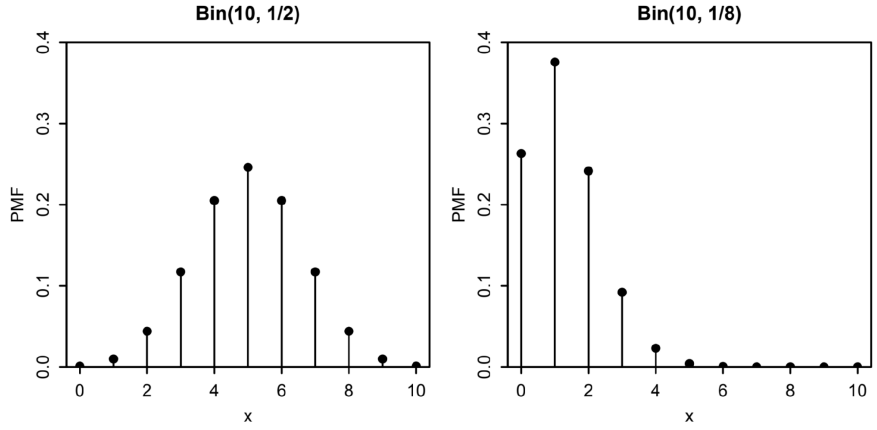


Figure B.1: Binomial PDFs

CDFs CDF of X, is the function $F_X(x) = P(X \leq x)$.

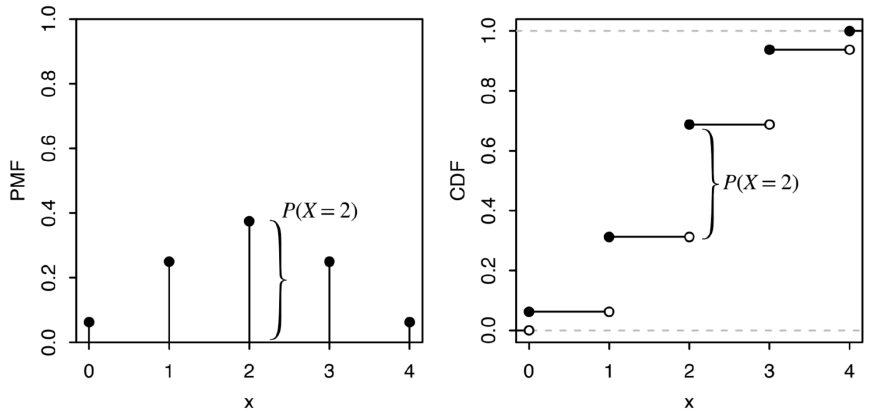


Figure B.2: PMF, CDF of $X \sim Bin(4; 1 = 2)$

B.3 Session 3: Joint r.v.s

- **Joint probability:** $P(A \cap B)$ or $P(A, B)$
- **Marginal (unconditional) probability:** $P(A)$
- **Conditional probability:** $P(A|B) = P(A, B)/P(B)$
- **Intersections via conditioning:** $P(A, B) = P(A)P(A|B)$
- **Unions via inclusion-exclusion:** $P(A \cup B) = (P(A) + P(B) - - - P(A \cap B)$

B.3.1 Independence of joint r.v.s

Continuous r.v.s

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$$

Discrete r.v.s

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

Conditional Independence

$$P(X \leq x, Y \leq y | Z = z) = P(X \leq x | Z = z) P(Y \leq y | Z = z)$$

B.3.2 Expectation

= weighted avg of possible values X can take:

$$E(X) = \sum_x x \cdot \underbrace{P(X = x)}_{\text{PMF at } x}$$

Eg 2x coin flip (heads)

$$E(X) = 0 \times \frac{1}{4} + 1 \times \frac{1}{2} + 2 \times \frac{1}{4} = 1$$

Eg Bernoulli: $X \sim Bern(p)$:

$$E(X) = 1p + 0(1-p) = p$$

Linearity of Expectation:

1. $E(cX) = cE(x)$
2. $E(X + Y) = E(X) + E(Y)$

B.3.3 Variance

$$Var(X) = E(X^2) - - - (EX)^2$$

Variance facts:

- $Var(c) = 0$ for any constant c
- $Var(X + c) = Var(X)$ for any constant c
- $Var(cX) = c^2 Var(X)$ for any constant c <— **NB**
- $Var(X + Y) = Var(X) + Var(Y)$ only if X and Y are independent.
  Caution: unlike expectation, variance is not linear
  - $Var(cX) \neq cVar(X)$
  - $Var(X + Y) \neq Var(X) + Var(Y)$ (in general)

  Except when the to r.v.s are independent! then $Var(cX) = c^2 Var(X)$

B.3.4 Marginal & Conditional Joint PMFs

Marginal PMF of X Sum over all y; marginalise out Y

$$P(X = x) = \sum_y P(X = x, Y = y)$$

if interested in $(Y = 1) \rightarrow$ sum over all Xs, that $(Y = 1, X = x)$

Conditional PMF of Y Joint divided by marginal.

$$P(Y = y | X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

Another way to think of it (same as above): Conditional for joint r.v.s is when

$$\begin{aligned} &P(Y=1|X=1) \\ &\quad = P(Y=1 \cap X=1)/P(X=1) \\ &\quad = (1/30)/5/30 \end{aligned}$$

$$\begin{aligned} &P(X=1|Y=2) \\ &\quad = P(X=1) \cap Y=2)/P(X=2) \\ &\quad = (4/30)/(24/30) \\ &\quad = 1/6 \end{aligned}$$

NB: where the Marginal X vs Marginal Y is... need to know which to make denominator for conditionals Contingency table is just another way to express PMF for joint variables; it expresses

	$Y=1$	$Y=0$	Marginal X
$X=1$	$\frac{5}{100}$	$\frac{20}{100}$	$\frac{25}{100}$
$X=0$	$\frac{3}{100}$	$\frac{72}{100}$	$\frac{75}{100}$
Marginal Y	$\frac{8}{100}$	$\frac{92}{100}$	1

Table B.1: Contingency table for X and Y with Marginal Distributions

how they move together
marginal is summing over the other thing
conditional is fixing the other thing at some value
joint is how they move together
Test for Independence:

$$\text{If independent: } P(X=x, Y=y) = P(X=x)P(Y=y)$$

so if the **cell value, is the product of the marginals**
for independence: every cell needs to be the product of the marginals.

B.4 Session 4: Calculus

1. **Rule 1: Powers**: $\frac{d}{dx}x^n = nx^{n-1}$
2. **Rule 2: Sum/Differences**:

$$\frac{d}{dx}(\mathrm{f}(x) \pm g(x)) = \frac{d}{dx}\mathrm{f}(x) \pm \frac{d}{dx}g(x)$$

3. **Rule 3: Constant Multiples**

$$\frac{d}{dx}[kf(x)] = k\frac{d}{dx}f(x)$$

4. **Rule 4: Products**

$$\frac{d}{dx}[g(x)f(x)] = g'(x) \cdot f(x) + g(x) \cdot f'(x)$$

5. **Rule 5: Quotients**

$$\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) = \frac{g(x)\cdot f'(x) - - - f(x)\cdot g'(x)}{g(x)^2}$$

6. **Rule 6: Chain**
If y is a function of u, and u is a function of x, then:

$$\frac{dy}{dx} = \frac{dy}{du}\cdot\frac{du}{dx}$$

ALWAYS REMEMBER TO PLUG THE U VALUE BACK N

7. **Rule 7: Natural Exponential**

$$y(x) = e^x \rightarrow \frac{dy}{dx} = e^x$$

8. **Rule 8: Natural Logarithms**

$$y = \ln(x) \rightarrow \frac{dy}{dx} = \frac{1}{x}$$

9. **General: Exponential Functions**
Power rule is for when x is a **constant**.
Exponential functions take x as the exponent itself.

$$\frac{d}{dx}a^x = \ln(a)\times a^x$$

$$\frac{d}{dx}a^{bx} = b\times\ln(a)\times a^{bx}$$

Example: a^x
10^x becomes $ln(10)$ $\times 10^x$:

$$f(x) = \frac{10^x}{\ln(10)}$$
$$f'(x) = \ln(10)\times 10^x \times \frac{1}{\ln(10)} = 10^x$$

Example: a^{bx}

$$f(x) = 2^{4x} + 4x^2$$
$$f'(x) = 4\ln(2)\times 2^{4x} + 8x$$

B.5 MLE

MLE steps:

1. (Identify the distribution: write out the PMF)
2. **Write the likelihood as a function of the data:**

   $$L(\theta) = P(x_1, x_2 \ldots . x_i)$$

   becomes:

   $$L(x_1, x_2 \ldots . x_i; \theta) = \prod_{j=1}^{n} \text{the PMF with } \theta \text{ substituted in for parameter of interest / } p???$$

3. **Expand as products**: we assume each event is i.i.d, so the likelihood is the product of each
   - First, write as a series of products for each r.v.:

     $$\lambda) = P(X = 1) \times P(X = 3) \times P(X = 1) \times \ \ldots$$

   - Then expand each using the relevant distribution with the data for the r.v plugged in:

     $$\frac{e^{-\lambda}\lambda^1}{1!} \times \frac{e^{-\lambda}\lambda^3}{3!} \times \frac{e^{-\lambda}\lambda^1}{1!} \times \ldots$$

   - collect terms and simplify as much as possible.
4. **Take the log-likelihood:**

   $$\ell(\lambda) = \log(L(\lambda))$$

   Use the properties of logs to break it up into its components parts to reshape it into nice +/- equation (ie without products or quotients... use these rules of logs!!)
5. **Derive the log with respect to parameter of interest:**
   - constants (ie not dependent on parameter of interest) drop out.
   - parameter terms differentiate as usual.
   - derivative of log =

     $$\frac{d}{dx}\ln(x) = \frac{1}{x}$$

     $$\frac{d}{dx}\ln(2x) = \frac{1}{2x} \times 2 = \frac{1}{x}$$

     NB chain rule here

     $$\frac{d}{d\lambda}7\log(\lambda) \ = 7 \times \frac{1}{\lambda} = \frac{7}{\lambda}$$

6. **Set to 0**
7. Solve for θ

Log rules:

$$\log(ab) = \log(a) + \log(b)$$
$$\log\left(\frac{a}{b}\right) = \log(a) - - - \log(b)$$
$$\log\left(\prod_{j=1}^{n} x_j\right) = \sum_{j=1}^{n} \log(x_j)$$
$$\log(a^b) = b \log$$
$$\log(0) = 1$$
$$\log(1) = 0$$
$$\log(e^x) = x$$
$$\log(a^b) = b \log(a)$$
$$\frac{d}{dx} \ln(x) = \frac{1}{x}$$

Exponent rules:

$$a^m \times a^n = a^{m+n}$$
$$\frac{a^m}{a^n} = a^{m-n}$$
$$(a^m)^n = a^{m \times n}$$
$$(ab)^n = a^n \times b^n$$
$$\left(\frac{a}{b}\right)^n = \frac{a^n}{b^n}$$
$$a^0 = 1 \quad (\text{where } a \neq 0)$$
$$a^{-n} = \frac{1}{a^n} \quad (\text{where } a \neq 0)$$
$$a^1 = a$$

B.6 Taylor Series Approximation

1. Find derivatives (n many, depending on polynonmial degree specified)
2. Evaluate them (a = {x})
3. Insert each of these into the Taylor formula...
4. ... Simultaneously: plug in a values with the given centring coordinate. This will leave various x values: this your line formula.

Appendix C

Cheat Sheet II

Finals Fall 2023 – Henry Baker

M4DS Finals Revision: Session 6
Calculus Meets Probability / Continuous Random Variables I

C.1 Wk 6 — Continuour r.v.s meets probability

C.1.1 Continuous r.vs

- r.v.s has continuous distribution if its CDF is differentiable
- PDF of X is derivative of CDF: $f(x) = F'(x)$
- CDF of X is integral of PDF (ie the area under the curve)
- unlike discrete r.v.s, for continuous: $P(X = x) = 0$ for all x
  - PDF of X gives probability density
  - BUT, CDF remains interpret able as P(X) (as it represents AUC under PDF)
  - The probability that a continuous random variable falls within a particular interval is given by area under the PDF curve over that interval.
- valid PDF conditions: 1) non-negative, 2) integrates to 1

C.1.2 expectation of continuous r.v

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx$$

= centre of mass / balance point
"To find the expected value of a continuous random variable, take every possible value that variable can have, multiply each by the probability of that value occurring, and then sum all these products together."

C.1.3 Uniform, continuous

PDF

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

Proof it is a valid PMF: it sums (integrates) to 1 : rectangle (b-a)*(1/b-a) = 1

CDF

$$F(x) = \begin{cases} 0 & \text{if } x \leq a \\ \frac{x---a}{b---a} & \text{if } a < x < b \\ 1 & \text{if } x \geq b \end{cases}$$

C.1.4 Normal

μ = location; σ^2 = scale (variance)

C.1.5 Standardisation

can transform any Normal-distributed r.v. into a Standard Normal (this is a z-score):

$$Z = \frac{X - - - \mu}{\sigma}$$

NB: r.v. itself does NOT have to be Normal distribution — can be ANY distribution; so long as it is i.i.d → you can take the sample mean, standardise it, and that will be normally distributed.

C.1.6 Exponential

- time to wait before first success
  - = continuous analog of the Geometric (number of failures until first success in a sequence of Bernoulli trials)
  - λ = rate of success for some unit of time
  - memorylessness

C.1.7 Joint Distributions of Continuous r.v.s

- Joint CDF: $F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$
- Joint PDF: $f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y)$ — first take the partial derivative of $F_{X,Y}(x, y)$ with respect to x, and then take the partial derivative of the result with respect to y
  - e.g. CDF: $F(x, y) = \frac{1}{2}x^2y^3$ → PDF: $f(x, y) = 3xy^2$
- Marginal Distribution of X from Joint PDF, integrate over all values of Y:
  - $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy$
- Conditional PDF (of Y, given $X = x$):
  - $f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$
  - Conditional = Joint PDF / Marginal

C.1.8 Bayes Rule and LOTP for continuous r.v.s

- Bayes rule: $f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)f_Y(y)}{f_X(x)}$
- LOTP: $f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y)f_Y(y)\,dy$

C.2 wk 7 — Continuous R.vs II

C.2.1 Covariance

- move together + ; move opp -; independent 0
- "expectation of the product, minus the product of the expectations"
- $\text{Cov}(X,Y) = \mathbb{E}((X - - - \mathbb{E}X)(Y - - - \mathbb{E}Y))$
- $\text{Cov}(X,Y) = \mathbb{E}[XY] - - - \mathbb{E}[X]\mathbb{E}[Y]$
- independence: covariance is 0: $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$

Some Covariance rules

1. $\text{Cov}(X,X) = \text{Var}(X)$
2. $\text{Cov}(X,Y) = \text{Cov}(Y,X)$
3. $\text{Cov}(X,c) = 0$ for any constant c
4. $\text{Cov}(aX,Y) = a \cdot \text{Cov}(X,Y)$ for any constant a
5. $\text{Cov}(X+Y,Z) = \text{Cov}(X,Z) + \text{Cov}(Y,Z)$
6. $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2 \cdot \text{Cov}(X,Y)$
7. **additional:** $E(EX)$: expectation of a constant is just a constant .

C.2.2 Correlation

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}}$$

Correlation coefficient: $\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$
corr between -1:1

Due to non-linearity in the data (think of parabola, $y = x^2$ — perfectly dependent, but uncorrelated):

- Independence $\rightarrow$ uncorrelated
- Uncorrelated $\rightarrow \times \rightarrow$ Independence

Independence defined as: $P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$
For Independent variables: $E(XY) = E(X)E(Y)$
For independent variables (= 0 covariance), their correlation is 0. But not necessarily visa versa.

Example of proving independence:

$$E(X) = \frac{1}{2}(1) + \frac{1}{2}(2) = 1.5$$
$$E(Y) = \frac{1}{2}(3) + \frac{1}{2}(4) = 3.5$$

$E(XY)$ can be calculated by considering all combinations of X and Y
We have four combinations: (1, 3), (1, 4), (2, 3), and (2, 4). Each combination occurs with a probability of 1/4, since the probabilities of X and Y are each 1/2.

$$E(XY) = \frac{1}{4}(1 \times 3) + \frac{1}{4}(1 \times 4) + \frac{1}{4}(2 \times 3) + \frac{1}{4}(2 \times 4) = 5.25$$
$$\text{Cov}(X, Y) = E(XY) - - - E(X)E(Y) = 5.25 - - - (1.5 \times 3.5) = 0$$

Since the covariance is 0, it suggests that X and Y are uncorrelated

C.2.3 Law of Large Numbers: as n grows large, the sample mean $\bar{X}$ converges to the true mean μ

- Sample mean (if i.i.d): $\bar{X}_n = \frac{X_1 + X_2 + ... + X_n}{n}$
- sample mean is itself r.v.:
  - Expectation = μ (i.e. the population mean — the sample mean and the population mean converge as n grows)
  - Variance = $\frac{\sigma^2}{n}$
  - Standard Deviation = $\frac{\sigma}{\sqrt{n}}$

C.2.4 Central Limit Theorem = that the standardised sample mean (standardised $\bar{X}$) converges in distribution to the standard Normal as $n \to \infty$

1. subtract expectation μ
2. dividing by standard deviation = $\frac{\sigma}{\sqrt{n}}$

So, regardless of underlying distribution: $\sqrt{n}\left(\frac{\bar{X}_n - - - \mu}{\sigma}\right) \xrightarrow{d} \mathcal{N}(0, 1)$

Also sum:

$$\sum_{i=1}^{n} X_i \xrightarrow{d} \mathcal{N}(n\mu, n\sigma^2) \quad \text{as} \quad n \to \infty$$

C.2.5 Example: Normal Approximation to the Binomial

Recalling that the Binomial (n, p) is the sum of n Bernoullis with probability p, we can even use the Normal distribution to approximate the Binomial.

$$\sum_{i=1}^{n} X_i \approx \mathcal{N}(n\mu, n\sigma^2)$$

and recalling that the mean of a Bernoulli is p and its variance is $p(1---p)$, we can use the CLT to say:

$$\sum_{i=1}^{n} X_i \approx \mathcal{N}(np, np(1---p))$$

NB:
Variance:

$$E[(X---\mu)^2] = \sum_i (x_i---\mu)^2 \cdot P(x_i)$$

for uniform:

$$\text{Var}(X) = \frac{(b---a)^2}{12}$$

C.3 Lab 7 — EM Algorithm

1. Initialise parameters
   - μ_1, μ_2 = means
   - σ_1, σ_2 = standard deviations
   - $\pi_1.\pi_2$ = mixing properties (the initial prob of being in one distribution)
2. E-step: Expectation — compute responsibilities of each data point (= calculate the γ of each data point: the prob that each data point belongs to each component given current parameter values)
3. M-step: Maximisation — update the parameters based on the responsibilities (= MLE of each parameter given the γ values for each data point (the parameters defined as some function involving sums of gamma-data point, so will give a single value).
4. Evaluate the new log-likelihood with new (i) parameter, (ii) responsibilities.
5. Check for convergence

NB: how Bayes rule used to obtain gammas:

$$\begin{aligned}\Pr(z_{1i} = 1|x_i) &= \frac{f(x_i|z_{1i} = 1)\Pr(z_{1i} = 1)}{f(x_i)} \\ &= \frac{\pi_1 \mathcal{N}(x_i|\mu_1, \sigma_1)}{\pi_1 \mathcal{N}(x_i|\mu_1, \sigma_1) + \pi_2 \mathcal{N}(x_i|\mu_2, \sigma_2)}\end{aligned}$$

C.4 Wk 8 — Matrix Algebra

- Matrix dimensions $m \times n$ = m rows, n columns
- row values → column, column values → row
- to add 2 matrices: need same dimensions
- $\mathbf{a}^T\mathbf{b} = [a_1, a_2, a_3] \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = a_1b_1 + a_2b_2 + a_3b_3 \ \ldots$ NB: output is scalar
- Matrix multiplication
  - conformable: $mn \times np$ (LHS columns = RHS rows) → output mp — the outer values

  - for each element a_{ij}: sum of the products of the elements of the corresponding **row of A** and the corresponding **column of B**.
  - so for item $a_1 2$:
    * all the elements of row 1 from A
    * all the elements of column 2 from B
    * multiplied by each other
    * summed

- Transpose Facts
  - $(A^T)^T = A$
  - $(A + B)^T = A^T + B^T$
  - $(AB)^T = B^T A^T$
  - $a^T b = b^T a$

- Identity matrix
  - $I_n x_n = x_n$
  - $I_m A_{m \times n} = A_{m \times n}$ and $A_{m \times n} I_n = A_{m \times n}$
  - $A^{-1} A = I_n$

- Vector Norms (measure of magnitude)
  - L1 $|x|_1 = \sum_i |x_i|$ = take absolute values of elements before summing
  - L2 $|x|_2 = \sqrt{\sum_i x_i^2}$ = square root of the sum of the squares of the vector's elements

C.5 Lab 8 — Regression

C.5.1 Linear Regression

The objective function for least squares regression is:

$$\min_{\beta_0, \beta_1, \ldots, \beta_p} \sum_{i=1}^{n} \varepsilon_i^2$$

$$= \min_{\beta_0, \beta_1, \ldots, \beta_p} \sum_{i=1}^{n} (y_i - - - \beta_0 - - - \beta_1 x_{i1} - - - \beta_2 x_{i2} - - - \ldots - - - \beta_p x_{ip})^2$$

Representing the squared error minimisation problem (the cost function) in Matrix form:

$$\varepsilon^T \varepsilon = Y^T Y - - - Y^T X \beta - - - \beta^T X^T Y + \beta^T X^T X \beta$$

This expanded **cost function** (expressed in marix notation) is what we will be minimising (by taking first derivative, set to zero)
(1) Setting to 0 $\rightarrow$ (2) taking first derivative with respect to $\beta \rightarrow$ (3) simplifying:

$$-2X^T Y + 2X^T X \beta = 0$$

Solving for beta:

$$\beta = (X^T X)^{-1} X^T Y$$

C.5.2 Penalised regression

The objective function in Lasso Regression becomes:

$$\text{Minimize} \left(\text{Residual Sum of Squares} + \lambda \sum_{i=1}^{n} |\beta_i| \right)$$

The objective function in Ridge Regression is:

$$\text{Minimize} \left(\text{Residual Sum of Squares} + \lambda \sum_{i=1}^{n} \beta_i^2 \right)$$

C.6 Wk 9 — Linear Algebra II

- Linear Dependence: $c_1 v_1 + c_2 v_2 + \cdots + c_n v_n = 0$
  - 1) set up the equation $c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 \cdot + c_n \mathbf{v}_n = \mathbf{0}$
  - 2) solve the system of equations: take one $c_n v_n$ over to the other side so you can write one of the $c_n v_n$ as a function of the others, then plug back in, etc.
    I STILL DONT GET HOW TO SOLVE THESE SYSTEMS?
- Span = all linear combinations of the vectors in the set.
  - S is a spanning set for V if all the dimensions of V can be represented by linear combinations of S. A Spanning Set S must contain at last as many elements as the linearly independent vectors from V.
  - There are exactly n orthogonal / linearly independent vectors in $\mathbb{R}^n$ — these are the basis vectors (another way to phrase: number of linearly independent vectors in V define dimension of its vector space)
- Determinant:
  - **non-zero (for square matrix) → linear independence → invertible**
  - determinant of a matrix reflects how the transformation changes the dimensions of the space
  - characteristic polynomial of a matrix, used to find its eigenvalues, is derived from its determinant
  - 2x2 Matrix: $\det(A) = ad - - - bc$
  - 3x3 matrix: $\det(A) = a(ei - - - fh) - - - b(di - - - fg) + c(dh - - - eg)$
- inverse conditions: $(AA^{-1} = A^{-1}A = I_n)$
  1. matrix is square
  2. columns are linearly independent (full rank)
  3. non-zero determinant
- eigenvalue-vector pairings:
  - $Av = \lambda v$
  - eigenvalue equation: $(\lambda I_n - - - A)v = 0$
  - characteristic equation: $\det(A - - - \lambda I) = 0 \rightarrow$ or $det(\lambda I_n - - - A) = 0$ finds eigenvalues

  - existence conditions: 1) square, 2) a lambda scalar that fulfils characteristic equation (i.e. linearly independent columns)
  - eigenspace of λ, E_λ, is all vectors v that satsify the condition $Av = \lambda v$

- Eigendecomposition: $A = Q\Lambda Q^{-1}$
  - A = sq matrix $n \times n$
  - Λ = diag matrix with eigenvalues of A $n \times n$
  - Q = matrix whose columns are eigenvectors of A, corresponding to eigenvalues in Λ
- SVD = $A = U\Sigma V^T$
  - $A \in \mathbb{R}^{m \times n}$
  - U is a matrix containing left singular vectors of A (contains the m eigenvectors of the square matrix AA^T) $(m \times n)$
  - V^T is a matrix containing the right singular vectors of A (contains the n eigenvectors of the square matrix A^TA $(n \times n)$
  - Σ is a matrix where the diagonal entries are the singular values of A (square roots of the eigenvalues of A^TA (or AA^T) on the diagonal. $(m \times n)$
  - ...
  - U & V^T: how much matrix A rotates an object
  - Σ: associated scaling
    - number of non-zero singular values gives rank of the matrix (number of linearly independent columns) — full rank: nothing redundant.

Dimensions:

- $A = m \times n$
- $U = m \times m$
- $\Sigma = m \times n$
- $V^T = n \times n$
- + NUMBER OF LINEARLY INDEPENDENT COLUMNS OR ROWS NEEDED TO FORM BASIS FOR COLUMN OR ROW SPACE OF A = RANK OF THE MATRIX A = THE NUMBER OF NON ZERO SINGULAR VALUES IN SIGMA

C.7 Lab 9 — PCA

C.7.1 As Variance Max

$$z_1 = \phi_{11}x_1 + \phi_{21}x_2 + \ldots + \phi_{p1}x_p$$

Maximising variance of the vector of linear combinations z_1 (the first principle component), i.e. maximising the variance across its elements (which represent linear combination associated with each observation, consisting of scaled variables, summed)

$$\max_{\phi_{11},\phi_{21},\ldots,\phi_{p1}} \text{Var}(z_1) = \max_{\phi_{11},\phi_{21},\ldots,\phi_{p1}} \frac{1}{n}\sum_{i=1}^{n}(z_{i1} - - - \bar{z}_1)^2$$
$$= \max_{\phi_{11},\phi_{21},\ldots,\phi_{p1}} \frac{1}{n}\sum_{i=1}^{n} z_{i1}^2$$

Constrain associated loadings:

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1$$

C.7.2 As Eigendecomposition

- derive a Variance-covariance matrix
- Eigendecompose that: $S\phi_1 = \lambda_1\phi_1$
- λ_1 = eigenvalue for first principle component
- ϕ_1 = loadings for first principle component
- i think you woild then multiply the original data matrix by the the loadings vector to get the first principle component.

C.8 Wk 10 — Optimisation

- 2nd deriv test: pos = min; neg = max
- Lagrangian: $L(x, y) = f(x, y) + \lambda g(x, y)$
- constrained opti: 1) set of Lagrangian, 2) Take the first derivatives $\frac{\partial L}{\partial x}$, $\frac{\partial L}{\partial y}$, and $\frac{\partial L}{\partial \lambda}$ and set them equal to 0, 3) solve
  PRACTICE SOLVING
- Hessian: each element a 2nd order partial derivative of a function. Has same dimensions as number of functions in the function:
  - all eigenvalues pos: local min
  - all eigenvalues neg: local max
  - mixed: saddle point
- Gradient (of a function)