Probability Density Functions

Week 2 - Lecture 3

Dr Matthew Ellis - COM2004/3004 Data Driven Computing 23/24

Goals of today's session

By the end of today's session you should be able to:

1. Define what is meant by the cumulative and probability density functions.

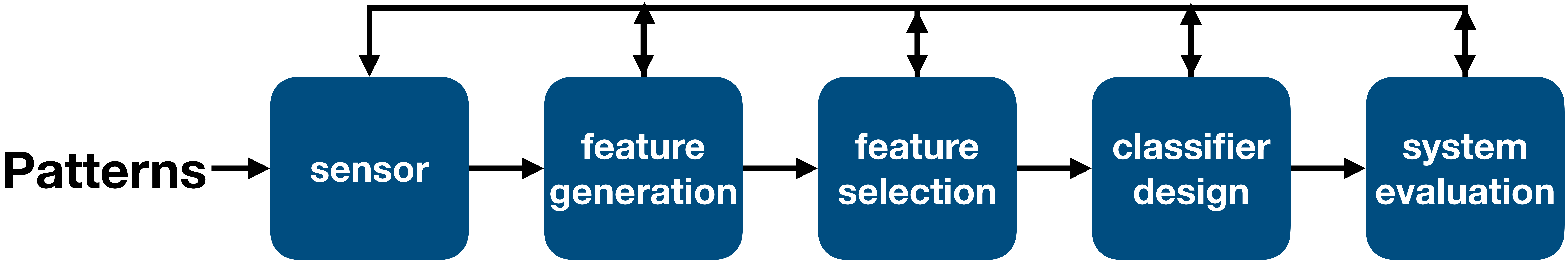
2. Define and sketch some simple probability distributions.

3. Understand the notations used to handle multivariate data.

4. Recall how the calculate the sample mean and covariances for the multivariate Gaussian distribution.

Lecture 2 Recap

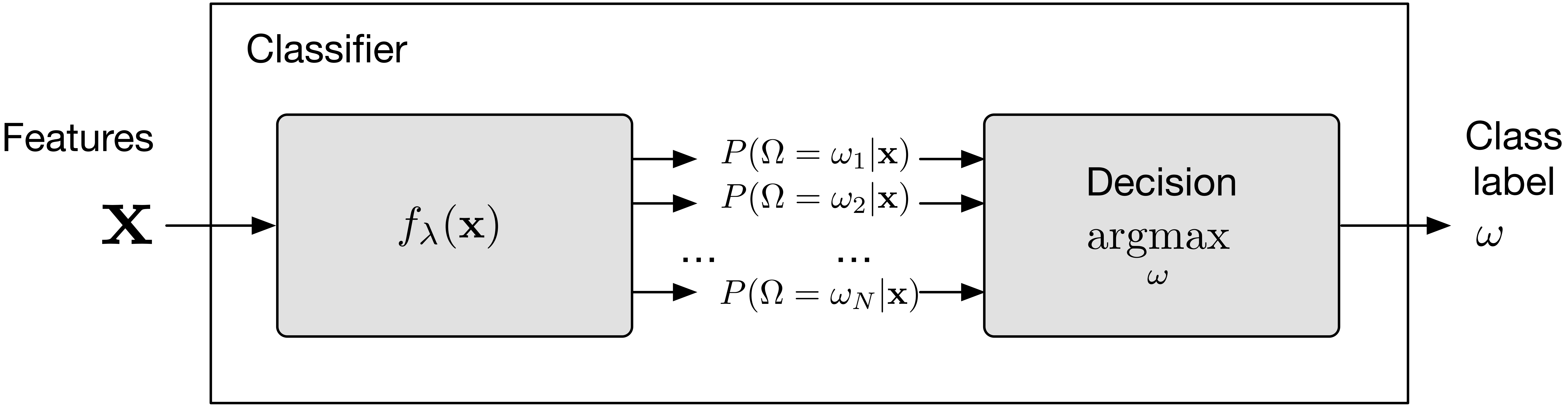
Designing a classifier



Bayes' Decision Rule

Bayes decision rule: label $\mathbf{x}$ with the class for which $P(\Omega = \omega \,|\, \mathbf{x})$ is the greatest. i.e., for a given input $\mathbf{x}$, the output label ω is given by

$$\omega^* = \underset{\omega \in \Omega}{\operatorname{argmax}}\, P\left(\Omega = \omega \,|\, \mathbf{x}\right)$$



e.g., if $P(\Omega = \text{cat} \,|\, \mathbf{x}) = 0.9$ and $P(\Omega = \text{dog} \,|\, \mathbf{x}) = 0.1$ then $\omega = \text{cat}$

Bayes' Rule

$$\underbrace{P\left(\omega_i \mid \mathbf{x}\right)}_{\text{Posterior}} = \frac{\overbrace{P\left(\mathbf{x} \mid \omega_i\right)}^{\text{Likelihood}} \overbrace{P\left(\omega_i\right)}^{\text{Prior}}}{\underbrace{P\left(\mathbf{x}\right)}_{\text{Evidence}}}$$

Prior: probability of the class **before** observing the data (evidence).

Posterior: probability of the class **after** observing the data (evidence).

Likelihood: probability of **observing** a (random) data point **given** the class label (**compatibility** of the data with the given class).

Evidence or marginal likelihood: probability of **observing** a (random) data point across **all classes.**

Bayes' Decision Theory

Consider a simple case with 2 classes, ω_1 and ω_2.

Our decision rules implies:

$$\mathbf{x} \text{ belongs to } \omega_1 \text{ if } P(\omega_1 \mid \mathbf{x}) > P(\omega_2 \mid \mathbf{x}),$$
$$\text{otherwise it belongs to } \omega_2$$

Using Bayes' rule, this criterion can be expressed as

$$P(\mathbf{x} \mid \omega_1)P(\omega_1) > P(\mathbf{x} \mid \omega_2)P(\omega_2)$$

Summary

We want to find a **class label** for a thing after observing its features

We can formalise this as picking the class ω which has the biggest **posterior probability**, $P(\omega \mid \mathbf{x})$

We can't compute posteriors directly so **we use Bayes' rule and compute**, $P(\mathbf{x} \mid \omega)P(\omega)$ instead because

"Posterior is proportional to likelihood times prior"

The class priors and the parameters of the distributions $P(\mathbf{x} \mid \omega)$ can be learnt during a **training stage using labelled data**.

Probability Density Functions

Continuous random variables

Our features, x, are typically continuous valued quantities not integers:

- We need to consider **continuous** random variables (RVs). e.g X is the weight of a Tiger, Y is the height of a person
- But what is the probability that one of these is exactly equal to a specific value, $P(X = x)$?
- It's actually infinitesimally small (I.e ≈ 0)

How can we conceptualise a continuous random variable?

- It's like having an **infinite amount of discrete events**.
- This is like having a histogram where the bin sizes decrease to 0. There will be an infinite number of bins with a height ≈ 0.
- We cannot tabulate all these infinite events, so how can we represent continuous probability distributions?

Continuous random variables

- The quantity may not be exactly equal to a value but…
- We can talk about the probability that it is **within a certain range**.
- For example, if X is the height (in metres) of a randomly selected person, then we can easily define:
    - $P(X < 1.5)$ i.e height less than 1.5 metres.
    - $P(X > 1.4 \cap X < 1.6)$ i.e height is greater than 1.4 metres but less than 1.6 metres.
    - These all have **non-zero probabilities**.

Probability density function

Definition:

The **cumulative distribution function** (CDF) is

$$F(x) = P(X \leq x)$$

where the notation $X \leq x$ represents the all the outcomes smaller than or equal to x.

We can then define a **probability density function** (PDF) which is related to the CDF by

$$F(x) = \int_{-\infty}^{x} p(u)du \quad \text{and similarly} \quad p(x) = \frac{dF(x)}{dx}$$

Properties of the PDF

There are 2 properties of all PDF's that arise from this definition:

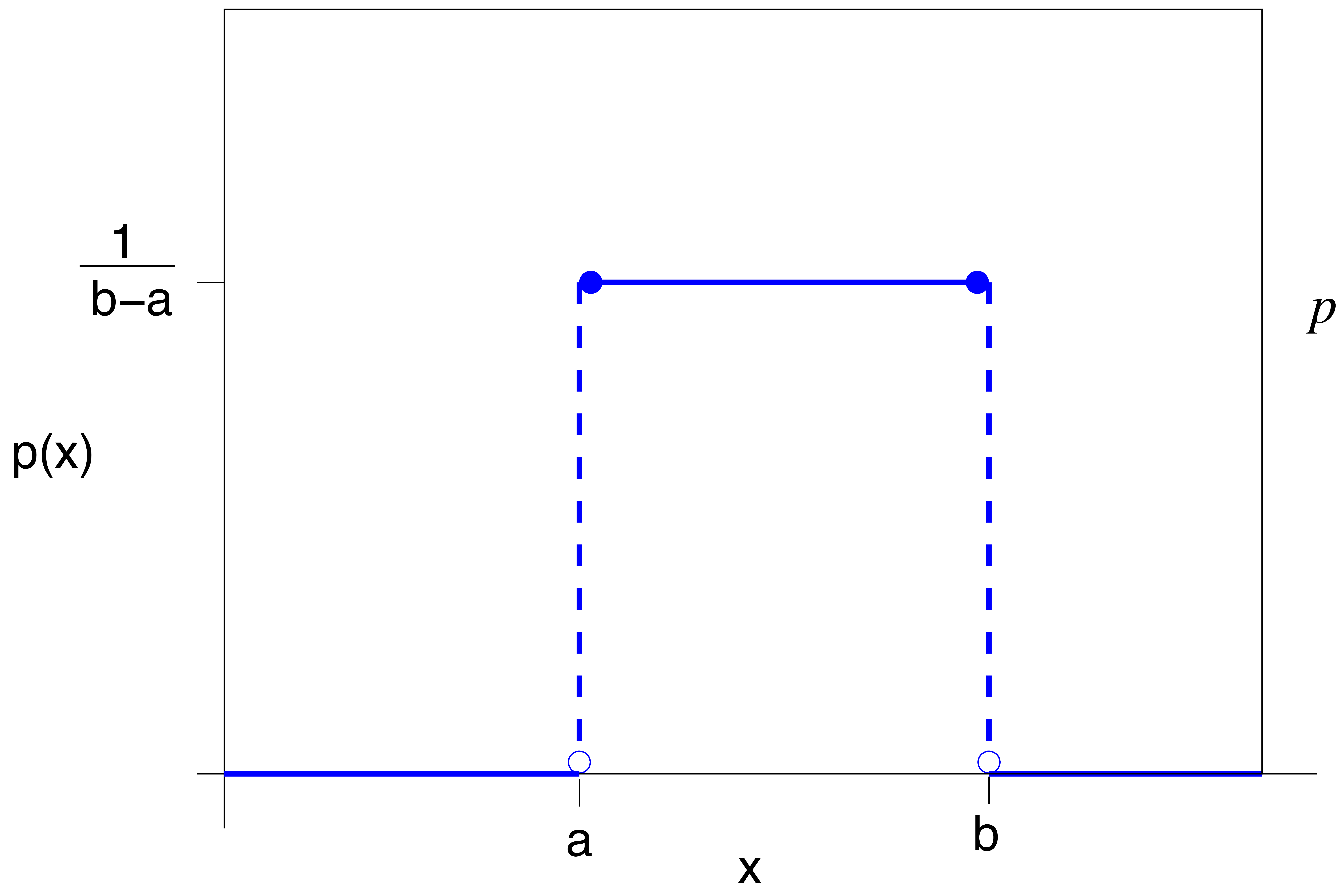
1. Its values can't be negative:

$$p(x) > 0 \text{ for all } x$$

2. The area under the curve must be equal to 1:

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

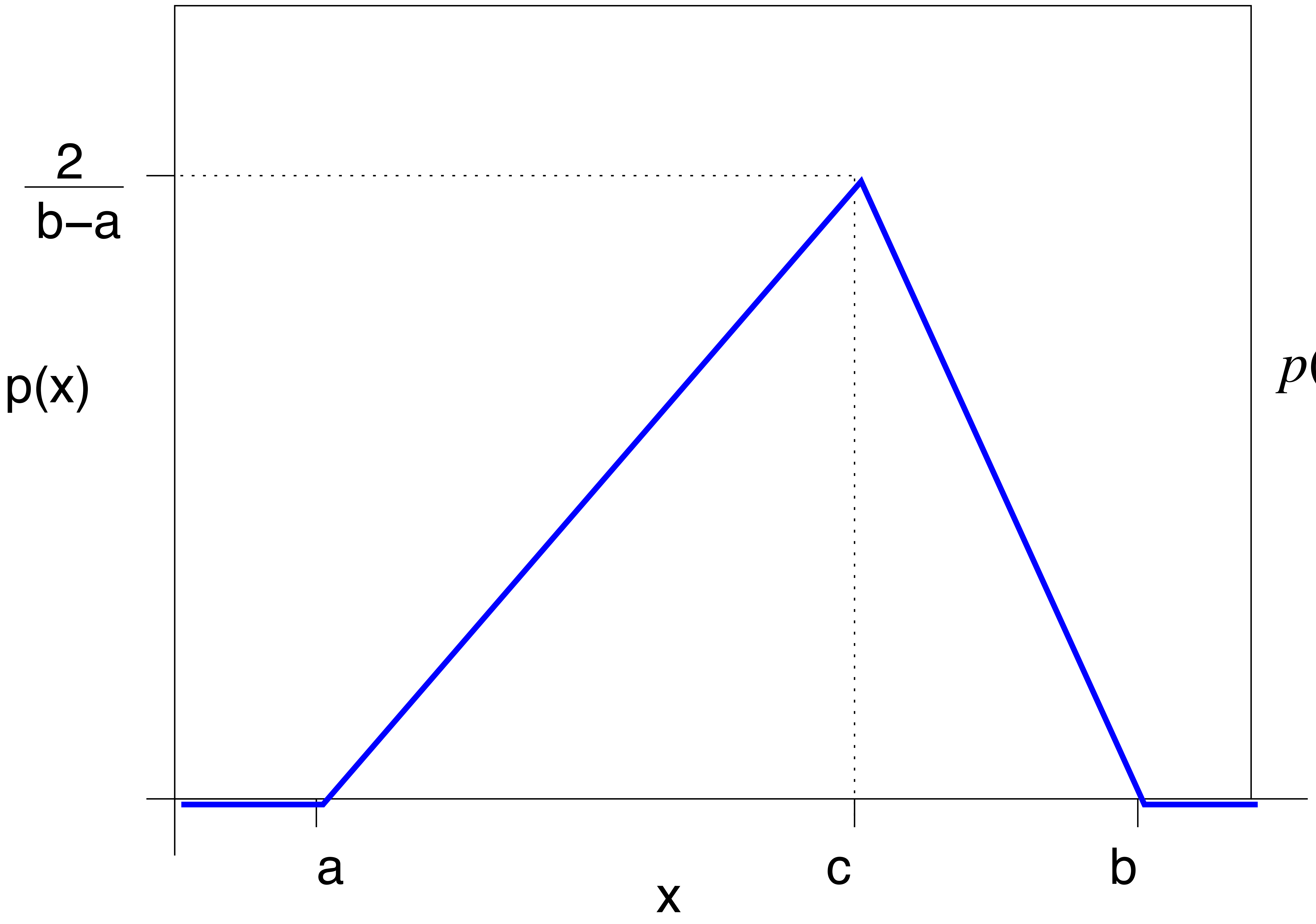
Uniform distribution



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

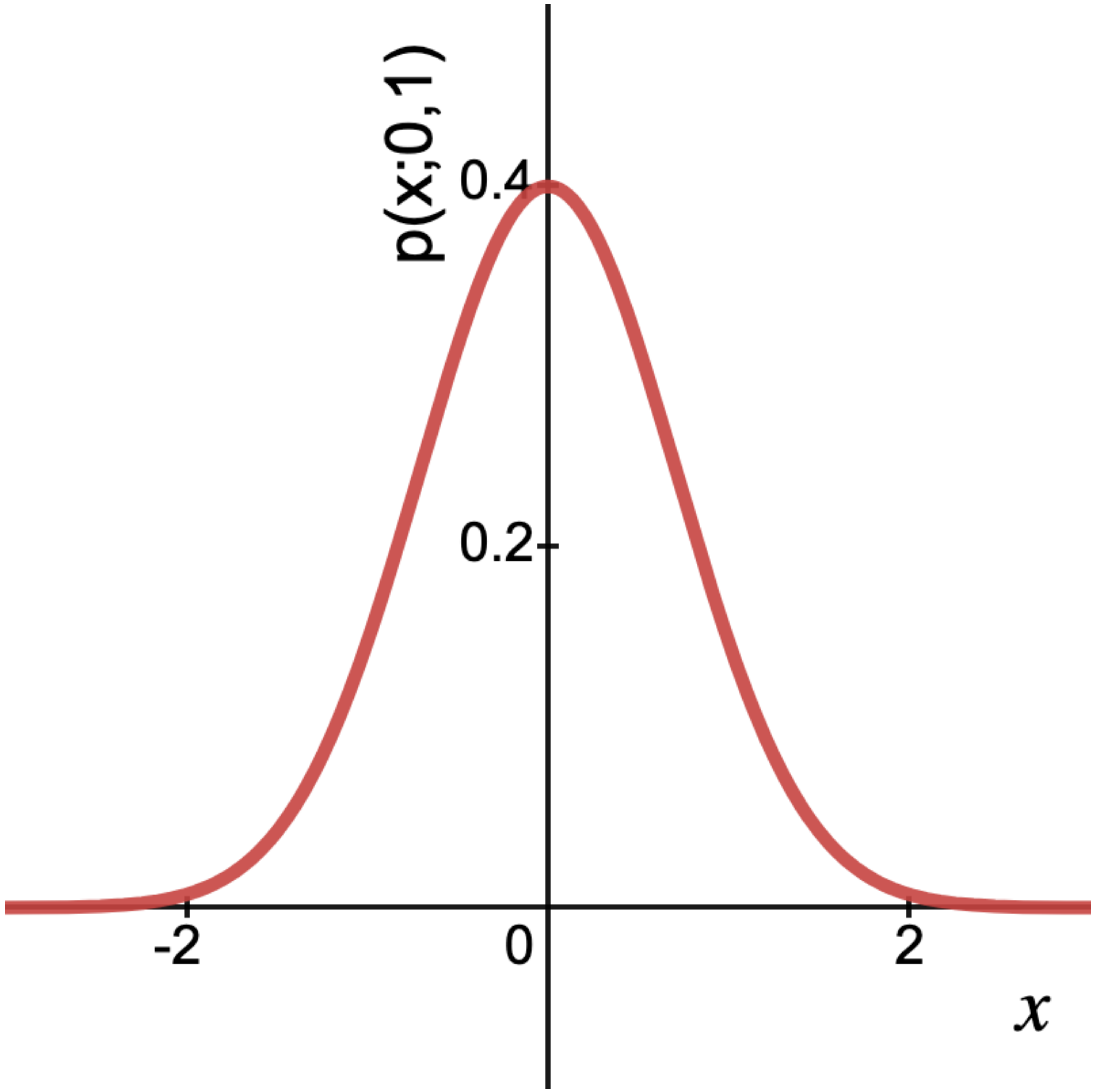
$$= U(a,b)$$

Triangle distribution



$$p(x;a,b,c) = \begin{cases} \frac{2}{(b-a)}\frac{x-a}{(c-a)} & a < x < c \\ \frac{2}{(b-a)}\frac{b-x}{(b-c)} & c < x < b \\ 0 & \text{otherwise} \end{cases}$$

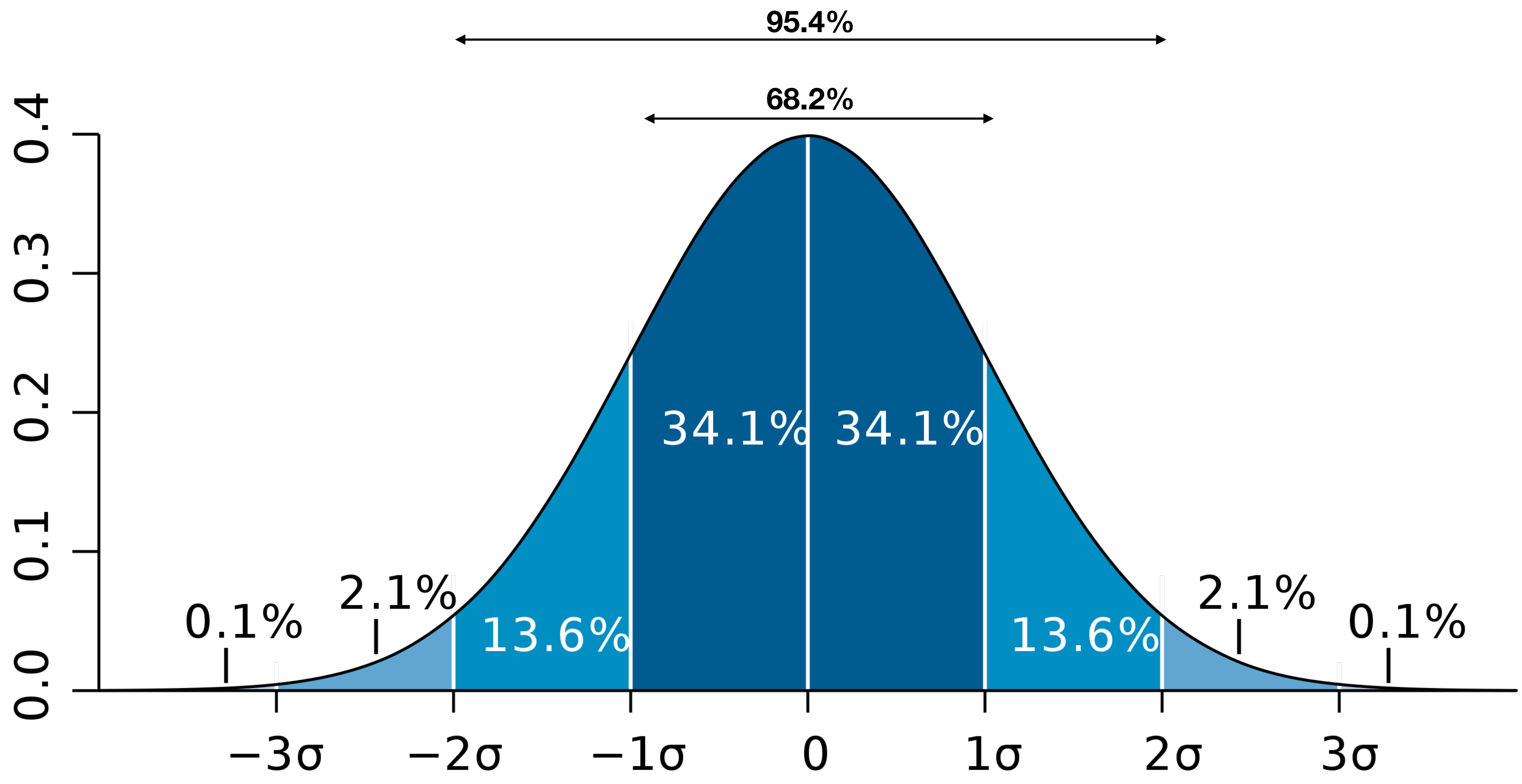
Univariate Gaussian distribution



$$p(x;\mu,\sigma) = \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}}_{\textbf{Normalisation}} \underbrace{\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}_{\textbf{Shape}}$$

- Parameter μ controls the location of the peak.
- Parameter σ controls the width of the peak.
- Defined for the range $-\infty < x < \infty$.
- Also called the **normal distribution**.

Gaussian distribution



Summary

Bayes decision rule requires that we know the distribution of our features, x, for each class, ω. i.e., we will need to **learn $p(x \mid \omega)$ for each class.**

x is a continuous quantity so $p(x \mid \omega)$ is a **continuous probability distribution**.

Continuous probability distributions are represented using the **probability density function (pdf)**.

The pdf is a function that is always **positive and has unit area under the curve**.

It can have an **arbitrarily complex shape**, but we will be assuming it has a specific form, e.g. a Gaussian, and then our learning algorithm will just **learn a small number of parameters** (e.g., the position and width of the peak)

Processing multivariate data

Multivariate data

Multivariate random sample

- Multivariate simply means that something is described by more than 1 variable

- Basically another way of saying the data is multidimensional

- A multivariate observation can be represented by a vector, $\mathbf{x}$. By convention, x will be a column vector:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_L \end{pmatrix} = \left(x_1, x_2, \cdots, x_L\right)^T$$

Multivariate data

If we have many samples, $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$, then we can store the data in a matrix:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1L} \\ x_{21} & x_{22} & \cdots & x_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NL} \end{pmatrix}$$

- $\mathbf{x}$ is formulated as a N by L matrix, having N samples (rows) and L features (columns).

- x_{nl} is the l-th feature of the n-th sample.

Multivariate data

e.g height, weight, age and IQ for five people (i.e 5 samples)

$$\mathbf{X} = \begin{pmatrix} 173 & 66 & 27 & 120 \\ 162 & 48 & 16 & 90 \\ 158 & 50 & 43 & 110 \\ 171 & 75 & 53 & 100 \\ 161 & 61 & 61 & 80 \end{pmatrix}$$

Calculating the mean for multivariate data

Mean - measurement of location/centre of feature l :

$$\bar{x}_l = \frac{1}{N}\sum_{n=1}^{N} x_{nl}$$

e.g average weight

$$\bar{x}_2 = \sum n = 1^N x_{n2} = \frac{66 + 48 + 50 + 75 + 61}{5} = 60$$

Calculating the mean for multivariate data

Multivariate sample mean is a vector of length L

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_L \end{pmatrix} = \begin{pmatrix} \frac{1}{N} \sum_{n=1}^{N} x_{n1} \\ \vdots \\ \frac{1}{N} \sum_{n=1}^{N} x_{nL} \end{pmatrix}$$

Or using vector addition

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n$$

Calculating the mean for multivariate data

e.g height, weight, age and IQ for five people (i.e 5 samples)

$$\mathbf{X} = \begin{pmatrix} 173 & 66 & 27 & 120 \\ 162 & 48 & 16 & 90 \\ 158 & 50 & 43 & 110 \\ 171 & 75 & 53 & 100 \\ 161 & 61 & 61 & 80 \end{pmatrix}$$

Sample mean vector:

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \\ \bar{x}_4 \end{pmatrix} = \begin{pmatrix} 165 \\ 60 \\ 40 \\ 100 \end{pmatrix}$$

Variance

Measure of the spread of the data:

$$s_{ll} = \frac{1}{N} \sum_{n=1}^{N} (x_{il} - \bar{x}_l)^2$$

e.g variance of the weight:

$$s_{22} = \frac{(66 - 60)^2 + (48 - 60^2) + (50 - 60)^2 + (75 - 60)^2 + (61 - 60)^2}{5}$$

$$s_{22} = 101.2$$

Covariance

Extension of the variance to multi-dimensions.

Measure of association/correlation **between two variables**:

$$s_{kl} = \frac{1}{N} \sum_{n=1}^{N} (x_{nk} - \bar{x}_k)(x_{nl} - \bar{x}_l)$$

This describes how the **two variables relate or change together**.

Calculating the mean for multivariate data

e.g height, weight, age and IQ for five people (i.e 5 samples)

$$\mathbf{X} = \begin{pmatrix} 173 & 66 & 27 & 120 \\ 162 & 48 & 16 & 90 \\ 158 & 50 & 43 & 110 \\ 171 & 75 & 53 & 100 \\ 161 & 61 & 61 & 80 \end{pmatrix}$$

Covariance between height (column 1) and weight (column 2)

$$s_{12} = \frac{1}{5}\big[(173 - 165)(66 - 60) + (162 - 165)(48 - 60) +$$

$$(158 - 165)(50 - 60) + (171 - 165)(75 - 60) + (161 - 160)(61 - 60)\big]$$

$$s_{12} = 48$$

Covariance matrix

Multivariate sample variances and covariances can be formed into a $L \times L$ matrix:

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1,L} \\ s_{21} & s_{22} & \cdots & s_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ s_{L1} & s_{L2} & \cdots & s_{LL} \end{pmatrix}$$

- Element s_{ij} is the covariance between feature i and feature j.
- The **diagonal** elements are the variances. The **off-diagonals** are the covariances.
- Symmetric about the diagonal, $s_{ij} = s_{ji}$

Population vs sample statistics

When we are using a sample to estimate the parameters of a whole population then we need to be careful about our estimates.

If $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$ are random samples drawn from a distribution with the **population mean** vector $\boldsymbol{\mu}$ and population covariance matrix $\boldsymbol{\Sigma}$. Then

- The sample mean $\bar{\mathbf{x}}$ is an **unbiased estimate** of the population mean $\boldsymbol{\mu}$.
- But the sample covariance $\mathbf{S}$ is a **biased estimate** of the population covariance $\boldsymbol{\Sigma}$.
- The **unbiased estimate** of the population covariance is $\dfrac{N}{N-1}\mathbf{S}$.

Unbiased covariance

Using this we can then define the **unbiased** estimate of the population covariance as

$$s_{kl} = \frac{1}{N-1} \sum_{n=1}^{N} (x_{nk} - \bar{x}_k)(x_{nl} - \bar{x}_l)$$

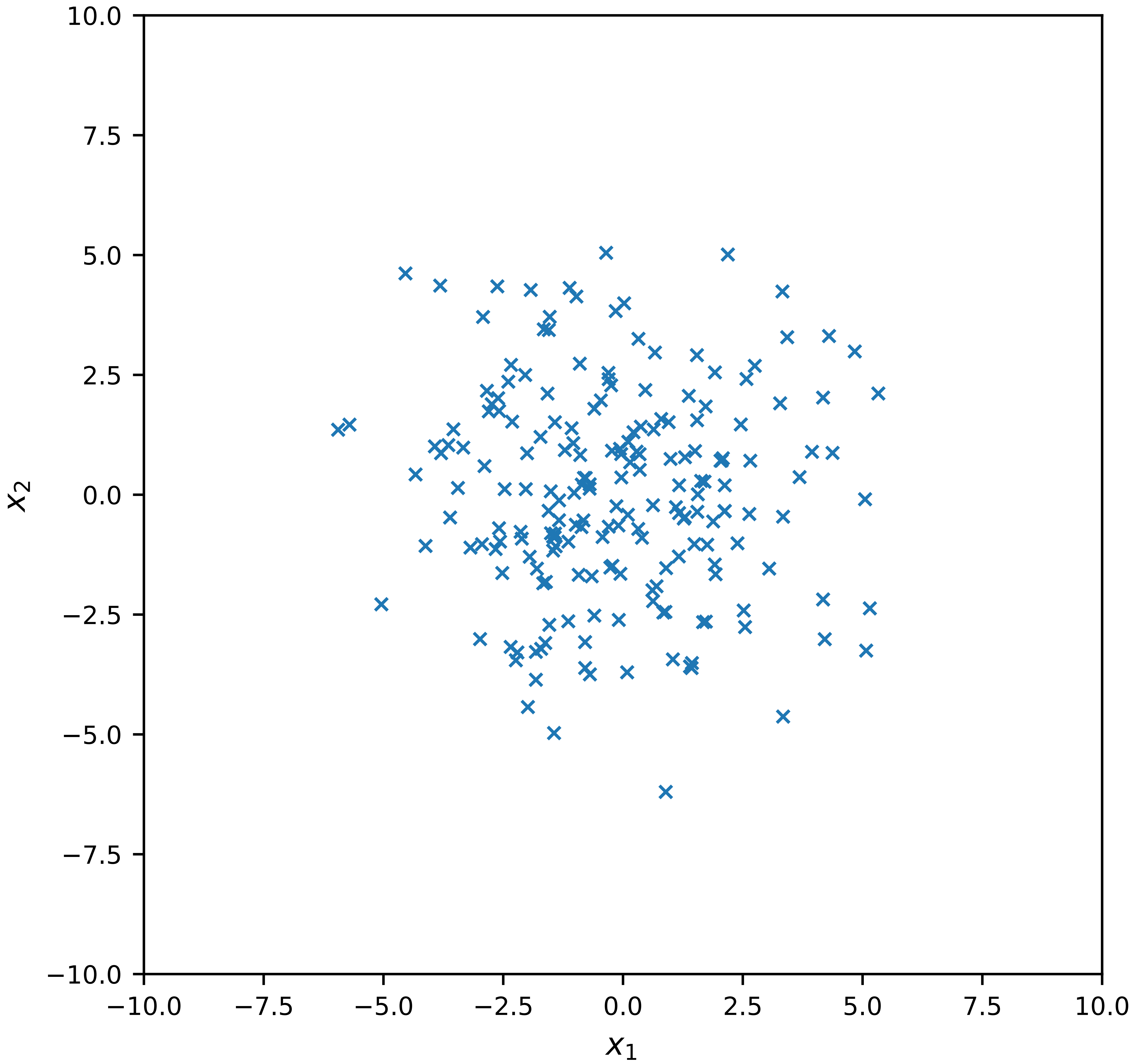
Don't worry about why this is for now - reasons will be explained later when we discuss 'parameter estimation'.

Visualising multivariate data

Visualising multivariate data

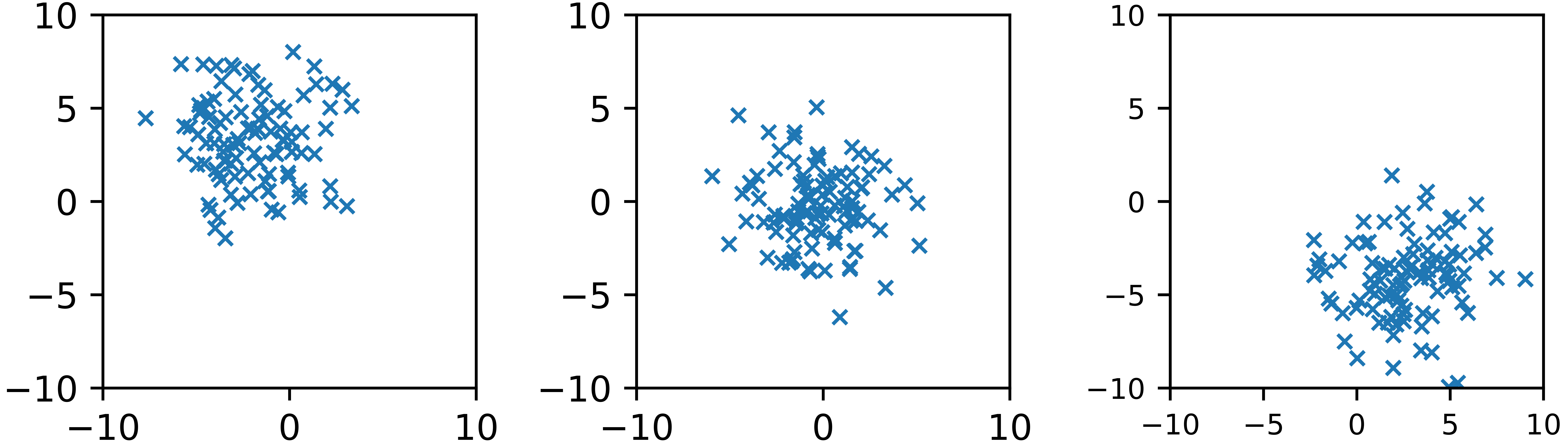
For 2d data we can plot is using a scatter plot.

Each sample, $\mathbf{x}_i$, is plotted at a location (x_1, x_2).



Varying the mean

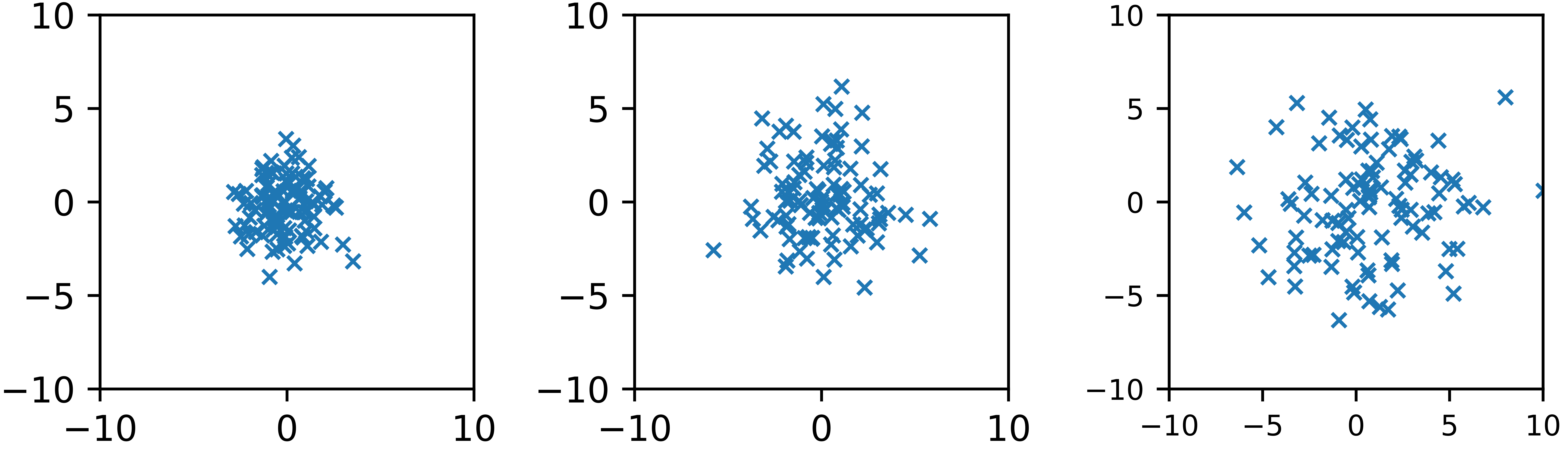
$$\bar{\mathbf{x}} = \begin{pmatrix} -2 \\ 3 \end{pmatrix} \qquad \bar{\mathbf{x}} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \bar{\mathbf{x}} = \begin{pmatrix} 3 \\ -4 \end{pmatrix}$$



Blob shape remains the same but moves.

Changing the variance

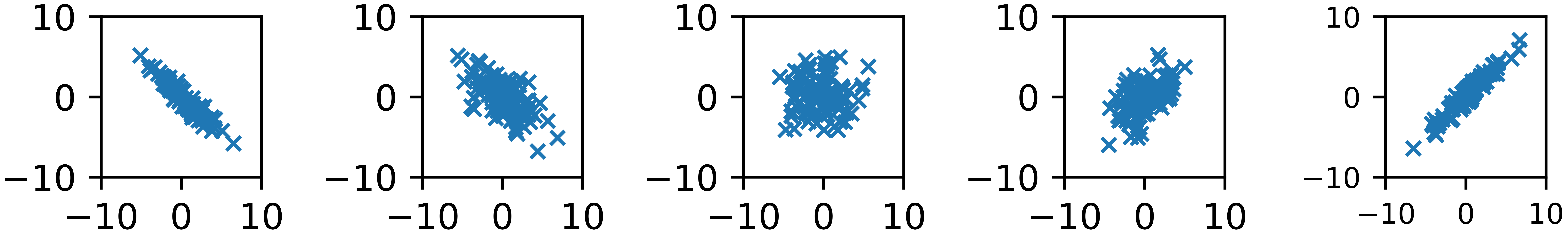
$$\mathbf{S} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \qquad \mathbf{S} = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix} \qquad \mathbf{S} = \begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}$$



Spreads wider with increasing variance but remains circular.

Changing the covariance (fixed variance)

$$\mathbf{S} = \begin{pmatrix} 5 & -4.8 \\ -4.8 & 5 \end{pmatrix} \quad \mathbf{S} = \begin{pmatrix} 5 & -3 \\ -3 & 5 \end{pmatrix} \quad \mathbf{S} = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix} \quad \mathbf{S} = \begin{pmatrix} 5 & 3 \\ 3 & 5 \end{pmatrix} \quad \mathbf{S} = \begin{pmatrix} 5 & 4.8 \\ 4.8 & 5 \end{pmatrix}$$



Spread elongated along the diagonal with increasing covariance.

Summary

When processing multivariate data each sample is represented as a **vector containing multiple features**.

An entire dataset is represented by building a **data matrix** with each **sample forming a separate row**.

We can compute **means** and **variances** of features by taking the mean or variance of the corresponding data matrix column.

We can also compute **covariances** between pairs of features. The full set of covariances (ie. between every pair of features) can be stored in a **covariance matrix**.

The distribution of the data can be understood by looking at its **mean vector and covariance matrix**.

For 2-D data we can visualise the distribution using a scatter plot.

Thanks for listening