

# Applied Statistics - Notes - v0.6.0

260236

June 2025

## Preface

Every theory section in these notes has been taken from two sources:

- The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. [2]
- An Introduction to Statistical Learning: with Applications in Python. [3]
- Applied Multivariate Statistical Analysis. [4]
- Course slides. [5]

About:

 [GitHub repository](#)



These notes are an unofficial resource and shouldn't replace the course material or any other book on applied statistics. It is not made for commercial purposes. I've made the following notes to help me improve my knowledge and maybe it can be helpful for everyone.

As I have highlighted, a student should choose the teacher's material or a book on the topic. These notes can only be a helpful material.

# Contents

<b>1</b>	<b>Business Data Analytics</b>	<b>5</b>
1.1	Multivariate Descriptive Statistics . . . . .	5
1.2	Dimensionality Reduction . . . . .	12
1.3	Principal Component Analysis . . . . .	14
1.4	PCA Reference System . . . . .	17
1.5	PCA as Optimization Problem . . . . .	21
1.6	Proportion of Variance Explained (PVE) . . . . .	26
1.7	Covariance vs. Correlation in PCA . . . . .	27
1.8	PCA via SVD - Computational Aspects . . . . .	29
<b>2</b>	<b>Clustering Methods</b>	<b>31</b>
2.1	Introduction . . . . .	31
2.2	Defining Similarity in Clustering . . . . .	33
2.3	Clustering Validity . . . . .	35
2.3.1	External Metrics . . . . .	35
2.3.2	Internal Metrics . . . . .	37
2.4	Hierarchical Clustering . . . . .	40
2.5	K-Means . . . . .	42
2.5.1	Introduction . . . . .	42
2.5.2	Definition . . . . .	43
2.5.3	Initialization Issues in K-Means . . . . .	49
2.5.4	K-Means as an Optimization Problem . . . . .	50
2.5.5	Algorithm . . . . .	52
2.5.6	Limitations . . . . .	53
2.6	Gaussian Mixture Models (GMMs) . . . . .	54
2.6.1	Introduction . . . . .	54
2.6.2	Mathematical Foundation . . . . .	61
2.6.3	Responsibilities . . . . .	63
2.6.4	Expectation-Maximization (EM) Algorithm . . . . .	67
2.6.5	Comparison between GMM and K-Means . . . . .	69
<b>3</b>	<b>Discriminant Analysis</b>	<b>71</b>
3.1	From Unsupervised to Supervised . . . . .	71
3.2	Introduction to Supervised Learning . . . . .	73
3.3	Linear Discriminant Analysis (LDA) . . . . .	76
3.4	Quadratic Discriminant Analysis (QDA) . . . . .	83
3.5	Comparison LDA vs QDA . . . . .	91
3.6	Bayes Classifier . . . . .	92
3.6.1	What is the Bayes Classifier? . . . . .	92
3.6.2	Bayes Error Rate . . . . .	95
3.6.3	Bayes vs LDA/QDA . . . . .	98
<b>4</b>	<b>Linear Regression</b>	<b>100</b>
4.1	Regression vs Classification . . . . .	100
4.2	Simple Linear Regression (SLR) . . . . .	101
4.2.1	Model Formulation and OLS Estimation . . . . .	101
4.2.2	Geometric Interpretation of Linear Regression . . . . .	106
4.3	Model Evaluation . . . . .	109

4.4	Statistical Inference . . . . .	112
4.4.1	Sampling Distributions & Central Limit Theorem (CLT) .	113
4.4.2	t-Test for Individual Coefficients . . . . .	115
4.4.3	Global F-Test . . . . .	116
4.4.4	Summary . . . . .	118
4.5	Uncertainty Intervals . . . . .	120
4.5.1	Confidence Interval (CI) for the Mean Response . . . . .	121
4.5.2	Prediction Interval (PI) for a New Observation . . . . .	123
4.6	Multiple Linear Regression (MLR) . . . . .	125
4.6.1	Model Expansion . . . . .	126
4.6.2	Matrix Formulation . . . . .	128
<b>5</b>	<b>Model Selection and Regularization</b>	<b>130</b>
5.1	Motivation & Problem Setup . . . . .	130
5.2	Collinearity . . . . .	131
5.2.1	Definition . . . . .	131
5.2.2	Variance Inflation Factor (VIF) . . . . .	134
	<b>Index</b>	<b>138</b>

# 1 Business Data Analytics

## 1.1 Multivariate Descriptive Statistics

When we move from **analyzing** a single variable (univariate analysis) to **multiple variables at once**, we enter the realm of **Multivariate (MV) analysis**. A natural question arises: *Is multivariate analysis just a replication of univariate analysis across several variables?*

The answer is no, multivariate analysis introduces new and fundamental questions that cannot be answered by simply analyzing variables individually. The **core focus** shifts to **understanding how these variables interact with each other**. Specifically, we are concerned with the **dependence** and **correlation between variables**.

### Covariance: Measuring Joint Variability

To capture how two variables vary together, we use **Covariance**. The **Sample Covariance** between variables  $x_j$  and  $x_k$  is calculated as:

$$\text{cov}_{jk} = \text{Cov}(x_j, x_k) = s_{jk} = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \quad (1)$$

- $s_{jk} = 0 \Rightarrow$  implies that there is **no linear relationship** between the two variables.
- $s_{jk} > 0 \Rightarrow$  suggests that **as one variable increases, the other tends to increase**.
- $s_{jk} < 0 \Rightarrow$  one **variable** tends to **decrease when the other increases**.

### Covariance Is Not Standardized

The **value of covariance is not standardized**, it depends on the units of measurement, which makes comparisons difficult. For example:

- Suppose we're measuring
  - Height in centimeters
  - Weight in kilograms
- The covariance between height and weight will be expressed in *centimeter-kilogram* units.

Now imagine we convert height to meters. The covariance value changes, because now we're multiplying meters  $\times$  kilograms instead of centimeters  $\times$  kilograms. Even though the **relationship between height and weight hasn't changed**, the **numerical value of covariance does change due to this unit change**. Because of unit dependency, it's hard to compare covariances between different variable pairs. Finally, it is hard to interpret the **magnitude of covariance in any absolute sense** (e.g., is 50 a large covariance or small? It depends on the units!).

### ✔ Correlation: Standardized Covariance

To standardize covariance and **measure the strength of a linear relationship** on a scale between  $-1$  and  $1$ , we use the **Correlation** coefficient, defined as:

$$\text{cor}_{jk} = r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}} = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \sqrt{\sum_{i=1}^n (x_{ik} - \bar{x}_k)^2}} \quad (2)$$

This formula divides the covariance by the product of the standard deviations of the two variables, giving a **unitless value**:

- $r = 0$ : No linear correlation
- $r > 0$ : Positive correlation (**both variables increase or decrease together**)
- $r < 0$ : Negative correlation (**one increases while the other decreases**)
- $|r| = 1$ : Perfect correlation (**exact linear dependence**)

Thus, correlation not only **reveals the direction of the relationship** but also its **strength**.

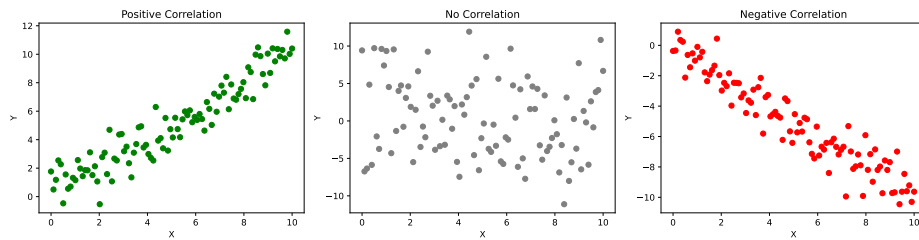


Figure 1: Direction of correlation: positive (left), none (center), negative (right).

### Describing MV Data: Vector and Matrices

When analyzing multivariate data:

- We compute the **vector of Sample Means**:

$$\bar{\mathbf{X}} = [\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p] \quad (3)$$

- And the **symmetric and positive semidefinite** (eigenvalues are non-negative) **Variance-Covariance matrix  $\mathbf{S}$** , which **summarizes the covariances between all pairs of variables**:

$$\mathbf{S} = \begin{bmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{pp} \end{bmatrix} \quad (4)$$

- Alternatively, we can use the **Correlation matrix  $\mathbf{R}$** , where **all diagonal elements are 1** (because each variable is perfectly correlated with itself) and **off-diagonal elements are correlation coefficients**:

$$\mathbf{R} = \begin{bmatrix} 1 & \cdots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{p1} & \cdots & 1 \end{bmatrix} \quad (5)$$

## 📊 Scatterplots - Visualizing Variable Pairs

One of the most intuitive and widely used tools in multivariate analysis is the **2D Scatterplot**. Each plot shows how two variables relate to each other:

- ✓ Clusters or linear trends can indicate correlation or dependence.
- ✓ Scatterplots are **ideal for spotting positive, negative, or no correlation** visually.

However, scatterplots have a **limitation**: they **only** allow us to **analyze two variables at a time**. When dealing with many variables, the **number of possible pairings becomes large**, making it **difficult to read or interpret** the scatterplots individually. This is where quantitative measures (like correlation matrices) and higher-dimensional graphics come into play.



Figure 2: Scatterplot matrix of four variables. This scatterplot matrix displays all pairwise relationships among four variables:

- Diagonal plots (top-left to bottom-right): Histograms showing the distribution of each variable.
- Off-diagonal plots: 2D scatterplots illustrating the relationship between each pair of variables.



### Rotated Plots in 3D - Capturing Complexity

When dealing with **three variables**, we can **extend scatterplots into 3D space** using **Rotated plots**. These visualizations allow us to:

- ✓ **Explore interdependencies** among three variables at once.
- ✓ **Gain spatial insight** into how data points spread in three-dimensional space.
- ✓ Observe **complex patterns that are invisible in 2D**.

Yet again, as we move **beyond three variables**, visualizing becomes **impractical**, our brains cannot easily comprehend 4D or higher dimensions. Hence, dimensionality reduction techniques like PCA are often used alongside visualizations to make high-dimensional data “digestible”.



Figure 3: A simple rotated plots in 3D.

### ☰ Star Plots - Shape-Based Comparison

**Star plots** offer a creative way to **represent multivariate data**:

- **Each variable is represented as a ray** (spoke) starting from a central point.
- The **length of each ray** corresponds to the **value** of that variable.
- When **rays are connected**, they form a “star-like shape” **unique** to each observation.

This method is **excellent for comparing patterns** between observations:

- ✓ **Similar shapes** suggest **similar data profiles**.
- ✓ **Differences in shape** can **quickly highlight outliers** or clusters.

However, star plots have limitations:

- ✗ They **do not quantify correlation**.
- ✗ The **direction** and **magnitude** of relationships between variables are **not explicit**.
- ✗ They are better for **visual pattern recognition** than for precise statistical analysis.



Figure 4: Star Plot (Radar Chart) of Multivariate Data.

### 📊 Chernoff Faces - Human-Centric Visualization

Chernoff faces [1] are an **innovative visualization method** where **multivariate data is represented as a human face**:

- Each **variable controls a facial feature** (e.g., mouth curvature, eye size, nose length).
- **People are naturally attuned to recognizing faces** and subtle differences in expressions.
- Hence, Chernoff faces **leverage human perception** for quickly comparing **multivariate observations**.

Despite being engaging, Chernoff faces also have **drawbacks**:

- ✗ They **do not provide numerical precision**.
- ✗ The **mapping of variables to facial features** can be **arbitrary**.
- ✗ They work best as a **qualitative summary tool** rather than for deep statistical inference.

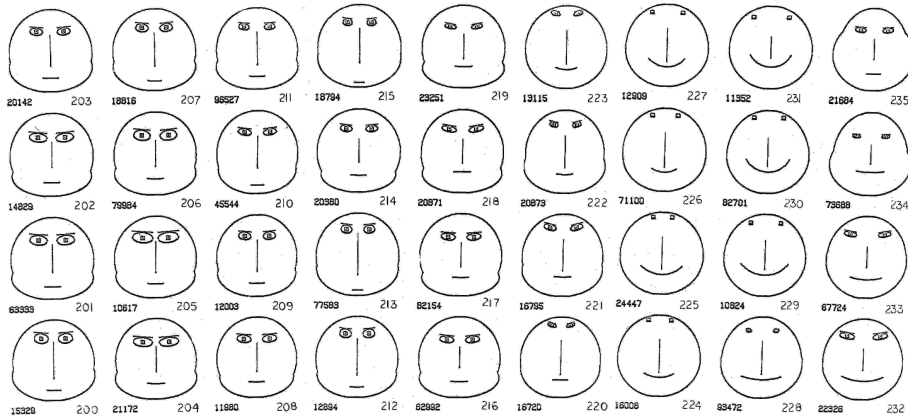


Figure 5: Some Chernoff faces. [1]

Graphic Type	Strengths	Limitations
Scatterplots	Clear view of pairwise relationships	Hard to scale beyond 2 variables
3D Plots	Visualizes 3-variable interaction	Limited to 3 dimensions, requires rotation
Star Plots	Quick shape-based comparison across variables	No quantification, poor at showing correlations
Chernoff Faces	Leverages facial perception for comparison	Subjective, lacks precision

Table 1: When and why to use graphics.

## 1.2 Dimensionality Reduction

### ⚠ The Challenge: Data in High Dimensions

In many real-world problems, we **collect multiple variables for each observation**. For example, in a medical study, a patient might be described by age, weight, blood pressure, and dozens of test results. This leads to **high-dimensional data**, where **each observation is a point in a complex, multi-dimensional space** (formally, a Euclidean space of dimension  $p$ ).

**The problem?** As the **number of variables ( $p$ ) increases**, the **data becomes harder to visualize, interpret, and model**:

- ❗ Some **variables** might be **redundant** or **highly correlated**.
- ❗ **Computations** become more **expensive**.
- ❗ **Patterns** become **obscured** by the complexity.

### 🎯 Goal of Dimensionality Reduction

We want to **summarize the data using fewer variables**, say  $k$  derived variables (with  $k < p$ ), that still **retain most of the information**. This process is a balancing act:

- **Clarity**: fewer variables make data **easier to understand and visualize**.
- **Risk of oversimplification**: reducing dimensions too much can cause **loss of important information**.

The key concept here is that in statistics, **information is variability**. The **more variability we retain from the original data, the more information we preserve**.

#### Example 1: Blood Cells

Imagine we measure thickness and diameter for a set of red blood cells:

- Each cell = one observation with two variables.
- We can represent this as a table (numbers) or as a 2D scatterplot.

Now we ask ourselves: *Can we describe these cells using only one feature instead of two?*

If we choose only diameter or only thickness, we'll lose detail:

- Some cells will appear more similar than they really are.
- We miss variability that distinguishes them.

So, we seek a better single feature, one that captures the most variation possible from both thickness and diameter combined.

### 📌 The Statistical Insight: Maximize Variability

Rather than randomly picking a feature, we **analyze the directions along which the data varies the most**.

1. First, we find the **direction of maximum spread**.
2. Then, the **second most spread direction**, orthogonal to the first.

This is the essence of **Principal Component Analysis (PCA)**, a **dimensionality reduction technique** that finds the best directions (linear combinations of variables) to **project the data**, while **maximizing retained variability**.

### 📌 Dimensionality Reduction in Practice

Let's formalize the idea:

- We start with a **data matrix**  $X$  of shape  $n \times p$  ( $n$  observations,  $p$  variables).
- The **goal** is to **obtain a new matrix**  $M$  of shape  $n \times k$ , with  $k < p$ , that **captures most of the variability**.
- The **difference** between  $X$  and  $M$  is **residual variation**, the **information lost**.

In summary, **Dimensionality Reduction** is about simplifying complexity: **transforming a large set of variables into a smaller**, more interpretable set **without losing the essence of the data**. It's central to data exploration, preprocessing, and modeling, especially when working with high-dimensional datasets.

## 1.3 Principal Component Analysis

### ✔ Why PCA?

Imagine we have a data set with **many variables**, such as measurements of people, products, or cities. Some **variables might be closely related** (like height and weight), **while others might carry similar information**. This creates two challenges:

1. It becomes **hard to analyze and interpret** the data.
2. **Redundancy** can lead to inefficiency and confusion.

**Principal Component Analysis (PCA)** helps solve this by providing a **simplified version of the dataset**, where we focus only on the **essential information**.

### ≡ The Main Idea

PCA works by **creating new variables**, called **principal components**, that are:

- **Combinations** of the original variables.
- **Ordered** so that the **first component captures the most variability** in the data.
- Each **subsequent component captures the next most variability**, but only from what's left over, and each **new component is uncorrelated with the previous** ones.

Imagine this like finding better “angles” or “perspectives” from which to view our data, ones that maximize how much we can see (i.e., variability) with as few perspectives as possible.

### ≡ Variability is Information

In statistics, **Variability**, how much values change from one observation to the next, is considered **information**. The more variability a component captures, the **more useful it is** in understanding the data. So, PCA's job is to **find the directions in which the data varies the most**, and to use those directions to summarize the dataset.

### 🔗 How PCA Changes the Dataset

Let's say we start with  $p$  variables (like height, weight, age, income...). **PCA gives us  $p$  principal components**, but the *magic* is that **we usually only need the first few** to understand most of what's going on.

So instead of working with the full, original dataset, we now have a **simpler version**:

- We still have the **same number of observations**.
- But each observation is now described by **fewer, more informative variables** (components).

This is called a **low-dimensional representation** of the data.

### 📊 Scores and Loadings - The Ingredients and the Result

In this new system:

- The **Scores** tell us where **each observation lies along the new axes** (the principal components).
- The **Loadings** tell us **how the original variables contribute to each component**.

Think of it like cooking:

- **Original variables** = ingredients
- **Loadings** = recipe instructions
- **Principal components** = final dishes
- **Scores** = ratings of the dishes for each person (observation)

### √\* Matrix Representation

Let's denote the data matrix as  $X$ , with  $n$  rows (observations) and  $p$  columns (variables). PCA is essentially about **transforming this matrix  $X$**  into a **new matrix**, where:

- The data is now described by **principal components** instead of the original variables.
- The **goal is to rotate and simplify** the data in a way that **emphasizes the most important directions** (maximum variability).

Two key matrices in PCA:

- **Loadings Matrix ( $V$ )**
  - This matrix contains the **weights** used to build the principal components.

- Each **column** in  $V$  corresponds to a **principal component**.
- Each **value** in  $V$  shows **how much each original variable contributes** to the component.

We can think of  $V$  as the recipe book: it tells us how to combine original variables to create the new components.

- **Scores Matrix ( $U$ )**

- This matrix contains the **projections of the data** onto the principal components.
- Each **row** in  $U$  represents an **observation in the new PCA space**.
- These are called scores, they tell us **where each observation lands along the new axes** (PC1, PC2...).

So  $U$  is the result: it shows **how our data looks in the new, simplified system**.

PCA can be **computed using Singular Value Decomposition (SVD)**:

$$X = U \cdot S \cdot V^T \quad (6)$$

PCA **factorizes the data matrix  $X$** . Here's what each matrix means:

- $X$ : Original data ( $n \times p$ )
- $U$ : Scores matrix ( $n \times p$ ), orthogonal (columns are independent)
- $S$ : Diagonal matrix of **singular values** (related to the variance explained)
- $V^T$ : Transposed loadings matrix ( $p \times p$ ), also orthogonal.

But for understanding, focus on this **simpler form**:

$$\text{PCA result} = \text{Scores} = X \cdot V$$

We **multiply the data  $X$  by the loadings matrix  $V$**  to obtain the **scores**.

However, both  $U$  and  $V$  are **orthogonal matrices**, meaning:

- Their columns are perpendicular (**no redundancy**).
- Principal **components are uncorrelated**, each new component captures new, non-overlapping information.

This is mathematically elegant and practically useful because it **removes multicollinearity** and makes downstream **analysis simpler and more robust**.

### 🔍 How Much Information Do We Keep?

Each principal component has a **percentage of variance explained**, this tells us how much of the original data's information it retains. Often, **the first 1 or 2 components explain so much that we can ignore the rest**.

In conclusion, **PCA helps us focus on what matters** in our data. It's like cleaning our glasses: everything becomes sharper, simpler, and more meaningful. We go from being overwhelmed by numbers to seeing clear patterns. It's a tool used everywhere, from finance to biology, marketing to engineering, whenever people need to make sense of complex data.



## 1.4 PCA Reference System

### 🧐 Why Do We Talk About Projections in PCA?

When we say that PCA projects data, we mean it **transforms the data points by placing them onto new axes**, the principal components, that better represent the structure of variability.

🧐 *But why are projections so important?* Because PCA has **two goals**:

1. **Maximize variance**: We want the **data to spread out as much as possible along the new axis**. This spread means we're capturing differences between observations.
2. **Minimize residuals**: We want to **minimize the error** when we approximate each point using the new axes. This is done by projecting each point perpendicularly onto the new axes, just like the shortest path between a point and a line.

This is why we “keep talking about projections”: they allow us to **retain the most information with the least distortion**.

### ≡ Generalizing to More Dimensions

In real datasets, we often have **more than two variables**. PCA scales to  $p$  dimensions, and the idea of projection still applies:

- PC1: The **direction in  $p$ -dimensional space** where data varies most.
- PC2: The next direction, **perpendicular to the first**, that captures remaining variability.
- This continues until we have  $p$  principal components, each orthogonal to the others.

So **PCA rotates the entire dataset into a new reference system** where the data structure is easier to understand.

### ≡ A New Reference System

After PCA, our data now lives in a **new coordinate system**:

- The **original variables** (e.g., height, weight) are no longer the axes.
- Instead, the axes are **principal components**, which are **combinations of the original variables**.

This new reference system has two main features:

1. **PCA results are rotation invariant**

🧐 *What does this mean?* It means that if we rotate our dataset (e.g., by changing the coordinate system), PCA still finds the same underlying structure, the same principal components (relative to the data).

- ❓ **Why?** PCA depends only on the relationships between the data points, specifically: the variances and covariances between variables. These are not affected by rotations of the data in space. Mathematically, PCA extracts eigenvectors of the covariance matrix that are invariant under rotation with respect to relative directions.

## 2. Principal Components are independent (uncorrelated)

- 📖 **What does this mean?** The principal components (PC1, PC2, ...) are uncorrelated with each other. Knowing the value of one PC tells us nothing about the value of the next.
- ❓ **Why?** Each new component is orthogonal to the previous one. And since the correlation is equal to the cosine of the angle between the variables ( $\cos(90) = 0$ ), PCA ensures that the correlation between the PCs is zero.
- ✅ In high dimensions, **collinearity** often indicates redundant information between variables and can cause problems in the analysis. **PCA solves** this problem by giving us independent, uncorrelated variables.



Figure 6: We have a bunch of scattered red dots. We have drawn a black line through the middle of the cloud of dots in the direction where the dots are most scattered. This line is our first principal component (PC1). For each red point, we drop a perpendicular line onto the black line. Where it lands, we place a blue point. This is called projection. Furthermore, the longer the black line, the more red points are taken into account (thus maximizing variance); on the other hand, the blue points must be close to the red points (shorter the distance) to minimize residuals and lose less information.



Figure 7: By drawing two lines, PC1 and PC2, we obtain a new reference system. It is a rotated system. The axes are principal components, which are combinations of the original variables.

### ✂ Principal Components: Linear Combinations

Let's now talk about **how principal components are constructed**. Each principal component is a **linear combination of the original variables**. This means:

$$PC_1 = \phi_{11} \cdot x_1 + \phi_{21} \cdot x_2 + \cdots + \phi_{p1} \cdot x_p$$

But in general:

$$PC_j = \phi_{1j} \cdot x_1 + \phi_{2j} \cdot x_2 + \cdots + \phi_{pj} \cdot x_p \quad (7)$$

Where:

- The  $\phi$ -values (phi) are called **Loadings**.
- They are the **weights assigned to each original variable**.
- The **higher the loading**, the **more that variable contributes** to that principal component.

For **each observation** (e.g., a person, product), we now **compute their position in the new PCA space**:

$$z_{ij} = \phi_{1j} \cdot x_{i1} + \phi_{2j} \cdot x_{i2} + \cdots + \phi_{pj} \cdot x_{ip} \quad (8)$$

These values  $z_{ij}$  are called **Scores**. They tell us:

- Where each observation lies **along a principal component axis**.
- How much that **component contributes** to describing this observation.

So now, instead of describing each person with  $p$  original variables, we describe them with  $k$  scores, where  $k < p$ , but we still capture the essence of their data.

## 1.5 PCA as Optimization Problem

### 🕒 What Is PCA Trying to Achieve?

PCA aims to find new variables (**principal components**) that are:

- **Linear combinations** of the original variables.
- Chosen so that the **first principal component** (PC1) captures the **maximum possible variance** in the data.
- **Second and further PCs** capture the **remaining variance**, while being uncorrelated with all previous ones.

In other words, **PCA tries to find a new axis (direction) along which the data varies the most**. This is the first principal component. Once this direction is found, each data point can be projected onto it to create a simplified representation of the data.

**To find this direction (PCA1), PCA solves a mathematical optimization problem.** This is because the **goal is to find the direction that maximizes the variance** of the data projected onto that direction (we want to maximize the variance because it captures the most variability, information, in the data).

### ❓ How Is This Formulated as an Optimization Problem? (PC1)

1. **Define a New Variable  $Z_1$ .** Let's construct a new variable  $Z_1$  which is a **linear combination** of the original variables:

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p$$

Where:

- $X_j$  is the  $j$ -th variable (all observations for that variable)
- $\phi_{j1}$  is the  $j$ -th weight (loading) assigned to the variable  $X_j$ .

The subscript 1 indicates that they belong to the first principal component.

2. **Objective - Maximize Variance of  $Z_1$ .** We want to maximize the variance to capture the most variability (information) in the data:

$$\text{Maximize } \text{Var}(Z_1) = \text{Var}(\phi_1^T X)$$

Where:

- $X$  is the **data matrix**  $n \times p$
- $\phi_1$  is the vector  $p \times 1$  of **weights (loadings)** used to build  $Z_1$ . It is transposed to allow multiplication with the data matrix  $X$ .

More specifically:

$$\text{Var}(Z_1) = \frac{1}{n} \sum_{i=1}^n z_{i1}^2 = \frac{1}{n} \sum_{i=1}^n \left( \sum_{j=1}^p \phi_{j1} \cdot x_{ij} \right)^2$$

Where  $z_{i1}$  is the **score** (see page 19) of the  $i$ -th observation on the **first principal component** (e.g.,  $z_{i2}$  score of observation  $i$  on PC2). In our case it is:

$$z_{i1} = \sum_{j=1}^p \phi_{j1} \cdot x_{ij}$$

3. **Constraint - Normalize the Loadings.** We need a **constraint** because we could scale the weights (loadings) infinitely. Therefore, we need **to fix the length (norm) of the loadings vector**.

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

Or, in vector notation:

$$\|\phi_1\|^2 = 1$$

The constraint says that the total energy or length of the loadings vector is fixed to 1 and PCA can only choose the direction of  $\phi_1$ , not its size.

Now we can write the **full optimization problem for PC1**:

$$\text{Maximize } \text{Var}(Z_1) \quad \text{subject to } \|\phi_1\|^2 = 1 \quad (9)$$

Specifically, the extended form:

$$\max_{\phi_{11}, \dots, \phi_{p1}} \frac{1}{n} \sum_{i=1}^n \left( \sum_{j=1}^p \phi_{j1} \cdot x_{ij} \right)^2 \quad \text{subject to } \sum_{j=1}^p \phi_{j1}^2 = 1 \quad (10)$$

### ❓ Okay, but how does PCA solve the optimization problem?

In previous steps, we posed PCA as:

- Maximize the variance of  $Z_1 = \phi_1^T X$
- Subject to  $\|\phi_1\|^2 = 1$

The solution to this problem: the **optimal loadings vector**  $\phi_1$  (for PC1) is **the first eigenvector of the sample covariance matrix  $S$** . Let's analyze this sentence to understand how and why.

❓ **Why Eigenvectors Help Solve This?** We need eigenvectors because PCA's optimization problem is mathematically equivalent to a linear algebra problem whose solution requires eigenvectors.

The **mathematical problem can also be written in quadratic form**<sup>1</sup>

$$\text{Var}(Z_1) = \phi_1^T S \phi_1$$

Where  $S$  is the covariance matrix (page 7). It follows the exact pattern of the quadratic form, because  $\phi_1$  is a vector of loadings,  $S$  is the covariance matrix (square), and the whole expression evaluates to a scalar: the variance of  $Z_1$ .

In linear algebra, the **quadratic form problem is solved by eigenvectors**. Using the **Rayleigh Quotient Theorem**, we can obtain the goal of PCA. Given a symmetric matrix  $S$  (like the covariance matrix in PCA), the *maximum value* of the quadratic form:

$$\phi_1^T S \phi_1 \quad \text{subject to} \quad \|\phi_1\|^2 = 1$$

Is achieved when  $\phi_1$  is the eigenvector corresponding to the largest eigenvalue of  $S$ . Therefore, the **solution must be the first eigenvector** (by the Rayleigh Quotient Theorem).

---

<sup>1</sup>A **Quadratic Form** is any expression that involves a vector, a matrix, and the same vector transposed. The general structure is:

$$Q(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x} \tag{11}$$

Where  $\mathbf{x}$  is a vector,  $\mathbf{A}$  is a square matrix, and the result is a scalar.

### 🔗 Optimization Problem for the Second Principal Component (PC2)

The following steps allow to obtain the PC2. We avoid a long explanation here, because we made the same logical steps of the optimization problem for PC1.

1. **Define a New Variable  $Z_2$ .** Just like with PC1, we define PC2 as a **linear combination** of the original variables:

$$Z_2 = \phi_{12}X_1 + \phi_{22}X_2 + \cdots + \phi_{p2}X_p$$

For a specific observation  $i$ :

$$z_{i2} = \sum_{j=1}^p \phi_{j2} \cdot x_{ij}$$

Where:  $\phi_2 = [\phi_{12}, \dots, \phi_{p2}]^T$  is loadings vector for PC2. Here the **goal is to find the best  $\phi_2$  to create PC2.**

2. **Objective - Maximize Variance of  $Z_2$ .** Just like PC1, we want to **maximize variance**:

$$\max_{\phi_2} \text{Var}(Z_2) = \frac{1}{n} \sum_{i=1}^n z_{i2}^2$$

3. **Constraints (there are two now)**

- (a) **First Constraint: Normalize the Loadings.** We must prevent arbitrary scaling of the weights:

$$\sum_{j=1}^p \phi_{j2}^2 = 1 \quad \text{or} \quad \|\phi_2\|^2 = 1$$

- (b) **Second Constraint: Ensure PC2 is Uncorrelated with PC1.** This is new compared to PC1. Here we want the covariance between PC1 and PC2 to be zero; in other words, we're saying *we don't want these two results to be correlated.*

$$\text{Cov}(Z_1, Z_2) = 0$$

Mathematically, this is equivalent to saying:

$$\phi_1^T \phi_2 = 0 \quad (\text{orthogonality})$$

If **two directions are orthogonal**, their **projections** (scores) are **uncorrelated**.

Finally, we can write the **full optimization problem for PC2**:

$$\max_{\phi_{12}, \dots, \phi_{p2}} \frac{1}{n} \sum_{i=1}^n \left( \sum_{j=1}^p \phi_{j2} x_{ij} \right)^2 \quad (12)$$



**Subject to:**

$$\sum_{j=1}^p \phi_{j2}^2 = 1 \quad \text{and} \quad \sum_{j=1}^p \phi_{j1} \cdot \phi_{j2} = 0 \quad (13)$$

**❓ And how to solve the PC2 optimization problem?** The solution to this constrained optimization problem is similar to the PC1:

$$\phi_2 = \mathbf{e}_2 \quad (14)$$

Where:

- $\mathbf{e}_2$  equal to the **second eigenvector** of the **covariance matrix**  $S$ .
- This eigenvector corresponds to the **second largest eigenvalue**  $\lambda_2$ .

## 1.6 Proportion of Variance Explained (PVE)

In the previous sections, we explained how to compute PC1, PC2, etc. as new variables. Each PC is a linear combination of the original variables, and each maximizes the variance under orthogonality constraints. But now the question becomes: “*okay, we computed these new components... but how useful are they?*”.

In PCA, variance equals information. So now our **goal** is: “how much **information (variance)** is explained by **PC1**? And by **PC2**? And by **PC3**? And so on”. This is where **Proportion of Variance Explained (PVE)** comes in.

We know that each principal component explains some variance:

- PC1 captures the **largest possible amount** of variance.
- PC2 captures the **next largest amount**, and so on.
- All PCs **together** explain **100% of the variance** in the data (no information is lost if we keep all of them).

If the data matrix  $X$  is centered (i.e., variables have mean 0), then the **total variance in the data** is:

$$\text{Var}_{\text{TOT}} = \sum_{j=1}^p \text{Var}(X_j) = \sum_{j=1}^p \left( \frac{1}{n} \sum_{i=1}^n x_{ij}^2 \right) \quad (15)$$

This is just the **sum of the variances** of the original variables.

To know the **variance** at the step  $k$ , so for the  $k$ -th **principal component**, we generalize:

$$\text{Var}_k = \frac{1}{n} \sum_{i=1}^n z_{ik}^2 \quad (16)$$

Where  $z_{ij}$  is the **score of observation**  $i$  on component  $k$ . In other words, it is the **spread of data along PC $k$** .

Since the **PVE** is a proportion, the equation is pretty obvious:

$$\text{PVE}_k = \frac{\text{Var}_k}{\text{Var}_{\text{TOT}}} \quad (17)$$

It is the **percentage of total variance** explained by PC $k$ . This tells us how informative PC $k$  is, and whether it’s worth keeping or can be discarded.

In practice, we **often keep only the first few PCs**. For example, if we find that  $PC1 + PC2$  explain 90% of the variance, we can keep only those two (dimensionality reduction).

## 1.7 Covariance vs. Correlation in PCA

In the previous sections, we discussed what PCA is and how we can use it. We know that PCA is based on the covariance matrix, that each principal component explains a portion of the total variance, and that PCA depends on the variance of the data. But “**what happens when the variables are on completely different scales?**” (Variable 1 is height, variable 2 is euros, variable 3 is just a count). If we apply PCA to the covariance matrix, the **variable with the largest variance will dominate the result**, even if it’s not more important!

More formally, **PCA is based on the spread (variance) and co-movement (covariance) between variables**. But this depends heavily on the units of measurement and scales of the variables. And the problem is, should **we use the covariance matrix or the correlation matrix** to perform PCA? The answer depends on the scales of our variables.

### ⚠ Covariance Matrix: use with caution

The **covariance matrix measures how variables change together**, in **absolute units**.

- ✓ This is fine only if **all variables are measured on comparable scales** (e.g., all in centimeters, or all in euros).
- ✗ If one variable has **much larger variance**, it will **dominate the PCA**, even if it’s not the most important feature.

For example, suppose height (meters) has a variance of 0.01 and income (euros) has a variance of 10’000. Income will control the direction of PC1 simply because its variance is greater, not because it’s more meaningful.

### ✓ Solution: Standardize the Data

To prevent PCA from being biased toward large-scale variables, we **standardize** each variable:

$$X'_{ij} = \frac{X_{ij} - \bar{X}_j}{SD_j} \quad (18)$$

Where:

- $\bar{x}_j$  is the mean of variable  $j$ .
- $SD_j$  is the standard deviation of variable  $j$ .

This standardization ensures that the **mean is zero** and the **standard deviation is one**. Now **all variables are on the same scale** and PCA treats them equally.

✔ **Alternative: use the correlation matrix instead of the covariance matrix**

Another way to achieve this is to run PCA on the **correlation matrix**, which is simply the **covariance matrix of the standardized variables**. This works because:

$$\text{Corr}(X_j, X_k) = \frac{\text{Cov}(X_j, X_k)}{\text{SD}(X_j) \cdot \text{SD}(X_k)} \quad (19)$$

So if we **standardize the variables first, the covariance becomes correlation**.

Case	Matrix to use	Why
All variables are on the same scale	Covariance matrix	Keeps units and variances as they are
Variables have different units/scales	Correlation matrix	Avoids bias toward large-variance variables
We standardized the variables	Covariance = correlation	Because standardization equalizes them

Table 2: When to use Covariance or Correlation.

## 1.8 PCA via SVD - Computational Aspects

Instead of computing PCA by:

1. Building the covariance matrix  $S = \frac{1}{n-1} X^T X$
2. Diagonalizing  $S$  to get eigenvectors and eigenvalues

We can instead **perform PCA directly using Singular Value Decomposition (SVD)** on the **centered data matrix**  $X$  itself.

### 📖 Singular Value Decomposition (SVD)

**Singular Value Decomposition (SVD)** method is a factorization of a matrix into three other matrices. If  $X$  is the centered data matrix (size  $n \times p$ ), then:

$$X = U \Sigma V^T \quad (20)$$

Where:

- **U**: matrix of **left singular vectors** (related to scores). It is orthogonal and  $n \times n$ .
- **$\Sigma$** : diagonal matrix of **singular values**  $\Sigma_1, \Sigma_2, \dots$ 
  - Non-negative real numbers on the diagonal (singular values)
  - Singular values sorted in descending order (from largest to smallest)
  - Number of eigenvalues guaranteed by the minimum between number of columns and number of rows:  $\min(n, p)$ .
- **V**: matrix of **right singular vectors**. It is orthogonal and  $p \times p$ . They are the main directions, so the loadings of PCA.

The eigenvalues of the covariance matrix can be calculated from the singular values  $\Sigma_i$ :

$$\lambda_i = \frac{\Sigma_i^2}{n-1}$$

This is because if we plug the SVD of  $X$  into this formula, the eigen decomposition of  $S$  naturally follows.

Concept	Meaning/Role in PCA
$X = U \Sigma V^T$	SVD decomposition of centered data matrix
<b>V</b>	Principal directions (same as eigenvectors of covariance)
$\Sigma_i$	Singular values
$\lambda_i = \frac{\Sigma_i^2}{n-1}$	Variance explained by each principal component

Table 3: Summary table of SVD for PCA.

**✓ Advantages**

- ✓ **Numerically more stable**: works better when the data matrix is large or ill-conditioned.
- ✓ **No need to compute and store the covariance matrix**: we can work directly on  $X$ .
- ✓ **More efficient** when  $p$  is large (many variables), especially in high-dimensional problems.

## 2 Clustering Methods

### 2.1 Introduction

In the previous section, we talked about PCA for compression or visualization. However, this topic falls under a type of technique called unsupervised learning.

- **Supervised Learning** is a type of machine learning where the algorithm is trained on a labeled dataset, meaning **each training example includes both the input data and the correct output**.

The **goal** is to learn a function that maps inputs to outputs, in order to make predictions on new, unseen data. Typical tasks include:

- Classification (e.g., spam detection, image recognition)
- Regression (e.g., predicting house prices or credit scores)

- **Unsupervised Learning** is a type of machine learning where the algorithm is given **only input data without any labeled output**.

The **objective** is to identify patterns, structures or groupings within the data. Common applications include:

- Clustering (this section)
- Dimensionality Reduction (e.g., PCA for compression or visualization, section 1.2, page 12)

In essence, while supervised learning is about *prediction*, unsupervised learning is about *discovery*.

#### 🔍 What is Clustering?

**Clustering** is the **process of grouping a set of objects** in such a way that **objects within the same group (cluster) are more similar** to each other than to those in other groups.

- No “true” labels are provided (unsupervised learning), the objective is to **uncover structure**.
- Often used in exploratory data analysis.
- The notion of “similarity” is fundamental and context-dependent (commonly based on **distance measures**).

The basic idea of clustering is:

1. **Minimize intra-cluster distances**: member of the **same cluster** should be **close to each other**.
2. **Maximize inter-cluster distances**: **different clusters** should be **well separated**.

### Example 1: Energy Consumption in Milan

A practical example:

- Consider  $n$  users, each described by their energy consumption across  $p$  time slots.
- The task is to identify **groups of users with similar consumption patterns**, potentially revealing roles like *residences vs offices* or *daytime vs nighttime* usage.
- Since there is no predefined label, this is a classic unsupervised learning task.

This example encapsulates the **essence of clustering**, discovering structure in data that wasn't explicitly labeled.

### ❓ How Many Clusters?

A fundamental **challenge** in clustering is **deciding the number of clusters** ( $k$ ). Visualizations show that data can often be reasonably clustered in 2, 4, or 6 groups, depending on the chosen definition or similarity.

The **clustering result depends on the similarity measure** (e.g., Euclidean distance), and the notion of a “true” cluster is often ambiguous. This ambiguity is a central theme in unsupervised learning: since there's no ground truth, **evaluating the “correctness” of a clustering is inherently difficult**.

### ≡ Types of Clustering

There are two main paradigms:

1. **Hierarchical Clustering**: builds a tree of clusters. Two main strategies:
  - **Agglomerative**: start from individual points and merge them.
  - **Divisive**: start from the whole dataset and split it recursively/
2. **Partitional Clustering** (also called flat clustering): divides data into  $k$  non-overlapping groups. Each data point belongs to exactly one cluster.

Hierarchical methods are more interpretable, while partitional methods (e.g., k-means) are often faster and scalable.



## 2.2 Defining Similarity in Clustering

At the heart of clustering lies a simple yet crucial idea: “objects in the same cluster should be **more similar** to each other than to objects in other clusters”. But how do we define “*similar*”? Clustering methods don’t work in a vacuum, they **need a distance (or similarity) function to determine how close two data points are**. This choice directly affects the clustering result.

### ≡ Types of Distance Metrics

Given two data points  $x = (x_1, x_2, \dots, x_p)$  and  $y = (y_1, y_2, \dots, y_p)$ , the distance between them can be computed in different ways, depending on the nature of the data and the goals of the analysis.

- **Euclidean Distance**

$$d_E(x, y) = \sqrt{\sum_{i=1}^p (x_i - y_i)^2} \quad (21)$$

- Measures straight-line distance.
- Works well when **features are on the same scale**.
- Variants: squared euclidean and standardized euclidean (normalize variables before computing).

- **Manhattan Distance**

$$d_M(x, y) = \sum_{i=1}^p |x_i - y_i| \quad (22)$$

Also called **L1 norm**, measures **grid-like distance** (think of navigating city blocks).

- **Chebyshev Distance**

$$d_C(x, y) = \max_i |x_i - y_i| \quad (23)$$

Takes the **maximum absolute difference** across dimensions. Sensitive to the **worst-case** coordinate difference.

- **Cosine Similarity (Angle-based)**

$$\text{similarity}(x, y) = \cos(\theta) \quad (24)$$

Based on the **angle between vectors**. The smaller the angle, the more similar the vectors. Common in **text mining** and **high-dimensional sparse data**.

- **Correlation-Based Distance**

$$d_R(x, y) = 1 - \text{corr}(x, y) \quad (25)$$

Measures **shape similarity** and ignores magnitude; focuses on patten of variation.

- **Mahalanobis Distance**

$$d_M(x, y) = \sqrt{(x - y)^T \Sigma^{-1} (x - y)} \quad (26)$$

Accounts for **correlation between variables**. Useful when features have **different variances** or **covariances**.

- **Minkowski Distance**

$$d_p(x, y) = \left( \sum_{i=1}^p |x_i - y_i|^p \right)^{\frac{1}{p}} \quad (27)$$

General formula that generalizes Euclidean and Manhattan. If  $p = 1$  is Manhattan, if  $p = 2$  is Euclidean.

### 🔍 Why Distance Matters

Clustering doesn't "know" what matter, we tell it via distance.

- The **choice of distance metric** can lead to **entirely different clusters**.
- Different applications require different definitions of similarity. For example, in customer segmentation, cosine similarity may be more appropriate than Euclidean distance, because we care about *purchasing trends*, not *absolute values*.

So instead of asking "*what's the best distance?*", ask: "*what kind of similarity is meaningful in our application?*"

## 2.3 Clustering Validity

Clustering can produce a result, a partition of our data, but **how do we know if it's good?** Unlike supervised learning, clustering lacks ground truth, so evaluating the quality of clustering is challenging. There are three main families of evaluation metrics, each answering a different kind of question: External Metrics, Internal Metrics, Relative Metrics.

---

### 2.3.1 External Metrics

**External metrics** assess the quality of a clustering by **comparing it to known class labels**. They answer the question: “how similar is the clustering result to the actual classification?”. This is **possible only when ground truth labels<sup>2</sup> are available**, which is often the case in benchmarking or simulated data. However, these metrics are **not available in real-world applications where true labels are unknown**.

We usually **compare clusters** to classes using a **contingency table**:

	Class 1	Class 2	Class 3	Total
Cluster 1	$m_{11}$	$m_{12}$	$m_{13}$	$m_1$
Cluster 2	$m_{21}$	$m_{22}$	$m_{23}$	$m_2$
Cluster 3	$m_{31}$	$m_{32}$	$m_{33}$	$m_3$
Total	$c_1$	$c_2$	$c_3$	$n$

- $m_{ij}$  number of points from class  $j$  assigned to cluster  $i$
- $m_i$  total points in cluster  $i$
- $c_j$  total points in class  $j$
- $n$  total number of points

From this, we compute **frequencies**:

$$p_{ij} = \frac{m_{ij}}{m_i} \quad (28)$$

---

<sup>2</sup>**Ground Truth Labels** are the true, correct categories or values assigned to data points. They represent the known answer. In clustering, ground truth refers to the real classification or grouping of our data, which is often manually annotated, observed from reality, or known from the context.

### Main External Metrics

- **Entropy Metrics.** Measures **class diversity within each cluster**.
  - A **pure cluster** (all elements from the same class) has entropy 0.
  - A **mixed cluster** has higher entropy (maximum when classes are equally mixed).

The formula for **cluster  $i$** :

$$e_i = - \sum_{j=1}^L p_{ij} \log(p_{ij}) \quad (29)$$

**Overall clustering entropy:**

$$e = \sum_{i=1}^K \frac{m_i}{n} e_i \quad (30)$$

Entropy decreases with better alignment. The **ideal value is zero**.

- **Purity Metrics.** Measures the **proportion of the dominant class in each cluster**. Like a “best guess” correctness. The formula for **cluster  $i$** :

$$p_i = \max_j (p_{ij}) \quad (31)$$

**Overall clustering purity:**

$$\text{purity} = \sum_{i=1}^K \frac{m_i}{n} p_i \quad (32)$$

Purity increases with better alignment. The **ideal value is one**.

- **Precision Metrics** and **Recall Metrics.** These are adapted from classification metrics:

- Precision (for cluster  $i$ , class  $j$ ):

$$\text{Prec}(i, j) = \frac{m_{ij}}{m_i} \quad (33)$$

Among the points in cluster  $i$ , **how many truly belong to class  $j$** ?

- Recall (for cluster  $i$ , class  $j$ ):

$$\text{Rec}(i, j) = \frac{m_{ij}}{c_j} \quad (34)$$

Among all points from class  $j$ , **how many are captured by cluster  $i$** ?

- **F-Measure Metrics.** Combines precision and recall into a **single score** using the harmonic mean.

$$F(i, j) = \frac{2 \cdot \text{Prec}(i, j) \cdot \text{Rec}(i, j)}{\text{Prec}(i, j) + \text{Rec}(i, j)} \quad (35)$$

We can aggregate these across all clusters to get a **global F-score** for the clustering.

Metric	Ideal	Interprets as...
Entropy	0	Purity of cluster (lower is better)
Purity	1	Dominance of single class (higher = better)
Precision	1	Cluster doesn't mix in wrong classes
Recall	1	Class is fully captured by a cluster
F-measure	1	Balanced precision and recall

Table 4: Summary External Metrics.

### 2.3.2 Internal Metrics

When class labels (ground truth) are not available, which is the most common case in unsupervised learning, we need **Internal Metrics** to assess:

- How **cohesive** each cluster is (tightness of points within clusters).
- How **well-separated the clusters are from each other**.

These metrics measure **structural quality** based on distances between two points.

- **Sum of Squared Errors (SSE)**. Also called **Within-Cluster Sum of Squares (WCSS)**.

#### 🔍 What is measures

- \* **How far the points** in each cluster are from their **cluster center** (or medoid)
- \* **Lower SSE means tighter clusters**<sup>3</sup>

#### 📖 Formula

$$SSE = \sum_{i=1}^K \sum_{x \in C_i} \|x - c_i\|^2 \quad (36)$$

Where:

- \*  $C_i$  is cluster  $i$
- \*  $c_i$  is the center of cluster  $i$

**Used to evaluate compactness:** a good clustering should have a small SSE.

<sup>3</sup>A tight cluster means that the data points inside the cluster are close to each other. They are packed together, not scattered.

- **Elbow Method**

- **What is measures**

- \* Used with SSE to find the **optimal number of clusters**  $K$ .

- **Procedure**

- \* Run clustering for various values of  $K$  (e.g., 1 to 30)
    - \* Plot  $SSE(K)$  vs  $K$ .
    - \* Find the “elbow” point: the value of  $K$  where the SSE stops decreasing sharply.

In other words, adding more clusters beyond the elbow gives **diminishing returns**.

- **Silhouette Coefficient**

- **What is measures**

- \* **Combines cohesion and separation** into a single score for each point.
    - \* For each data point:
      - $a$ : average distance to points in **same cluster**
      - $b$ : average distance to points in **nearest other cluster**

- **Formula**

$$s = \frac{b - a}{\max(a, b)} \quad s \in [-1, 1] \quad (37)$$

- ✓ +1 well clustered
    - \* 0 on the border
    - ✗ -1 misclassified

The **average silhouette score** across all points is **used to evaluate the whole clustering**.

- **Between-cluster Sum of Squares (BSS)**. Measures **cluster separation**:

$$BSS = \sum_i m_i \|c_i - \bar{c}\|^2 \quad (38)$$

Where:

- $m_i$ : size of cluster  $i$
- $c_i$ : center of cluster  $i$
- $\bar{c}$ : global centroid

A **good clustering** has **low WCSS** (or SSE) and **high BSS**.

### ⚠ Limitation of Internal Metrics

Most clustering algorithms **don't explicitly optimize internal metrics**. As a result, the **best-looking clustering under one metric may not be optimal in another**. For example, K-Means minimizes WCSS but may produce poor separation between clusters.

We can **design custom clustering algorithms that directly optimize an internal metric**, but this is **not always practical or generalizable**. A famous quote from Jain and Dubes (1988): “The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, clustering remains a black art accessible only to those true believers who have experience and great courage.”. This reflects **how tricky and interpretation-heavy clustering evaluation can be**.

## 2.4 Hierarchical Clustering

**Hierarchical Clustering** is a method of clustering that builds a hierarchy of clusters, either by **merging** smaller clusters into larger ones (bottom-up) or **splitting** a large cluster into smaller ones (top-down).

- **Agglomerative Clustering (Bottom-Up)**

1. Starts with **each data point as its own cluster**
2. At each step:
  - (a) Find the **two closest clusters**
  - (b) **Merge** them into one
  - (c) **Repeat** until there's one single cluster or a predefined number  $k$

This is the most common form of hierarchical clustering.

- **Divisive Clustering (Top-Down)**

1. Starts with **all points in one large cluster**
2. At each step:
  - (a) **Split** one cluster into two
  - (b) **Continue recursively** until each point is in its own cluster (or until  $k$  is reached)

This method is less commonly used due to higher complexity.

### ✂ Implementation Agglomerative Hierarchical Clustering

Hierarchical clustering works with a **distance matrix**, which **contains the distance between every pair of observations**. We don't need to specify  $k$  beforehand, but we need a **stopping criterion** (e.g., stop when there are  $k$  clusters, or use a threshold on distance).

1. Compute the **proximity matrix** (distances between all points)
2. Let each point be its **own cluster**
3. **Repeat**:
  - (a) Find and **merge the two closest clusters**
  - (b) **Update** the distance matrix
4. **Stop** when only one cluster remains (or another stopping condition is met)

The behavior of step 3, how we defined “closest clusters”, depends on the **linkage method**.



### 🔗 Linkage Methods: How we Merge Clusters

The **linkage method** defines the distance between two clusters, based on the distances between points in those clusters.

#### 1. Single Linkage Method

- ❓ **Distance.** It is the **minimum distance** between any two points (nearest neighbors)
- ✅ Can handle **non-globular shapes**.
- ⚠️ Sensitive to **noise and chaining effects**.

#### 2. Complete Linkage Method

- ❓ **Distance.** It is the **maximum distance** between any two points (furthest points).
- ✅ Less sensitive to noise, but biased toward **spherical or compact clusters** (we mean that the method tends to favor or naturally produces clusters that have a specific shape or property).

#### 3. Average Linkage Method

- ❓ **Distance.** It is the **average distance** between all points pairs.
- ✅ Uses **centroids** as reference and offers a **balance** between single and complete.

#### ✅ Pros

- ✅ Does not require predefining the number of clusters  $k$ .
- ✅ **Produces a dendrogram**<sup>4</sup>, a tree-like diagram showing how clusters are formed.
- ✅ Can reveal structure at **multiple levels of granularity**.

#### ❌ Limitations and Challenges

- ❌ **Computational complexity:** distance matrix is  $O(n^2)$ , problematic for large datasets.
- ❌ Once a merge/split is made, it **cannot be undone**.
- ❌ There's **no direct optimization** of a global objective (unlike k-means minimizing SSE) .
- ❌ **Results depend heavily on the linkage method used**, can lead to very different clusterings.

---

<sup>4</sup>The dendrogram is a tree where the leaves are individual data points and branches are merges between clusters. To choose  $k$ , “cut” the dendrogram at the level where the structure is stable (e.g., large vertical gaps suggest good splits).

## 2.5 K-Means

### 2.5.1 Introduction

**K-Means** is one of the most popular **clustering method**, specifically a type of **partitional clustering** (page 32). It's used to **group data into distinct clusters based on similarity**.

K-Means divides a dataset into  $K$  groups, where:

- Each group is called a **cluster**.
- Each **cluster** has a **center** called **centroid**.
- Every data point belongs to the cluster with the **nearest centroid**.

#### ❓ How does it work?

The K-means method can be divided into **five steps**:

1. **Choose  $K$**  (the number of clusters).
2. **Randomly assign** each data point to a cluster.
3. **Compute the centroid** of each cluster (the mean of all its points).
4. **Reassign each point to the cluster** with the closest centroid.
5. **Repeat** steps 3 and 4 **until** the points **no longer change** clusters (convergence criterion).

The algorithm tries to **minimize the distance** between each point and the center of its cluster. This means we want the **clusters to be as compact and well-separated as possible**.

Note that this is only a simple, quick introduction to K-Means clustering. In the following chapters, we will delve deeply into each topic.

### 2.5.2 Definition

**K-Means** is a **unsupervised learning algorithm** used to partition a dataset into  $K$  **distinct non-overlapping groups** called **clusters**, where each observation belongs to the cluster with the nearest mean (called the **centroid**).

#### 🕒 Main Goal

The goal of K-Means is to find a partition of the data that **minimizes the total within-cluster variation**, usually measured as the **sum of squared Euclidean distances** between each point and the centroid of its assigned cluster.

In other words, the K-means algorithm aims to **partition a dataset of  $n$  observations into  $K$  distinct clusters**:

- Each observation belongs to **one and only one** cluster.
- Each cluster has a **centroid**, i.e., a central point.
- Each point is assigned to the cluster with the **nearest centroid**.
- The goal is to **minimize the total distance** (squared Euclidean distance) from each point to its cluster's centroid.

We can express the previous goals in raw mathematical terms as follows:

$$\min_{C_1, \dots, C_K} \sum_{k=1}^K \sum_{\mathbf{x}_i \in C_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 \quad (39)$$

Where:

- $C_k$  is the **set of points** in the  $k$ -th cluster. Let  $C_1, C_2, \dots, C_K$  be the **index sets** of the clusters; these must form a **partition** of the dataset:
  - $C_1 \cup C_2 \cup \dots \cup C_K = \{1, 2, \dots, n\}$ , every observation is included in some cluster.
  - $C_i \cap C_j = \emptyset$  for  $i \neq j$ , no observation can belong to two clusters at once.

So, if the index  $i \in C_k$ , that means the  $i$ -th observation is assigned to the  $k$ -th cluster.

- $\boldsymbol{\mu}_k$  is the **centroid** (mean vector) of cluster  $C_k$ .
- $\|\mathbf{x}_i - \boldsymbol{\mu}_k\| = \|\cdot\|$  denotes **Euclidean distance**. It is the quantity that K-Means tries to minimize over all clusters.

Concept	Meaning
<b>Clustering type</b>	Partitional (flat).
<b>Centroid</b>	Mean of points in a cluster.
<b>Assignment rule</b>	Assign each point to the <b>closest centroid</b> .
<b><math>K</math> (number of clusters)</b>	Must be <b>chosen in advance</b> .
<b>Objective function</b>	Minimize total within-cluster distance (e.g., sum of squared distances).
<b>Partition conditions</b>	Clusters are <b>disjoint and exhaustive</b> (no overlap, no omission).

Table 5: K-Means summary

### ❓ How Many Clusters? (choosing $K$ )

One of the main limitations of K-Means is that the number of clusters  $K$  must be **specified in advance**. However, in real-world data, the true number of natural groupings is **usually unknown**. Choosing the wrong  $K$  can lead to:

- **Underfitting**: too **few** clusters  $\Rightarrow$  **merging** different groups
- **Overfitting**: too **many** clusters  $\Rightarrow$  **splitting** coherent groups

The common methods to choose  $K$  are:

- **Elbow Method** looks at the **total within-cluster variation** (also called inertia or distortion), defined as:

$$W(K) = \sum_{k=1}^K \sum_{\mathbf{x}_i \in C_K} \|\mathbf{x}_i - \mu_k\|^2 \quad (40)$$

- $W(K)$  is the total within-cluster variability (it is explained in depth on page 50). It measures **how compact the clusters are**, the smaller  $W(K)$ , the tighter the points are around their centroids.
- When  $K = 1$ , all points are in a single cluster, so  $W(1)$  is very high. As  $K$  increases, **clusters** become **smaller and tighter**, then  $W(K)$  **decreases**. In the extreme case  $K = n$ , each point is its own cluster, then  $W(n) = 0$ .

So,  $W(K)$  **always decreases as  $K$  increases**, but not linearly.

We compute Elbow Method for different values of  $K$ , and plot  $W(K)$  vs  $K$ . The reason we plot the **within-cluster sum of squares**  $W(K)$  against the **number of clusters**  $K$  is to **visualize how the clustering quality improves** as we increase the number of clusters.

Plotting  $W(K)$  against  $K$  helps us **identify the point where adding more clusters gives diminishing returns**. It is called the “*elbow method*” because of the shape of the plot. When plotting  $W(K)$  versus  $K$ , the curve usually drops sharply at first because adding more clusters

quickly improves clustering. After a certain point, however, adding more clusters only yields small improvements. The resulting plot often resembles a bent arm. The “*elbow*” is the point of maximum curvature, where the curve begins to flatten out. This point is important because **before the elbow**, we see **significant gains in reducing** within-cluster variance. **After the elbow**, however, the **gains are minimal**, indicating that increasing model complexity is not worthwhile.



Figure 8: This is an example of the *elbow method* applied to K-means clustering. In the plot, we can see how the within-cluster sum of squares ( $W(K)$ ) changes as the number of clusters  $K$  increases from 1 to 10.

- When  $K = 1$ :  $W(K)$  is very large (around 2500), since all points belong to a single, poorly-fitting cluster.
- From  $K = 1$  to  $K = 3$ :  $W(K)$  decreases rapidly, showing that adding clusters significantly improves the clustering.
- From  $K = 4$  onward: the curve starts to flatten, meaning adding more clusters gives only small improvements.

The elbow is visible at  $K = 4$ . This is where the rate slows down sharply. Adding more clusters doesn’t significantly reduce  $W(K)$ . It balances clustering quality with model simplicity.

- **Silhouette Score** (or **Silhouette Coefficient**) evaluates **how well each point fits within its cluster** versus others. The Silhouette Coefficient for a point is:

$$s = \frac{b - a}{\max(a, b)} \quad (41)$$

Where:

- $a$  is the average distance to points in the same cluster

- $b$  is the average distance to points in the nearest other cluster

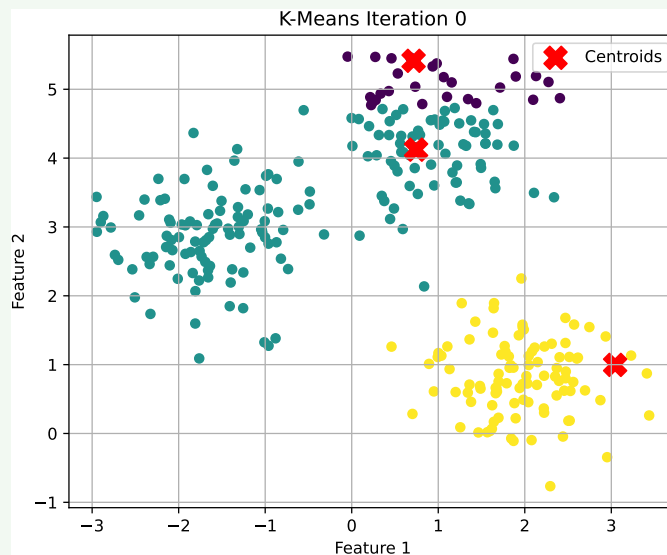
Values close to 1 mean well-clustered points. We compute the average Silhouette score for all points and pick the  $K$  that maximizes it.

- **Gap Statistics** (too advanced) compares the total within-cluster variation for the real data to that for randomly generated data with no structure. A large “gap” means the real data forms more distinct clusters. Choose the smallest  $K$  for which the gap is maximal or within 1 standard deviation of the maximum.

### Example 2: K-Means

Below is a simple run of the K-means algorithm on a random dataset.

- Iteration 0 - **Initialization**



This is the starting point of the K-Means algorithm. **Three centroids are randomly placed in the feature space.** At this point, no data points are assigned to clusters yet, or all are assumed to be uncolored/unclustered. The positions of the centroids will strongly influence how the algorithm proceeds.

The goal here is to start with some guesses. The next step will use these centroids to form the initial clusters.

- Iteration 1 - **First Assignment and Update**



Each data point is assigned to the closest centroid, forming the first version of the clusters. New centroids are computed by taking the average of the points in each cluster. We can already see structure forming in the data, as points begin grouping around centroids.

This step is the first real clustering, and centroids begin to move toward dense regions of data.

- **Iteration 2 - Re-Assignment and Refinement**



Clusters are recomputed based on updated centroids. Many points remain in the same clusters, but some may shift to a new cluster if a centroid has moved. Centroids continue moving closer to the center of their respective groups.

The algorithm is now refining the clusters and reducing the total distance from points to centroids.

- Iteration 3 - Further Convergence



At iteration 3, the K-Means algorithm reached convergence. The centroids no longer moved, and no points changed cluster. This means:

- The algorithm has found a locally optimal solution.
- Further iterations would not improve or change the clustering.
- The final configuration is considered the result of the algorithm.

In practice, this is how K-Means stops: it checks whether the centroids remain unchanged, and if so, it terminates automatically.



### 2.5.3 Initialization Issues in K-Means

#### ❓ Why initialization matters

The K-Means algorithm starts by choosing  $K$  **initial centroids**. These **starting positions are crucial** because K-Means builds the entire clustering process based on them.

Since K-Means is an **iterative algorithm**, it gradually improves the clusters from where it begins. However, if it starts from a bad configuration (i.e. bad initial centroids), it can easily get stuck in a **local minimum**, a clustering solution that isn't optimal but looks good from that starting point.

This means that:

- **Different initial centroids** can lead to **different final results**.
- Some clusters may be poorly formed or even empty.
- The quality of the final clustering can vary, even on the same data.

#### ⚠️ What can go wrong

- **Centroids too close**: if initial centroids are placed near each other, they might all end up **capturing the same group of points**, leaving **other areas underrepresented**.
- **Unbalanced clusters**: a poor start can lead to **clusters with very different sizes or shapes**, even if the data has nicely separated groups.
- **Empty clusters**: it's possible that **some centroids end up with no points assigned**, which breaks the algorithm's assumptions.

#### ✅ How to improve initialization (conceptually)

To avoid these problems, two common strategies are used:

1. **Multiple random initializations**. Instead of relying on just one random start, K-Means is **run several times with different initial centroids**. The **result with the lowest total distance is selected**. This increases the chance of finding a good clustering.
2. **Smart initialization**. Instead of choosing all centroids randomly, a better approach is to choose the **first centroid randomly**, then pick the **next one in a way that they are as far apart as possible**. This increases the chances of covering different regions of data from the beginning.

This smarter approach significantly reduces the risk of bad clustering and makes the algorithm more stable and reliable.

### 2.5.4 K-Means as an Optimization Problem

Given:

- A set  $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  of  $n$  data points in  $\mathbb{R}^d$
- A chosen number of clusters  $K$

The goal of K-Means is to **partition the data into  $K$  clusters**:

$$C = \{C_1, C_2, \dots, C_K\}$$

So that a measure of **within-cluster variability** is **minimized**.

#### ❓ What is Within-Cluster Variability?

In simple terms:

- **Cluster**: a group of points that are similar (close together).
- **Variability**: how spread out the points are.
- **Within-Cluster Variability** how spread out the points are **inside each cluster**.

So within-cluster variability measures **how close the points in the same cluster are to each other**. We have **low variability** when the **points are tightly** packed and **high variability** when the **points are spread out**.

#### ❓ Okay, but why do we need to minimize within-cluster variability?

Because a **good cluster** should contain **points that are close together**, not randomly scattered. So K-Means algorithm tries to form clusters where the **internal “spread”** is as small as possible.

#### ❓ Objective Function: what are we minimizing?

For each cluster  $C_k$ , define its **centroid**  $\mathbf{c}_k$  as the **mean of the points in the cluster**. The within-cluster variability is measured by the following:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^d (x_{ij} - x_{i'j})^2 \quad (42)$$

- Each observation is a point:

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{id}) \in \mathbb{R}^d$$

That means  $\mathbf{x}_i$  is a vector with  $d$  features.

- $\sum_{i, i' \in C_k}$  means we are looking at **all pairs of points  $i$  and  $i'$**  inside the same cluster  $C_k$ .

- $\sum_{j=1}^d (x_{ij} - x_{i'j})^2$ , for each feature  $j$ , we compute the **squared distance** between point  $i$  and point  $i'$ .
- $\sum_{i, i' \in C_k} \sum_{j=1}^d (x_{ij} - x_{i'j})^2$  measures the **total squared distance** between **all point pairs inside the cluster**.
- $\frac{1}{|C_k|}$ , we divide by the number of points to get an **average distance**.

So  $W(C_k)$  is the **average squared distance between all pairs of points inside cluster  $C_k$** .

### 🔗 Optimization Criterion

The full K-Means objective is to **find the clustering that minimizes the total within-cluster variability**:

$$\min_{C_1, \dots, C_K} \sum_{k=1}^K W(C_k) \quad (43)$$

This means we are searching for the best possible partition  $\{C_1, \dots, C_K\}$  that results in clusters where **points are as close as possible to each other**, i.e., the clusters are compact.

In other words, we want **every cluster to be compact**. So we add up the within-cluster variability for all clusters. We **minimize the problem while keeping it as compact as possible**.

### ✅ Interpretation

A **good clustering** has **low within-cluster variability**: points in the same cluster are close together. Also, K-Means solves an optimization problem where it tries to find such a clustering by minimizing a specific cost function.

### 2.5.5 Algorithm

The K-Means algorithm is an iterative method that alternates between assigning data points to clusters and updating the cluster centroids, with the goal of minimizing within-cluster variability.

0. **Choose the number of clusters  $K$ .** Before running the algorithm, we must decide the number of clusters  $K$ . Each cluster will have:

- A **set of points**  $C_1, \dots, C_K$
- A **centroid**  $\mathbf{v}_1, \dots, \mathbf{v}_K$
- A **partition matrix**  $\Gamma = [\gamma_{ij}]$  where:

$$\gamma_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_j \in C_i \\ 0 & \text{otherwise} \end{cases}$$

1. **Initialization.** Randomly assign each observation to one of the  $K$  clusters. Then compute the **initial centroids** for each cluster:

$$\mathbf{v}_k = \frac{1}{|C_k|} \sum_{\mathbf{x}_i \in C_k} \mathbf{x}_i$$

These are the **mean vectors** of the assigned points.

2. **Assignment.** For each point  $\mathbf{x}_i$ , compute the distance to all centroids, and assign it to the cluster with the **closest centroid**. This updates the **partition matrix**  $\Gamma$ .
3. **Update Centroids.** After the new assignments, **recompute the centroid** of each cluster using the updated points:

$$\mathbf{v}_k = \frac{1}{|C_k|} \sum_{\mathbf{x}_i \in C_k} \mathbf{x}_i$$

Repeat **Step 2** and **Step 3** until:

- ✓ No points change cluster (**convergence**), or
- ✓ A **stopping criterion** is met (e.g., max iterations)

#### ☰ Properties of the algorithm

##### ✓ Good

- The algorithm **always converges** (it stabilizes after a finite number of iterations)
- Simple and fast

##### ✗ Bad

- It does **not always find the global optimum**
- The **results depends** heavily on the **initialization**

In practice, K-means clustering is **typically run multiple times with different random initializations**, and the **result with the lowest cost is kept**.

### 2.5.6 Limitations

While K-Means is simple and efficient, it has several important limitations:

1. **We must choose the number of clusters  $K$  in advance.** Unlike **hierarchical clustering**, which builds a full tree of clusterings, K-Means requires we to **fix  $K$  before the algorithm starts**.
  - ✗ Choosing the wrong  $K$  can lead to poor results.
  - ✗ There's no universal rule for selecting  $K$ ; methods like the Elbow or Silhouette must be used.
2. **Assumes clusters are globular (spherical) in shape.** K-Means tends to form **round-shaped clusters** due to the use of **Euclidean distance**.
  - ✗ It **struggles with elongated, irregular, or non-convex shapes**.
  - ✗ Works best when clusters are **well-separated, equally sized, and dense**.
3. **Fails with clusters of different sizes or densities.** If one cluster is very dense and another is sparse, K-Means might:
  - ✗ Split one true cluster into several pieces
  - ✗ Merge multiple small clusters into one

This happens because Euclidean distance doesn't account for cluster **spread or shape**.
4. **Sensitive to outliers.** K-Means uses **centroids (means)** to define clusters. Since the **mean is sensitive to extreme values**, an outlier can significantly shift a centroid and distort the clustering.

As an alternative, we could use **K-Medoids**, which is **based on medians rather than means** and is more robust to outliers.
5. **Sensitive to feature scaling and representation.** Because K-Means relies on distance, the algorithm is:
  - ✗ Affected by **unscaled or poorly represented data**
  - ✗ Sometimes improved by **dimensionality reduction** (e.g., PCA)

## 2.6 Gaussian Mixture Models (GMMs)

### 2.6.1 Introduction

A **Gaussian Mixture Model (GMM)** is a **probabilistic model** that assumes the data is generated from a **mixture of several Gaussian distributions** (also called normal distributions), each representing a different **latent group or cluster**.

Each **data point** is not assigned to one specific cluster (as we have seen in the K-means algorithm), but is instead modeled as **having a probability of belonging to each cluster**.

#### ❓ Why is it called a “*mixture*”?

Because the model represents the **overall data distribution** as a **weighted sum of several Gaussian components**. Mathematically, the probability density function is:

$$f(x) = \sum_{k=1}^K \pi_k \cdot \mathcal{N}(x \mid \mu_k, \Sigma_k)$$

Where:

- $K$  is the number of **Gaussian components** (clusters)
- $\pi_k$  is the **mixing coefficient** of component  $k$  (like a *weight*), with  $\sum \pi_k = 1$
- $\mathcal{N}(x \mid \mu_k, \Sigma_k)$  is the **multivariate Gaussian distribution** with:
  - Mean  $\mu_k$
  - Covariance matrix  $\Sigma_k$

A more detailed explanation can be found on page 61.

#### Remark: What is the Multivariate Gaussian Distribution?

##### ❓ What is a 1D Gaussian (Normal) Distribution?

The **1D Gaussian** (also called **Normal Distribution**) is the classic bell curve and it tells us how likely we are to get values near the **mean**  $\mu$ , and how unlikely values far from the mean are. The spread is controlled by the **standard deviation**  $\sigma$ .

In other words, it describes a **continuous** variable whose values are **most likely to occur near a central point**, and **less likely the farther we move away**.

The **Probability Density Function (PDF)** of a 1D Gaussian is:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (44)$$

- $x$  is the value of the random variable.

- $\mu$  is the **mean**, the center of the distribution.
- $\sigma$  is the **standard deviation**, it controls the “spread” of the curve.

The bell-shaped curve is symmetric around the mean  $\mu$ . Most values fall within:

- $\mu \pm \sigma$ :  $\approx 68\%$
- $\mu \pm 2\sigma$ :  $\approx 95\%$
- $\mu \pm 3\sigma$ :  $\approx 99.7\%$

For example, if we have a 1D Gaussian with  $\mu = 0$ ,  $\sigma = 1$ , it is called the **standard normal distribution**, where the curve peaks at 0 and spreads out smoothly in both directions.

### ? What is a 2D Gaussian Distribution?

The **2D Gaussian** (or **Bivariate Normal Distribution**) is the natural extension of the 1D Gaussian to **two dimensions**. It tells us how likely a point  $(x, y)$  is to appear in a **2D space**, just like how the 1D Gaussian tells us about single values on a line.

A **2D Gaussian distribution** is completely defined by:

1. **Mean Vector**  $\mu \in \mathbb{R}^2$

This is the **center** (or peak) of the distribution.

$$\mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \quad (45)$$

It tells us the **average position** of the points, the location where the distribution is highest.

2. **Covariance Matrix**  $\Sigma \in \mathbb{R}^{2 \times 2}$

This tells us how the data is **spread out and oriented** in 2D space.

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} \quad (46)$$

The diagonal values control the **spread** along each axis, and the off-diagonal values ( $\rho$ ) represent **correlation** between  $x$  and  $y$ .

- $\sigma_x^2, \sigma_y^2$  is the variances along the X and Y axes.
- $\rho$  is the correlation between  $x$  and  $y$ .
  - If  $\rho = 0$ : the axes are independent.
  - If  $\rho > 0$ : the ellipse tilts upward (positive correlation).
  - If  $\rho < 0$ : the ellipse tilts downward (negative correlation).

In short, a 2D Gaussian is a “bell-shaped hill” in 2D space, centered at  $\mu$ , with shape and tilt determined by  $\Sigma$ .



Figure 9: 1D Gaussian and 2D Gaussian (as Contours). In the **1D plot** (left), we see the classic **bell-shaped curve**. This shows the probability density of a variable  $x$  under a Gaussian distribution centered at  $\mu = 0$  with standard deviation  $\sigma = 1$ .

- The peak occurs at the mean.
- The density decreases symmetrically as we move away from the mean.
- Most of the probability mass lies within  $\pm 1$  or  $\pm 2$  standard deviations.

This curve describes how **likely** different values of  $x$  are under the distribution. The second plot (right) shows a **2D Gaussian distribution using contour lines**. This represents a Gaussian defined over two variables,  $x$  and  $y$ .

- The **concentric ellipses** indicate regions of equal probability density (like topographic maps).
- The ellipses are **tiled**, showing that the variables  $x$  and  $y$  are **positively correlated**.
- The closer to the center, the higher the density.
- A 2D Gaussian distribution models probabilities over two variables  $\mathbf{x} = (x, y)$ . At each point  $(x, y)$ , the distribution gives a **density value**, that is, how likely it is that a data point would occur at that location. This means the 2D Gaussian is actually a function from 2D to 1D:

$$f(x, y) = \text{density}$$

So even though it's defined in 2D, it **outputs a single number** (the density).

This view helps visualize the **shape and orientation** of the distribution in 2D space.





Figure 10: A 3D plot showing the same 2D Gaussian as a surface in three dimensions.

- The horizontal plane is the  $x$ - $y$  space.
- The vertical axis shows the **density** at each point.
- The result is a **bell-shaped hill**, centered at the mean.
- The ridge direction and slope reflect the **correlation and spread** between the two variables.

We use  $x$  and  $y$  as the horizontal plane (input space), and  $z$ -axis as **density value**  $f(x, y)$ . Now, we are plotting the **full shape of the probability surface**: a bell-shaped hill where the height indicates the likelihood of a given point. We add a **third axis to represent the function output**, the density. While **2D contours** are useful for seeing **direction, correlation, and structure**, a **3D surface** provides a **better sense of the bell shape**.

### ✔ What is the Multivariate Gaussian Distribution?

The **Multivariate Gaussian Distribution** is simply a **generalization of the 1D and 2D Gaussian to any number of dimensions**. If we have a random vector  $\mathbf{x} \in \mathbb{R}^d$ , then: a **multivariate Gaussian** models the probability of observing different values of that vector, assuming that all components jointly follow a Gaussian distribution.

**✓ Formal Definition**

A random vector  $\mathbf{x} = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$  follows a multivariate normal distribution with:

- **Mean vector:**  $\boldsymbol{\mu} \in \mathbb{R}^d$
- **Covariance matrix:**  $\Sigma \in \mathbb{R}^{d \times d}$

We write this as:

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \quad (47)$$

Its **probability density function** is:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \cdot \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right) \quad (48)$$

- $\boldsymbol{\mu}$  is the **center** (expected value) of the cloud of points.
- $\Sigma$  is the **shape and orientation** of that cloud
  - The diagonal entries  $\sigma_{ii}$  control how spread out the data is in each direction.
  - The off-diagonal entries  $\sigma_{ij}$  represent **correlation** between different variables.

So a multivariate Gaussian forms an **ellipsoidal distribution** in high-dimensional space.

**❓ Why GMM uses Multivariate Gaussian Distribution?**

In Gaussian Mixture Models, each component is a **multivariate Gaussian**:

- It models a cluster in **multidimensional space**.
- The full GMM is a **weighted sum** of these multivariate Gaussians.

This allows GMM to **model complex, overlapping, and anisotropic (non-spherical) clusters** much better than K-Means.

**☰ Relation to K-Means**

K-Means is a **special case** of GMM:

- K-Means assigns each point to **only one cluster**, called **Hard Clustering**.
- GMM assigns each point a **probability of belonging** to each cluster, called **Soft Clustering**.

In fact, if we assume that:

- All clusters have **equal, spherical** covariance.

- Each point is assigned to the **closest cluster center**.

Then *GMM reduces to K-Means*.

### ❓ But why should we use GMM instead of K-means?

Four reasons:

#### 1. Soft Clustering vs Hard Clustering

- **Hard Assignment:** K-Means assigns each point to **only one cluster**.
- **Soft Assignment:** GMM assign each point a probability of belonging to each cluster.

GMM is useful when clusters overlap, we're uncertain or want probabilistic interpretation, or we want to model real-world uncertainty (e.g., "60% likely to be in Cluster 1").

#### 2. Flexible Cluster Shapes

K-Means assumes clusters are **spherical and equally sized**. Instead, GMM allows:

- Elliptical shapes;
- Different sizes and orientations;
- Correlations between features.

This is because GMM uses **covariance matrices**, whereas K-Means just uses Euclidean distance.

#### 3. Statistical Foundation

K-Means is an algorithm that minimizes distance. Instead, GMM is a **probabilistic model** with a well-defined likelihood function. Then, we can estimate the uncertainty and use model selection to determine the number of clusters.

#### 4. K-Means is a Special Case of GMM

If we assume:

- All clusters have **equal and diagonal covariance matrices** (no correlation);
- Use hard assignments.

Then GMM becomes K-Means. So GMM is a **more general, more powerful model**.

However, GMM is not always the best method to consider. It depends on the data and the final goal:

- Use **K-Means** when **clusters** are clearly **separated, round, and fast results** are needed.
- Use **GMM** when we want a **flexible, probabilistic, and realistic model** of how our data is generated.

Feature	K-Means	GMM
Assignment	Hard	Soft (probabilistic)
Cluster shape	Spherical	Elliptical, flexible
Model type	Geometric	Probabilistic (generative)
Speed	Fast	Slower (EM algorithm)
Interpretability	Simple centroids	Full statistical model

Table 6: Trade-offs between K-Means and GMM.

### 2.6.2 Mathematical Foundation

Even though we have already introduced the GMM formula, we provide a thorough explanation here.

#### 🔍 What is a Mixture Model?

A **mixture model** assumes that our data was generated from **multiple probability distributions**, not just one. For GMMs, we assume: each data point  $\mathbf{x}_i \in \mathbb{R}^d$  was **generated by one of several multivariate Gaussian distributions**. But we don't know:

- Which point came from which component.
- What the component parameters are.

#### 📖 GMM Definition

The probability density function of a Gaussian Mixture Model with  $K$  components is:

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \cdot \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \Sigma_k) \quad (49)$$

Where:

- $\mathbf{x} \in \mathbb{R}^d$  is the observed data point.
- $\pi_k$  is the **mixing coefficient** for component  $k$ , with  $\sum_{k=1}^K \pi_k = 1$ , and  $\pi_k \geq 0$ .
- $\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \Sigma_k)$  is the **multivariate Gaussian density** of the  $k$ -th component, with:
  - Mean  $\boldsymbol{\mu}_k \in \mathbb{R}^d$
  - Covariance matrix  $\Sigma_k \in \mathbb{R}^{d \times d}$

This is the **probability density** of the Multivariate Gaussian Distribution, also called the **likelihood of observing the point  $\mathbf{x}$ , given the parameters of component  $k$**  (mean  $\mu$  and covariance  $\Sigma$ ).

For example, imagine we're standing at a point  $\mathbf{x}$  in space, and asking: *how likely is it that a point like this came from Gaussian cluster  $k$ ?* The answer is:

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \Sigma_k)$$

The value is **higher** if  $\mathbf{x}$  is **closer to the center** of cluster  $k$ , and **lower** if it's farther away.

The meaning of each term is as follows:

Symbol	Meaning
$\pi_k$	Prior probability of choosing cluster $k$ .
$\boldsymbol{\mu}_k$	Mean (center) of the $k$ -th Gaussian component.
$\Sigma_k$	Covariance (shape and orientation) of the $k$ -th component.
$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \Sigma_k)$	Likelihood of $\mathbf{x}$ under the $k$ -th Gaussian.
$p(\mathbf{x})$	Overall probability of observing $\mathbf{x}$ under the full model.

To generate a data point  $\mathbf{x}$ , first choose a component  $k$  with probability  $\pi_k$ , then draw  $\mathbf{x}$  from the Gaussian  $\mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$ . This is why GMM is a generative model, **it models how data could have been generated**.

### 2.6.3 Responsibilities

Unlike **K-Means**, where each point is assigned to exactly **one cluster** (hard clustering), **GMM** performs **soft clustering**: each point has a **probability of belonging to every cluster**. These probabilities reflect *how likely* it is that a point came from each component of the mixture.

#### ❓ Why do we need to introduce Responsibilities in GMM?

⚠️ We have a **problem** at the moment. When using a Gaussian Mixture Model, we assume that each data point  $\mathbf{x}_i$  was generated by **one of the  $K$  Gaussian components**. But, we **don't know which one**. That information is **hidden** or **latent**. We only observe:

- The data point  $\mathbf{x}_i$
- The parameters of the Gaussian  $(\pi_k, \mu_k, \Sigma_k)$

We don't observe which component  $k$  generated  $\mathbf{x}_i$ .

***I still don't understand what the problem is.*** In a Gaussian Mixture Model, we assume that each data point  $\mathbf{x}_i$  is generated by **one of the  $K$  Gaussian components**. So the true process that generates each point is:

1. Randomly choose a component  $k$  with probability  $\pi_k$ .
2. Generate a point  $\mathbf{x}_i$  from the Gaussian  $\mathcal{N}(\mu_k, \Sigma_k)$

But we don't know which Gaussian generated each point and this is the core issue:

- We **observe only the data points**  $\mathbf{x}_1, \dots, \mathbf{x}_n$ .
- We **do not observe the “label”** of which component (cluster) generated each point.

That **label is missing!** This is problematic because, if we want to **train a model** (i.e., estimate the  $\pi_k$ ,  $\mu_k$ , and  $\Sigma_k$ ), we would **ideally want to group the data by component**. However, we can't do that because we don't know which point belongs to which component.

***Okay, but if we're using the GMM formula, then we know everything about each cluster that produces each point.*** That's a good point, but the practice is slightly different. In the real world, especially in machine learning, we **don't know the origin of each point because we only have the observed data**. So, we are trying to **infer the underlying clusters, without knowing** which one generated each point.

#### Example 3: analogy of the GMM problem

Imagine a factory that has 3 machines (A, B, C) making cookies. Each machine:

- Adds a different amount of sugar and chocolate

- Sometimes makes similar-looking cookies, but not exactly the same
- Has a different production rate (e.g., Machine A makes 50% of all cookies, B makes 30%, C makes 20%)

Now, someone gives us a box of mixed cookies and says: “*These were made using machines A, B, and C, but I’m not telling you which cookie came from which machine.*” We taste each cookie and write down how sweet and chocolatey it is. That’s our data.

**⚠ The Problem.** We don’t know:

- Which machine made each cookie.
- How sweet/chocolatey each machine’s recipe really is.
- How often each machine is used.

In GMM terms:

- We have the data  $\mathbf{x}_1, \dots, \mathbf{x}_n$  (cookie features).
- But the label “*this came from machine A*”, is hidden.

**✅ The Solution.** Since we can’t know for sure where each cookie came from, we say: “*Let me estimate how likely it is that this cookie came from each machine.*”. So for Cookie #1, we might say:

- 70% Machine A
- 20% Machine B
- 10% Machine C

This set of numbers is the responsibility vector for Cookie #1. We do this for every cookie. Now, even though we don’t know the true labels, we have soft guesses, and that’s enough to:

- Estimate what each machine’s recipe looks like (mean and variance).
- Estimate how many cookies each machine probably made (mixing proportions).

The core idea is that we need “responsibilities” because the cluster label is missing. Rather than guessing the labels directly, we estimate the probability that each point belongs to each cluster. These probabilities are called “responsibilities”.



### ✔ Solution: Responsibilities = Posterior Probabilities

We come back to GMM. We assume that:

1. Each point  $\mathbf{x}_i$  was generated by **one of  $K$  Gaussians**.
2. But we **don't observe** which one.

So we introduce a **latent variable**  $z_i \in \{1, \dots, K\}$  that **represents the unknown cluster** of  $\mathbf{x}_i$  (hidden cluster assignments).

We introduce the idea of a **latent variable**  $z_i$ , which says: “let's pretend each point  $\mathbf{x}_i$  has a hidden label  $z_i \in \{1, \dots, K\}$ ”. But since we don't observe  $z_i$ , we'll estimate the **probability** that it takes each value.

A **Latent Variable** is a variable that:

1. **Exists conceptually**, and
2. **Affects the data**, but
3. Is **not directly observed** in our dataset

In other words, a latent variable is something we don't see, but it explains patterns in the things we do see.

#### Example 4: continuation of the example on page 63

Let's go back to our cookie story.

- We **see the cookie**  $\Rightarrow$  that's observed data.
- We **don't see the machine** that made it  $\Rightarrow$  that's a **latent variable**.

The **machine label** (e.g. A, B, or C) is a latent variable  $z$  because:

1. It determines the cookie's sweetness/chocolate level.
2. But we aren't told what it is, we must infer it.

Since we cannot observe  $z_i$  directly, we can **estimate the probability** that  $z_i = k$  for each component  $k$ . This estimate is called **Responsibility** and is mathematically expressed as follows:

$$\gamma_{ik} = P(z_i = k \mid \mathbf{x}_i) \quad (50)$$

We are trying to *approximate* the unknown values of  $z_i$ . This enables us to **learn from incomplete data** without needing the true labels.

Latent Variable	Responsibility
Missing truth	Out best guess of that truth.

The responsibility is computed using **Bayes' Rule**:

$$\gamma_{ik} = \frac{\pi_k \cdot \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \cdot \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \Sigma_j)} \quad (51)$$

In simple terms, we have:

- The numerator is: how well component  $k$  explains point  $\mathbf{x}_i$ , weighted by how likely that component is in general.
- The denominator ensures that the total probability sums to 1 (normalization across all components).

The **main idea** behind GMM is very simple. Rather than stating that “point  $i$  is in cluster 2”, GMM says that point  $i$  is 80% likely to be in cluster 2, 15% likely to be in cluster 1, and 5% likely to be in cluster 3. These values are the **responsibilities**  $\gamma_{ik}$ . The word *responsibility* reflects that **each component is partially responsible** for explaining each point.

### 2.6.4 Expectation-Maximization (EM) Algorithm

The **Expectation-Maximization (EM) Algorithm** is how we **train** a Gaussian Mixture Model (GMM). Meaning, how we **estimate** the unknown parameters:

- Mixing coefficients  $\pi_k$
- Means  $\mu_k$
- Covariances  $\Sigma_k$

We use the EM algorithm **due to the unknown cluster labels**,  $z_i$ . The data is incomplete, so we cannot group the points by component. The EM algorithm usually works when some variables are latent (hidden).

#### ✂ How EM Works (Big Picture)

EM is an **iterative algorithm** with two main steps:

1. **E-step (Expectation)**. In this step, we **compute responsibilities**:

$$\gamma_{ik} = P(z_i = k \mid \mathbf{x}_i) = \frac{\pi_k \cdot \mathcal{N}(\mathbf{x}_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \cdot \mathcal{N}(\mathbf{x}_i \mid \mu_j, \Sigma_j)}$$

This is the **expected value of the latent variable**  $z_i$ , given the current parameter estimates. In simpler terms: we update our guess of how much each cluster  $k$  is responsible for each point  $\mathbf{x}_i$ .

2. **M-step (Maximization)**. In this step, we **update the model parameters** using the responsibilities:

- (a) **Update weights**  $\pi_k$ :

$$\pi_k = \frac{1}{n} \cdot \sum_{i=1}^n \gamma_{ik}$$

- (b) **Update means**  $\mu_k$ :

$$\mu_k = \frac{\sum_{i=1}^n \gamma_{ik} \cdot \mathbf{x}_i}{\sum_{i=1}^n \gamma_{ik}}$$

- (c) **Update covariances**  $\Sigma_k$ :

$$\Sigma_k = \frac{\sum_{i=1}^n \gamma_{ik} \cdot (\mathbf{x}_i - \mu_k) \cdot (\mathbf{x}_i - \mu_k)^T}{\sum_{i=1}^n \gamma_{ik}}$$

We're **maximizing the likelihood** of the data under the current soft assignments.

We alternate E and M steps **until the log-likelihood stops improving significantly**, or we **reach a maximum number of iterations**. This gives us the best estimate of the parameters, despite not knowing the cluster labels.

### ❓ What EM is really doing

Each iteration:

- Uses the **E-step** to “*guess*” hidden information (soft labels).
- Uses the **M-step** to “*retrain*” the model based on those guesses.

That's why it's called **Expectation-Maximization**:

- We **expect** the **missing values**.
- Then **maximize** the **likelihood given that expectation**.

### 2.6.5 Comparison between GMM and K-Means

Even though both Gaussian Mixture Models (GMM) and K-Means are used for **clustering**, they differ significantly in their assumptions, outputs, and flexibility.

- **Conceptual Difference**

Concept	K-Means	GMM
Type	Geometric algorithm	Probabilistic generative model
Goal	Minimize distances to centroids	Maximize likelihood of data under Gaussian model
Output	Hard assignments (1 cluster)	Soft assignments (probabilities per cluster)

- **Shape and Structure**

Feature	K-Means	GMM
Cluster Shape	Spherical (equal radius)	Elliptical (any orientation/size via covariance)
Covariance Info	Not modeled	Fully modeled via $\Sigma_k$ for each cluster
Equal size	Assumes clusters are similar in size	Allows different size and shape per cluster

- **Assignment Method**

- **K-Means** assigns each point to the **nearest centroid** (based on Euclidean distance).
- **GMM** assigns each point a **probability** of belonging to each cluster (based on likelihood).

This makes GMM better for:

- Overlapping clusters.
- Uncertain or fuzzy boundaries.
- Data with correlated features.

- **Optimization Strategy**

- **K-Means** minimizes **within-cluster sum of squared distances**.
- **GMM** maximizes the **log-likelihood** of the observed data via the **EM algorithm**.

- **Computational Cost**

Aspect	K-Means	GMM
Speed	Faster	Slower (EM is iterative)
Convergence	Guaranteed	Also guaranteed, but may find local optima
Initialization	Important	Even more sensitive to initialization

- **Summary Table**

Feature	K-Means	GMM
Assignments	Hard	Soft (probabilities)
Cluster shape	Spherical	Elliptical (via covariance)
Uses probability model?	✗ No	✓ Yes
Optimization objective	Distance	Log-likelihood
Suitable for overlapping?	✗ No	✓ Yes

- **When to Use What?**

Use Case	Choose
We need fast, simple partitioning	K-Means
We want a statistical model or density info	GMM
Clusters overlap or vary in shape	GMM
We have very large datasets	K-Means (or approximate GMM)

## 3 Discriminant Analysis

### 3.1 From Unsupervised to Supervised

So far, we've been working with **GMMs** and **clustering**, which are **unsupervised** methods. This meaning:

- We had no labels.
- We tried to discover patterns (e.g., clusters) **just from the data**.

Now we're moving to **Supervised Learning** where:

- Each observation  $\mathbf{x}_i$  **comes with a label**  $y_i \in \{1, \dots, K\}$
- We **want to learn a function** that can classify new inputs based on past labeled data.

#### 🔗 Why This Transition Is Natural (GMM $\rightarrow$ LDA)

Let's revisit how we modeled things in GMM. In GMM, we assumed:

$$\mathbf{x} \sim \sum_{k=1}^K \pi_k \cdot \mathcal{N}(\mu_k, \Sigma_k)$$

But we **didn't know** which  $k$  each  $\mathbf{x}$  came from, hence, EM algorithm. In Supervised Learning (like Linear Discriminant Analysis, LDA), we **do know** which class  $k$  each  $\mathbf{x}$  comes from. This lets us directly estimate the parameters  $\pi_k, \mu_k, \Sigma_k$  from the **labeled data**. Thus:

GMM (Unsupervised)	LDA/QDA (Supervised)
No labels	Labels $y_i \in \{1, \dots, K\}$
Estimate hidden assignments (EM)	Directly estimate class parameters
Responsibilities: soft membership	Class probabilities from Bayes rule
Soft clustering	Hard classification

Term	Meaning
<b>Latent Variable</b>	In GMM, the unknown class membership $Z$ ; in supervised learning, it's known.
<b>Generative Model</b>	A model for how data is generated: first sample label $y \sim \pi_k$ , then $\mathbf{x} \sim \mathcal{N}(\mu_k, \Sigma_k)$ .
<b>Supervised Learning</b>	Learn a classifier from labeled data — i.e., a function $f(\mathbf{x}) = \hat{y}$ .
<b>Discriminant Analysis</b>	Use Bayes' rule with class-conditional distributions to assign classes.

**✂ Key Formulation Shift**

In GMM (unsupervised), we care about:

$$\mathbb{P}(\text{class} = k \mid \mathbf{x}) = \frac{\pi_k \cdot \phi(\mathbf{x} \mid \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \cdot \phi(\mathbf{x} \mid \mu_j, \Sigma_j)}$$

Now in **supervised learning**, we do the same, but:

- We estimate  $\mu_k, \Sigma_k$  **per class** from labeled data.
- Then use Bayes' rule to classify a new point  $\mathbf{x}$ .

This is the basis for LDA and QDA (Quadratic Discriminant Analysis).

We're not throwing away the GMM mindset, we're now using it **with supervision**:

- Same assumptions: Gaussian class-conditional densities.
- But now, we **know the true class**, so we don't need EM.

This is why **LDA** is sometimes called the **supervised version of GMM with equal covariances**.



### 3.2 Introduction to Supervised Learning

**Supervised Learning** is about learning a mapping from inputs to outputs using **labeled data**. We're given a dataset:

$$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\} \quad (52)$$

Where:

- $\mathbf{x}_i \in \mathbb{R}^p$ : vector of features (e.g., age, salary, pixel values, etc.).
- $y_i \in \{1, 2, \dots, K\}$ : label.

Our goal is to **learn a function**:

$$f(\mathbf{x}) = \hat{y} \quad (53)$$

That can predict the correct  $y$  for a new input  $\mathbf{x}$ .

#### ≡ Types of Supervised Learning

Problem Type	Output $y$	Examples	Algorithms
Classification	Categorical/Discrete	Tumor type	LDA, QDA, logistic reg
Regression	Real-valued	Temperature	Linear regression

#### 🔗 Supervised vs. Unsupervised

	Unsupervised	Supervised
Input	$\mathbf{x}_i$	$(\mathbf{x}_i, y_i)$
Goal	Discover structure (clusters)	Predict outcome $y$ from $\mathbf{x}$
Methods	GMM, K-Means	LDA, QDA

So in GMM, we tried to **infer the labels**. In Supervised Learning, the labels are **given**, so the task becomes more about **learning rules**.

#### ❓ What is Discriminant Analysis, and how is it linked to Supervised Learning?

Because **Discriminant Analysis** is one of the **oldest and most classic forms of supervised learning**, specifically for **classification tasks**.

In Supervised Learning, we want to **learn a rule that predicts the class  $y$  given an input  $\mathbf{x}$** . Discriminant Analysis **does exactly that**, by:

1. Using the labeled data  $(\mathbf{x}_i, y_i)$

2. Estimating how each class looks in feature space, with a **probability model** for  $\mathbb{P}(\mathbf{x} \mid y = k)$
3. Applying **Bayes' rule** to invert that: estimate  $\mathbb{P}(y = k \mid \mathbf{x})$
4. Choosing the class with the highest probability

So we use the label  $y_i$  to:

- Group the data by class;
- Estimate a distribution for each class;
- Then **discriminate** new points by which class distribution they're most likely to belong to.

 **More in-depth. Discriminant Analysis** is a method to **classify observations into categories** (like “spam” vs “not spam”, or species A vs species B) by:

1. **Modeling how each class generates data**, using a probability distribution (usually multivariate normal).
2. Then using **Bayes' Rule** to compute the probability that a new point belong to each class.
3. Finally, assigning the new point to the **most probable class**.

The word *discriminant* refers to the idea of **finding rules or boundaries that discriminate** between classes based on the features  $\mathbf{x}$ . So:

- LDA: uses linear boundaries
- QDA: uses quadratic boundaries

These are both types of discriminant functions.

Discriminant Analysis **is called generative because it is a generative model** that assumes we can model:

- $\mathbb{P}(\mathbf{x} \mid y = k)$ : the distribution of features **given** the class.
- $\mathbb{P}(y = k)$ : the prior probability of each class.

Using Bayes' theorem:

$$\mathbb{P}(y = k \mid \mathbf{x}) = \frac{\mathbb{P}(\mathbf{x} \mid y = k) \cdot \mathbb{P}(y = k)}{\mathbb{P}(\mathbf{x})}$$

This is discriminant analysis, we **use these class-conditional densities to make predictions**.

**Definition 1: Generative Model**

A **Generative Model** models **how the data is generated**, i.e., it learns the **joint distribution**:

$$\mathbb{P}(\mathbf{x}, y)$$

From which we can **generate new data**, or compute the conditional:

$$\mathbb{P}(y \mid \mathbf{x})$$

More specifically, in classification tasks, a generative model models:

$$\mathbb{P}(\mathbf{x} \mid y = k) \quad (\text{class-conditional density})$$

And:

$$\mathbb{P}(y = k) \quad (\text{class prior})$$

Then it uses **Bayes' Rule** to compute:

$$\mathbb{P}(y = k \mid \mathbf{x}) = \frac{\mathbb{P}(\mathbf{x} \mid y = k) \cdot \mathbb{P}(y = k)}{\mathbb{P}(\mathbf{x})}$$

**Example 1: LDA as a Generative Model**

LDA assumes:

- Each class  $k$  has a Gaussian distribution:

$$\mathbf{x} \mid y = k \sim \mathcal{N}(\mu_k, \Sigma)$$

- We learn:
  - $\mu_k$ , the mean of class  $k$
  - $\Sigma$ , shared covariance matrix
  - $\pi_k$ , prior probability of class  $k$

With those, we can:

- **Generate data**: sample a class  $y \sim \pi_k$ , then sample  $\mathbf{x} \sim \mathcal{N}(\mu_k, \Sigma)$
- **Classify new points** using **Bayes' Rule**

**✂ Practical Importance**

- **Data**: we often have real-world datasets where class labels are known (e.g., in medicine, finance, spam detection).
- **Goal**: we want to *classify* new observations as accurately as possible.
- **Approach**: fit models like LDA, logistic regression, or modern ML classifiers (SVM, Trees, etc.).

### 3.3 Linear Discriminant Analysis (LDA)

**Linear Discriminant Analysis (LDA)** is a **supervised classification method** that models the probability distribution of each class as a multivariate **Gaussian with the same covariance matrix**, and uses **Bayes' rule** to assign a new observation to the class with the **highest posterior probability**.

Because it assumes equal covariances across classes, the resulting **decision boundaries are linear**, hence the name *Linear* Discriminant Analysis.

#### 🔍 What LDA Tries to Do

We're given data from multiple classes (say  $K = 2$  or more), and we want to **assign a new point  $\mathbf{x}$  to the most likely class**. LDA does this by:

1. Modelling **how each class generates data** (generative model).
2. Using **Bayes' rule** to compute  $\mathbb{P}(y = k \mid \mathbf{x})$ .
3. Assigning  $\mathbf{x}$  to the **class with the highest probability**.

#### Example 2: LDA

Suppose we want to classify fruit by weight and color:

- Oranges: heavy, orange.
- Lemons: light, yellow.

LDA will:

- Estimate the **mean** and **spread** (covariance) of orange and lemon data.
- Find a **line** that best separates the two, that's the linear discriminant.

#### 📋 Assumptions Behind LDA

1. **The data in each class looks like a cloud of points shaped like a Gaussian**. Imagine we have two groups of data:

- Group A, orange points
- Group B, blue points

Each group is made of points in 2D (like height and weight). We **assume** that in each group, the points:

- (a) Are **spread out in a bell-shaped way**.
- (b) Cluster around a mean (center).
- (c) Are more dense in the middle, less dense on the sides.

That's called a **multivariate normal distribution** (like a 2D Gaussian blob).

🔍 **Why?** Because this makes the math nice. We can use formulas to describe where the data is most likely to appear in each class.

2. **All the groups are spread out in the same way.** This means:

- Even though Group A and Group B may be **centered in different places** (different means),
- They are **equally wide, equally tilted, and equally stretched** (same covariance).

Two round clouds of points, one at the left, one at the right, then same shape, just different centers.

🔍 **Why?** Because this gives us **linear boundaries**. Meaning, the separator between groups will be a **straight line** (or a plane in higher dimensions). That's why it's called *Linear* Discriminant Analysis.

3. **We roughly know how frequent each class is.** This is called the **prior probability**: if we know that 70% of our training data is oranges and 30% lemons, we should take that into account when classifying.

🔍 **Why?** Because we don't want to treat all classes as equally likely if they aren't. Priors help us adjust for that.

### 📖 Derivation via Bayes Rule

To really understand how LDA works, not just *what it does* but *why it works*, we need to know that LDA classifies a point  $\mathbf{x}$  by computing "*which class is most likely to have generated this point?*". And the way to compute that is:

- Estimate how likely  $\mathbf{x}$  is *under each class* (the likelihood).
- Multiply it by how common that class is (the prior).
- Then use Bayes' Rule to get the final answer: the **posterior probability**.

🎯 **Goal.** Specifically, we want to **assign a new observation,  $\mathbf{x}$** , to one of  $K$  classes by computing the following:

$$\mathbb{P}(y = k \mid \mathbf{x}) \quad (\text{posterior probability}) \quad (54)$$

It is the **Posterior Probability**, which answers the question, "*given an observation  $\mathbf{x}$ , what is the probability that it belongs to class  $k$ ?*" It's called the **posterior** probability because it's what we know **after** seeing the data  $\mathbf{x}$ . In other words, this is the **probability that  $\mathbf{x}$  belongs to class  $k$** , given the observed features.

1. **Use Bayes' Rule.** Bayes' Rule helps us to **compute the posterior**:

$$\mathbb{P}(y = k \mid \mathbf{x}) = \frac{\mathbb{P}(\mathbf{x} \mid y = k) \cdot \mathbb{P}(y = k)}{\mathbb{P}(\mathbf{x})}$$

Where:

- $P(y = k | \mathbf{x})$  is the posterior, what we want to compute.
- $P(\mathbf{x} | y = k)$  is the **likelihood**, how likely  $\mathbf{x}$  is if it comes from class  $k$ .
- $P(y = k)$  is the **prior**, how likely class  $k$  is overall.
- $P(\mathbf{x})$  is the **evidence**, same for all classes, just a normalizing constant.

2. **What LDA Assumes.** LDA **assumes** that:

- (a) Each class  $k$  generates data using a **Gaussian distribution**:

$$\mathbf{x} | y = k \sim \mathcal{N}(\mu_k, \Sigma)$$

- (b) All classes share the same covariance  $\Sigma$ .

- (c) The prior  $P(y = k) = \pi_k$  is known or estimated from the data.

3. **Simplify the Formula.** Because  $\mathbb{P}(\mathbf{x})$  is the same for every class (it doesn't depend on  $k$ ), we can **ignore it when choosing the best class**. So the posterior is **proportional** to:

$$\mathbb{P}(y = k | \mathbf{x}) \propto \pi_k \cdot \phi(\mathbf{x} | \mu_k, \Sigma)$$

Where:

- $P(y = k | \mathbf{x})$  is the posterior, what we want to compute.
- $\propto$  is the proportional symbol.
- $\phi(\mathbf{x} | \mu_k, \Sigma)$  is the Gaussian probability density for class  $k$ .

To make this easier to work with, we **take the logarithm** of this expression (log doesn't change which class is biggest). This gives the **Discriminant Function**:

$$\delta_k(\mathbf{x}) = \log(\pi_k) - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \mathbf{x}^T \Sigma^{-1} \mu_k \quad (55)$$

Then we classify  $\mathbf{x}$  to the class  $k$  with the **highest**  $\delta_k(\mathbf{x})$ .

The final rule is that we need to **assign  $\mathbf{x}$  to the class  $k$  for which  $\delta_k(\mathbf{x})$  is the largest**. This is the **LDA decision rule**.

### ❓ What is the Decision Rule?

The **LDA Decision Rule** assigns a new observation  $\mathbf{x}$  to the class  $k$  with the **highest posterior probability**:

$$\hat{y} = \arg \max_k \mathbb{P}(y = k | \mathbf{x}) \quad (56)$$

But since **computing** that posterior **directly is messy**, we used **Bayes' Rule to rewrite it as**:

$$\mathbb{P}(y = k | \mathbf{x}) \propto \pi_k \cdot \phi(\mathbf{x}; \mu_k, \Sigma)$$

So instead, we define a **discriminant function**:

$$\delta_k(\mathbf{x}) = \log(\pi_k) - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \mathbf{x}^T \Sigma^{-1} \mu_k$$

Then the decision rule becomes:

$$\hat{y} = \arg \max_k \delta_k(\mathbf{x}) \quad (57)$$

### 🔍 Geometry: Why It's Called Linear

LDA is called Linear Discriminant Analysis because the decision boundary it creates between classes is a **straight line** (in 2D), or a **hyperplane** (in higher dimensions).

🔍 **What's a decision boundary?** It's the set of points  $\mathbf{x}$  where:

$$\delta_k(\mathbf{x}) = \delta_l(\mathbf{x})$$

i.e., where the classifier is **undecided** between class  $k$  and class  $l$ . This boundary tells us: *“here, both classes are equally likely, it's the tipping point between them.”*.

📖 **Remark: LDA Discriminant Function.** The discriminant function:

$$\delta_k(\mathbf{x}) = \log(\pi_k) - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \mathbf{x}^T \Sigma^{-1} \mu_k$$

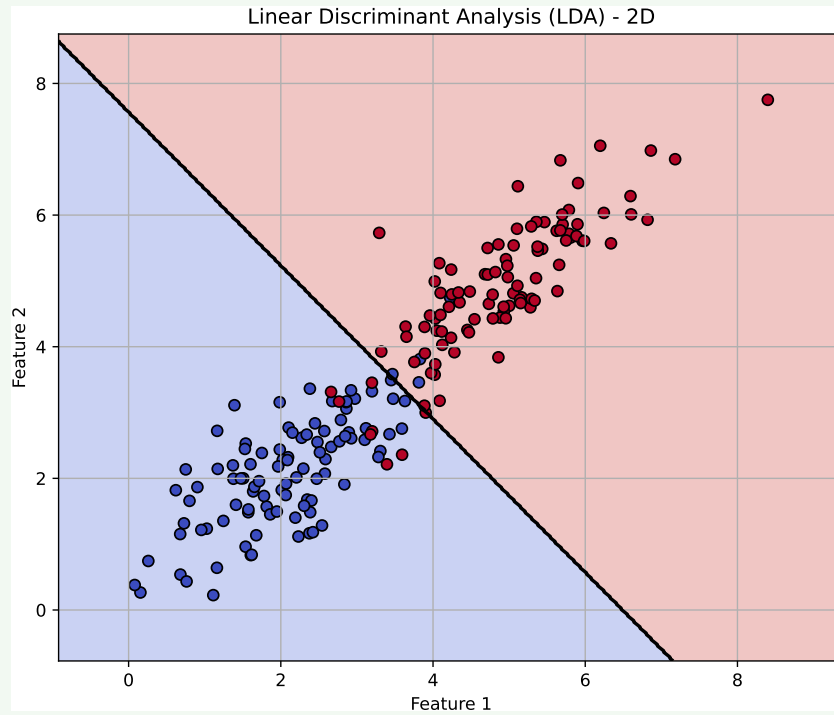
It's **linear in  $\mathbf{x}$** , there's no square, no exponent, just a dot product of  $\mathbf{x}$  with some vector. So if we compare two classes  $k$  and  $l$ , their difference:

$$\delta_k(\mathbf{x}) - \delta_l(\mathbf{x})$$

Is also **linear in  $\mathbf{x}$** .

**Example 3: Plot of a 2D LDA**

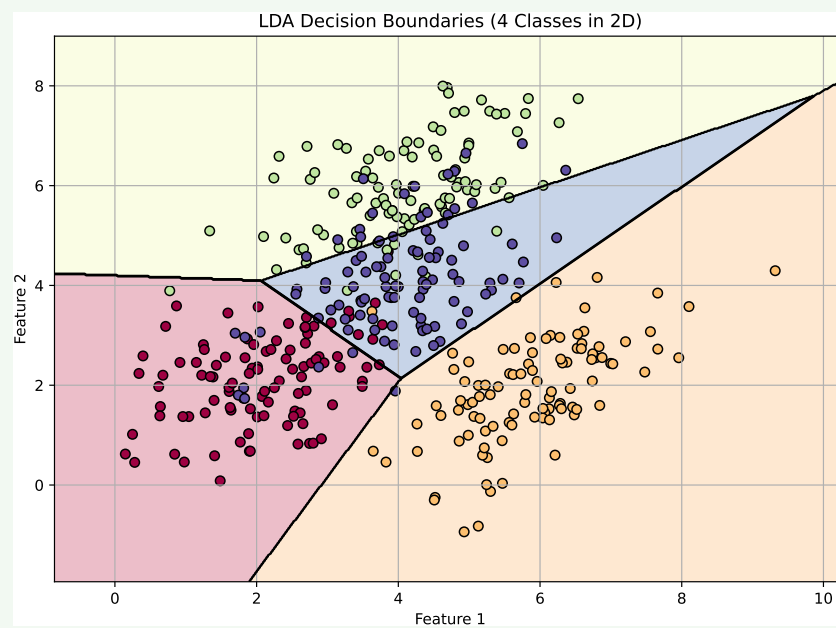
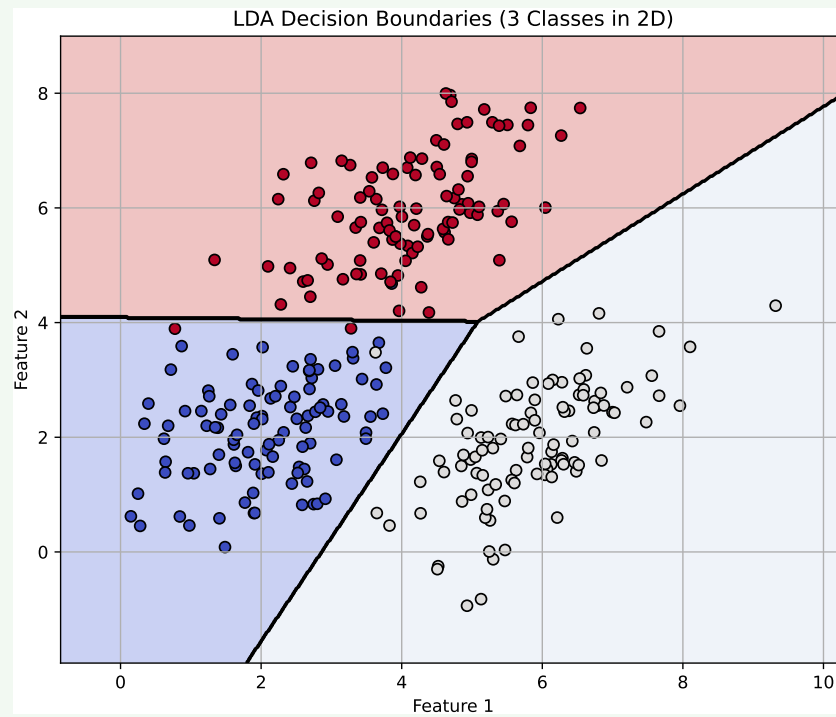
Let's say we have 2 features: height (cm), weight (kg). Then the LDA boundary between class A and class B will look like a straight line in the 2D plane:



All the points on the blue side belong to class A, and all the points on the other side belong to class B.

However, LDA support any number of classes  $K \geq 2$ . With 2 classes, the decision boundary is a single line; with 3 or more classes, LDA creates multiple regions, each with its own boundary between class pairs.







### 3.4 Quadratic Discriminant Analysis (QDA)

**Quadratic Discriminant Analysis (QDA)** is a **supervised classification method** that models each class as a multivariate Gaussian distribution **with its own covariance matrix**.

It uses **Bayes' Rule** to classify new observations by computing posterior probabilities, like LDA, but allows **more flexible class shapes**.

#### Intuition

QDA is just like LDA, **but more flexible**:

- In **LDA**, all classes share the same shape (same covariance, then same spread, same orientation).
- In **QDA**, each class has its **own shape and orientation** in feature space.

This means we can model curved boundaries and non-symmetric class regions. But the **price is more parameters**, and **more chance of overfitting** on small datasets.

#### What is Overfitting?

**Overfitting** occurs when a **model learns** not just the **true patterns** in the data, **but also the random noise**. It's like memorizing answer for an exam instead of understanding the subject; we do well on the questions we saw before (training data), but fail on new ones (test data).

Imagine fitting a curve to data points:

Fit Type	Description
<b>Underfitting</b>	The model is too simple (e.g., a straight line when the pattern is curved), it misses important structure.
<b>Good Fit</b>	The model captures the true pattern without being too complex.
<b>Overfitting</b>	The model bends and twists to go exactly through every training point, even if those points contain random noise.

Table 7: Fitting type.

Our model is overfitting when we observe the following:

- **Training error:** *very low*
- **Test error:** *high*

The model performs **too well** on the training set because it's **memorized it**, not generalized it.

## Overfitting Demonstration: Polynomial Regression

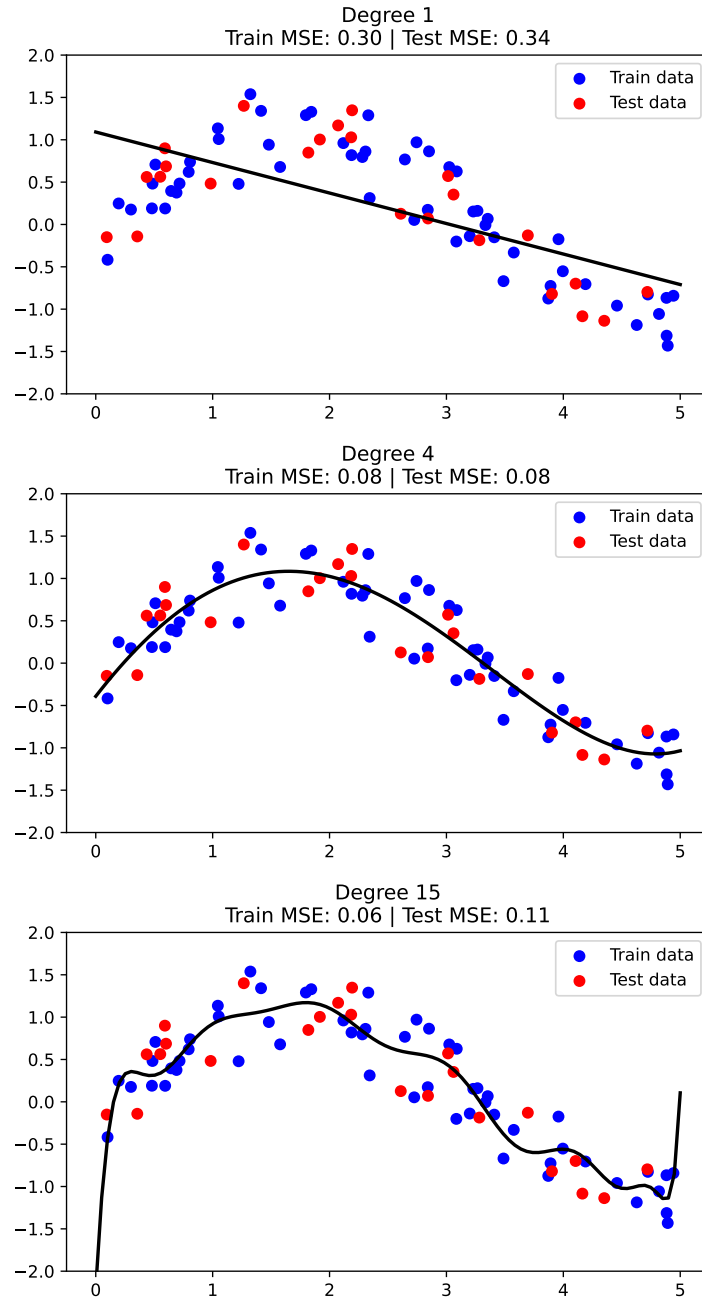


Figure 11: An example of overfitting. This figure shows polynomial regression models with increasing complexity. Although polynomial regression itself hasn't been explained, the purpose of this figure is to illustrate the effect of overfitting as model complexity increases (degree 1 underfits, too simple, degree 4 fits well, good generalization, degree 15 overfits, captures noise).

### 🔗 How does QDA differ from LDA?

- **Covariance Matrix**

LDA	QDA
Shared across all classes, $\Sigma$	Unique per class, $\Sigma_k$ for each class

Covariance Matrix.

- **Decision Boundary**

LDA	QDA
<b>Linear</b> (straight lines / hyperplanes)	<b>Quadratic</b> (curved lines / surfaces)

Decision Boundary.

- **Model Complexity**

LDA	QDA
Simpler, fewer parameters to estimate	More complex, more parameters (one full covariance per class)

Model Complexity.

- **Bias vs. Variance**

LDA	QDA
Lower variance, higher bias (more stable)	Lower bias, higher variance (risk of overfitting)

Bias vs. Variance.

- **Shape of Class Regions**

LDA	QDA
All classes modeled with the <b>same shape &amp; orientation</b>	Each class can have a <b>different shape &amp; orientation</b>

Shape of Class Regions.

- **Interpretability**

LDA	QDA
Easier to interpret decision boundaries	Harder to visualize, especially in higher dimensions

Interpretability.

- **Performance with Few Data**

LDA	QDA
Works well, robust	Needs more data to estimate separate covariances reliably

Performance with Few Data.

### ? What's the key difference?

LDA and QDA both model each class as a multivariate Gaussian, but the key difference is:

- **LDA assumes all classes share the same covariance matrix.** The data clouds (ellipses) all have the same shape and orientation.
- **QDA allows each class to have its own covariance matrix.** The data clouds (ellipses) can have different shapes, sizes, and orientations.

### ? What this means in practice

Concept	LDA	QDA
<b>Boundary shape</b>	Straight lines (linear)	Curved lines (quadratic)
<b>Flexibility</b>	Less, all classes “look similar”	More, each class can “look unique”
<b>Math form</b>	Linear discriminant function	Quadratic discriminant function



Figure 12: We have two clouds of points. In LDA, they are both ellipses of the same shape, just located in different places. In QDA, one ellipse might be wide and flat, the other tall and narrow, and the model adapts to those shapes.

### 🔗 The Decision Rule and Why It's Quadratic

We want to classify a new point  $\mathbf{x}$  by computing the **posterior probability** (same as page 77):

$$\mathbb{P}(y = k \mid \mathbf{x})$$

And assigning  $\mathbf{x}$  to the class  $k$  with the **highest posterior**. This is the same idea as LDA, but now, each class has **its own covariance matrix**, which changes the math.

1. **Bayes' Rule** (same as page 77).

$$\mathbb{P}(y = k \mid \mathbf{x}) = \frac{\mathbb{P}(\mathbf{x} \mid y = k) \cdot \mathbb{P}(y = k)}{\mathbb{P}(\mathbf{x})}$$

Again, since the denominator  $\mathbb{P}(\mathbf{x})$  is the same for all classes, we only need to **maximize the numerator**:

$$\mathbb{P}(y = k \mid \mathbf{x}) \propto \pi_k \cdot \phi(\mathbf{x} \mid \mu_k, \Sigma_k)$$

- $\pi_k$  is the class prior.
  - $\phi(\mathbf{x} \mid \mu_k, \Sigma_k)$  is the Gaussian density with **class-specific**  $\Sigma_k$
2. **Log of the Gaussian Density**. Taking the **log** of this expression (just like in LDA, page 78), we get a **Discriminant Function**:

$$\delta_k(\mathbf{x}) = -\frac{1}{2} \cdot \log |\Sigma_k| - \frac{1}{2} \cdot (\mathbf{x} - \mu_k)^T \Sigma_k^{-1} (\mathbf{x} - \mu_k) + \log(\pi_k) \quad (58)$$

The **quadratic part** comes from:

$$(\mathbf{x} - \mu_k)^T \Sigma_k^{-1} (\mathbf{x} - \mu_k)$$

Which is a **quadratic form**. A **Quadratic Form** is just a fancy way to describe any expression where variables are multiplied together, especially with squares and cross-terms.

The **decision boundary** between class  $k$  and  $l$  is:

$$\delta_k(\mathbf{x}) = \delta_l(\mathbf{x})$$

This becomes a **quadratic equation** in  $\mathbf{x}$ . Hence, the boundary is a **curve**, not a straight line (in 2D curved lines, in 3D curved surfaces, and in  $p$ -D quadratic hypersurfaces).



### ✔ When to Prefer QDA Over LDA

✔ **Class Covariances Are Clearly Different.** If our data shows that:

- Class A is tightly clustered.
- Class B is widely spread out.
- Or they have very different **orientations**.

Then QDA is the better choice because it allows:

$$\Sigma_1 \neq \Sigma_2 \neq \dots \neq \Sigma_K$$

This leads to **more accurate boundaries**, especially when classes have **non-linear separation**.

✔ **We Have Enough Data.** QDA estimates a **full covariance matrix for each class**, which means:

- More parameters to estimate.
- Higher risk of overfitting if sample size is small

So we **use QDA when our dataset is large enough** to support all those estimates reliably.

✔ **We Suspect Curved Boundaries.** If we visually or intuitively suspect:

1. The data cannot be separated with a straight line.
2. There's curvature in how the classes divide.

Then QDA is more likely to model the real structure.

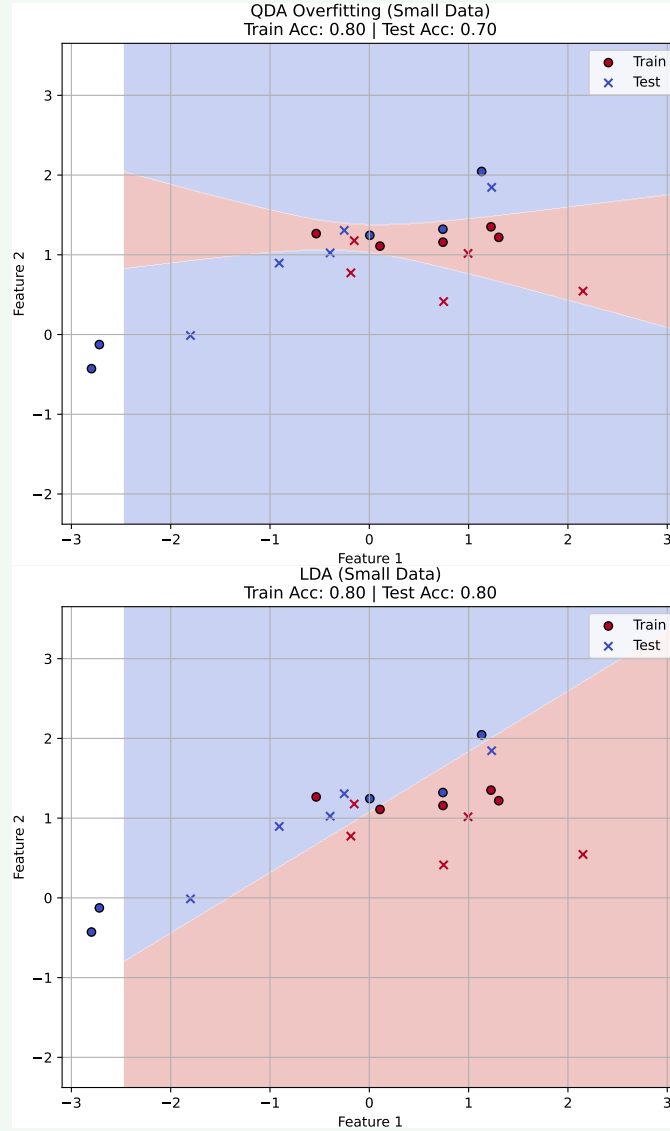
### ⚠ When to Avoid QDA

✗ **Small sample size.** Estimating separate covariance matrices can become unstable.

✗ **High-dimensional data.** In  $p$ -dimensional space, each  $\Sigma_k$  has  $\frac{p(p+1)}{2}$  parameters, QDA becomes data-hungry.

✗ **Risk of overfitting.** More flexibility, then higher variance, especially with noise or unbalanced classes.

Use <b>LDA</b> when...	Use <b>QDA</b> when...
We want simplicity and robustness	We want flexibility and accuracy
Covariances are similar	Covariances are different
Data is limited	We have enough samples per class
Boundaries are approximately linear	Boundaries appear curved

**Example 4: Overfitting**

In this example, both QDA and LDA were trained on a very small dataset with only 10 training samples per class. QDA achieved 80% accuracy on the training set by fitting a flexible, curved decision boundary that tightly adapted to the few observed points. However, this resulted in poor generalization, with test accuracy dropping to just 70%, a possible case of overfitting. In contrast, LDA used a simpler, linear decision boundary, achieving same training accuracy (80%) but significantly better performance on the test set (80%). This demonstrates that LDA, being more constrained, generalizes better in low-data regimes, whereas QDA requires more data to avoid overfitting its more complex model.

### 3.5 Comparison LDA vs QDA

#### ✂ Core Similarities

Aspect	Description
<b>Both are generative</b>	They model how each class generates data using a multivariate Gaussian.
<b>Both use Bayes' rule</b>	Classification is based on computing posterior probabilities.
<b>Both are supervised</b>	They require labeled training data.
<b>Both work for multiple classes</b>	Not limited to binary classification.

#### ⚖ Key Differences

Feature	LDA	QDA
<b>Covariance assumption</b>	Same for all classes: $\Sigma$	Unique for each class: $\Sigma_k$
<b>Decision boundary</b>	Linear (straight lines/planes)	Quadratic (curved boundaries)
<b>Model complexity</b>	Fewer parameters, then simpler	More parameters, then more flexible
<b>Training data needed</b>	Works well with small datasets	Needs more data (per class)
<b>Risk of overfitting</b>	Lower	Higher
<b>Bias-Variance tradeoff</b>	Higher bias, lower variance	Lower bias, higher variance
<b>Shape of decision region</b>	All classes have same shape	Each class can have different shape/size/orientation

#### ✓ Rule of Thumb

Situation	Recommended
We want a <b>stable model</b> with limited data	Use <b>LDA</b>
We have <b>enough data</b> and suspect curved decision boundaries	Try <b>QDA</b>
Our classes look like <b>different clouds</b> (e.g. rotated, stretched)	Use <b>QDA</b>
Our classes are well-separated by a <b>straight line/plane</b>	<b>LDA</b> is likely sufficient

## 3.6 Bayes Classifier

### 3.6.1 What is the Bayes Classifier?

The **Bayes Classifier** is a **decision rule** that assigns each observation to the class with the **highest posterior probability**, based on the conditional probability distribution of the features given the class.

#### ✓ Formal Definition

Let:

- $\mathbf{x} \in \mathbb{R}^p$  be a new observation (feature vector).
- $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_K$  be the possible **classes**.
- $\pi_k = P(Y = \mathcal{C}_k)$  be the **prior probability** of class  $k$ .
- $f_k(\mathbf{x}) = P(\mathbf{X} = \mathbf{x} \mid Y = \mathcal{C}_k)$  be the **class-conditional density**.

Then the **posterior probability** of class  $\mathcal{C}_k$  given  $\mathbf{x}$  is:

$$P(Y = \mathcal{C}_k \mid \mathbf{X} = \mathbf{x}) = \frac{\pi_k f_k(\mathbf{x})}{\sum_{j=1}^K \pi_j f_j(\mathbf{x})} \quad (59)$$

The **Bayes classifier** assigns  $\mathbf{x}$  to the class with the **highest posterior**:

$$\hat{y} = \arg \max_k P(Y = \mathcal{C}_k \mid \mathbf{X} = \mathbf{x}) \quad (60)$$

#### ✓ Math Behind the Bayes Classifier

We will try to explain it in the simplest terms possible. Imagine we're a doctor trying to decide if a patient has **flu** or **allergy**, based on their **temperature**. We know from past experience:

- People with the flu usually have higher temperatures.
- People with allergies usually have normal temperatures.

Now, a new patient walks in with a temperature of 37.8°C. We ask: “*given this temperature, what is the most likely diagnosis?*”. The Bayes Classifier answer this by saying “*pick the class (flu or allergy) that is most likely, based on what we know about temperatures in each class.*”. So:

- If flu patients often have temperatures around 37.8°C, and allergy patients don't;
- Then the Bayes Classifier says: “*assign this patient to the flu class.*”.

Now that the intuition is clear, let's explain the math behind the Bayes Classifier.

We are trying to **classify a new observation  $\mathbf{x}$** , such as someone's temperature, into one of  $K$  classes (e.g., flu vs allergy). We want to choose the class  $\mathcal{C}_k$  that **maximizes the probability that it's the correct class**, given the data. This is the **posterior probability** (page 77):

$$P(Y = \mathcal{C}_k \mid \mathbf{X} = \mathbf{x})$$

To compute this, we use **Bayes' theorem**:

$$P(Y = \mathcal{C}_k \mid \mathbf{X} = \mathbf{x}) = \frac{\pi_k f_k(\mathbf{x})}{\sum_{j=1}^K \pi_j f_j(\mathbf{x})} \quad (61)$$

Let's explain each term:

- $\pi_k = P(Y = \mathcal{C}_k)$  is the **prior**, *how common is class  $k$  in general?* For example, if 40% of patients have the flu, then  $\pi_{\text{flu}} = 0.4$ .
- $f_k(\mathbf{x}) = P(\mathbf{X} = \mathbf{x} \mid Y = \mathcal{C}_k)$  is the **Class-conditional density**. For example, how likely is temperature 37.8°C *if* someone has the flu?
- The **numerator**  $\pi_k f_k(\mathbf{x})$  tells us: how likely is it that someone is in class  $k$  and has this data  $\mathbf{x}$ ?
- The **denominator** just makes sure that all the probabilities sum to 1.

Although the formula may be complex, we **don't need to compute the entire formula**. Since the denominator is the same for all classes, we only care about comparing:

$$\pi_k f_k(\mathbf{x})$$

We choose the class  $k$  that **maximizes** this value:

$$\hat{y} = \arg \max_k \pi_k f_k(\mathbf{x})$$

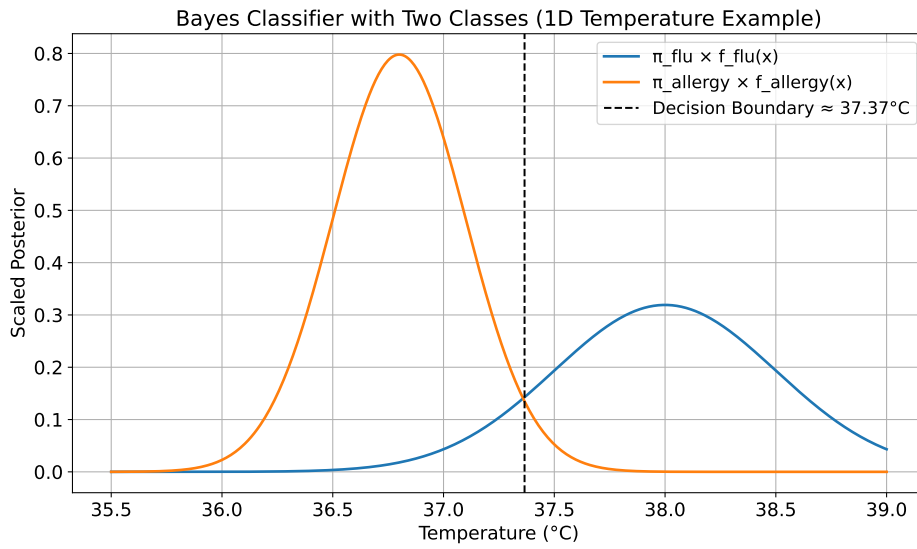


Figure 13: This plot shows how the Bayes classifier works in a simple case with two classes (flu and allergy), and one feature (body temperature). The blue curve  $\pi_{\text{flu}} \cdot f_{\text{flu}}(x)$  is the likelihood of flu at each temperature, and the orange curve  $\pi_{\text{allergy}} \cdot f_{\text{allergy}}(x)$  is the likelihood of allergy. At any temperature  $x$ , we compare the two curves; whichever is higher, we choose that class. The black dashed line shows the Bayes decision boundary, around  $37.37^\circ\text{C}$ : for  $x < 37.37$  choose allergy, otherwise for  $x > 37.37$  choose flu.

### 3.6.2 Bayes Error Rate

Even the **best possible classifier** (the Bayes classifier) can make mistakes. *Why?* Because **sometimes different classes overlap**: for some values of  $\mathbf{x}$ , it's just not clear which class it came from, even with perfect knowledge of the distributions. This unavoidable error is called the **Bayes Error Rate**.

The **Bayes Error Rate** is the **lowest possible classification error achievable by any classifier**. It's the error made by the Bayes classifier, which is optimal. Mathematically, it's the probability that the Bayes classifier **predicts the wrong class**:

$$\text{Bayes Error Rate} = \mathbb{E}_{\mathbf{X}} \left[ 1 - \max_k P(Y = C_k \mid \mathbf{X} = \mathbf{x}) \right] \quad (62)$$

This means:

- For each observation  $\mathbf{x}$ , compute the **posterior probabilities** for each class.
- Keep the **highest one**, that's our confidence in the prediction.
- Subtract it from 1, that's the chance of being wrong at that point.
- Average this over all possible  $\mathbf{x}$ , gives us the total Bayes error.

#### Binary Case (2 classes)

Since the previous formula works for any number of classes, *why focus on the binary case?* Because in the 2-class case, everything simplifies nicely:

- We only need to compare two posterior probabilities.
- The error at each  $\mathbf{x}$  is just the smaller of the two posteriors.
- The total Bayes error becomes the integral of  $\min(\pi_1 f_1(\mathbf{x}), \pi_2 f_2(\mathbf{x}))$ , which is easy to compute or visualize.

So the binary case gives a clean formula and is ideal for examples.

For two classes  $C_1$  and  $C_2$ , the error at any point  $\mathbf{x}$  is:

$$\text{Error at } \mathbf{x} = \min(P(Y = C_1 \mid \mathbf{x}), P(Y = C_2 \mid \mathbf{x})) \quad (63)$$

So:

$$\text{Bayes Error Rate} = \int \min(\pi_1 f_1(\mathbf{x}), \pi_2 f_2(\mathbf{x})) \, d\mathbf{x} \quad (64)$$

That is: **integrate over the overlap** of the two class distributions, where the “wrong” class is more likely.

- If the class distributions are **well separated**, the Bayes error is **very low** or even 0.
- If they **overlap**, the Bayes classifier will still make **errors**, no matter how our model is.

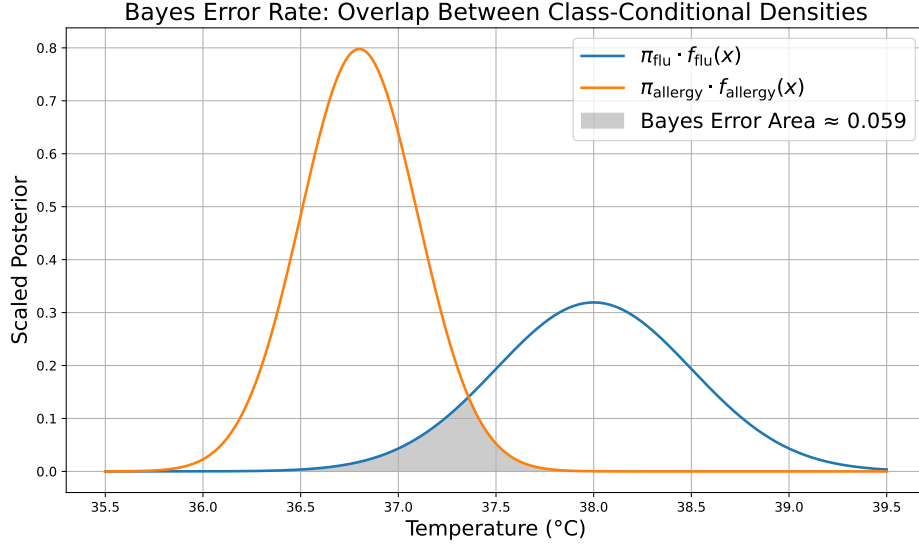


Figure 14: This plot shows the Bayes error rate visually as the gray shaded area. The value (from the integral of the shaded region) is  $\approx 0.059$ , meaning even the best classifier will make an error 5.9% of the time in this setting.

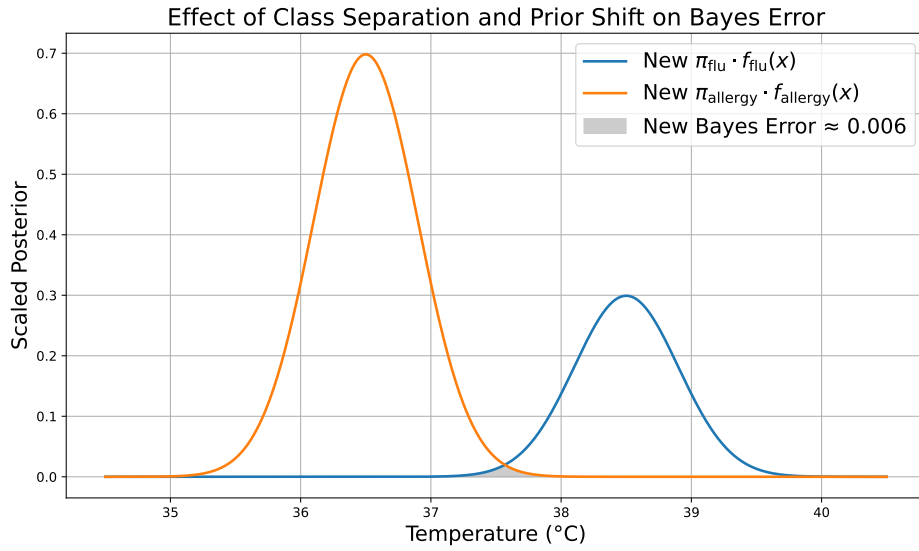


Figure 15: This plot is similar to the previous one, but here we increase the separation between the class means and change the priors. The two curves overlap much less, then the shaded area is smaller. The Bayes error rate drops significantly (now around 0.006, 0.6%) compared to  $\approx 5.9\%$  before.



### 🔗 Performance Measures

First, it's important to know that the **Oracle Classifier** is another name for the **Bayes classifier**. It emphasizes that the classifier knows the true class-conditional distributions and gives the lowest possible error.

In simulation studies or theoretical analysis, we can:

- Generate data from known distributions;
- Then simulate the Bayes/Oracle classifier (since we know  $f_k$ );
- And evaluate its **true error rate**.

This gives us a **reference point** to compare real classifiers like LDA, QDA, or logistic regression.

Once we have a classifier (e.g., LDA, QDA, Bayes), we **need to evaluate how well it performs**. We use two key error measures:

- **Apparent Error Rate (APER)**. It is the **training error**. It's calculated by applying the classifier to the **same data it was trained on**. Often **optimistic**, it **underestimates the real error**, especially if the model is overfitting.

$$\text{APER} = \frac{1}{n} \cdot \sum_{i=1}^n \mathbb{I}(\hat{y}_i \neq y_i) \quad (\text{on training data}) \quad (65)$$

- **Actual Error Rate (AER)** (or **True Error Rate**). The **true classification error**, measured on **independent test data**. Gives a **more honest estimate of model performance**. In simulations, we can also compare to the **true labels generated from known distributions**.

$$\text{AER} = \frac{1}{m} \cdot \sum_{j=1}^m \mathbb{I}(\hat{y}_j \neq y_j) \quad (\text{on test data or true labels}) \quad (66)$$

### 3.6.3 Bayes vs LDA/QDA

As we've seen in previous sections, the **Bayes classifier** is the gold standard because it yields the **lowest possible error** when the true distributions of the data are known.

But in practice, we rarely know those distributions. That's where LDA (Linear Discriminant Analysis) and QDA (Quadratic Discriminant Analysis) come in. They are **approximations** of the Bayes classifier under some assumptions.

As we've seen, the Bayes classifier is:

$$\hat{y} = \arg \max_k \pi_k \cdot f_k(\mathbf{x})$$

Where  $f_k(\mathbf{x})$  is the **true class-conditional density that is not known in practice** because we only have data. **That's why we can't use the Bayes classifier in practice, it's only a theoretical benchmark.** Instead, LDA and QDA assume that the shape of  $f_k$  is known. However, they do not know the shape of  $f_k$ ; they estimate it from the data.

- **LDA**: assumes all classes share the same covariance matrix.

#### Assumptions

- Class-conditional distributions: Multivariate Gaussian.
- **Same covariance matrix**  $\Sigma$  for all classes.
- Different means  $\mu_k$ .

This leads to a **linear decision boundary**. It is linear because when we plug the Gaussian formula into the Bayes rule and simplify under equal  $\Sigma$ , we get a linear function of  $\mathbf{x}$ .

- **QDA**: allows each class to have its own covariance matrix.

#### Assumptions

- Class-conditional distributions: Multivariate Gaussian.
- **Different covariance matrix**  $\Sigma_k$  for each class.

This leads to a **quadratic decision boundary**.

Property	Bayes Classifier	LDA	QDA
<b>Requires true <math>f_k</math></b>	✓ Yes	✗ No, assumes Gaussian	✗ No, assumes Gaussian
<b>Covariance</b>	Anything	Shared across classes	One per class
<b>Decision boundary</b>	Arbitrary (optimal)	Linear	Quadratic
<b>Flexibility</b>	Highest	Low	Medium
<b>Risk of overfitting</b>	None (theoretical)	Low	Higher
<b>When to use</b>	Only in theory	When classes are well-separated, simple	When boundaries are curved, and more data is available

### ❓ What Is the Purpose of the Bayes Classifier?

Although the Bayes classifier cannot be used in practice, it plays a fundamental role in machine learning and statistics:

- ✓ **It Defines the Best Possible Classifier.** It is the **gold standard**: the classifier with the **lowest possible error rate**. This error is called the Bayes error rate. All other classifiers are trying to get as close as possible to this theoretical optimum.
- ✓ **It Helps Us Understand the Role of Assumptions.** By comparing real classifiers to the Bayes rule, we can say:
  - *How much error is due to **model approximation**?*
  - *How much is due to **data limitations**?*

For example:

- LDA  $\approx$  Bayes **only** when its assumptions are correct (Gaussian, equal covariance).
- When assumptions are wrong (e.g., non-Gaussian data), LDA will do worse than Bayes.
- ✓ **It Provides a Foundation for Deriving Other Classifiers.** Many practical classifiers (LDA, QDA, Naive Bayes, even some neural networks) are **inspired by Bayes' rule**. They make simplifying assumptions or estimate  $f_k$ , but the **form of the classifier** is based on Bayes.
- ✓ **It Lets Us Theoretically Analyze the Limits of Learning.** It helps define concepts like bias-variance tradeoff, irreducible error, and Bayes consistency. We can measure how “good” a classifier is by how close it comes from to **Bayes-optimal**.

#### Example 5: Bayes Analogy

Think of the Bayes classifier like a **perfect GPS** that always knows the best route, because it sees the **entire traffic map of the world**. In practice, we only get local signals (data), so our actual GPS (LDA, QDA, etc.) is trying to guess the best route, and we measure its performance **by how close it is to the perfect one**.

## 4 Linear Regression

### 4.1 Regression vs Classification

In the previous chapter, we examined classification methods such as the Bayes classifier, QDA, and LDA. Now, we will examine regression methods.

- **Classification**

✓ **Goal:** predict a **category or class label**.

❓ **Output:** variable  $Y_i \in \{1, 2, \dots, G\}$ .

✓\* **Function**

$$f : \mathcal{X} \rightarrow \{1, 2, \dots, G\}$$

For example, email spam detection (spam, not spam), disease diagnosis (positive, negative).

- **Regression**

✓ **Goal:** predict a **continuous numerical value**.

❓ **Output:** variable  $Y_i \in \mathbb{R}$ .

✓\* **Function**

$$f : \mathcal{X} \rightarrow \mathbb{R}$$

For example, predicting house prices, forecasting temperature, estimating blood pressure.

Feature	Regression	Classification
Output type	Continuous	Categorical (Discrete labels)
Loss function	Usually MSE: $(y - \hat{y})^2$	Usually 0-1 loss or log-loss
Model types	Linear regression, Ridge, etc.	Logistic regression, SVM, etc.
Evaluation	RMSE, $R^2$	Accuracy, Precision, AUC

Table 8: Main Differences.

## 4.2 Simple Linear Regression (SLR)

### 4.2.1 Model Formulation and OLS Estimation

The **Simple Linear Regression (SLR)** is a **statistical method** that models the relationship between one **independent variable** (also called predictor or input)  $X$  and one **dependent variable** (also called response or output)  $Y$ , using a **straight line**. The model assumes:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (67)$$

Where:

- $Y_i$ : **observed response** (output) for sample  $i$ . It is the value we want to predict.
- $X_i$ : **predictor** (input). It is the value we use to predict.
- $\beta_0$ : **intercept** (value of  $Y$  when  $X = 0$ ).
- $\beta_1$ : **slope** (change in  $Y$  for a unit increase in  $X$ ). It indicates how much  $Y$  increases when  $X$  increases by 1.
- $\varepsilon_i$ : **random error** (assumed to be  $\sim \mathcal{N}(0, \sigma^2)$ , in other words normally distributed).

It can also be viewed in matrix form:

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

It's called **simple** because there is only one predictor ( $X$ ) and if there were multiple predictors  $X_1, X_2, \dots$ , we would call it **multiple linear regression**.

✔ **Goal.** We want to **predict a continuous outcome**  $Y \in \mathbb{R}$  based on a single input variable  $X \in \mathbb{R}$  using a **linear relationship**. For example, predicting house price  $Y$  based on house size  $X$ .

✔ **Key Idea**

We want to find the **best straight line**:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

That minimizes the **total squared error** between the actual  $Y$  and the predicted  $\hat{Y}$ .



Figure 16: This is an example of an SLR model in action. The X-axis represents the independent variable, hours studied, the Y-axis represents the dependent variable, exam score, the orange dots represent the actual observed data points, and the black line represents the best-fit regression line  $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ . The OSL finds the straight line that best represents the trend in the data; here, students who study more tend to get higher exam scores. The line summarizes that linear relationship with just 2 numbers: intercept and slope.

### ❓ What is OSL Estimation (Ordinary Least Squares)?

**Ordinary Least Squares Estimation (OSL Estimation)** is a method to find the best-fitting line for our data by minimizing the total squared errors between observed values and predicted values.

Given the SLR model, we don't know  $\beta_0$  and  $\beta_1$ . We want to **estimate them** using data  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ . We define the prediction as:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

And the **error (residual)** for point  $i$  is:

$$e_i = Y_i - \hat{Y}_i \quad (68)$$

### ❓ What does OLS Minimize?

OLS finds  $\hat{\beta}_0$  and  $\hat{\beta}_1$  that **minimize the sum of squared residuals**:

$$\text{RSS} = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \quad (69)$$

This is called the **Residual Sum of Squares (RSS)**.

### Core Idea of RSS

We want our regression line to be **as close as possible** to the observed data points. But *how do we measure “closeness”*? By the **vertical distance** between each true point  $(X_i, Y_i)$  and the predicted point  $(X_i, \hat{Y}_i)$ , that is:

$$e_i = Y_i - \hat{Y}_i \quad (\text{residual})$$

But some residuals are negative, some positive. So **we square them to avoid cancellation and emphasize large errors**:

$$\text{RSS} = \sum e_i^2$$



Figure 17: In this plot, we highlight the residual and the residual sum of squares (RSS). The blue dots are the actual data point  $(X_i, Y_i)$ , the red line is the regression line  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ , the gray dashed lines are the residuals  $e_i = Y_i - \hat{Y}_i$ , i.e., vertical distances from each point to the line, and the RSS value is displayed in the title, it's the sum of the squares of these residuals.

OLS chooses the red line that **minimizes the total gray distances squared (RSS)**. A different line (e.g., tilted more or less) would **increase the sum of squared errors** (see Figure 18, page 104). This is **why** OLS minimizes RSS, it leads to the line that best balances all errors.

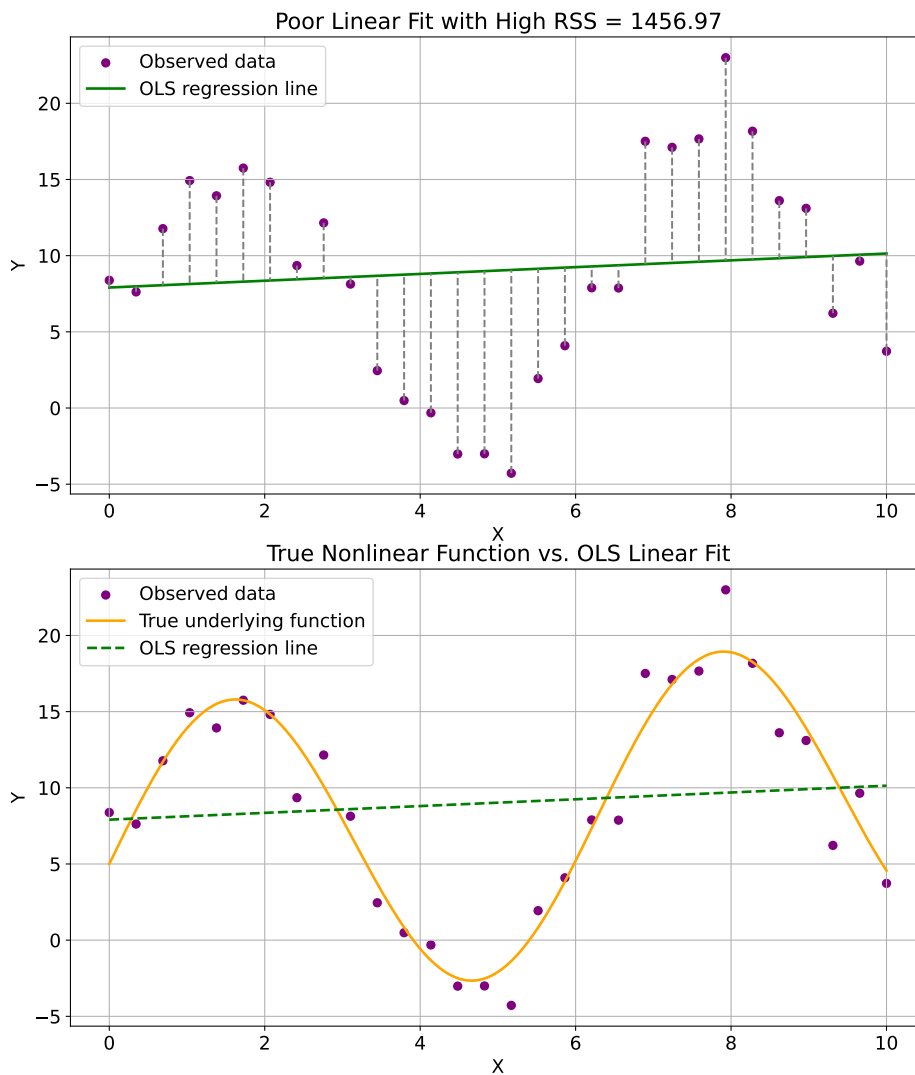


Figure 18: This is an example of a poor linear fit, which is also highlighted by the RSS. The purple dots are the true data points, which follow a **nonlinear (sinusoidal) trend**, green line is the best-fitting **straight line** found by OLS, the gray dashed lines are the **residuals** (the vertical errors between data points and the regression line), and the RSS is very high because the straight line fails to capture the curve in the data.

OLS always minimizes the **sum of squared residuals**, but if the model is misspecified (e.g., fitting a line to curved data), the **RSS will still be high**.

✔ Good fit  $\Rightarrow$  **low RSS**

✘ Bad fit  $\Rightarrow$  **high RSS**

OLS chooses the line that makes RSS as small as possible, even if the fit is poor due to model limitations.



### ☰ Ordinary Least Squares (OLS) Estimators

Solving for the minimum yields two formulas for the best-fitting line, called OLS estimators. **Ordinary Least Squares estimators** are the **values** of  $\hat{\beta}_0$  (intercept) and  $\hat{\beta}_1$  (slope) **that minimize the Residual Sum of Squares (RSS)** in simple linear regression (SLR):

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

We want to find  $\hat{\beta}_0, \hat{\beta}_1$  such that:

$$\text{RSS} = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Is **minimized**.

OLS gives **explicit formulas** for the best values:

- **Slope**

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X}) (Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (70)$$

Measures how much  $Y$  changes for a 1-unit increase in  $X$ . It's the **slope** of the best-fitting line.

- **Intercept**

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad (71)$$

Value of  $Y$  when  $X = 0$ .

These two formulas define the **Regression Line**:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \quad (72)$$

For example, suppose we compute  $\hat{\beta}_0 = 3, \hat{\beta}_1 = 2$ , then our model says: “*each time  $X$  increases by 1,  $Y$  increases by 2. If  $X = 0$ , predicted  $Y = 3$* ”.

### 4.2.2 Geometric Interpretation of Linear Regression

Understanding OLS **geometrically** gives deep intuition about what's really happening during the model fitting process.

✂ **Setup.** Suppose we have a dataset:

- We observe  $n$  responses and save them in a vector of observed responses  $\mathbf{Y} \in \mathbb{R}^n$ :

$$\mathbf{Y} = [Y_1, Y_2, \dots, Y_n]^T$$

- We have one predictor  $X$ , but the model includes a constant (intercept) too. So the **design matrix**  $\mathbf{X} \in \mathbb{R}^{n \times 2}$  has:

$$\mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$$

The column of ones is fundamental because, to calculate the regression line  $\hat{\mathbf{Y}}$  in matrix form, we need to  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ , which translates to:

$$\hat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \boldsymbol{\varepsilon} = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots + \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix} + \boldsymbol{\varepsilon}$$

So the column of ones ensures the **intercept**  $\beta_0$  is part of every prediction. Without the column of ones, we would be **forcing the line to pass through the origin**, i.e., no intercept!

Now, the column space of  $\mathbf{X}$  is all combinations of the two vectors: the all-ones vector (for the intercept), and the values of  $X$ :

$$\text{Col}(\mathbf{X}) = \{\beta_0 \cdot \mathbf{1} + \beta_1 \cdot X\}$$

This is a **2D plane inside**  $\mathbb{R}^n$ .

✔ **Goal.** The goal of OLS is to **project** the vector  $\mathbf{Y}$  onto the **space spanned by the predictors** in  $\mathbf{X}$ . In other words, since it is impossible to fit all our observed values ( $\mathbf{Y}$ ) perfectly with a line, **OLS selects a vector  $\hat{\mathbf{Y}}$  in the column space of  $\mathbf{X}$  that is as close as possible to  $\mathbf{Y}$** . This is done by **orthogonal projection**:

$$\begin{aligned} \hat{\mathbf{Y}} &= \text{projection of } \mathbf{Y} \text{ onto the space spanned by the columns of } \mathbf{X} \\ &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \text{projection of } \mathbf{Y} \text{ onto } \text{Col}(\mathbf{X}) \end{aligned}$$

It is important because the **projection minimizes the distance** from  $\mathbf{Y}$  (our observed values) to  $\hat{\mathbf{Y}}$  (regression line to find), exactly what OLS does.

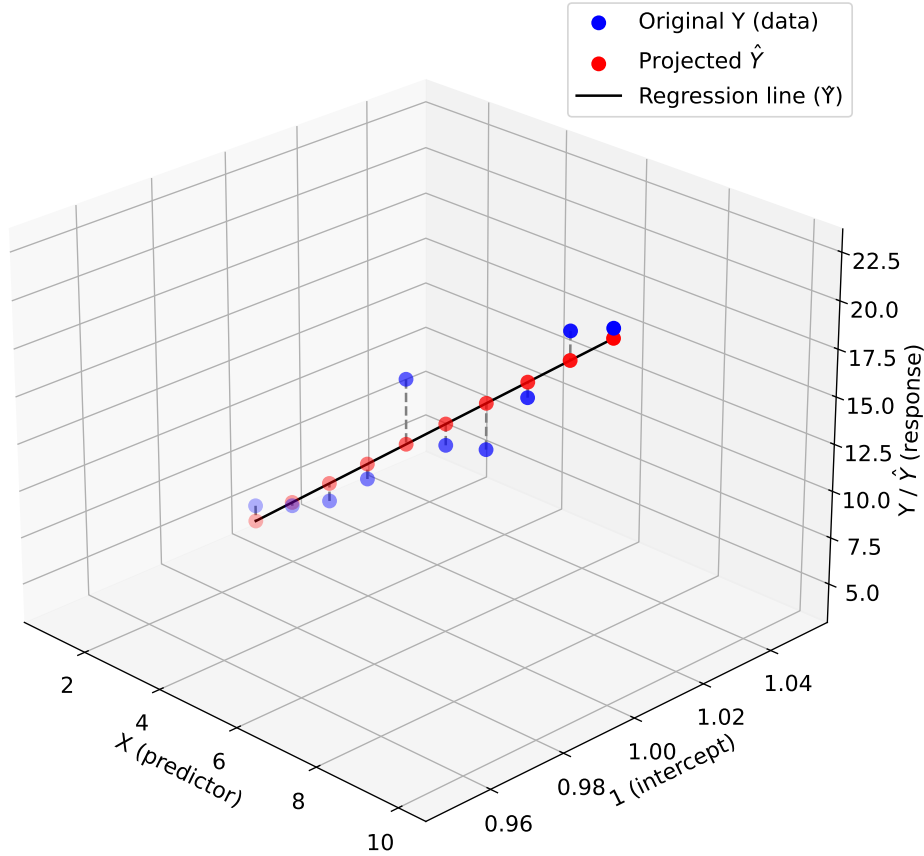
Geometric View: Projection of  $\mathbf{Y}$  onto Column Space of  $\mathbf{X}$ 

Figure 19: Blue dots are the actual observed  $Y_i$  values, the full response vector  $\mathbf{Y}$ . The red dots are the projected values  $\hat{Y}_i$ , they lie in the space spanned by the columns of  $\mathbf{X}$ . The gray dashed lines are the residuals, the vertical distance from each observed  $Y_i$  to its projected  $\hat{Y}_i$ . The black line is the fitted regression line (through the red points), this lies in the **column space** of the design matrix.

The design matrix  $\mathbf{X}$  has two columns: one column of ones (to include the intercept) and one column with the values of  $X$ . The red points from the best approximation  $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$ . The black line shows the **vector space of all possible**  $\hat{Y}$  values using linear combinations of the intercept and  $X$ . **OLS chooses the projection of  $\mathbf{Y}$  onto that line**, minimizing the **sum of squared residuals**.

### √\* Vector Space View

In Simple Linear Regression, the design matrix is, as we have seen, as follows:

$$\mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$$

OLS finds the **orthogonal projection** of  $\mathbf{Y}$  onto this space:

$$\hat{\mathbf{Y}} = \mathbf{P}_X \mathbf{Y} \quad (73)$$

Where  $\mathbf{P}_X$  is called **Projection Matrix**, or **Hat Matrix** because it maps  $\mathbf{Y}$  to its fitted values  $\hat{\mathbf{Y}}$ . It is defined as follows:

$$\mathbf{P}_X = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \quad (74)$$

It **projects** any vector in  $\mathbb{R}^n$  (like  $\mathbf{Y}$ ) onto the **column space of  $\mathbf{X}$** , that is, onto the space of all possible predicted values.

Finally, the residual vector is:

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} \quad (75)$$

That is **orthogonal** to the fitted vector  $\hat{\mathbf{Y}}$ .

### 4.3 Model Evaluation

After fitting the regression line, the next question is: *how good is the model at explaining the data?* That's what this section is about.

#### ✓ Goal of Model Evaluation

We want to understand:

1. How much of the variation in  $Y$  is **explained** by the model?
2. How much remains **unexplained** (random noise or poor fit)?

This leads to the **decomposition of variance** and the **coefficient of determination**  $R^2$ .

#### ≡ Variance Decomposition

We break down the total variability in the output  $Y$  into:

- **Regression Sum of Squares (SSR)** (or explained sum of squares)

$$\text{SSR} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \quad (76)$$

Measures how much of the variability in  $Y$  is **explained by the model**. It's the variability **captured by the regression line**.

- **Residual Sum of Squares (RSS)** (or error sum of squares)

$$\text{RSS} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Measures the remaining **unexplained error**, how far off the predictions are from the actual values. This was explained on page 102.

- **Total Sum of Squares (TSS)**

$$\text{TSS} = \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (77)$$

Measures the **total variability** in the data. It's how much the actual values  $Y_i$  vary around the mean  $\bar{Y}$ . In other words, how “*spread out*” the data is.

In Ordinary Least Squares (OLS) regression, we always have the following **Fundamental Identity**:

$$\text{TSS} = \text{SSR} + \text{RSS} \quad (78)$$

OLS finds  $\hat{\beta}$  to **minimize RSS**, and it guarantees that the residuals  $\mathbf{e} = Y - \hat{Y}$  are **orthogonal** to the fitted values. Because of this orthogonality, the squared norms satisfy the Pythagorean theorem:

$$\|Y - \bar{Y}\|^2 = \|\hat{Y} - \bar{Y}\|^2 + \|Y - \hat{Y}\|^2 \Rightarrow \text{TSS} = \text{SSR} + \text{RSS}$$

### ☰ Coefficient of Determination

The **Coefficient of Determination** is **one of the most important metrics** to evaluate how well our linear regression model fits the data. It is defined as:

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{\text{SSR}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \quad (79)$$

It tells us the **proportion** of the variance in  $Y$  that is **explained by our regression model**.

✓ If our model predicts the values **perfectly**, then:

$$\hat{Y}_i = Y_i \Rightarrow \text{RSS} = 0 \Rightarrow R^2 = 1$$

⚖ If our model does **not better than predicting the mean**, then:

$$\hat{Y}_i = \bar{Y} \Rightarrow \text{SSR} = 0 \Rightarrow R^2 = 0$$

✗ If our model is **worse than predicting the mean**, then:

$$\text{RSS} > \text{TSS} \Rightarrow R^2 < 0$$

For example, if  $\text{TSS} = 100$ , and  $\text{RSS} = 20$ , then:

$$R^2 = 1 - \frac{20}{100} = 0.80 \Rightarrow \text{Our model explains 80\% of the variation in } Y$$

$R^2$  always **increases** as we add **more predictors** (even useless ones), and for multiple regression, we often use Adjusted  $R^2$  to penalize unnecessary complexity.



Figure 20: A side-by-side comparison of a Good Model vs. a Bad Model using the same dataset.

Metric	What it means	Good Model	Bad Model
TSS	Total variability in $Y$	3821.1 Same for both, this is fixed by the data	3821.1
RSS	Unexplained error	325.96 Low, model fits well	3751.18 High, model fits poorly
SSR	Explained variation	3495.13 High, captures most of the structure	69.92 Very low
$R^2$	Proportion of explained variance	0.91 91% explained	0.02 Only 2% explained

Table 9: Model evaluation of Figure 20.

This illustrates how  $R^2$  measures **how well ur model captures the true pattern in the data.**

## 4.4 Statistical Inference

**Statistical Inference** is the process of using data from a sample to learn something about the **true population parameters**, and to **quantify uncertainty**.

In linear regression, it means: we've estimated  $\hat{\beta}_0, \hat{\beta}_1$  from a sample, but the true values  $\beta_0, \beta_1$  are unknown. So we ask:

- *Are these estimates reliable?*
- *Are they statistically significant?*
- *How much could they vary if we repeated the data collection?*

🔗 **So what do we do in regression inference?** We perform tests and intervals on the model coefficients:

- **Sampling Distributions & CLT** (page 113). Under the hood, regression estimates come from random samples. The **Central Limit Theorem** (CLT) tells us these **estimates** follow an **approximate normal distribution**. This gives us the **foundation** to build confidence intervals and test hypotheses.
- **t-Test for Individual Coefficients** (page 115). We test each  $\hat{\beta}_j$  (like slope or intercept) against 0: *is this variable useful?* If  $\hat{\beta}_j$  is far from 0 (statistically), we say it's **significant**.
- **Global F-Test** (page 116). This is a test on the **entire model**: *does any variable in the regression help explain the response?* It compares the model with predictors vs. a model with just the intercept.

Question	Answered by
Is this predictor significant?	t-test
Is the model better than nothing?	F-test
How much do my estimates vary?	CLT and confidence intervals



### 4.4.1 Sampling Distributions & Central Limit Theorem (CLT)

#### 🔗 Why do we care?

In the real world we don't know the true coefficients  $\beta_0, \beta_1$ . We **estimate them** from a sample data. But if we collected a **different sample**, we'd get **different estimates**.

So the question is: *How variable are the estimates  $\hat{\beta}_0$  and  $\hat{\beta}_1$ ?* In other words: *"if we repeated the data collection and regression **many times**, how much would  $\hat{\beta}_1$  change from sample to sample?"*. This is where **sampling distributions** and the **Central Limit Theorem (CLT)** come in.

#### 🔗 What is a Sampling Distribution?

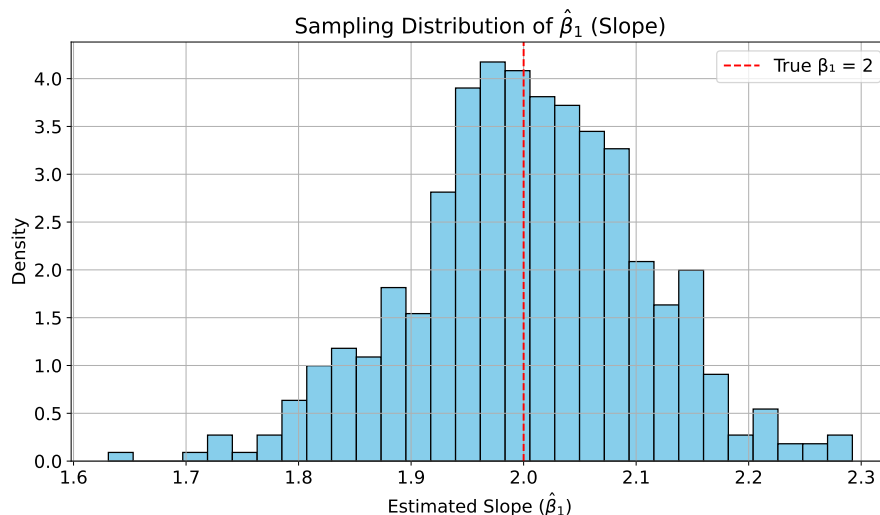
A **Sampling Distribution** is the **distribution of a statistic** (like  $\hat{\beta}_1$ ) that we get when we **repeat the sampling process many times**.

Imagine we have a population where:

$$Y = 3 + 2X + \varepsilon$$

We randomly collect 50 data points from this population, and we get  $\hat{\beta}_1 = 1.96$ . We collect another 50 data points from the same process, and now we get  $\hat{\beta}_1 = 2.15$ .

Each time we get a different dataset (due to random noise), so we get a different estimate of  $\hat{\beta}_1$ . Now imagine we plot a histogram of all those  $\hat{\beta}_1$  values. That histogram is the **sampling distribution** of  $\hat{\beta}_1$ .



### What is the Central Limit Theorem (CLT)?

If we take **many random samples** from any population (no matter its shape), and compute the **mean** (or other statistics) for each sample, then the **sampling distribution** of that statistic will be **approximately normal**, as long as the sample size is large enough.

**Theorem 1 (Central Limit Theorem).** Let  $X_1, X_2, \dots, X_n$  be independent random variables with:

- Any distribution (not necessarily normal)
- Mean  $\mu$  and finite variance  $\sigma^2$

Then the sample mean:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

Has a sampling distribution that **converges** to:

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

In short, **means behave like normals**, even if the data itself doesn't!

### Why does CLT matter in Regression?

Because it allows us to:

1. **Quantify Uncertainty.** When we run regression, we get estimates like  $\hat{\beta}_1 = 2.1$ . But it's just from one sample, *what if the true slope is different?* The CLT says: "if we took many samples, the estimates  $\hat{\beta}_1$  would follow a normal distribution around the true value  $\beta_1$ ". So now we know **how much those estimates vary**.
2. **Build Confidence Intervals.** CLT tells us we can say: "with 95% confidence, the true slope lies between 1.85 and 2.35". Without the CLT, we couldn't say that. We'd have no way to define what "95% confidence" even means.
3. **Perform Hypothesis Tests.** The CLT lets us test:
  - *Is the slope really different from zero?*
  - *Is this variable important in predicting the outcome?*

That's done using the fact that  $\hat{\beta}_1$  is **approximately normal**, thanks to the CLT.

4. **Trust the Shape of Our Estimators.** Even if our data isn't perfect (e.g., not normally distributed), CLT tells us: "*it's okay, our estimated slope is still behaving normally if our sample is large enough*".

The CLT gives us the theoretical foundation to measure uncertainty, build intervals, and test if our model is meaningful, even when we only have one dataset.

### 4.4.2 t-Test for Individual Coefficients

**t-Test for Individual Coefficients** is a **hypothesis test** to check whether an individual variable in our regression is **statistically significant**, i.e., whether it **truly affects the response variable**  $Y$ .

#### ✔ Why do we do this?

We run a regression and get:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

But we want to know: “*is this slope  $\hat{\beta}_1$  **actually meaningful**, or did we just get lucky with the data?*”. That’s exactly what the **t-test** answers.

#### ✂ How does it work?

The core logic behind the t-test is to determine **whether there is sufficient evidence to conclude that  $\beta_1 \neq 0$**  in the population. In other words, we need to establish whether the estimated slope is significantly different from zero.

##### 1. Set up the hypotheses.

- Null Hypothesis  $H_0$ :  $\beta_1 = 0$  (no effect)
- Alternative Hypothesis  $H_1$ :  $\beta_1 \neq 0$  (some effect)

##### 2. Compute the t-Statistic.

This tells us how many standard errors away is our slope from 0:

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)} \quad (80)$$

It tells us **how far** our estimated slope  $\hat{\beta}_1$  is from the null value (usually 0), in units of standard error.

##### 3. Compare to t-Distribution.

Use a t-distribution with  $n - 2$  degrees of freedom. The **t-distribution** looks like a **normal distribution**, but it has fatter tails. We use the **t-distribution instead of the normal distribution** because we are **estimating the standard deviation from the data** rather than using the true population standard deviation.

Finally, we compute the p-value. The **p-value** is the probability, **under the null hypothesis**  $H_0$ , of observing a t-statistic as extreme or more extreme than the one we got.

$$p = P(|T| > |t|) \quad \text{where } T \sim t_{n-2} \quad (81)$$

This is calculated using the **cumulative distribution function (CDF)** of the t-distribution.

and compute the p-value.

- If p-value  $< 0.05$ , reject  $H_0$
- If p-value  $> 0.05$ , not significant

### 4.4.3 Global F-Test

While the t-test checks if an individual coefficient (e.g.,  $\beta_1$ ) is significant, the F-test asks: “*does the **entire regression model** explain the data **better than just using the mean**?*”. In other words, do **any** of the predictors explain  $Y$ ? Or are all the  $\beta_j = 0$ , meaning the model is useless?

#### ✔ Hypotheses

- **Null Hypothesis  $H_0$ :**

$$\beta_1 = \beta_2 = \dots = \beta_p = 0$$

No predictors matter, model has no explanatory power.

- **Alternative Hypothesis  $H_1$ :**

$$\text{At least one } \beta_j \neq 0$$

At least one predictor helps explain  $Y$ .

#### ☰ How is the F-Statistic computed?

$$F = \frac{\frac{(\text{TSS} - \text{RSS})}{p}}{\frac{\text{RSS}}{(n - p - 1)}} = \frac{\text{Explained variance per predictor}}{\text{Unexplained variance per residual dof}} \quad (82)$$

Where:

- $p$  is the number of predictors
- $n$  is the number of observations
- TSS is the total sum of squares
- RSS is the residual sum of squares

We compare this F-statistic to an F-distribution with  $p$  and  $n - p - 1$  degrees of freedom.

#### ✔ What does it tell us?

- **Large F, small p-value:** the model is **statistically significant**.
- **Small F, large p-value:** the model doesn't explain the data better than a flat mean.

#### ❓ Why use the Global F-Test?

Because it answers: “*is this model **useful at all**, or should we just predict  $Y$  using its average?*”. This is especially useful when we have **multiple predictors**, or we want to test the **entire model at once**.

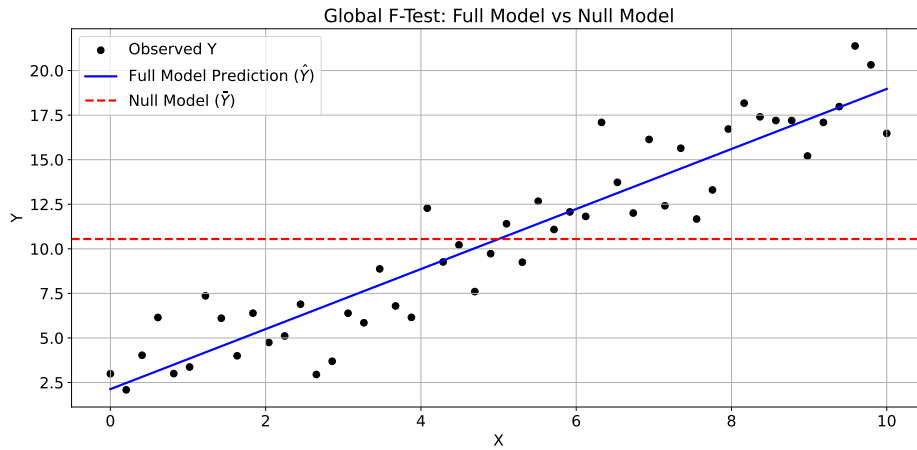


Figure 21: A visual explanation of the Global F-Test.

- Black points: the actual observed  $Y_i$  values
- Blue line: predictions from the **full regression model** ( $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ )
- Red dashed line: prediction from the **null model**, simply the mean of  $Y$ , i.e.,  $\hat{Y}_i = \bar{Y}$

The **null model** (red line) says “Forget  $X$ , just guess everyone’s output is the average”. The **full model** (blue line) says “Use  $X$  to predict  $Y$ ”. The Global F-Test checks “does the blue line explain significantly more than the red one?”.

$$F = \frac{\text{Improvement (Red} \rightarrow \text{Blue)}}{\text{Leftover Error}}$$

If:

- Blue is **much closer** to the points than red  $\Rightarrow$  large  $F \Rightarrow$  **significant model**.
- Blue is **barely better** than red  $\Rightarrow$  small  $F \Rightarrow$  **not useful**.

#### 4.4.4 Summary

##### ✓ Goal

After estimating the regression line:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

We want to answer:

- Is this model **trustworthy**?
- Are the coefficients **significantly different from 0**?
- What's the **uncertainty** in our predictions?

To do that, we use **Statistical Inference**.

##### ≡ 1. Sampling Distributions & the Central Limit Theorem (CLT)

- Our coefficients  $\hat{\beta}_0, \hat{\beta}_1$  come from a **sample**
- If we collected different data, we'd get different values
- The CLT tells us:

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \text{Var}(\hat{\beta}_j)) \quad (\text{approximately})$$

This normality enables confidence intervals and hypothesis testing.

##### ≡ 2. Standard Errors and Confidence Intervals

- The **standard error (SE)** measures how much  $\hat{\beta}_j$  would vary across samples
- A **95% confidence interval** for  $\beta_1$  is:

$$\hat{\beta}_1 \pm t_{n-2} \cdot \text{SE}(\hat{\beta}_1)$$

- Interpretation: “*we’re 95% confident the true slope lies in this range.*”

##### ≡ 3. t-Test for Individual Coefficients

- Tests whether a specific  $\beta_j$  is **statistically significantly different from 0**
- Hypotheses:

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0$$

- Test statistic:

$$t = \frac{\hat{\beta}_j - 0}{\text{SE}(\hat{\beta}_j)}$$

- Compare this to a **t-distribution** to compute the **p-value**
- ✔ If the p-value is small  $\Rightarrow$  **predictor matters**
- ✖ If the p-value is large  $\Rightarrow$  **no strong evidence**

#### 4. Global F-Test (Model-Wide)

- Tests whether **any** of the predictors have a significant effect
- Hypotheses:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0 \quad (\text{model does nothing})$$

- F-statistic:

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n - p - 1)}$$

- Compare to an **F-distribution** with  $(p, n - p - 1)$  degrees of freedom
- ✔ Small p-value  $\Rightarrow$  the model is **useful**
- ✖ Large p-value  $\Rightarrow$  the model might not be better than guessing the mean

Tool	Tests	Interpretation
CLT	All coefficients	Allows inference via normal approximation
SE + CI	Estimate accuracy	How much estimates vary
t-Test	Each coefficient	Is this predictor significant?
F-Test	Entire model	Is the model useful at all?

## 4.5 Uncertainty Intervals

After fitting a regression model, we often want to **predict** the response  $Y$  at a new value of  $X$ . But point predictions like:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

*Only* give a **single number**. That's not enough. In real life, we care about:

- **How uncertain is this prediction?**
- “*What range of outcomes should we expect?*”

✔ **Goal.** To build **intervals** around predictions that account for uncertainty. We study two types:

- **Confidence Interval (CI) for the Mean Response** (page 121). It calculates the range of the **average**  $Y$  value when the experiment is repeated many times at the same  $X$  value. In other words, it **captures uncertainty in the mean prediction**.

Used when:

- We're interested in the **mean trend**.
- We're estimating the regression line itself.

- **Prediction Interval (PI) for a New Observation** (page 123). It calculates the range of a new  $Y$  value at a given  $X$  value, taking randomness into account. In other words, it **captures the uncertainty in individual outcome**.

Used when:

- We want to **predict an individual outcome**.
- We care about real-world variability.



### 4.5.1 Confidence Interval (CI) for the Mean Response

Suppose we have fit a regression model:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Now we pick a specific  $X = x_0$ , and ask: “*what’s the range of values for the mean of  $Y$  at this  $x_0$ ?*”. This is the **Confidence Interval (CI) for the Mean Response** and the formula is:

$$\hat{Y}_0 \pm t_{n-2} \cdot \text{SE}(\hat{Y}_0) \quad (83)$$

Where:

- $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$  is the predicted mean.
- $\text{SE}(\hat{Y}_0)$  is the standard error of that prediction.
- $t_{n-2}$  comes from a t-distribution with  $n - 2$  degrees of freedom.

#### ✔ Why do we need it?

Even if we know  $x_0$ , we don’t know the true  $\beta_0, \beta_1$ , so our prediction  $\hat{Y}_0$  is uncertain. This CI tells us: “*if we repeated this experiment many times and fit a new line each time, what range would the average  $Y$  fall in at  $x_0$ ?*”.

#### ❓ How to compute the Standard Error

$$\text{SE}(\hat{Y}_0) = \sqrt{\hat{\sigma}^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right)} \quad (84)$$

Where:

- $\hat{\sigma}^2$  is the MSE (mean squared error).
- The first term  $\frac{1}{n}$  accounts for uncertainty in the intercept.
- The second term captures how far  $x_0$  is from the center of the data.

Note: the CI is **narrowest at  $\bar{X}$**  and gets wider as  $x_0$  moves away.

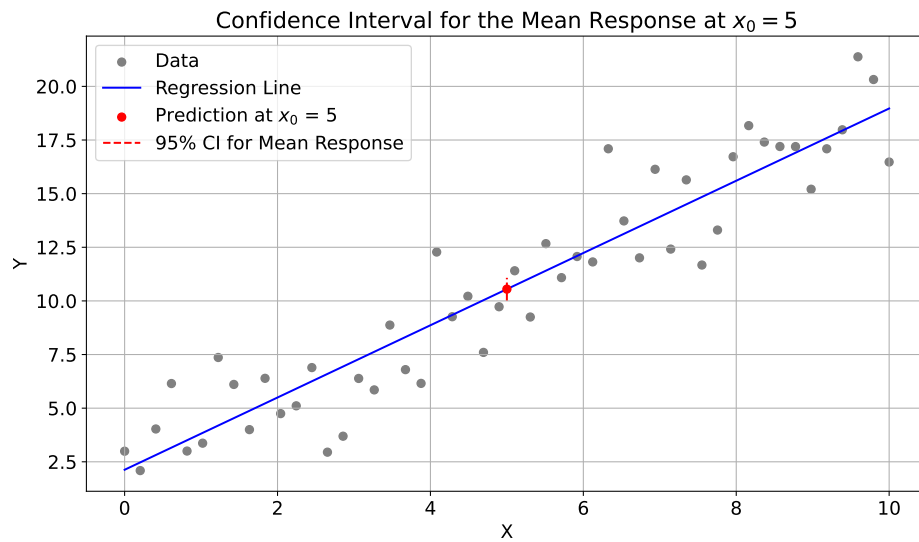


Figure 22: This plot shows a visual of the Confidence Interval (CI) for the Mean Response at  $x_0 = 5$ . The red dashed vertical line indicates the 95% confidence interval for the **average**  $Y$  at  $x_0 = 5$ , even if it is small. So, if we repeated this regression many times, we're 95% confident the average value of  $Y$  when  $X = 5$  would fall in that vertical interval.

### 4.5.2 Prediction Interval (PI) for a New Observation

We've fit a regression model:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Now we pick a specific value  $X = x_0$  and ask: “if we get a **new data point** with  $X = x_0$ , what range of values could the actual  $Y$  fall into?”. That's the **Prediction Interval (PI) for a New Observation** and the formula is:

$$\hat{Y}_0 \pm t_{n-2} \cdot \text{SE}_{\text{pred}}(x_0) \quad (85)$$

Where:

$$\text{SE}_{\text{pred}}(x_0) = \sqrt{\hat{\sigma}^2 \left( 1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right)} \quad (86)$$

Compared to the Confidence Interval (page 121), this has an **extra +1** inside the square root, which accounts for the **random error in a single new observation**.

#### ✔ Why do we need it?

Because real-world observations aren't always on the line, they **scatter around** due to noise  $\varepsilon$ . So we want a range that captures **both**:

- The **uncertainty in the estimated mean** at  $x_0$ , and
- The **random variation in a new individual outcome**

#### 🔍 Difference vs. Confidence Interval

Interval Type	Captures	Width
<b>CI</b>	Uncertainty in <i>average</i> $Y$	Narrower
<b>PI</b>	Uncertainty in <i>single</i> $Y$	<b>Wider</b> , includes noise!

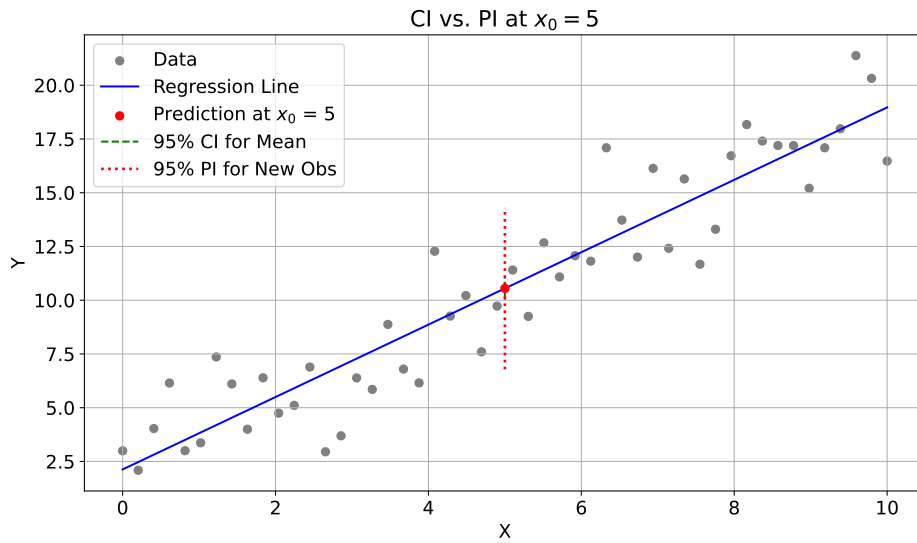


Figure 23: This plot illustrates the visual comparison of the two types of uncertainty interval at  $x_0 = 5$ . The red point is the predicted value  $\hat{Y}_0$  at  $x_0 = 5$ .

Unlike Figure 22 on page 122, we also draw the prediction interval here.

- Green dashed line is the Confidence Interval (CI), for the mean of  $Y$  at  $X = 5$ . It is **narrow** because it only accounts for uncertainty in estimating the mean. It seems invisible because it is smaller than PI.

The CI tells us: “*where would the **average**  $Y$  be at  $X = 5$ ?*”.

- Red dotted line is the Prediction Interval (PI), for a **new individual** with  $X = 5$ . It is **wider** because it also includes randomness (noise) in future observations.

The PI tells us: “*where would a **single new**  $Y$  likely fall at  $X = 5$ ?*”.

## 4.6 Multiple Linear Regression (MLR)

### 🔍 Why Move Beyond Simple Linear Regression?

Up until now, we've modeled a response variable  $Y$  as a function of **just one predictor**  $X$ :

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

But in real-world problems, the outcome usually depends on **more than one factor**.

For example, let's say we're predicting a house price.

- One predictor might be: square footage
- But what about: number of bedrooms? location? year built?

That's where **Multiple Linear Regression** comes in.

**Multiple Linear Regression (MLR)** generalizes the model to include **multiple independent variables**:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p \quad (87)$$

Where:

- $Y$ : outcome.
- $X_1, X_2, \dots, X_p$ : predictors.
- $\beta_0$ : intercept.
- $\beta_j$ : effect of predictor  $j$ .

### 🔍 Why use MLR?

Use Case	MLR Benefit
We want better predictions	More predictors, more information
We want to isolate effects	Estimate each variable's impact, <b>controlling for others</b>
We want realism	Real systems depend on <b>many factors</b> , not just one

### 4.6.1 Model Expansion

So far, we've worked with:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

This is a **simple linear regression (SLR)** with only **one predictor** (page 101).

We now expand this to:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \quad (88)$$

This is **Multiple Linear Regression (MLR)** with  $p$  predictors.

#### ✔ What's new here?

We now have **multiple input variables**, not just one  $X$ , but many contributions. Each gets its own coefficient  $\beta_j$ , representing its **independent contribution** to predicting  $Y$ .

#### ☰ Intuition

Each  $\beta_j$  answers: “*if we increase  $X_j$  by 1 unit, how much does  $Y$  change, holding all the other  $X$ 's constant?*”. The last part is critical because, in MLR, we **isolate the effect of each variable while controlling for the effects of the others**.

For example, let's say we model student grades:

$$\hat{Y} = \beta_0 + \beta_1 \cdot \text{StudyTime} + \beta_2 \cdot \text{Sleep} + \beta_3 \cdot \text{Caffeine}$$

Then:

- $\beta_1$ : how grades change **per hour of study**, keeping sleep and caffeine constant.
- $\beta_3$ : how grades change **per mg of caffeine**, keeping study and sleep constant.

#### ⚠ Beware: Correlation Between Predictors

If predictors (e.g., Sleep and StudyTime) are **highly correlated**, it becomes harder to isolate the effect of one from the other. This is known as **multi-collinearity**, and it can:

- ✗ Make coefficients unstable.
- ✗ Inflate standard errors.
- ✗ Confuse interpretation.

More formally, **Multicollinearity** occurs when two or more predictor variables in a multiple regression model are **highly correlated** with each other. This means they carry **overlapping information** about the outcome  $Y$ , making it hard for the model to distinguish their individual effects.

**❓ Why Multicollinearity is a Problem?** If  $X_1$  and  $X_2$  are strongly correlated, the model can't **tell who's responsible** for changes in  $Y$ . The estimated coefficients  $\hat{\beta}_1, \hat{\beta}_2$  become **unstable**:

- They can change drastically with small changes in data.
- They may have **unexpected signs** (positive/negative).
- Their **standard errors** become large, then **t-tests become unreliable**.

For example, imagine this:  $X_1 = \text{StudyTime}$ ,  $X_2 = \text{TimeAtLibrary}$ . If students who study more also spend more time at the library, then  $X_1 \approx X_2$ . So the regression model says: *"wait... both variables kind of say the same thing. Who gets the credit?"*. The answer is that the model doesn't know, and the estimates become inaccurate.

### ❓ How to detect Multicollinearity?

Method	Description
Correlation matrix	Check if predictors are highly correlated
Variance Inflation Factor (VIF)	Measures how much a variable is explained by other predictors
Unstable coefficients	Coefficients flip sign or have huge standard errors

### ✅ When is it *not* a problem?

Multicollinearity **does not hurt prediction accuracy** (if all predictors are used). But it does hurt Interpretability, and Statistical Inference (t-test and p-values).

### 4.6.2 Matrix Formulation

#### 🔍 Why Use Matrix Form?

Matrix notation gives us:

- **Compact representation** of the model.
- A way to handle **any number of predictors** with one formula.
- Clean derivation of **OLS estimators**.

#### 📊 Model in Matrix Form

We rewrite:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

As:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (89)$$

Where:

- $\mathbf{y}$ :  $n \times 1$  column vector of observed responses.
- $\mathbf{X}$ :  $n \times (p+1)$  **design matrix**, including a column of ones for the intercept.
- $\boldsymbol{\beta}$ :  $(p+1) \times 1$  vector of coefficients.
- $\boldsymbol{\varepsilon}$ :  $n \times 1$  vector of errors (residuals)

For example, if we have two predictors (plus intercept):

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix}$$

#### ✅ OLS Solution Using Matrix Algebra

To estimate the coefficients that minimize residual sum of squares:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (90)$$

This generalizes the simple regression formula to **any number of predictors**. It is the **Ordinary Least Squares estimators in matrix form**.

- $\mathbf{X}$ : design matrix (rows are the data points, columns are the predictors +1).
- $\mathbf{X}^T$ : transpose of  $\mathbf{X}$ .
- $\mathbf{X}^T \mathbf{X}$ : matrix of sums of products, square and invertible if predictors are independent.



- $(\mathbf{X}^T \mathbf{X})^{-1}$ : inverse matrix, handles “division” in multidimensional space.
- $\mathbf{X}^T \mathbf{y}$ : vector of dot products between predictors and output.
- $\hat{\beta}$ : vector of estimated coefficients  $[\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p]$ .

### Predictions and Residuals

- **Fitted values**

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} \quad (91)$$

Where:

- $\hat{\mathbf{y}}$  is the vector of **fitted (predicted) values** for all our data points.
- $\mathbf{X}$  is the **design matrix** with all our predictors and a column of ones (for the intercept).
- $\hat{\beta}$  is the vector of **estimated coefficients**.

This gives us all the predicted outcomes using the regression line we just fit.

- **Residuals**

$$\hat{\epsilon} = \mathbf{y} - \hat{\mathbf{y}} \quad (92)$$

Where:

- $\hat{\epsilon}$  is the vector of **residuals**, the errors the model made.
- The residuals is the difference between:
  - \*  $\mathbf{y}$  is the **true observed values**.
  - \*  $\hat{\mathbf{y}}$  is the **predicted values** from your model. It is used for evaluating how well the model fits.

In other words: “*how far off was our prediction from the actual result?*”.

## 5 Model Selection and Regularization

### 5.1 Motivation & Problem Setup

✔ **The Goal.** We learned *how* to evaluate a model and which key factors to consider. Now, we want *to build* a model that is as simple as possible, but still accurate when predicting new data. In other words, to **find a linear model that is accurate, stable, and easy to interpret**. More precisely, we want to:

- ✔ **Make Good Predictions.** Not just on training data, but on new unseen data. This is why we care about generalization and not just high  $R^2$  (coefficient of determination, page 110).
- ✔ **Avoid Overfitting.** Including too many predictors, our model fits noise. We want to choose only the useful variables.
- ✔ **Reduce Variance of Coefficients.** If variables are correlated, coefficients jump around a lot. We want stable estimates even if the data changes slightly.
- ✔ **Interpret the Model Easily.** Simpler models are easier to explain. Sparse models (with some  $\beta_j = 0$ ) highlight the important variables.

✂ **Setup: Linear Regression.** We have data:

- $Y$ : outcome (like house price, disease risk, exam score)
- $X_1, X_2, \dots, X_p$ : features (like square footage, age, sleep hours)

We want to model:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon$$

Using Ordinary Least Square (OLS, page 102), we estimate:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

It gives us the “best” coefficients to fit the data (minimize squared error).

⚠ **Problems.** However, this approach has some problems:

- **Too Many Variables?** The more variables we add, the higher  $R^2$  goes. But adding variables blindly causes **overfitting**, and the model fits noise, not the true signal.
- **Multicollinearity** (page 127). If two variables  $X_1 \approx X_2$ , the OLS estimate becomes unstable. Interpretation of  $\beta_j$  becomes meaningless. The **variance** of the coefficients explodes:

$$\text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

If  $X^T X$  is close to singular (not invertible),  $\hat{\beta}$  becomes extremely variable!

- **High Variance, Low Bias.** OLS has **low bias** (accurate on average), but if predictors are correlated or too many, **variance** is high, then **poor generalization**.

## 5.2 Collinearity

### 5.2.1 Definition

**Collinearity** means that one predictor variable is **highly correlated with one or more** of the other predictors.

- The **Perfect Collinearity** is when one **variable is an exact linear combination of others**:

$$X_j = \alpha_0 + \sum_{k \neq j} \alpha_k X_k \quad (93)$$

- The **High Collinearity** (but not perfect), is when variables are almost linearly related (like  $X_1 \approx X_2$ ).

### ⚠ Why is this a problem?

1. **OLS becomes unstable**: The matrix  $X^T X$  becomes **ill-conditioned**<sup>5</sup> or **non-invertible**. The variance of  $\hat{\beta}_j$  becomes **very large**.
2. **Interpretation breaks down**: It's unclear which variable is responsible for the effect on  $Y$ . We might see strange signs or inflated coefficients.
3. **Poor prediction**: Even if our training error is low, generalization on new data suffers.

### ❓ Impact on Variance of $\hat{\beta}$

Collinearity **doesn't hurt prediction** much (model can still fit  $Y$  well). But it makes our  $\hat{\beta}_j$  estimates:

- **Unstable**: they swing wildly if we change data a little.
- **Hard to interpret**: the model is unsure whether to assign effect to  $X_1$  or  $X_2$ .

It happens because the **variance of  $\hat{\beta}$  becomes large** (*effect*) **when predictors are collinear** (*cause*). This is **not just a side effect**, it is the mechanism by which collinearity causes damage.

The variance of  $\hat{\beta}$  is:

$$\hat{\beta} \sim \mathcal{N}_p\left(\beta, \sigma^2 (X^T X)^{-1}\right)$$

<sup>5</sup>An **Ill-Conditioned matrix** is a matrix that has a high condition number, indicating that small changes in the input can cause large changes in the output. This makes the matrix sensitive to errors and can lead to significant inaccuracies in numerical computations.

This formula means that the **distribution** of our OLS coefficients  $\hat{\beta}$  is a **multivariate normal** with:

- Mean = true coefficients  $\beta$
- Covariance (variance) =  $\sigma^2(X^T X)^{-1}$

🔗 **Okay, so what's the point of this variance?** The interesting part is  $(X^T X)^{-1}$  because it is the inverse of a **matrix that summarizes how our predictors relate to each other**:

- If predictors are **independent**, this matrix is *nice*, so inverse is stable and the **variance** is **small**.
- If predictors are **collinear**, this matrix is *close to singular*, so inverse is huge and the **variance explodes**.

#### Example 1: Analogy 1D

In 1D:

$$\text{Var}(\hat{\beta}) = \frac{\sigma^2}{\sum x_i^2}$$

- If all  $x_i$  values are **spread out**, then the **denominator** is **large**, and the **variance** is **small**.
- If all  $x_i$  values are **clustered** (close together), then the **denominator** is **small**, and the **variance** is **large**.

In multivariable OLS that denominator becomes  $X^T X$  and the variance becomes:

$$\text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

🔗 **Why Variance “explodes” with Collinearity?** If  $X_1 \approx X_2$ , then the matrix  $X^T X$  becomes almost **non-invertible**. Its inverse  $(X^T X)^{-1}$  contains **huge numbers**. That means that even a small change in the data can cause a large change in  $\hat{\beta}$ . This is why our model is “*wiggly*”, it’s jumping around because it’s uncertain which variable to credit.

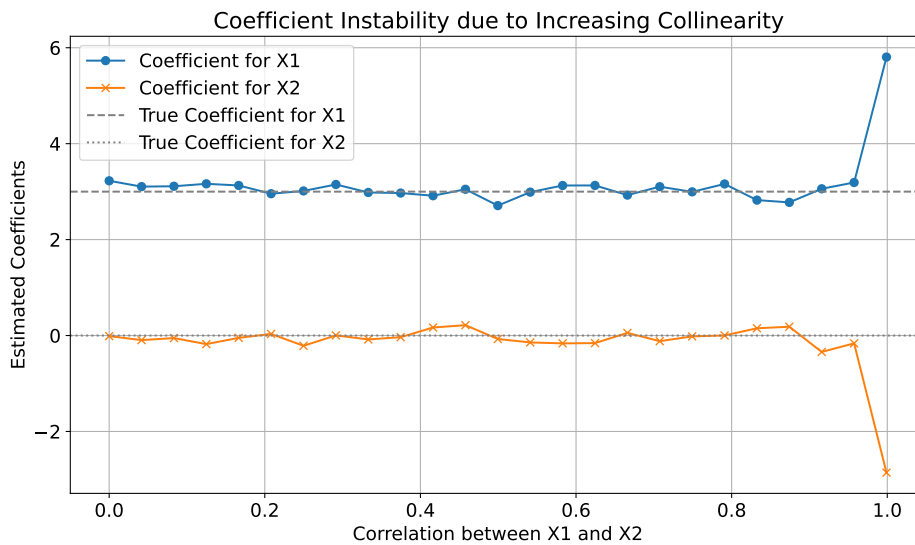


Figure 24: Coefficient Instability due to Increasing Collinearity.

- **X-Axis:** shows how strongly the two predictor variables  $X_1$  and  $X_2$  are related. We vary correlation from 0 (completely independent) to nearly 1 (almost identical).
- **Y-Axis:** these are the OLS estimates  $\hat{\beta}_1$  and  $\hat{\beta}_2$ . The **true model** is:

$$Y = 3X_1 + \text{noise}$$

So the “truth” is  $\beta_1 = 3$ ,  $\beta_2 = 0$ .

- **Blue line ( $X_1$ ):** Estimated  $\hat{\beta}_1$  (should be near 3).
- **Orange line ( $X_2$ ):** Estimated  $\hat{\beta}_2$  (should be near 0).
- **Gray dashed line (3):** True value for  $\beta_1$ .
- **Gray dotted line (0):** True value for  $\beta_2$ .

### 🔍 What Happens as Correlation Increases?

- ✅ Correlation near 0: **Low**. OLS correctly identifies  $X_1$  as important, and  $\hat{\beta}_1 \approx 3$ ,  $\hat{\beta}_2 \approx 0$ .
- 🔧 Correlation: **Medium**. The model gets slightly confused: some effect is “shared” between both.
- ⚠️ Correlation near 1: **High**. The model becomes unstable:  $\hat{\beta}_1$  and  $\hat{\beta}_2$  swing wildly, but still together approximate 3.

This is the essence of **multicollinearity**. The model can’t distinguish the true source of variation. It becomes **numerically unstable** even though prediction may still seem “okay”.

### 5.2.2 Variance Inflation Factor (VIF)

**Variance Inflation Factor (VIF)** measures **how much the variance** of a regression coefficient is **inflated due to multicollinearity**. It quantifies the severity of multicollinearity in an OLS regression.

#### 🔍 How is VIF defined?

For a coefficient  $\hat{\beta}_j$ , the VIF is:

$$\text{VIF}_j = \frac{1}{1 - R_j^2} \quad (94)$$

Where  $R_j^2$  is the **coefficient of determination** when **regressing  $X_j$  on all other predictors**.

Usually in regression, we do this:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

But for VIF, we switch things up. **Instead of using  $X_j$  to predict  $Y$** , we do:

$$X_j = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \cdots + \gamma_{j-1} X_{j-1} + \gamma_{j+1} X_{j+1} + \cdots + \gamma_p X_p + \eta$$

So we make  $X_j$  the dependent variable, and we use all other  $X_k$  (except  $X_j$ ) as independent variables. Then we compute the  $R^2$  of this regression.

- If  $R_j^2 = 0$ :  $X_j$  is independent, then no multicollinearity and  $\text{VIF}_j = 1$ .
- If  $R_j^2 = 0.9$ : 90% of  $X_j$ 's variance is explained by others, then  $\text{VIF}_j = \frac{1}{1 - 0.9} = 10$ , danger!

VIF value	Interpretation
$\text{VIF} = 1$	✅ No correlation with other variables.
$1 < \text{VIF} < 5$	🟡 Moderate correlation, usually okay.
$\text{VIF} > 5$	⚠️ High correlation, potential multicollinearity.
$\text{VIF} > 10$	💀 Very high correlation, serious problem.

Table 10: Interpretation of the VIF.

### ✂ Why it works?

The variance of the OLS estimate  $\hat{\beta}_j$  is:

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{(1 - R_j^2) \cdot S_{jj}}$$

Where:

- $\sigma^2$  is the noise variance.
- $S_{jj}$  is the variance of  $X_j$  itself (scaled sum of squares).
- $R_j^2$  is how well  $X_j$  is explained by the other variables.

If  $R_j^2$  is high ( $R_j^2 \rightarrow 1$ ), then  $1 - R_j^2$  is small (denominator  $\rightarrow 0$ ). This is the **mechanism** behind multicollinearity, the more predictable  $X_j$  is from the others, the more **uncertain** our model is about how much  $X_j$  actually contributes to predicting  $Y$ .

So Variance Inflation Factor (VIF) is a **multiplier**, it tells us **how much larger the variance of  $\hat{\beta}_j$  is due to collinearity**. For example, if  $R_j^2 = 0.95$ , then  $\text{VIF}_j = \frac{1}{0.05} = 20$ , and the variance of  $\hat{\beta}_j$  is  $20\times$  worse than it would be with independent variables.

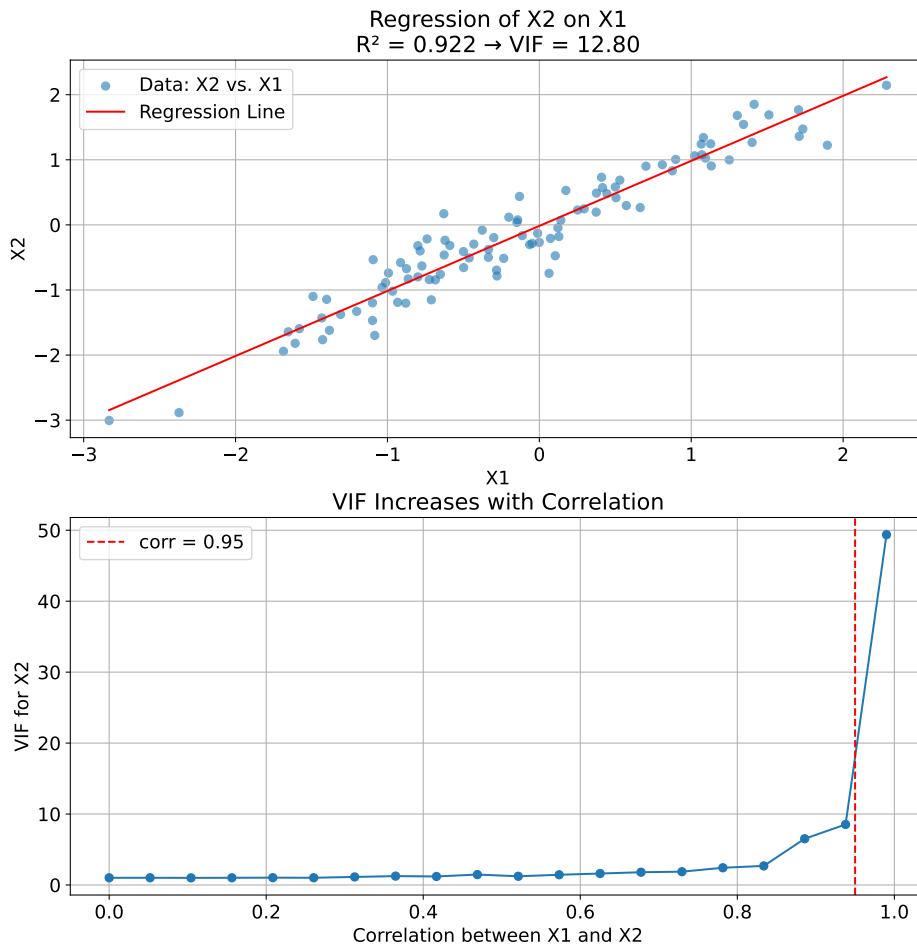


Figure 25: Top Plot: Regression Line. The **line fits the data well visually**, it looks like a strong relationship. Indeed, the  $R^2 = 0.903$ : 90% of the variability in  $X_2$  is explained by  $X_1$ . But this is **not a good thing in this context**.

Bottom Plot: VIF Curve. The correlation of 0.95 (marked in red) leads to  $\text{VIF} \approx 15$ . Even though the fit looks great, the **usefulness of  $X_2$**  in a regression model with  $X_1$  is highly questionable.

A regression line might **look clean and convincing**, but if the predictor is **highly collinear**, our model is **mathematically unstable**. The **VIF reveals what the plot hides**: danger beneath the surface.



## References

- [1] Herman Chernoff. The use of faces to represent points in k-dimensional space graphically. *Journal of the American statistical Association*, 68(342):361–368, 1973.
- [2] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York, 2009.
- [3] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in Python*. Springer Texts in Statistics. Springer New York, 2013.
- [4] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Applied Multivariate Statistical Analysis. Pearson Prentice Hall, 2007.
- [5] Beraha Mario. Applied statistics. Slides from the HPC-E master’s degree course on Politecnico di Milano, 2024-2025.

## Index

### Symbols

1D Gaussian	54
2D Gaussian	55
2D Scatterplot	8

### A

Actual Error Rate (AER)	97
Agglomerative Clustering (Bottom-Up)	40
Apparent Error Rate (APER)	97
Average Linkage Method	41

### B

Bayes Classifier	92
Bayes Error Rate	95
Between-cluster Sum of Squares (BSS)	38
Bivariate Normal Distribution	55

### C

Central Limit Theorem	114
Chebyshev Distance	33
Clustering	31
Coefficient of Determination	110
Collinearity	131
Complete Linkage Method	41
Confidence Interval (CI) for the Mean Response	121
Correlation	6
Correlation matrix	7
Correlation-Based Distance	33
Cosine Similarity (Angle-based)	33
Covariance	5

### D

Dimensionality Reduction	13
Discriminant Analysis	73, 74
Discriminant Function - LDA	78
Discriminant Function - QDA	88
Divisive Clustering (Top-Down)	40

### E

Elbow Method	38, 44
Entropy Metrics	36
Euclidean Distance	33
Expectation-Maximization (EM) Algorithm	67
External metrics	35

### F

F-Measure Metrics	36
Fundamental Identity	109

**G**

Gap Statistics	46
Gaussian Mixture Model (GMM)	54
Generative Model	75
Ground Truth Labels	35

**H**

Hard Clustering	58
Hat Matrix	108
Hierarchical Clustering	40
High Collinearity	131

**I**

Ill-Conditioned matrix	131
Internal Metrics	37

**K**

K-Means	42, 43
K-Means centroid	42, 43
K-Means Overfitting	44
K-Means Underfitting	44
K-Medoids	53

**L**

L1 norm	33
Latent Variable	65
LDA Decision Rule	78
Linear Discriminant Analysis (LDA)	76
Loadings	15, 19
Loadings Matrix ( $V$ )	15

**M**

Mahalanobis Distance	34
Manhattan Distance	33
Minkowski Distance	34
Multicollinearity	127
Multiple Linear Regression (MLR)	125, 126
Multivariate (MV) analysis	5
Multivariate Gaussian Distribution	57

**N**

Normal Distribution	54
---------------------	----

**O**

Oracle Classifier	97
Ordinary Least Squares Estimation (OSL Estimation)	102
Ordinary Least Squares estimators	105
Ordinary Least Squares estimators in matrix form	128
Overfitting	83

**P**

p-value	115
---------	-----

Perfect Collinearity	131
Posterior Probability	77
Precision Metrics	36
Prediction Interval (PI) for a New Observation	123
Principal Component Analysis (PCA)	13, 14
Projection Matrix	108
Proportion of Variance Explained (PVE)	26
Purity Metrics	36

## Q

Quadratic Discriminant Analysis (QDA)	83
Quadratic Form	23, 88

## R

Rayleigh Quotient Theorem	23
Recall Metrics	36
Regression Line	105
Regression Sum of Squares (SSR)	109
Residual Sum of Squares (RSS)	102, 109
Responsibility	65
Rotated plots	9

## S

Sample Covariance	5
Sample Means	7
Sampling Distribution	113
Scores	15, 20
Scores Matrix ( $U$ )	16
Silhouette Coefficient	38, 45
Silhouette Score	45
Simple Linear Regression (SLR)	101
Simple Linear Regression - Intercept	105
Simple Linear Regression - Slope	105
Single Linkage Method	41
Singular Value Decomposition (SVD)	29
Soft Clustering	58
Star plots	10
Statistical Inference	112
Sum of Squared Errors (SSE)	37
Supervised Learning	31, 71, 73

## T

t-distribution	115
t-Test for Individual Coefficients	115
Total Sum of Squares (TSS)	109
True Error Rate	97

## U

Unsupervised Learning	31
-----------------------	----

**V**

Variability	14
Variance Inflation Factor (VIF)	134
Variance-Covariance matrix	7

**W**

Within-Cluster Sum of Squares (WCSS)	37
Within-Cluster Variability	50