Department of Mathematics
and Computer Science

Advanced Machine Learning

Background

Lukas Galke

Spring 2026

SDU

Slides based on Richard Röttger's material

Keywords

- **Linear algebra**
- Probabilities & Standard Distributions
- Eigendecomposition & PCA

Scalar and Vector

Scalar

- Single number
- Normally: $x \in \mathbb{R}$ or $x \in \mathbb{N}$

Vector

- An array of numbers
  - Arranged in order
  - Each no. identified by an index
- $\boldsymbol{x} = \begin{bmatrix} x_1 \\ \cdots \\ x_d \end{bmatrix}$ and $\boldsymbol{x}^T = [x_1, \ldots, x_d]$, $\boldsymbol{x} \in \mathbb{R}^d$
- We think of vectors as points in space
  - Each element gives coordinate along an axis

Matrix

- 2-D array of numbers
- Each element identified by two indices
- Denoted by bold typeface $\boldsymbol{A}$
- Elements indicated as $A_{m,n}$

$$\boldsymbol{A} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}$$

- $A_{i:}$ is ith row of $\boldsymbol{A}$, $A_{:j}$ is jth column
- $\boldsymbol{A}$ has m rows and n columns, then $A \in \mathbb{R}^{m \times n}$

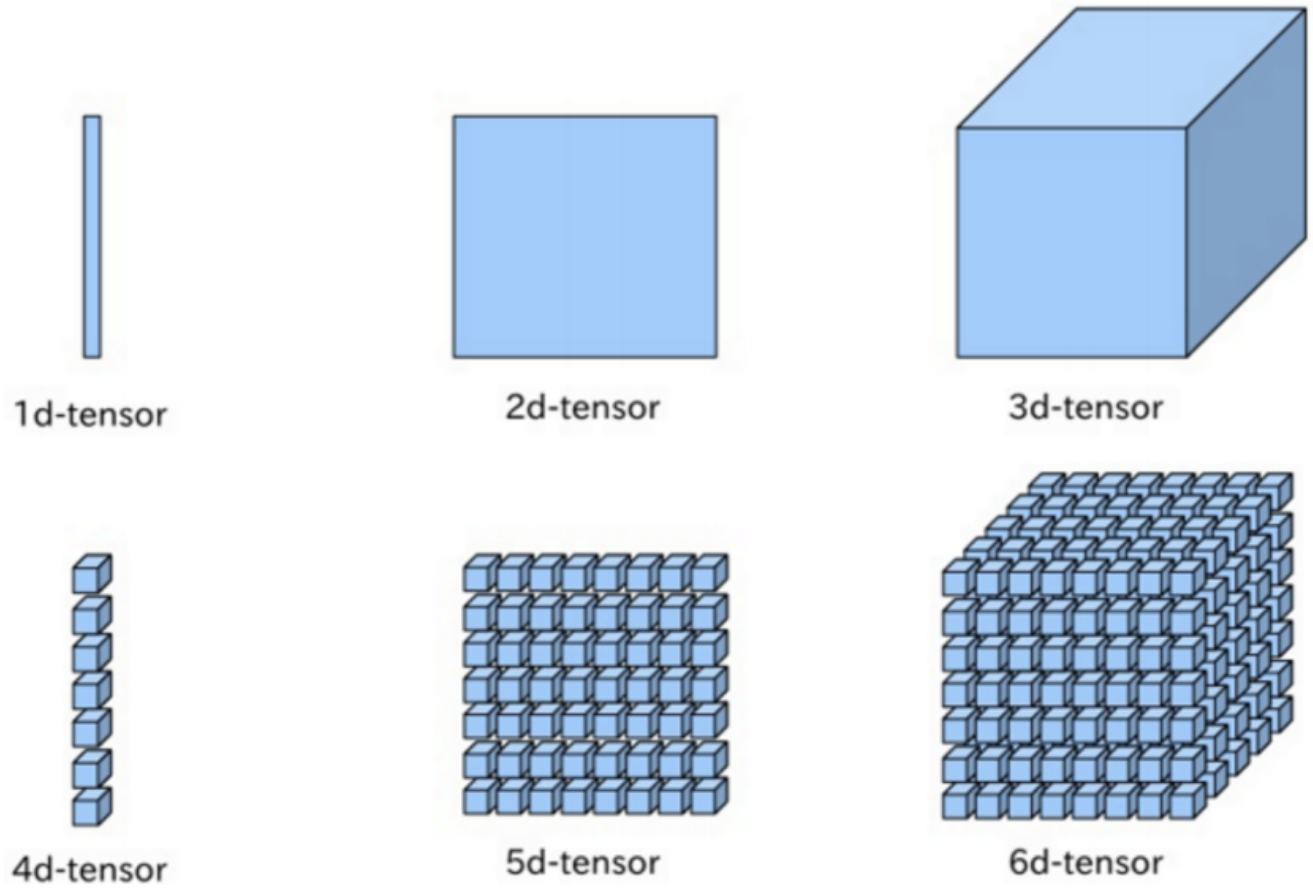
Tensor

- Sometimes need an array with more than two axes
- An array arranged on a regular grid with variable number of axes is referred to as a **Tensor**

- We denote a tensor with bold non-italic typeface: A in comparison to a normal Matrix $\boldsymbol{A}$
- I try to keep the notation consistent, but double-check ☺
- An Element (i, j, k) of tensor denoted by $\mathrm{A}_{i,j,k}$

- Example: RGB image is a 3D Tensor (height x width x color channel)

Shape of Tensors



Terminology:

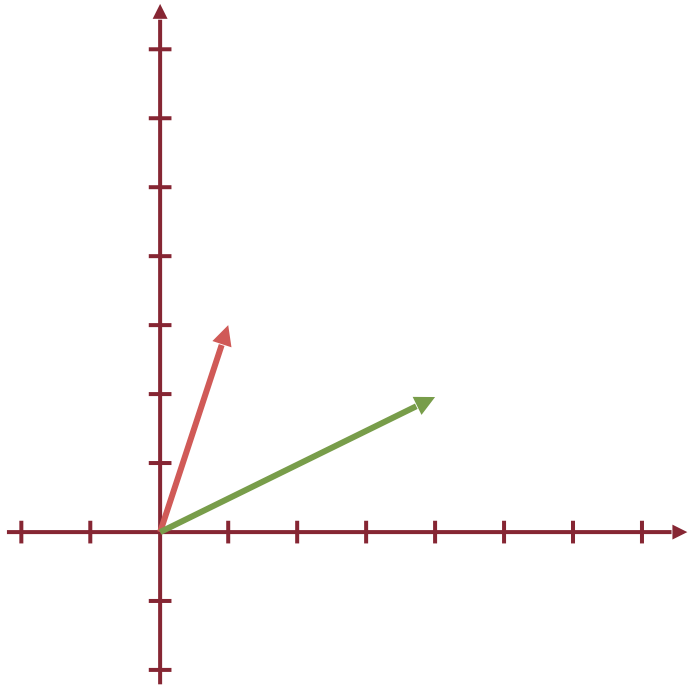
The number of axes is also denoted as **the rank** of a tensor.

Vector Interpretation

➢ Vectors can are often depicted as arrows in Euclidean space.

➢ Let assume two Vectors:

$$\boldsymbol{a} = \begin{pmatrix} 1 \\ 3 \end{pmatrix} \qquad \boldsymbol{b} = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$$



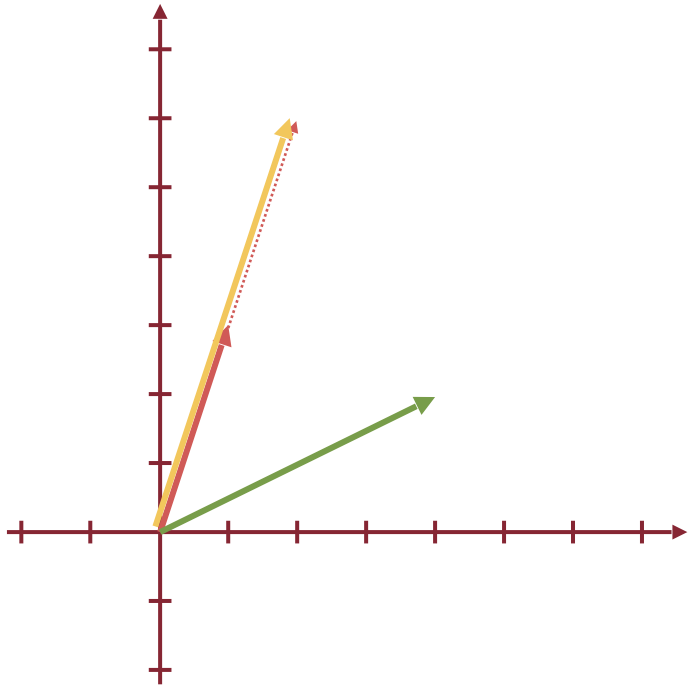
Vector Operations: Multiply with a Scalar

➢ Vectors can be multiplied with a scalar by elementwise multiplying the entries:

$$\lambda \cdot x = \lambda \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} \lambda \cdot x_1 \\ \vdots \\ \lambda \cdot x_d \end{pmatrix}$$

➢ Example:

$$2 \cdot \boldsymbol{a} = 2 \cdot \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 2 \cdot 1 \\ 2 \cdot 3 \end{pmatrix} = \begin{pmatrix} 2 \\ 6 \end{pmatrix}$$



Vector Operations: Addition

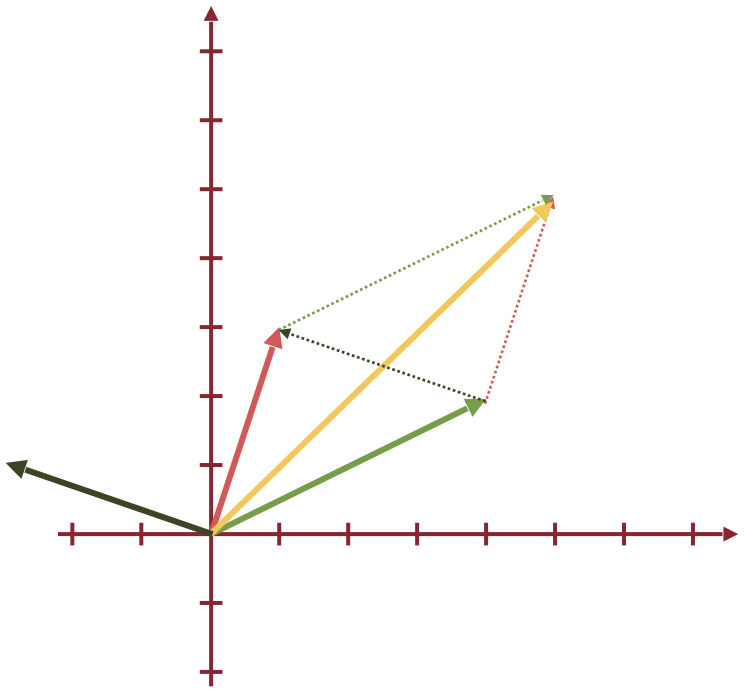
➢ Vectors can be added by element-wise adding the components:

$$x + y = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} + \begin{pmatrix} y_1 \\ \vdots \\ y_d \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_1 + y_1 \end{pmatrix}$$

➢ Examples:

$$\boldsymbol{a} + \boldsymbol{b} = \begin{pmatrix} 1 \\ 3 \end{pmatrix} + \begin{pmatrix} 4 \\ 2 \end{pmatrix} = \begin{pmatrix} 5 \\ 5 \end{pmatrix}$$

$$\boldsymbol{a} - \boldsymbol{b} = \begin{pmatrix} 1 \\ 3 \end{pmatrix} - \begin{pmatrix} 4 \\ 2 \end{pmatrix} = \begin{pmatrix} -3 \\ 1 \end{pmatrix}$$



Vector Operations: The Dot Product

➢ The scalar product of two vectors is the sum of the element-wise multiplication of the entries:

$$\boldsymbol{x} \cdot \boldsymbol{y} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ \vdots \\ y_d \end{pmatrix} = x_1 y_1 + \ldots + x_d y_d = \sum_{i=1}^{d} x_i y_i = \boldsymbol{x}^T \boldsymbol{y}$$

➢ Note: The result is a real number, not a vector

➢ For this reason, the dot product is sometimes called the scalar product (or inner product).

$$\boldsymbol{a} \cdot \boldsymbol{b} = \begin{pmatrix} 1 \\ 3 \end{pmatrix} \cdot \begin{pmatrix} 4 \\ 2 \end{pmatrix} = 1 \cdot 4 + 3 \cdot 2 = 10$$

Properties of the Scalarproduct

➢ The dot product obeys many of the laws that hold for ordinary products of real numbers.

➢ Let $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ be vectors, λ is a scalar.

➢ Then:

1. $\boldsymbol{a} \cdot \boldsymbol{b} = \boldsymbol{b} \cdot \boldsymbol{a}$
2. $\boldsymbol{a} \cdot \left(\boldsymbol{b} + \boldsymbol{c}\right) = \boldsymbol{a} \cdot \boldsymbol{b} + \boldsymbol{a} \cdot \boldsymbol{c}$
3. $(\lambda \boldsymbol{a}) \cdot \boldsymbol{b} = \lambda\left(\boldsymbol{a} \cdot \boldsymbol{b}\right) = \boldsymbol{a} \cdot \left(\lambda \boldsymbol{b}\right)$
4. $0 \cdot a = 0$

Vector Operations: Length of Vector

➢ The length of a vector is normally calculated as the square root of the summed squares of the entries:

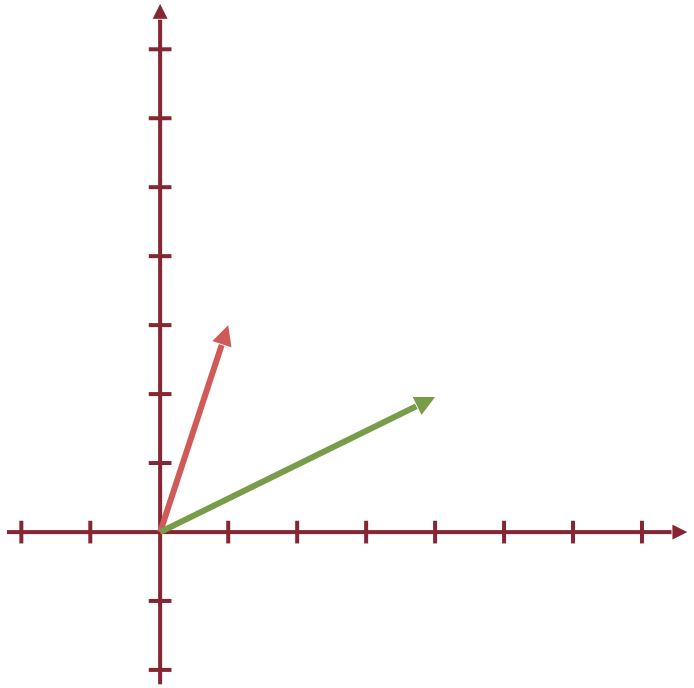
$$|x| = \left| \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \right| = \sqrt{x_1^2 + \ldots + x_d^2} = \left(\sum_{i=1}^{d} x_i^2 \right)^{\frac{1}{2}}$$

➢ Relation to the dot-product:

$$x \cdot x = |x|^2$$

➢ Example:

$$|a| = \left| \begin{pmatrix} 1 \\ 3 \end{pmatrix} \right| = \sqrt{1^2 + 3^2} = \sqrt{10} \approx 3.16$$



L^p Norms for Vectors

- What we have seen is just a member of a family of norms:

$$L^p = \left(\sum_i |x_i|^p \right)^{\frac{1}{p}}$$

- L^2 Norm
  - Most commonly used
  - If nothing is explicitly stated, the Euclidean norm is used
- L^1 Norm
  - Useful when 0 and non-zero have to be distinguished (since L^2 increases slowly near origin, e.g., $0.1^2 = 0.01$)
- L^∞ Norm
  - $\left|\left|\boldsymbol{x}\right|\right|_\infty = \max|x_i|$
  - Called max norm

Angle between Vectors

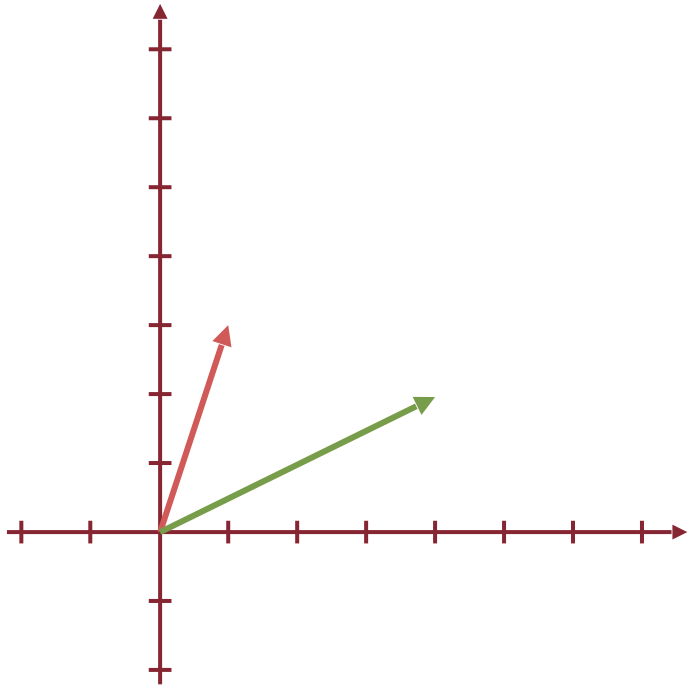
➢ Dot product of two vectors can be written in terms of their L^2 norms and angle θ between them

$$\boldsymbol{x} \cdot \boldsymbol{y} = |\boldsymbol{x}| \left|\boldsymbol{y}\right| \cdot \cos\theta$$

➢ Example:

$$\theta = \cos^{-1}\left(\frac{\boldsymbol{a} \cdot \boldsymbol{b}}{|a| \cdot |b|}\right) =$$

$$= \cos^{-1}\left(\frac{10}{\sqrt{10} \cdot \sqrt{20}}\right) = 45°$$



Names around Matrices

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

- **Order of matrix**: It represent the number of rows and number columns of a matrix. For above matrix, the order is 3×3
- **Square matrix**: If a matrix has same number of row and columns then it is called square matrix.
- Special Shapes:
  - Row Matrix: If a matrix only has one row: a transposed Vector
  - Column Matrix: Similarly, if a matrix only has one column: a Vector
  - 1×1 Matrix: A scalar

Matrix Operations: Addition

➢ Two Matrices can be added by simply adding the entries element-wise:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} + \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} a+e & b+f \\ c+g & d+h \end{pmatrix}$$

➢ Example:

$$\begin{pmatrix} 1 & 2 \\ 5 & 6 \end{pmatrix} + \begin{pmatrix} 3 & 4 \\ 2 & 5 \end{pmatrix} = \begin{pmatrix} 1+3 & 2+4 \\ 5+2 & 6+5 \end{pmatrix} = \begin{pmatrix} 4 & 6 \\ 7 & 11 \end{pmatrix}$$

➢ The matrices necessarily need to be of the same order (or dimensionality)

Matrix Operations: Multiplication with a Scalar

➢ A Matrix can be multiplied by a scalar by multiplying each element of the matrix with the scalar:

$$\lambda \cdot \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \lambda \cdot a & \lambda \cdot b \\ \lambda \cdot c & \lambda \cdot d \end{pmatrix}$$

➢ Example:

$$2 \cdot \begin{pmatrix} 3 & 4 \\ 2 & 5 \end{pmatrix} = \begin{pmatrix} 2 \cdot 3 & 2 \cdot 4 \\ 2 \cdot 2 & 2 \cdot 5 \end{pmatrix} = \begin{pmatrix} 6 & 8 \\ 4 & 10 \end{pmatrix}$$

Matrix Operations: Multiplication

- To multiply two matrices, we perform the “Dot Product” between rows of the first matrix and columns of the second matrix:

$$\boldsymbol{A} \cdot \boldsymbol{B} = \boldsymbol{C}$$

$$c_{ij} = \sum_{k} a_{ik} b_{kj}$$

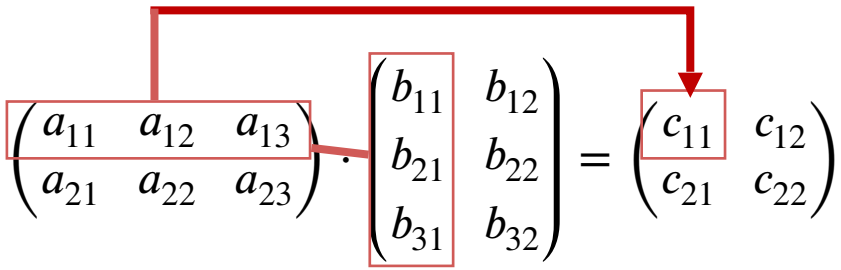
  - Two matrices can only be multiplied if number of columns in the first matrix is same as the number of rows of the second matrix.
  - If matrix $\boldsymbol{A}$ is of order $m \times n$ and matrix $\boldsymbol{B}$ is of order $r \times s$ the $\boldsymbol{A} \cdot \boldsymbol{B}$ will be possible if $n = r$.
  - The resulting matrix will be of order $m \times s$.

Matrix Operations: Multiplication

➢ Two multiply two matrices, we perform the “Dot Product” between rows of the first matrix and columns of the second matrix:

$$\boldsymbol{A} \cdot \boldsymbol{B} = \boldsymbol{C}$$

$$c_{ij} = \sum_{k} a_{ik} b_{kj}$$



$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \cdot \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$$

Matrix Product Properties

➢ Distributivity over addition:

$$\boldsymbol{A}(\boldsymbol{B}+\boldsymbol{C}) = \boldsymbol{A}\boldsymbol{B}+\boldsymbol{A}\boldsymbol{C}$$

➢ Associativity:

$$\boldsymbol{A}(\boldsymbol{B}\boldsymbol{C}) = (\boldsymbol{A}\boldsymbol{B})\boldsymbol{C}$$

➢ Not commutative: $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{A}$ is not always true

➢ Dot product between vectors is commutative:

$$\boldsymbol{x}^T\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{x}$$

➢ Transpose of a matrix product has a simple form:

$$\left(\boldsymbol{A}\boldsymbol{B}\right)^T = \boldsymbol{B}^T\boldsymbol{A}^T$$

Transpose of a Matrix

➢ Mirror image across principal diagonal

$$\boldsymbol{A} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \\ A_{3,1} & A_{3,2} \end{bmatrix}, \quad A^T = \begin{bmatrix} A_{1,1} & A_{2,1} & A_{3,1} \\ A_{1,2} & A_{2,2} & A_{3,2} \end{bmatrix}$$

➢ Vectors are then either a Matrix with a single column or a single row; For convenience often written in a line: $\boldsymbol{x}^T = [x_1, x_2, x_3]$

➢ Scalar is a Matrix with just one element, thus:

$x^T = x$

Norm of a Matrix

➢Frobenius norm

$$\|\boldsymbol{A}\| = \left(\sum_{i,j} A_{i,j}^2 \right)^{\frac{1}{2}}$$

➢It is analogous to L^2 norm of a vector

Square Matrices

➢ Remember, square matrices have the same number of rows as columns

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1d} \\ \vdots & \ddots & \vdots \\ a_{d1} & \cdots & a_{dd} \end{pmatrix}$$

➢ Some operations are special to square matrices and are not defined for arbitrary matrices

Identity Matrix

- An identity or unit matrix of size d is square matrix of order $d \times d$ where all the diagonal elements are ‘1' and all the other elements are ‘0'.
- It is denoted by $\boldsymbol{I}$.

$$\boldsymbol{I}_1 = (1) \quad \boldsymbol{I}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \boldsymbol{I}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

- Important Property ($\boldsymbol{A}$ being a square matrix):

$$\boldsymbol{A} \cdot \boldsymbol{I} = \boldsymbol{I} \cdot \boldsymbol{A} = \boldsymbol{A}$$

Inverse of a Matrix

➢ For a square matrix, the inverse matrix will be such that the product of the original matrix and the inverse matrix will produce an identity matrix.

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

➢ Note: Not all square matrices have inverses. A square matrix which has an inverse is called invertible or nonsingular matrix and a square matrix which does not has an inverse is called noninvertible or singular matrix.

➢ A square matrix is singular if and only if its **determinant** is 0

Square Matrices: Determinant of a Matrix

- Determinant of a square matrix det(A) is a mapping to a scalar
- It is equal to the product of all eigenvalues of the matrix
- Measures how much multiplication by the matrix expands or contracts space

Square Matrices: How To Calculate a Determinant

➢ For a 1x1 matrix:

$$\det(a_{11}) = a_{11}$$

➢ For a 2x2 matrix:

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

➢ For a 3x3 matrix:

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} =$$

$$a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31} =$$

$$a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31})$$

➢ (There exists a rather simple way to do this with arbitrarily large matrices, but we omit this in this course).

Special Matrices & Vectors

- **Unit Vector**
  - A vector with unit norm

$$\left|\left|\boldsymbol{x}\right|\right|_2 = 1$$

- **Orthogonal Vectors**
  - A vector x and a vector y are orthogonal to each other if

$$\boldsymbol{x}^T\boldsymbol{y} = 0$$

  - Vectors are at 90 degrees to each other

- **Orthogonal Matrix**
  - A square matrix whose columns and rows are orthogonal unit vectors

$$\boldsymbol{A}^{-1} = \boldsymbol{A}^T$$

Special Matrices & Vectors

➢ Diagonal Matrix

- ➢ Mostly zeros, with non-zero entries in diagonal
- ➢ $diag(\boldsymbol{v})$ is a square diagonal matrix with diagonal elements given by entries of vector $\boldsymbol{v}$
- ➢ Multiplying $diag(\boldsymbol{v})$ by vector $\boldsymbol{x}$ only needs to scale each element x_i by v_i

➢ Symmetric Matrix

- ➢ Is equal to its transpose:

$$\boldsymbol{A} = \boldsymbol{A}^T$$

- ➢ E.g., a distance matrix is symmetric with $A_{ij} = A_{ji}$

Matrices as Linear Transformation

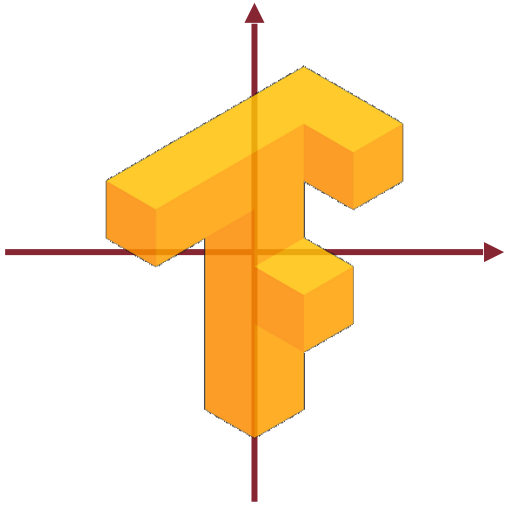
➢ At the end of the day, a Matrix defines a linear transformation

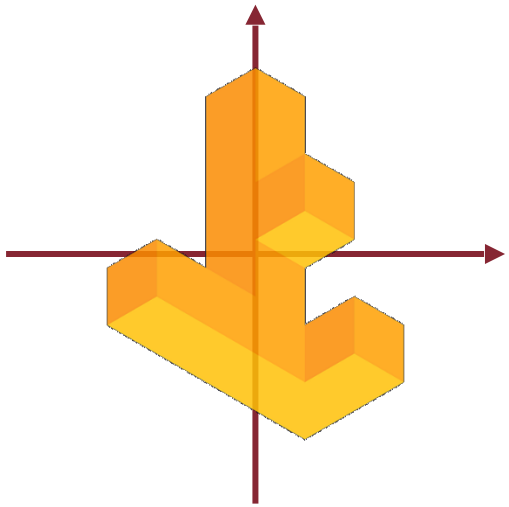
➢ A Matrix $A \in \mathbb{R}^{n \times m}$ projects points from $\mathbb{R}^n$ to their image in $\mathbb{R}^m$

➢ This is later exactly what we will do with the hidden layers

➢ Let's have a look at some examples for $\mathbb{R}^{2 \times 2}$ since we can visualize them

Examples: Mirror

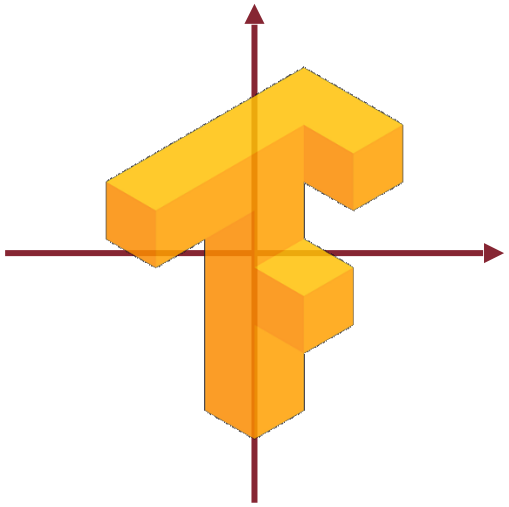


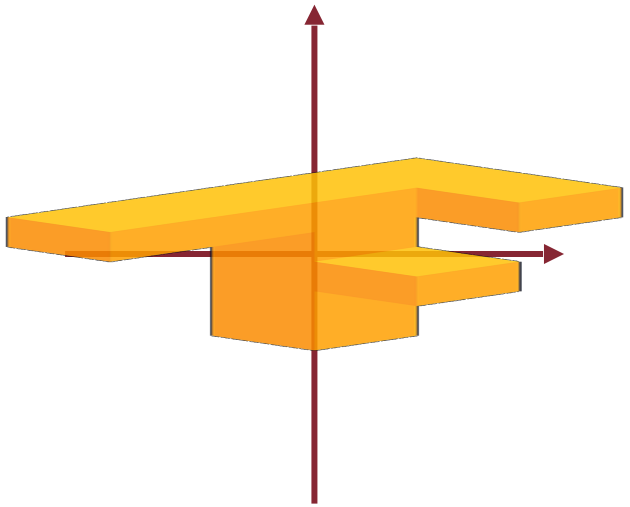


➢ Reflection on the x-Axis:

$$Mx = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x \\ -y \end{bmatrix}$$

Examples: Stretching





➢ Stretching along the Axes:

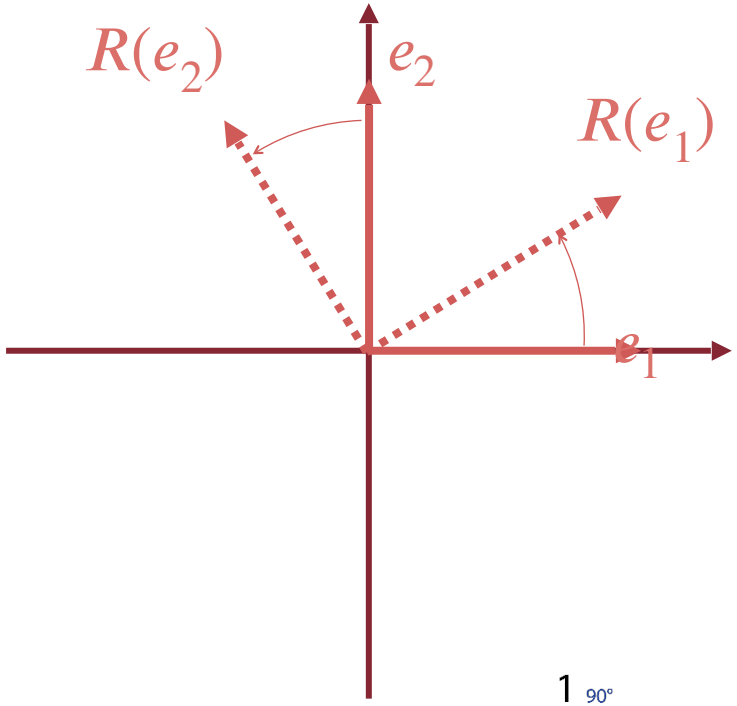
$$Sx = \begin{bmatrix} 2 & 0 \\ 0 & 0.5 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2x \\ 0.5 \cdot y \end{bmatrix}$$

Example: Rotation

- We are looking for the image of the standard basis e_1, e_2
- We need the following images:

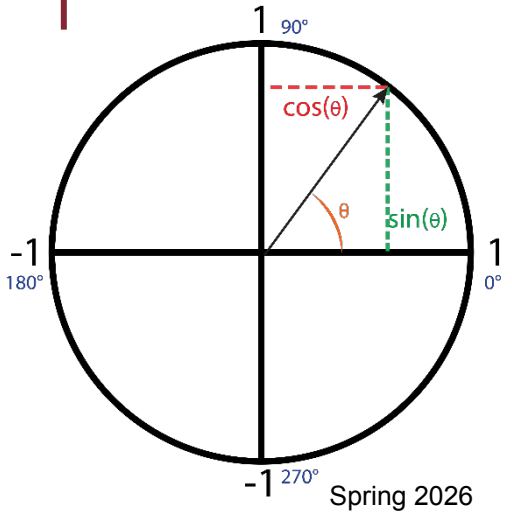
$$R(e_1) = R\begin{bmatrix}1\\0\end{bmatrix} = \begin{bmatrix}\cos\theta\\ \sin\theta\end{bmatrix}$$

$$R(e_2) = R\begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}-\sin\theta\\ \cos\theta\end{bmatrix}$$

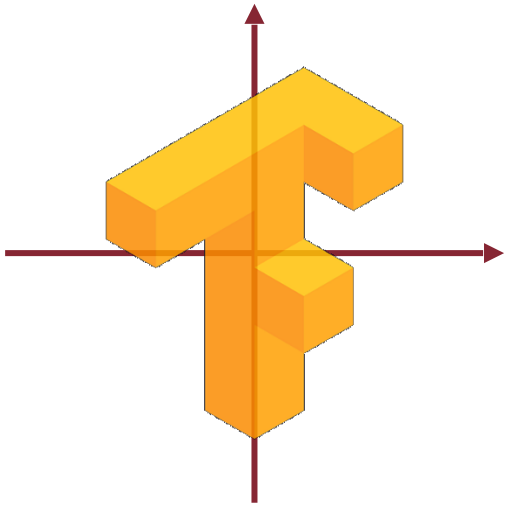


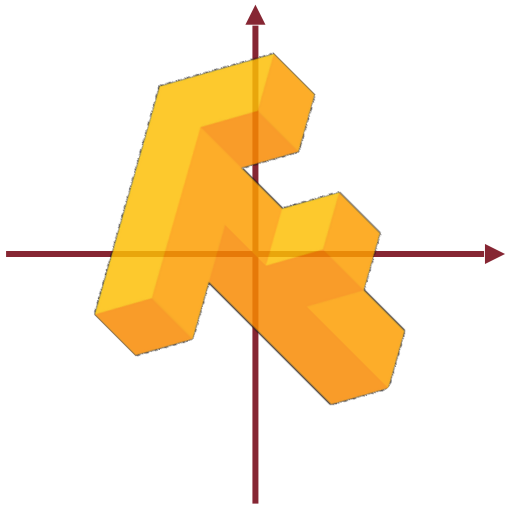
- Together for a rotation of 45 degree:

$$R = \begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix} = \begin{bmatrix}\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{bmatrix}$$



Examples: Rotation

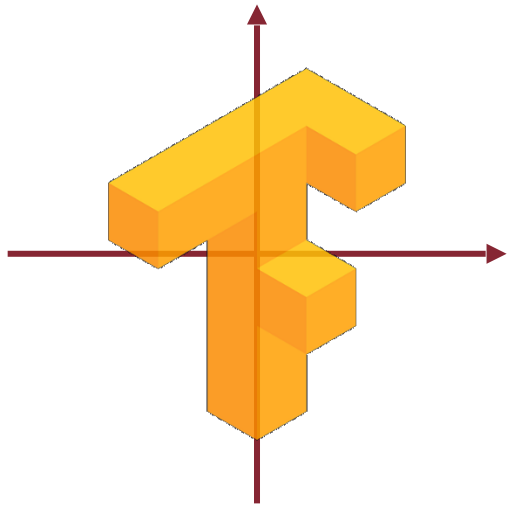


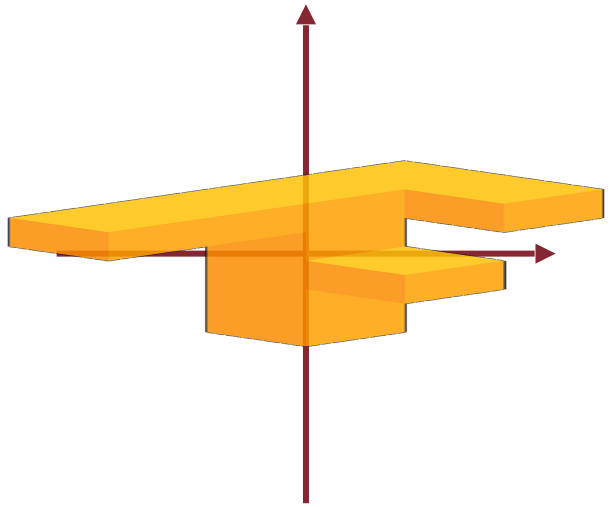


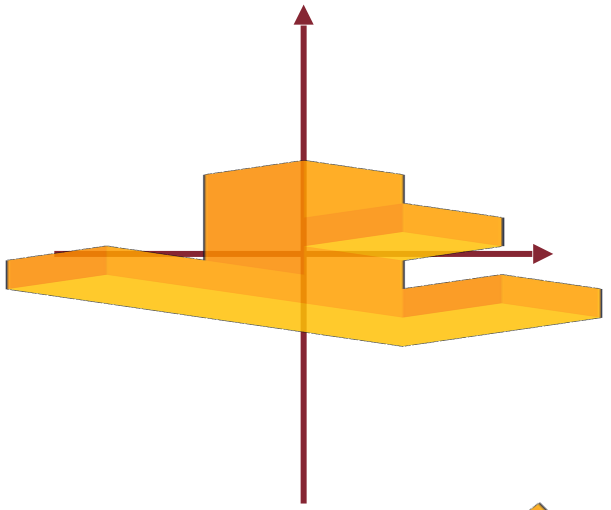
➢ Rotation on the x-Axis:

$$Rx = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}}x - \frac{1}{\sqrt{2}}y \\ \frac{1}{\sqrt{2}}x + \frac{1}{\sqrt{2}}y \end{bmatrix}$$

Example: Combinations

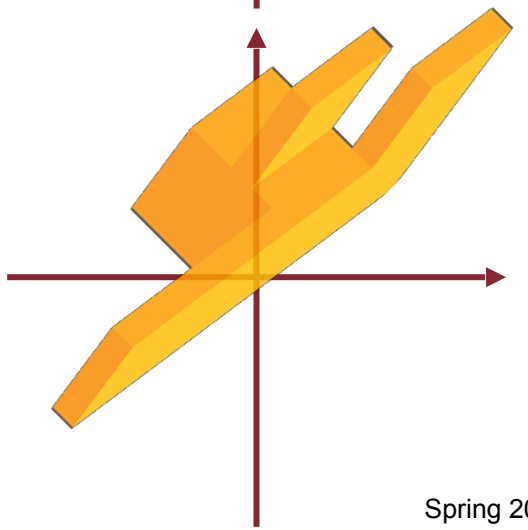




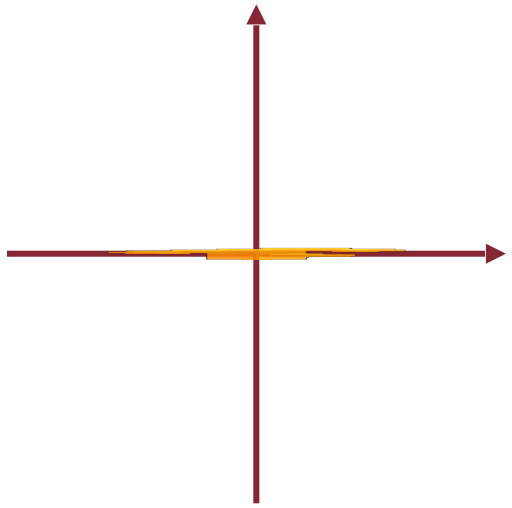


➢ Simply combine the transformations

$$M = RMS = \begin{bmatrix} 2\frac{1}{\sqrt{2}} & -2\frac{1}{\sqrt{2}} \\ -\frac{1}{2}\frac{1}{\sqrt{2}} & -\frac{1}{2}\frac{1}{\sqrt{2}} \end{bmatrix}$$



On the invertibility of the projections



- It becomes clear that not every transformation is invertible.
- For instance, $Mx = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = x$
  - Every point (x, y) gets projected onto the same value x
  - Impossible to "reconstruct" y
- Non square matrices project points into higher/lower dimensional spaces (as the example above does)

Keywords for this session

- Linear algebra
- **Probabilities & Standard Distributions**
- Eigendecomposition & PCA

Why Probability?

- Much of CS deals with entities that are certain
  - CPU executes flawlessly
    - At least almost … there are CPU bugs and CPUs can also be broken
  - CS and software engineers work in clean and certain environment
  - Surprising that ML heavily uses probability theory

- Reasons for ML use of probability theory
  - Must always deal with uncertain quantities
    - Also with non-deterministic (stochastic) quantities
  - Many sources for uncertainty and stochasticity

Sources of Uncertainty

1. Inherent stochasticity of system being modeled
   - ➢ Subatomic particles are probabilistic
   - ➢ Cards shuffled in random order

2. Incomplete observability
   - ➢ Deterministic systems appear stochastic when not all variables are observed

3. Incomplete modeling
   - ➢ Discarded information results in uncertain predictions

Practical to use uncertain rule

- Simple rule “Most birds fly” is cheap to develop and broadly useful
- Rules of the form “Birds fly, except for very young birds that have not learned to fly, sick or injured birds that have lost ability to fly, flightless species of birds...” are expensive to develop, maintain and communicate
  - Also still brittle and prone to failure

Tools of Probability

- Probability theory was originally developed to analyze frequencies of events
  - Such as drawing a hand of cards in poker
  - These events are repeatable
  - If we repeated experiment infinitely many times, proportion of p of outcomes would result in that outcome

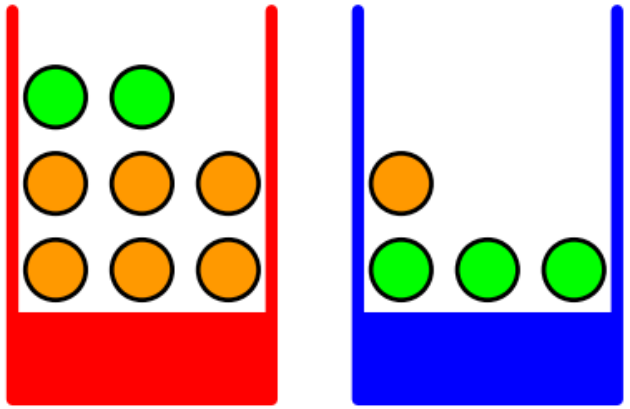
- Is it applicable to propositions not repeatable?
  - Patient has 40% chance of flu
    - Cannot make infinite replicas of the patient
  - We use probability to represent degree of belief

- Former is frequentist probability, latter Bayesian, just fyi.

Definition of Probability

➢ To begin with, we shall define the probability of an event to be the fraction of times that event occurs out of the total number of trials, in the limit that the total number of trials goes to infinity.

➢ Let's look at the following simple Experiment:
- ➢ Two boxes, red (r) and blue (b)
- ➢ Each boxes contain either apples (a) or oranges (o)
- ➢ The box that will be chosen is a random variable B. It takes the values r or b
- ➢ The fruit sampled out of the selected box is donated F and takes the values a and o



➢ Taken from Christopher M. Bishop, "Pattern Recognition and Machine Learning"

Random Variables

- Let's assume we select the blue box 60% of the time
- Repeating the box selection infinite number of times, we say that the probability of selecting the red box is 4/10 and the blue box 6/10.
- We formally write this as follows:

$$p(B = r) = \frac{4}{10} \qquad p(B = b) = \frac{6}{10}$$

  - Note that, by definition, probabilities must lie in the interval [0, 1]
  - Also, if the events are mutually exclusive and if they include all possible outcomes, then we see that the probabilities for those events must sum to one.

- Taken from Christopher M. Bishop, "Pattern Recognition and Machine Learning"

Joint Probability Distributions

- We can now ask questions such as:
  - "what is the overall probability that the selection procedure will pick an apple?"
  - "given that we have chosen an orange, what is the probability that the box we chose was the blue one?"

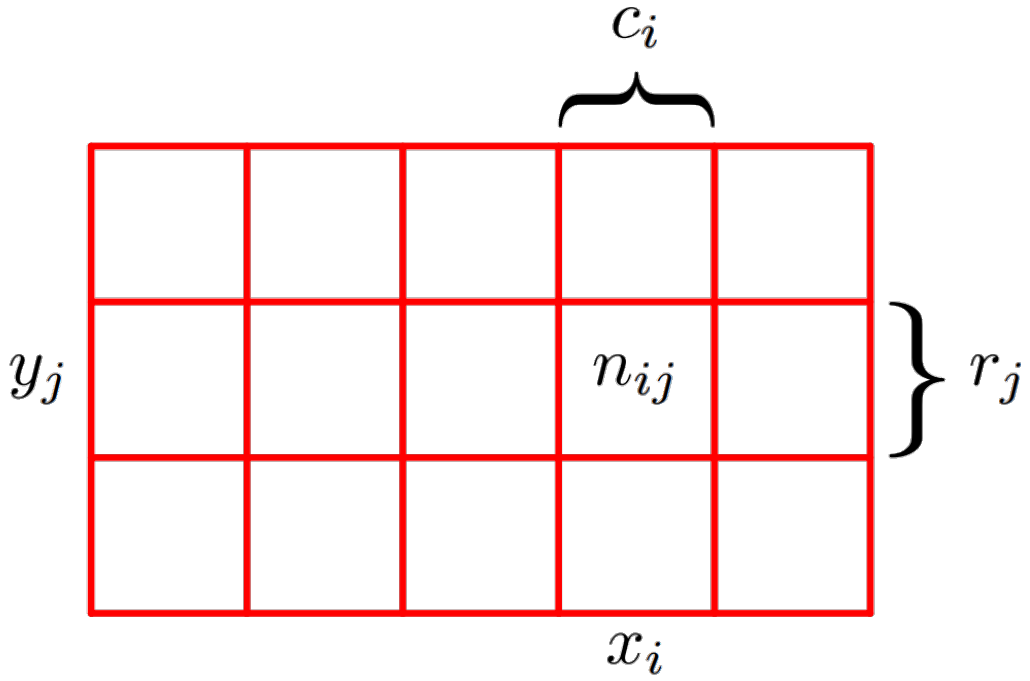
- To answer such questions, we need the elementary rules of probability:
  - the **sum rule**
  - the **product rule**

- Taken from Christopher M. Bishop, "Pattern Recognition and Machine Learning"

Joint Probability Distributions

- Consider we have two random variables X and Y
  - X can take any of the values x_i where $i = 1,\ldots, M$
  - Y can take the values y_j where $j = 1,\ldots, L$

- Consider a total of N samples
  - The number of such trials in which $X = x_i$ and $Y = y_j$ be n_{ij}
  - The number of trials in which X takes the value x_i is c_i
  - The number of trials in which Y takes the value y_j is r_j

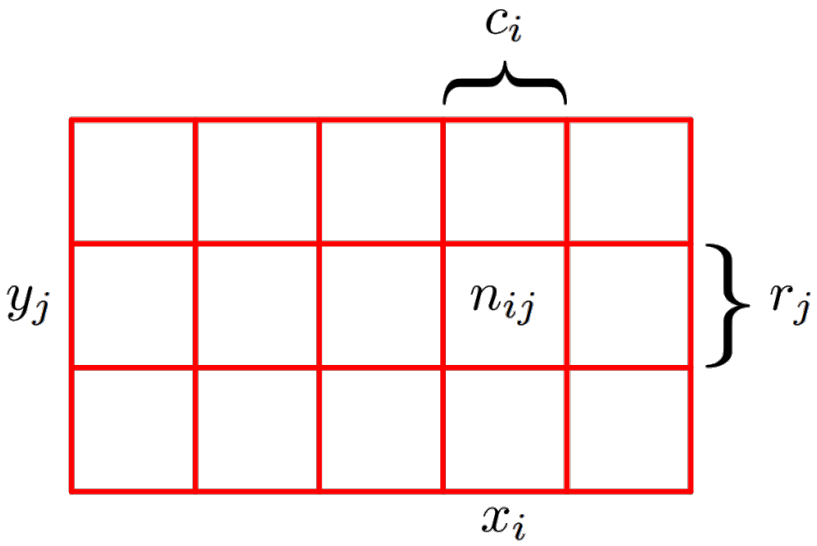


- Taken from Christopher M. Bishop, "Pattern Recognition and Machine Learning"

Joint Probability Distributions

➢ The probability that $X = x_i$ and $Y = y_i$ is written as:

$$p\Big(X = x_i, Y = y_j\Big)$$



➢ This is called the joint probability distribution

➢ It is given by the fractions of point in cell i, j

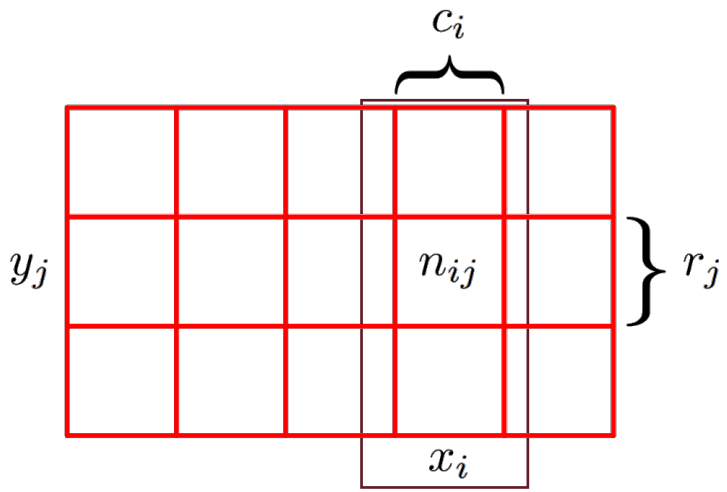
$$p\Big(X = x_i, Y = y_j\Big) = \frac{n_{ij}}{N}$$

➢ Taken from Christopher M. Bishop, "Pattern Recognition and Machine Learning"

Sum Rule

➢ The probability that X takes the value x_i irrespective of the value of Y is written as $p(X = x_i)$ and is given by the fraction of points in column i

$$p(X = x_i) = \frac{c_i}{N}$$



➢ Since c_i is just the sum of the cells of the column, we can also write:

$$p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

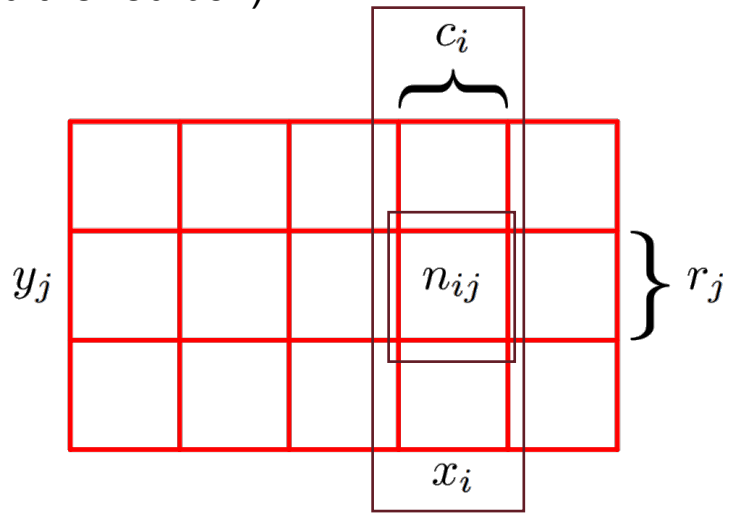
Note that $p(X = x_i)$ is sometimes called the marginal probability, because it is obtained by marginalizing, or summing out, the other variables (in this case Y).

➢ Taken from Christopher M. Bishop, "Pattern Recognition and Machine Learning"

Product Rule or Conditional Probabilities

- Now, let's only consider those cases, where $X = x_i$.
- We are now interested in the probability that $Y = y_j$ if X is already fixed to x_i
  (For example: What is the probability of sampling an apple, when we have selected the red box)
- We write this as:

$$p\left(Y = y_j \middle| X = x_i\right) = \frac{n_{ij}}{c_i}$$



- With the results from before, we get:

$$p\left(X = x_i, Y = y_j\right) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} = p\left(Y = y_j \middle| X = x_i\right) p\left(X = x_i\right)$$

- Taken from Christopher M. Bishop, "Pattern Recognition and Machine Learning"

Rules of Probability

➢ **Sum Rule**

$$p(X) = \sum_{Y} p(X, Y)$$

➢ **Product Rule**

$$p(X, Y) = p(Y \mid X) p(X)$$

(Note the more compact notation: We simply write $p(B)$ to denote a distribution over the random variable B, or $p(r)$ to denote the distribution evaluated for the particular value r, provided that the interpretation is clear from the context.)

➢ Taken from Christopher M. Bishop, "Pattern Recognition and Machine Learning"

Bayes theorem

➢ From the rules, we can directly derive Bayes theorem

$$\overbrace{P(Y \mid X)}^{\text{posterior}} = \frac{\overbrace{P(X \mid Y)}^{\text{likelihood}} \overbrace{P(Y)}^{\text{prior}}}{\underbrace{P(X)}_{\text{evidence}}}$$

➢ **Prior**: Our assumptions about Y before observing the data.

➢ **Likelihood**: The effect of the observed data X.

➢ **Posterior**: The uncertainty in Y after we have observed X.

"A Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule."

Bayes theorem in Action

- Consider a SPAM filter which classifies mail into:
  - Good: H (Ham)
  - Bad: S (Spam)
- We observe a message containing the word W "replica"
- We use Bayes theorem

$$p(S \mid W) = \frac{p(W \mid S)p(S)}{p(W \mid S)p(S) + p(W \mid H)p(H)} = \frac{p(W \mid S)p(S)}{p(W)}$$

  - $p(S \mid W)$ the probability that this message is SPAM
  - $p(S)$ the probability that any given message is SPAM (our **prior**)
  - $p(W \mid S)$ the probability that "replica" appears in SPAM (our **likelihood**)
  - $p(W)$ the probability that this word appears in any message (our **evidence**)

How does it work?

- We have trained the classifier, i.e., we have scanned through emails we know are SPAM or HAM and have calculated the following this:
  - 80% of the mail are SPAM, i.e. our prior $p(s) = 0.8$
    - That means, without any evidence, we are 80% certain, a message is SPAM
  - The probabilities of words occurring in SPAM and HAM

- Here, our filter again, in all its glory:

$$p(S \mid W) = \frac{p(W \mid S)p(S)}{p(W \mid S)p(S) + p(W \mid H)p(H)}$$

How does it work?

- We observe a completely irrelevant word: "the"
  - It appears in every singe message
  - $p(W \mid S) = 1$
  - $p(W \mid H) = 1$

- When we now apply the filter, we get:

$$p(S \mid W) = \frac{p(W \mid S)p(S)}{p(W \mid S)p(S) + p(W \mid H)p(H)}$$

$$= \frac{1 \cdot 0.8}{1 \cdot 0.8 + 1 \cdot 0.2} = 0.8$$

- That means, this useless evidence has neither strengthen nor weakened our prior believe

How does it work?

- Now we observe a typical SPAM word: "replica"
  - It appears in every fourth SPAM message, but only in every 100th HAM:
  - $p(W \mid S) = 0.25$
  - $p(W \mid H) = 0.01$

- When we now apply the filter, we get:

$$p(S \mid W) = \frac{p(W \mid S)p(S)}{p(W \mid S)p(S) + p(W \mid H)p(H)}$$

$$= \frac{0.25 \cdot 0.8}{0.25 \cdot 0.8 + 0.01 \cdot 0.2} = 0.99$$

- That means, the observed evidence, lead our posterior to be stronger than the prior!

How does it work?

- Now we observe a HAM, my name is spelled correct: "Galke"
  - It appears as follows:
  - $p(W \mid S) = 0.05$
  - $p(W \mid H) = 0.75$

- When we now apply the filter, we get:

$$p(S \mid W) = \frac{p(W \mid S)p(S)}{p(W \mid S)p(S) + p(W \mid H)p(H)}$$

$$= \frac{0.05 \cdot 0.8}{0.05 \cdot 0.8 + 0.75 \cdot 0.2} = 0.21$$

- That means, the observed evidence, lead our posterior to less leaning towards SPAM!

Random Variables

➢ A **random variable** X is a variable that can take on different values randomly

➢ On its own, a random variable is just a description of the states that are possible;

➢ It must be coupled with a probability distribution that specifies how likely each of these states are.

➢ Random variables may be **discrete** or **continuous**

Probability Distributions

- A probability distribution is a description of how likely a random variable or a set of random variables is to take each of its possible states

- The way to describe the distribution depends on whether it is discrete or continuous

Probability Mass Functions

- The probability distribution over discrete variables is given by a probability mass function
- PMFs of variables are denoted by P and inferred from their argument, e.g., $P(x)$, $P(y)$
- They can act on many variables and is known as a joint distribution, written as $P(x, y)$

- To be a PMF it must satisfy:
  - Domain of P is the set of all possible states of x
  - $\forall x \in X,\ 0 \leq P(x) \leq 1$
  - $\sum_{x \in X} P(x) = 1$ (normalization)

Continuous Variables and PDFs

- When working with continuous variables, we describe probability distributions using probability density functions

- To be a pdf p must satisfy:
  - The domain of p must be the set of all possible states of X
  - $\forall x \in X,\ p(x) \geq 0$. Note, there is no requirement for $p(x) \leq 1$.
  - $\int p(x)dx = 1$

Expectation

➢ Expectation or expected value of $f(x)$ w.r.t. $P(X)$ is the average or mean value that f takes on when x is drawn from P

➢ For discrete variables:

$$E[X] = \sum_{x \in X} P(x) \cdot x$$

➢ For continuous variables:

$$E[X] = \int_{x} p(x)x \; dx$$

Properties of Expectation

- **Linearity of expectation**
  - $E[X+Y] = E[X] + E[Y]$
  - $E[\alpha X] = \alpha E[X]$
- **Expectation is also defined for functions on random variables**
  - **Discrete:** $E[f] = \sum_x p(x) \cdot f(x)$
  - **Continuous**: $E[f] = \int_{-\infty}^{+\infty} p(x) \cdot f(x) dx$

Variance

$$Var(X) = E\left[(x - E[X])^2\right]$$

- Measures how much the value of $f(x)$ vary from the expectation
- Low variance means values cluster around its expectations
- Square root of the variance is the standard deviation

Bernoulli Distribution

- Distribution over a single binary random variable
- It is controlled by a single parameter $\phi \in [0,1]$
  - Which gives the probability a random variable being equal to 1
- It has the following properties (to the right)

$$P(\mathrm{x} = 1) = \phi$$

$$P(\mathrm{x} = 0) = 1 - \phi$$

$$P(\mathrm{x} = x) = \phi^{x}(1 - \phi)^{1-x}$$

$$\mathbb{E}_{\mathrm{x}}[\mathrm{x}] = \phi$$

$$\mathrm{Var}_{\mathrm{x}}(\mathrm{x}) = \phi(1 - \phi)$$

Multinoulli Distribution

- Distribution over a single discrete variable with k different states

- It is parameterized by a vector $\boldsymbol{p} \in [0,1]^{k-1}$
  - where p_i is the probability of the ith state
  - The final kth state's probability is implicitly given by $1 - \mathbf{1}^T \boldsymbol{p}$
  - We must constrain $\mathbf{1}^T \boldsymbol{p} \leq 1$

- Multinoullis refer to distributions over categories
  - So we don't assume state 1 has value 1, etc.
  - For this reason we do not usually need to compute the expectation or variance of multinoulli variables since the states are not necessarily ordered

Gaussian Distribution

- Probably one of the most commonly used distributions

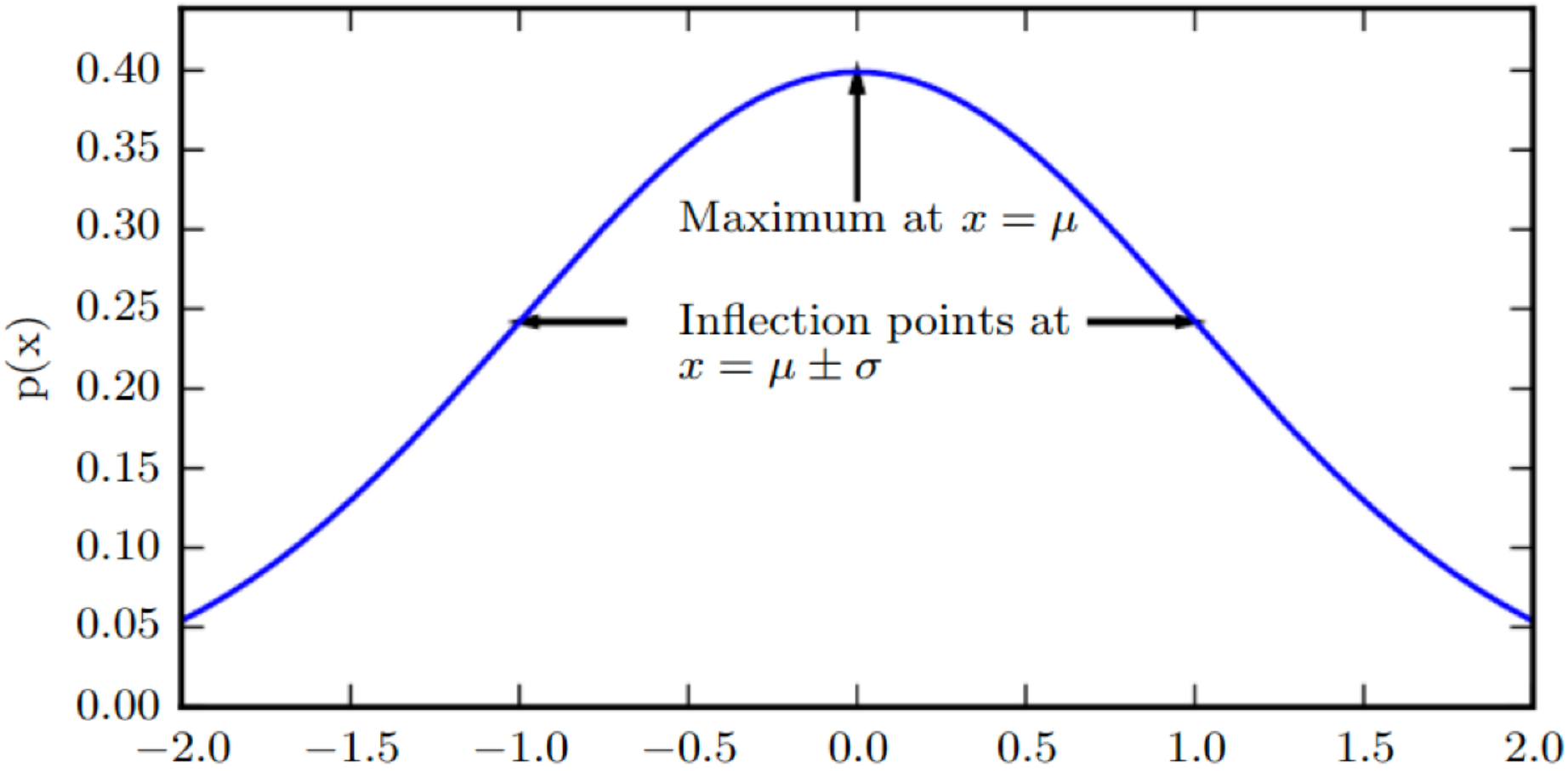
$$N(x;\mu,\sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

- Two parameters
  - μ gives the location of the central peak, which is also the mean of the distribution
  - The standard deviation is given by σ and variance by σ^2

- If this is evaluated frequently, sometimes parameterized with the inverse variance (or precision) β:

$$N(x;\mu,\beta^{-1}) = \sqrt{\frac{\beta}{2\pi}}\exp\left(-\frac{1}{2}\beta(x-\mu)^2\right)$$

Gaussian Distribution, $\mu = 0, \sigma = 1$



Justifications for Normal Assumption

- 1. **Central Limit Theorem**
  - Many distributions we wish to model are truly normal
  - Sum of many independent distributions is normal
  - Can model complicated systems as normal even if components have more structured behavior

- 2. **Maximum Entropy**
  - Of all possible probability distributions with the same variance, normal distribution encodes the maximum amount of uncertainty over real numbers
  - Thus, the normal distributions inserts the least amount of prior knowledge into a model

Multidimensional Gaussian Distributions

➢ The Gaussian Distribution can easily be extended to the multivariate case.

➢ Now, $\boldsymbol{x}$ and $\boldsymbol{\mu}$ are a vector, $\boldsymbol{\Sigma}$ a positive semidefinite symmetric matrix (the covariance matrix):

$$N(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sqrt{\frac{1}{(2\pi)^2|\boldsymbol{\Sigma}|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

➢ Analogously, with the precision Matrix

$$N(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\beta}^{-1}) = \sqrt{\frac{|\boldsymbol{\beta}|}{(2\pi)^2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\beta}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

The *variance* of $f(x)$ is defined by

$$\text{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] \tag{2.44}$$

and provides a measure of how much $f(x)$ varies around its mean value $\mathbb{E}[f(x)]$. Expanding out the square, we see that the variance can also be written in terms of the expectations of $f(x)$ and $f(x)^2$:

Exercise 2.8

$$\text{var}[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2. \tag{2.45}$$

Keywords for this session

- Linear algebra
- Probabilities & Standard Distributions
- **Eigendecomposition & PCA**

Matrix Decomposition

- Matrices can be decomposed into factors to learn universal properties about them not discernible from their representation
  - E.g., from decomposition of integer into prime factors 12=2x2x3 we can discern that
    - 12 is not divisible by 5 or
    - any multiple of 12 is divisible by 3

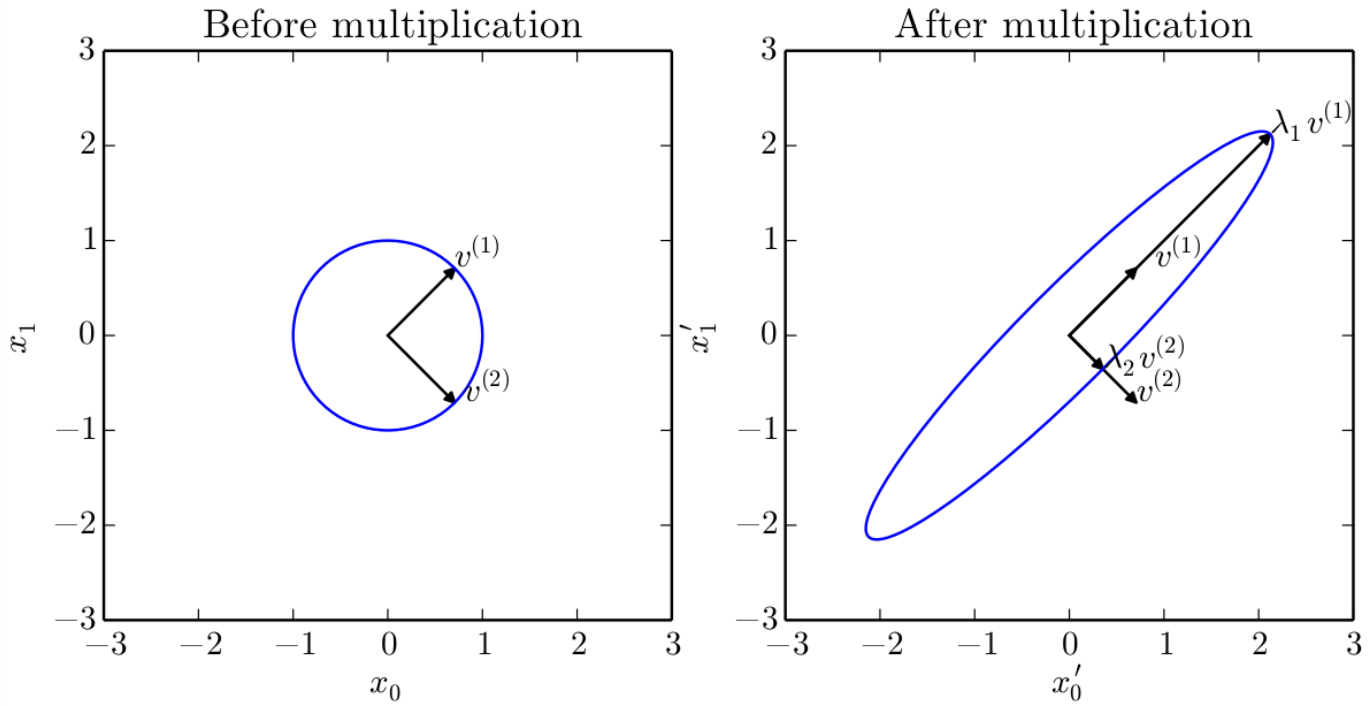
- Analogously, a matrix is decomposed into Eigenvalues and Eigenvectors to discern universal properties

Eigenvector

➢ An eigenvector of a square matrix $\boldsymbol{A}$ is a non-zero vector $\boldsymbol{v}$ such that multiplication by $\boldsymbol{A}$ only changes the scale of $\boldsymbol{v}$

$$\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$$

➢ The scalar λ is known as eigenvalue
- ➢ If $\boldsymbol{v}$ is an eigenvector of $\boldsymbol{A}$, so is any rescaled vector $s\boldsymbol{v}$.
- ➢ $s\boldsymbol{v}$ still has the same eigen value.
- ➢ Thus, the unit Eigenvector is used



Eigenvalue and Characteristic Polynomial

➢ Consider $\boldsymbol{Av} = \boldsymbol{w}$

➢ If $\boldsymbol{v}$ and $\boldsymbol{w}$ are scalar multipliers, then

$$\boldsymbol{Av} = \lambda \boldsymbol{v}\left(\boldsymbol{A} - \lambda \boldsymbol{I}\right)\boldsymbol{v} = 0$$

➢ This has a non-zero solution if

$$\left|\boldsymbol{A} - \lambda \boldsymbol{I}\right| = 0$$

➢ The nulls of the polynomial of degree n are the eigenvalues of $\boldsymbol{A}$

Example

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

➢ Then, the characteristic polynomial is

$$\left| \boldsymbol{A} - \lambda \boldsymbol{I} \right| = \begin{bmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix} = 3 - 4\lambda + \lambda^2$$

➢ The polynomial has the roots $\lambda = 1$ and $\lambda = 3$ which are the eigenvalues of $\boldsymbol{A}$.

➢ The eigenvectors can be found by solving the equation $Av = \lambda v$ using the different eigenvectors:

$$v_{\lambda=1} = \begin{bmatrix} 1 \\ -1 \end{bmatrix},\ v_{\lambda=3} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Eigendecomposition

- Suppose that matrix $\boldsymbol{A}$ has n linearly independent eigenvectors $\left\{v^{(1)}, \ldots, v^{(n)}\right\}$ with the eigenvalues $\left\{\lambda_1, \ldots, \lambda_n\right\}$
- Concatenate eigenvectors to form matrix $\boldsymbol{V}$
- Concatenate eigenvalues to form vector $\lambda = \left[\lambda_1, \ldots, \lambda_n\right]$ (normally in descending order)
- The Eigendecomposition of $\boldsymbol{A}$ is given by

$$\boldsymbol{A} = \boldsymbol{V}\mathrm{diag}\left(\lambda\right)\boldsymbol{V}^{-1}$$

Decomposition of Symmetric Matrix

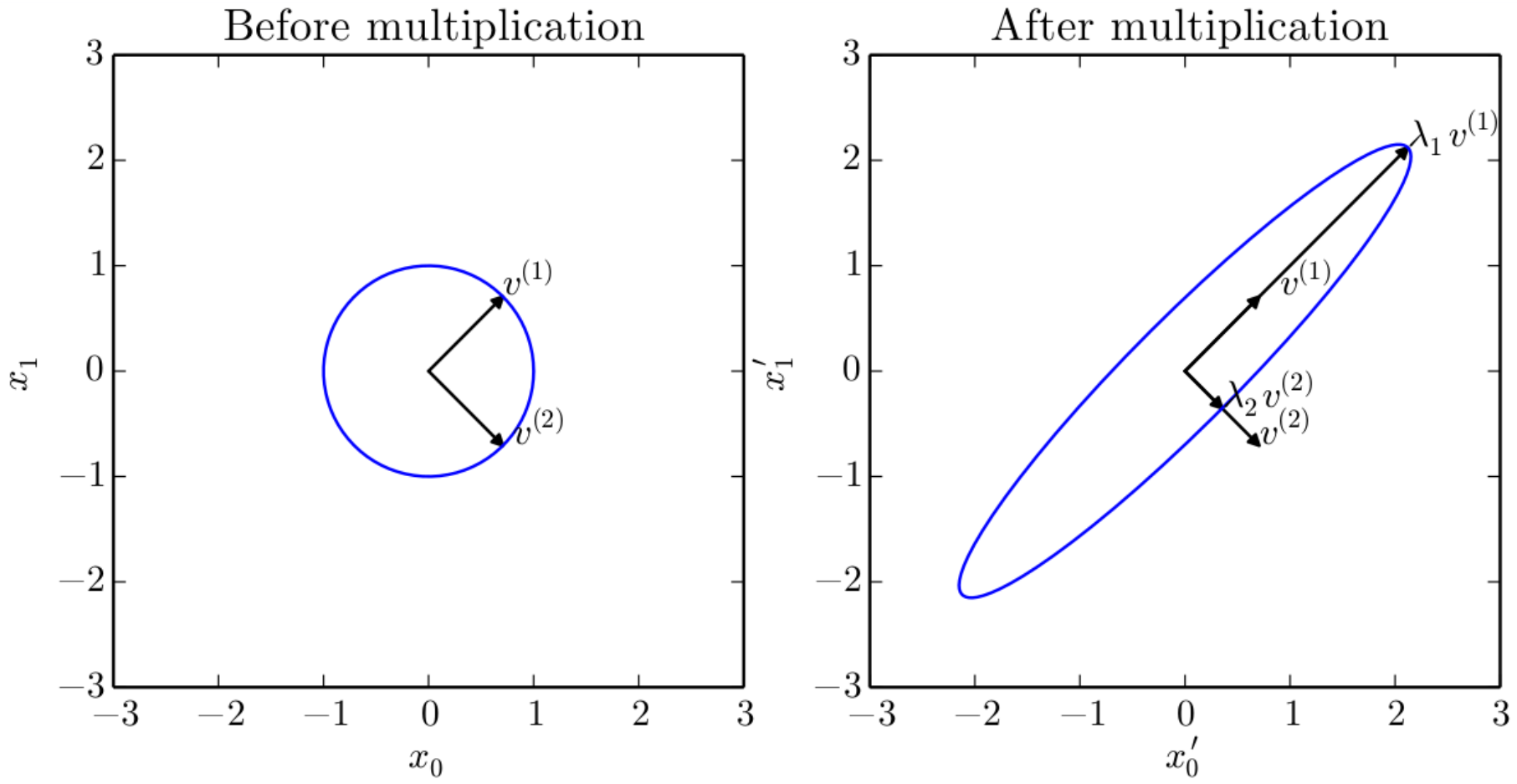
➢ Every real symmetric matrix $\boldsymbol{A}$ can be decomposed into real-valued eigenvectors and eigenvalues

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$$

➢ where $\boldsymbol{Q}$ is an orthogonal matrix composed of the eigenvectors and $\boldsymbol{\Lambda}$ the diagonal matrix of eigenvalues.

➢ We can think of $\boldsymbol{A}$ as scaling space by λ_i in direction $v^{(i)}$

Effect of Eigenvectors and Eigenvalues



Effect of Eigenvectors and Eigenvalues

- A matrix whose eigenvalues are
  - all positive is called **positive definite**
  - all positive or zero-valued is called **positive semidefinite**
  - all negative is called **negative definite**
  - all negative or zero-valued is called **negative semidefinite**

- Semidefinite matrices are interesting because they guarantee that

$$\forall \boldsymbol{x},\ \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \geq 0$$

- Positive definite matrices additionally guarantee that

$$\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} = 0 \Rightarrow \boldsymbol{x} = 0$$

Application: Principal Component Analysis

- We have observed the usefulness of the decomposition of matrices to understand the effect of the mapping

- The PCA does follow a similar idea

- There are many interpretations to PCA
  - We look at it as a dimensionality reduction technique
  - Later, we will see the correspondence to layers in deep neural networks

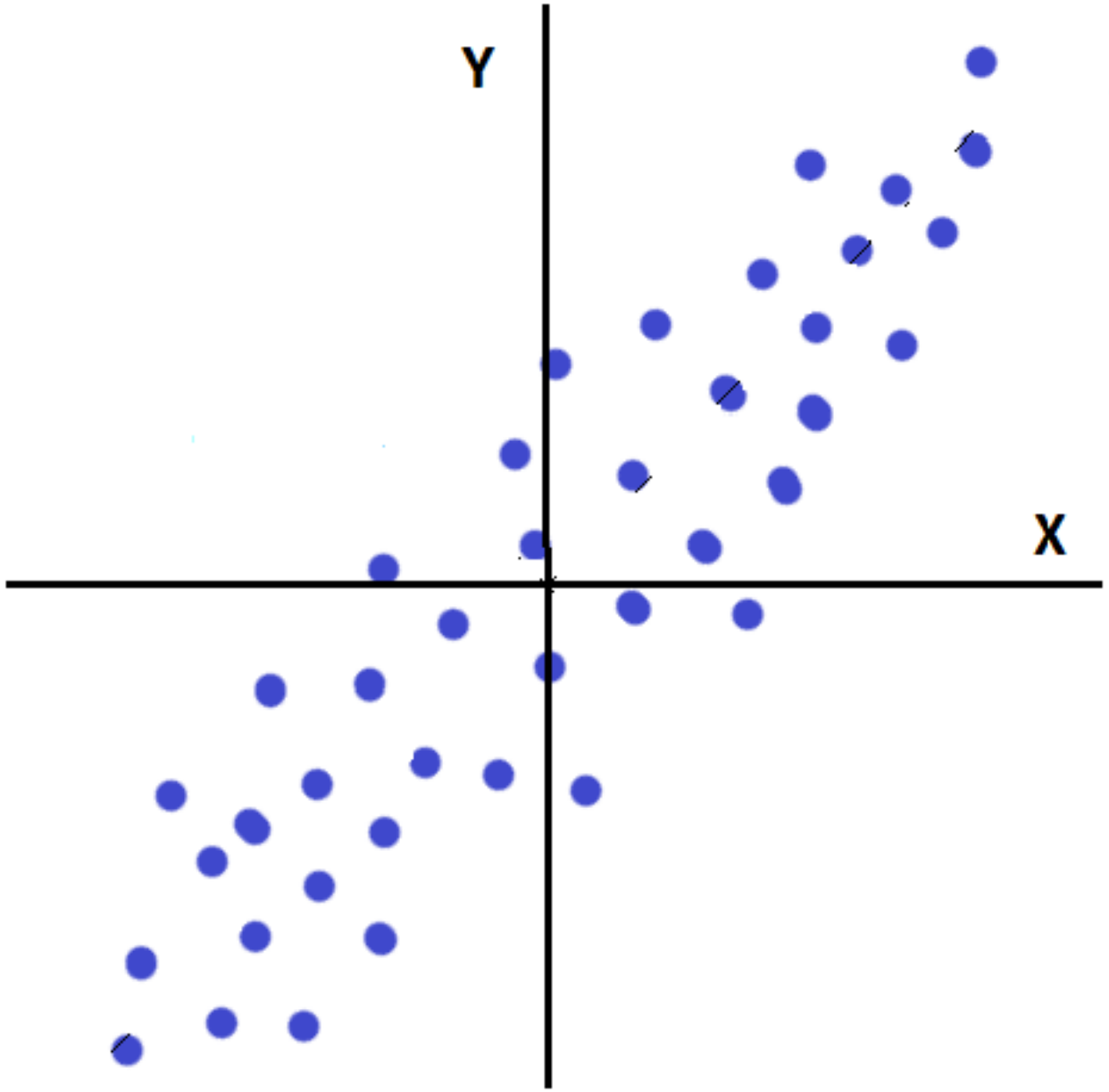
Dimensionality Reduction

- There are basically two ways to reduce dimensionality:

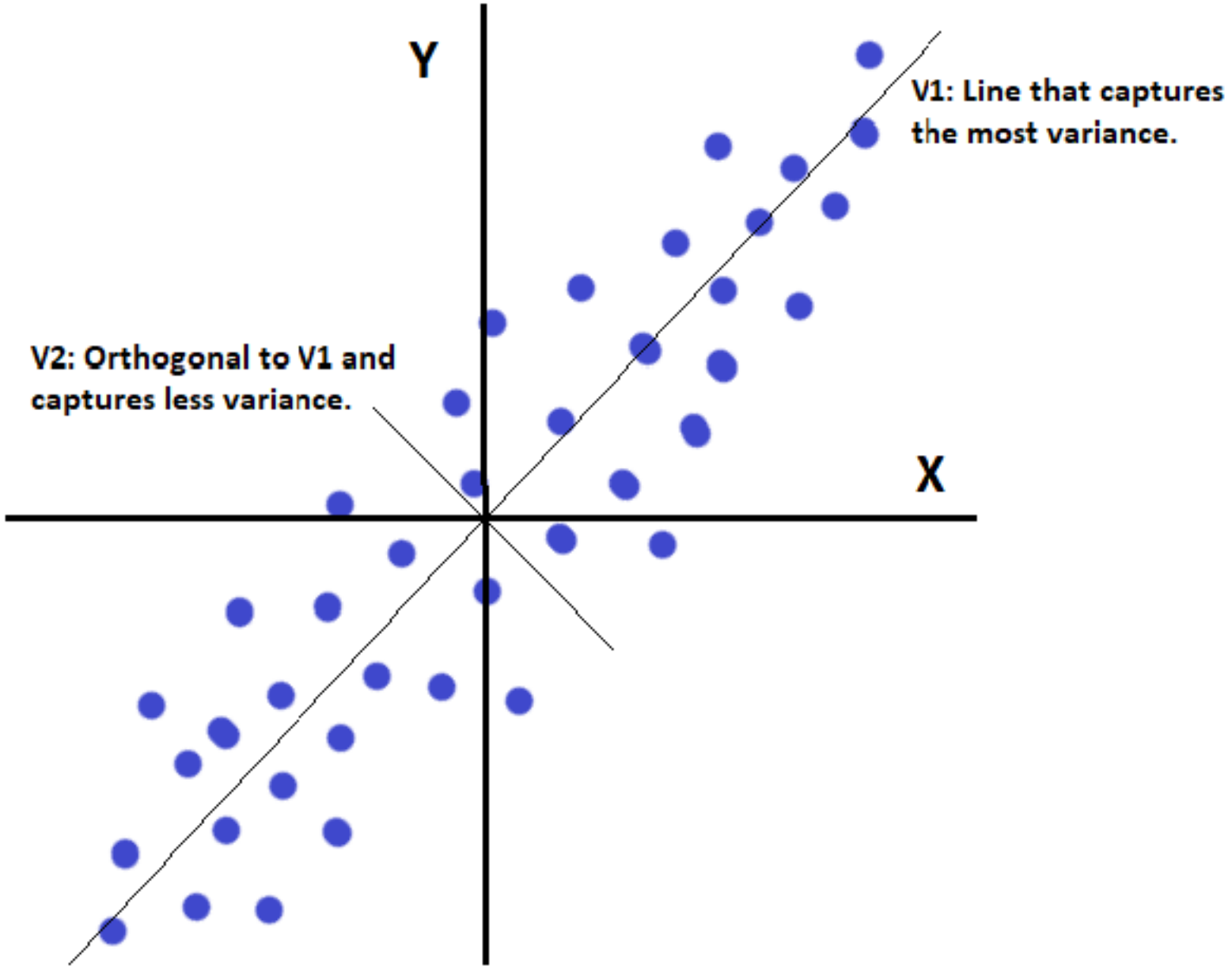
- **Feature extraction**: creating a subset of new features by combinations of the existing features
- **Feature selection**: choosing a subset of the available features

- What properties do we wish to retain by dimensionality reduction:
  - We aim to retain as much information and structure as possible in the reduced dimensions
  - In case of dimensionality, we want to capture most of the variance of the dataset
  - The dimension should be independent

Example



Example


Y
X
V1: Line that captures the most variance.
V2: Orthogonal to V1 and captures less variance.

Covariance Matrix

- The PCA tries to retain most of the variance, so we have to assess the variance within our dataset
  - Variance and Covariance are a measure of the “spread“ of a set of points around their center of mass (mean)
  - Variance – measure of the deviation from the mean for points in one dimension e.g. heights
  - Covariance as a measure of how much each of the dimensions vary from the mean with respect to each other
  - Covariance is measured between 2 dimensions to see if there is a relationship between the 2 dimensions
    - e.g. number of hours studied & marks obtained
  - The covariance between one dimension and itself is the variance

Covariance Matrix

➢ The covariance between to variables A and B is defined as

$$\mathrm{Cov}(A,B) = \frac{1}{n-1}\sum_{i=1}^{n}\left(a_i - \bar{A}\right)\left(b_i - \bar{B}\right)$$

➢ For all dimensions, e.g., A,B,C in Matrix-form:

$$Q = \begin{pmatrix} \mathrm{Cov}(A,A) & \mathrm{Cov}(A,B) & \mathrm{Cov}(A,C) \\ \mathrm{Cov}(B,A) & \mathrm{Cov}(B,B) & \mathrm{Cov}(B,C) \\ \mathrm{Cov}(C,A) & \mathrm{Cov}(C,B) & \mathrm{Cov}(C,C) \end{pmatrix}$$

- ➢ Diagonal is the variance of A, B, C
- ➢ The matrix is symmetric
- ➢ d-dimensional data will result in a $d \times d$ covariance matric

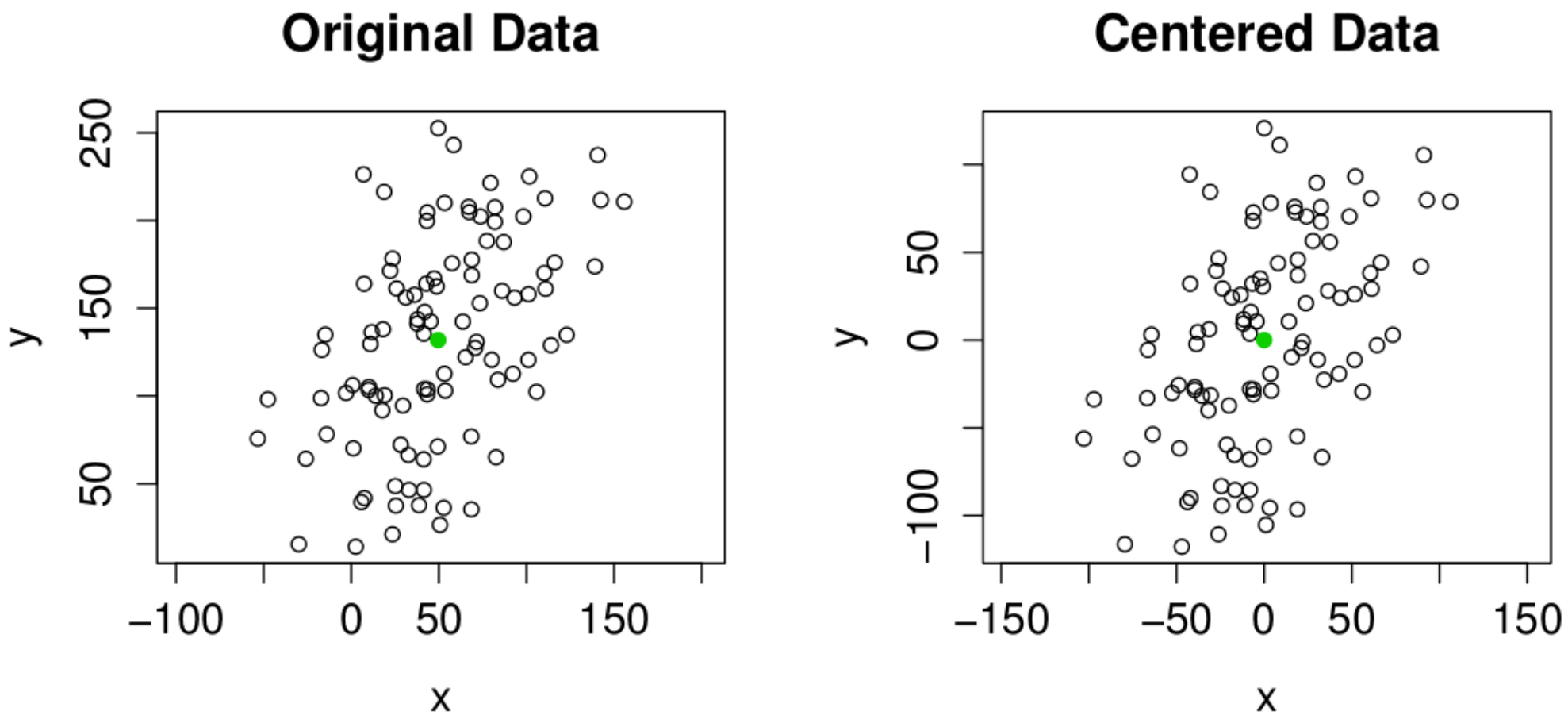
Decomposition

- Let X be our d-dimensional dataset which is centered
- Then we can calculate the co-variance simply by

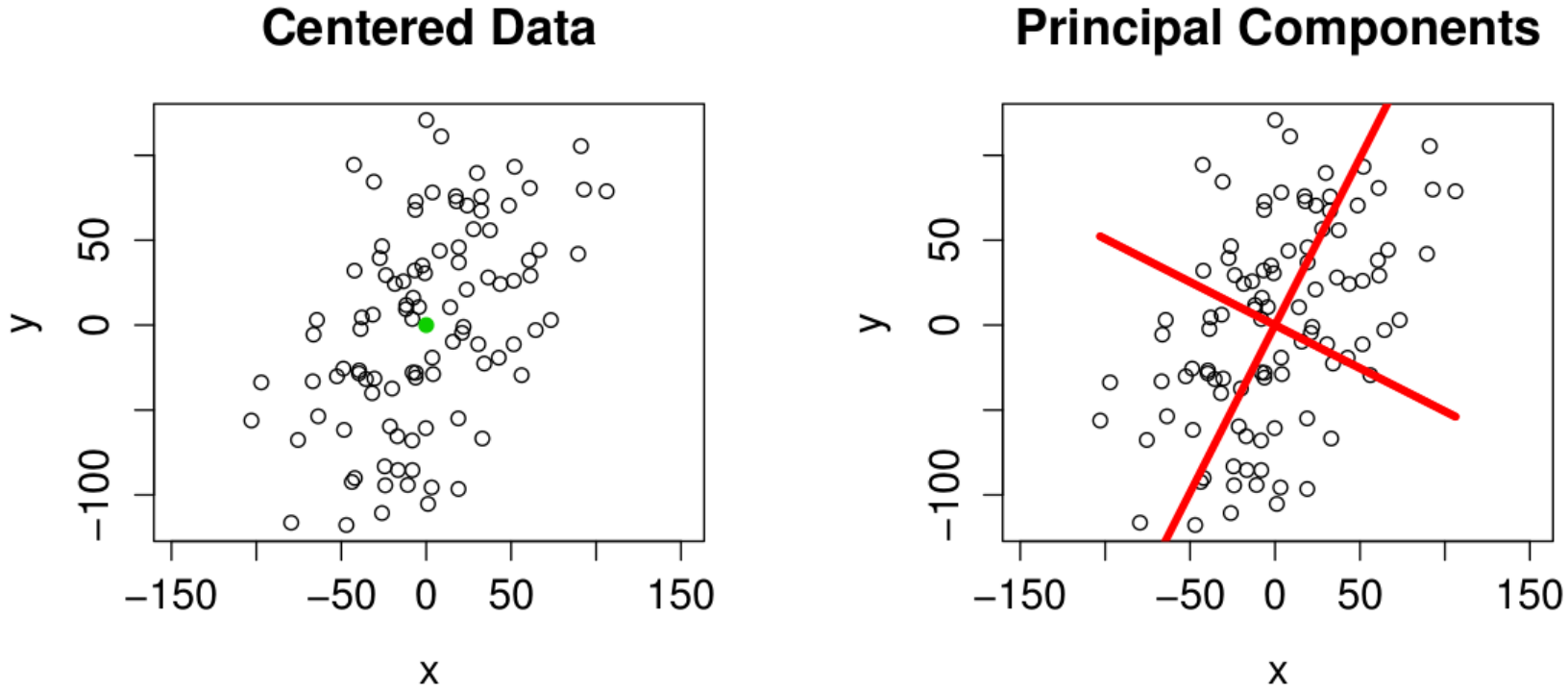
$$Q = \frac{1}{n-1} X X^T$$

- If we now perform an Eigen-Decomposition of Q, we will learn about the spread of the Variance like we did before
- If we order the Eigenvectors according to their Eigenvalue, we will see that along the first eigenvector, the most variance is captured, along the second, the second most, and so forth
- The eigenvectors are called the **principal components**

Step 1: Center Data



Step 2: Eigendecomposition of Q



➢ The Eigenvectors are:

$$e_1 = \begin{pmatrix} 0.45 \\ 0.89 \end{pmatrix} \qquad e_2 = \begin{pmatrix} -0.89 \\ 0.45 \end{pmatrix}$$

The Eigenbasis

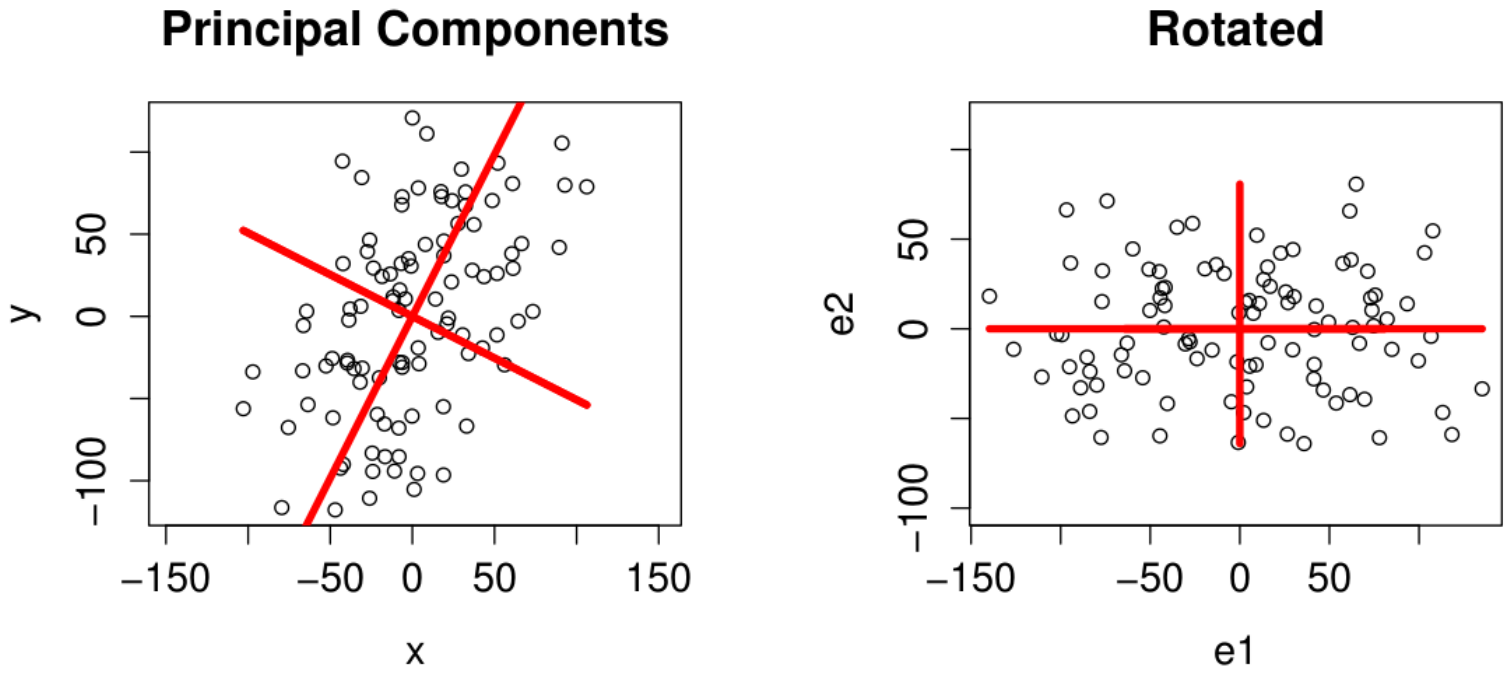
- The eigenvectors form a basis of the space
- **The eigenvectors are normalized**, i.e., $\|e_i\| = 1$

- That means, each vector x_j can be expressed by means of the eigenvectors:

$$x_j = \sum_{i=1}^{d} g_{ij} e_i$$

- The scalars are now the coordinates with respect to the eigenspace

Step 3: Rotate Data



➢ The rotation can be calculated by

$$R = E^T X^T$$

with E being the matrix of the eigenvectors

What is this good for?

- Fantastic, all that stuff in order to rotate some data?
- Expressing X in terms of $e_1, \ldots, e_d$ has not changed the size of the data at all, we just performed a basis transformation

- **BUT**: Hopefully, most of the new coordinates have values close to zero (as there is almost no variance " left" when calculating the PCAs)
- That means in turn, the data lie in a lower-dimensional linear subspace ...
- Thus, we don't use all of the eigenvectors to transform the data
- Which ones should we take?

Lower Dimensional Projection

➢ Let λ_i be the eigenvalue to the eigenvector e_i. Further, the eigenvalues are sorted, s.t.:

$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$

➢ Assume $\lambda_i \approx 0 \ \forall \ i > k$, then

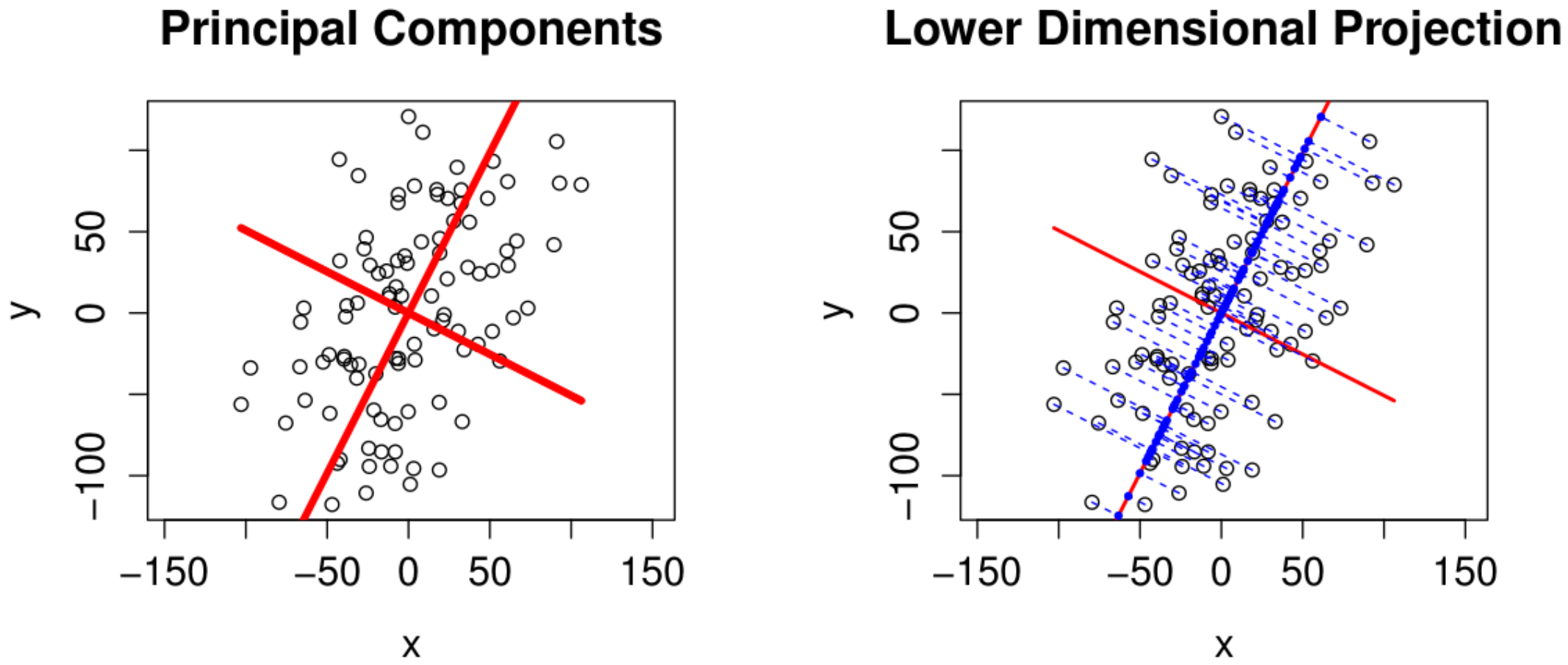
$$x_j \approx \sum_{i=1}^{k} g_{ij} e_i$$

➢ That means, we can represent x_j with $k < d$ components

➢ Variance captured in the lower dimensional eigenspace:

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{d} \lambda_i}$$

Lower Dimensional Projection



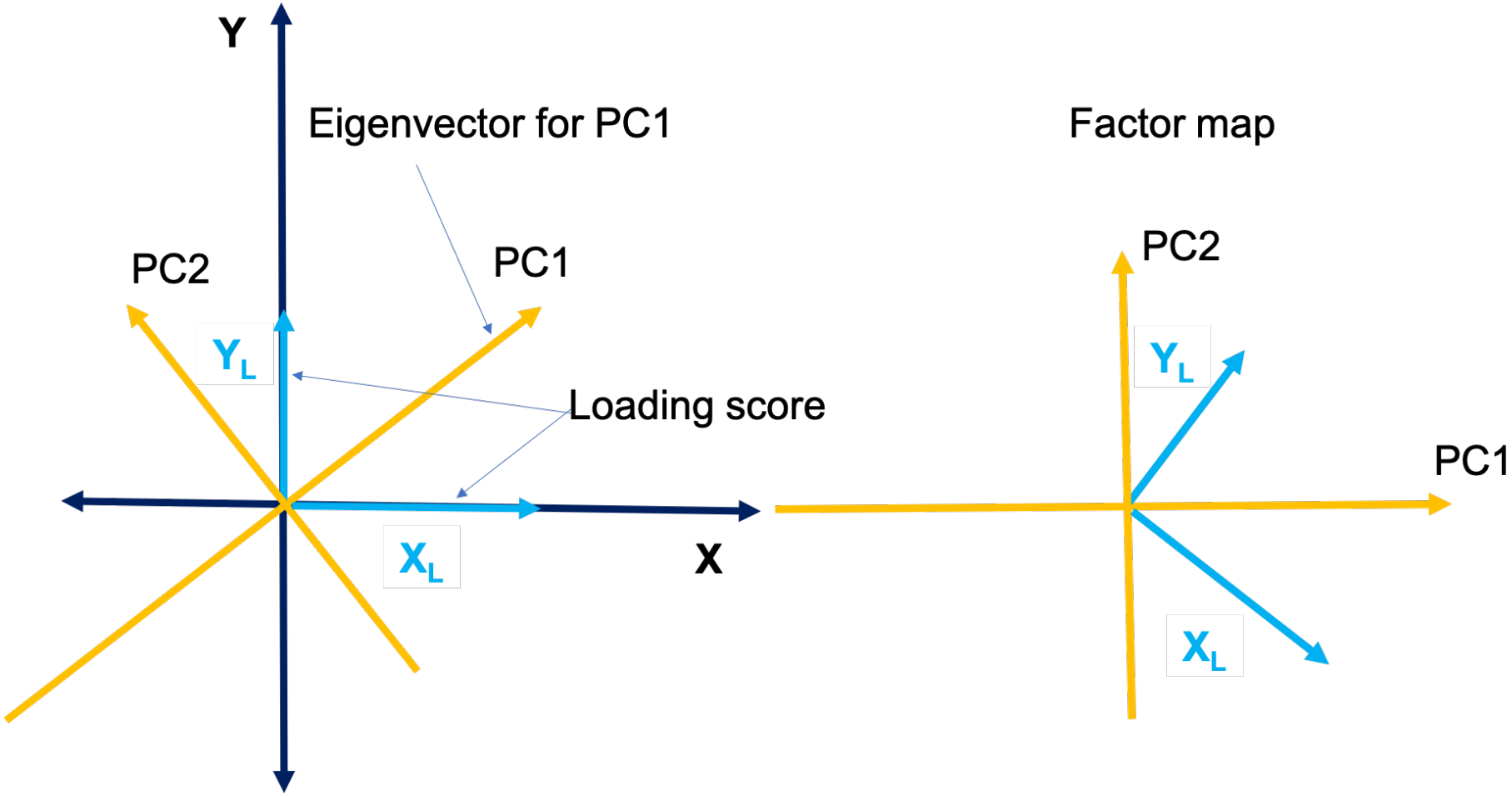
Interpretation of the Eigenvectors

➢ Each eigenvector is a linear combination of the original features $X_1, \ldots, X_d$:

$$e_i = \begin{pmatrix} \phi_{i1} \\ \phi_{i2} \\ \vdots \\ \phi_{id} \end{pmatrix} = \phi_{i1} X_1 + \phi_{i2} X_2 + \ldots + \phi_{id} X_d$$

➢ ϕ_{ij} Are called the loadings of eigenvector e_i

➢ Having a large loading means that the original feature has a great influence on the projection along this eigenvector


Y
Eigenvector for PC1
PC2
PC1
Y_L
Loading score
X_L
X
Factor map
PC2
Y_L
PC1
X_L

Keywords

- Linear algebra
- Probabilities & Standard Distributions
- Eigendecomposition & PCA