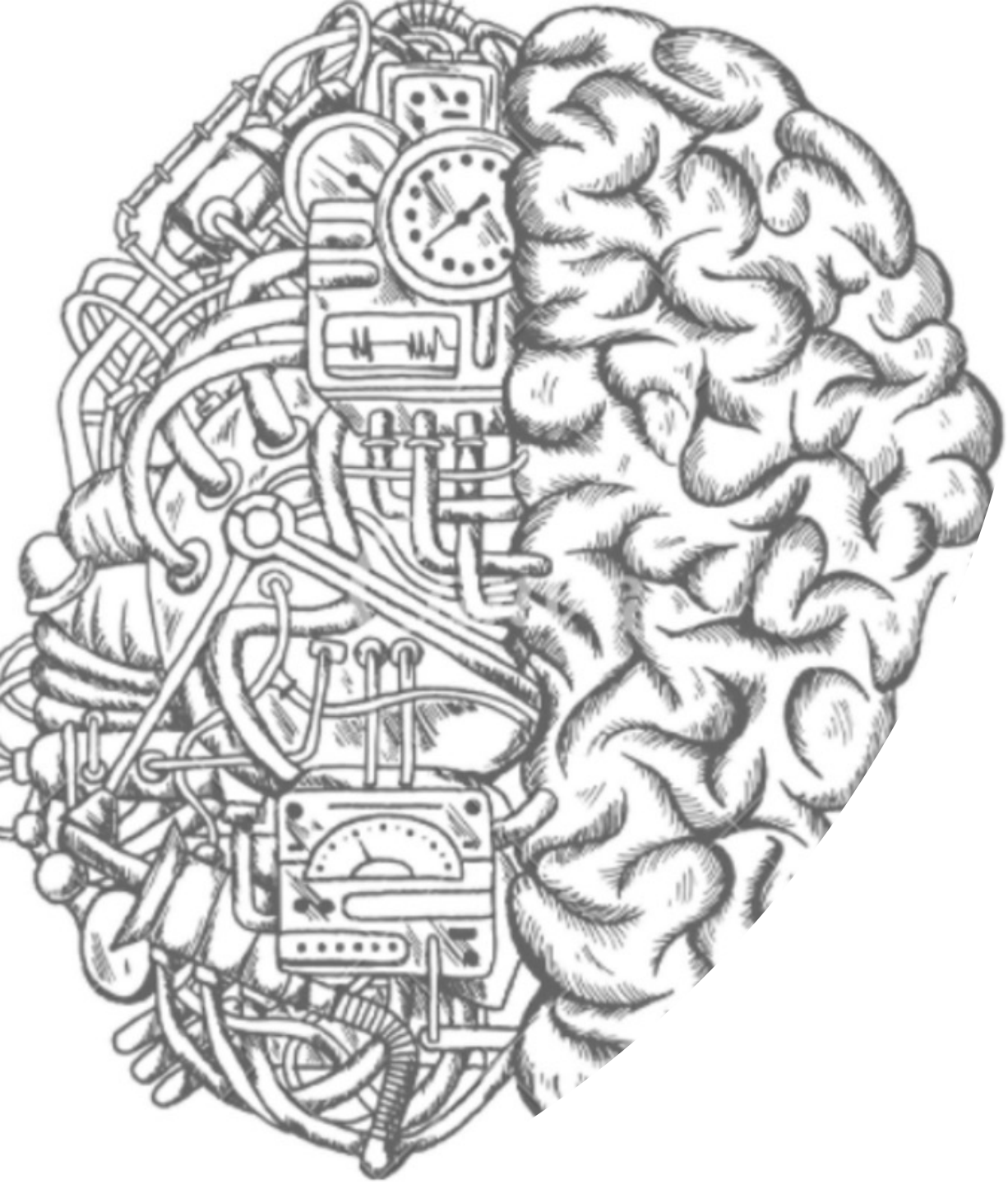


# Machine Learning: Models and Applications

## Lecture 2

Lecturer: Xinchao Wang

xinchao@nus.edu.sg

National University of Singapore

# Overview

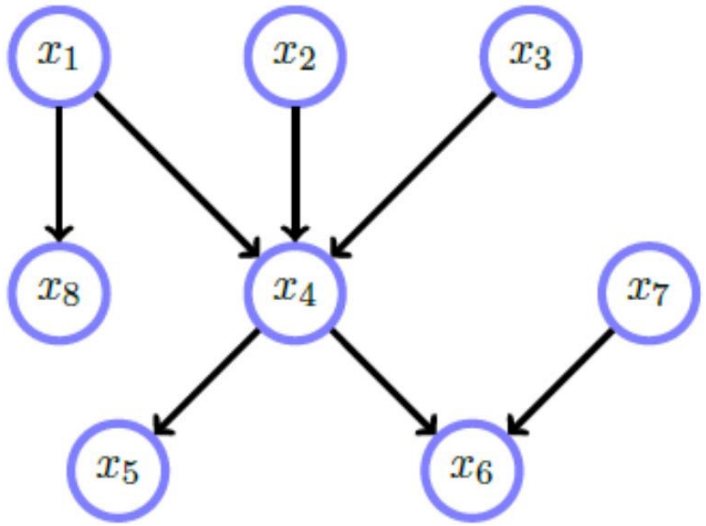
- Introduction to Graphical Model
- Belief Networks
- Linear Algebra Review

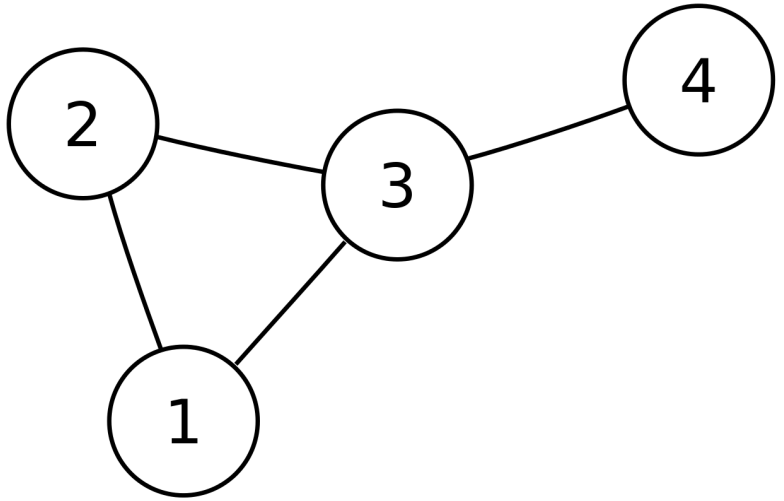
# Graphical Models

- GMs are graph-based representations of various factorization assumptions of distributions
  - These factorizations are typically equivalent to **independence** statements amongst (sets of) variables in the distribution

# Definitions

- A graph G contains:
  - Nodes, also known as vertices
  - Edges, also know as links between nodes
- Edges may be directed or undirected
  - They may also have associated weights
- Directed Graphs:
  - All edges are directed
- Undirected Graphs:
  - All edges are undirected



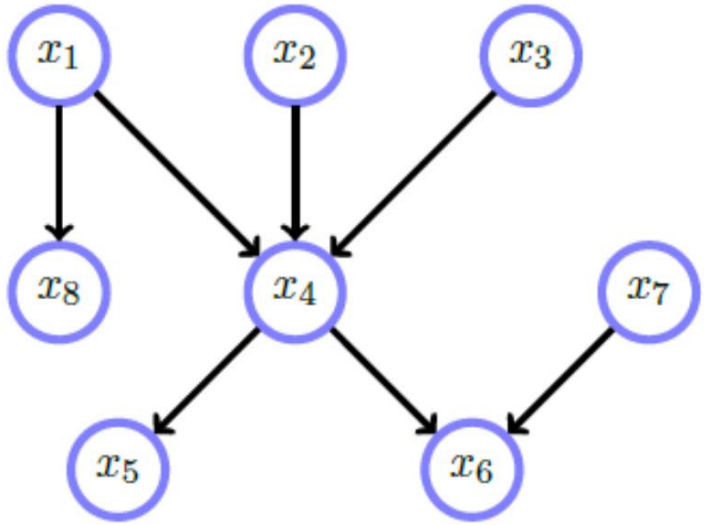
# More Definitions

- Path:
  - The Path A->B from node A to node B is a sequence of nodes that connects A to B
- Cycle:
  - A cycle is a directed path that starts and returns to the same node
- Directed Acyclic Graph (DAG):
  - A DAG is a graph G with directed edges (arrows on each link) between the nodes such that by following a path of nodes from one node to another along the direction of each edge, no path will revisit a node

# More Definitions

- Parent
  - Parent of $x_4$ are $pa(x_4) = \{x_1, x_2, x_3\}$
- Children
  - Children of $x_4$ are $ch(x_4) = \{x_5, x_6\}$

- Graphs can be represented using:
  - The edge list L={(1,8),(1,4),(2,4) …}
  - The adjacency matrix

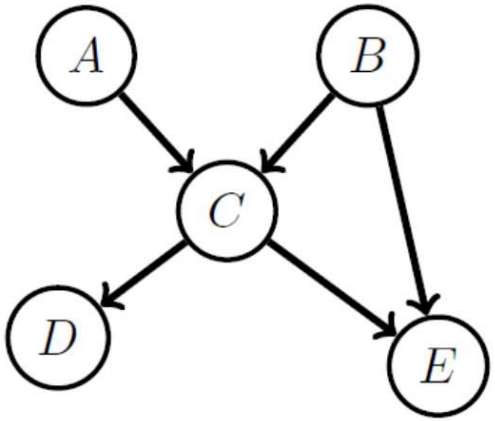


# Belief Networks

# Belief Networks (Bayesian Networks)

- A belief network is a **directed acyclic graph** in which each **node** has associated the **conditional probability** of the node given its parents
- The joint distribution is obtained by taking the product of the conditional probabilities:

$$p(A, B, C, D, E) = p(A)p(B)p(C|A, B)p(D|C)p(E|B, C)$$



# Belief Networks (Bayesian Networks)

- We study two tasks
  - Task 1: Given the graph, how to do factorization
  - Task 2: How to derive the graph at all

# Task 1: Given the graph, how to do factorization

# Task 2: How to derive the graph (Example 1)

- Sally‘s burglar **Alarm (A)** is sounding
- Has she been **Burgled (B)**, or was the alarm triggered by an **Earthquake (E)**?
- She turns the car **Radio (R)** on for news of earthquakes

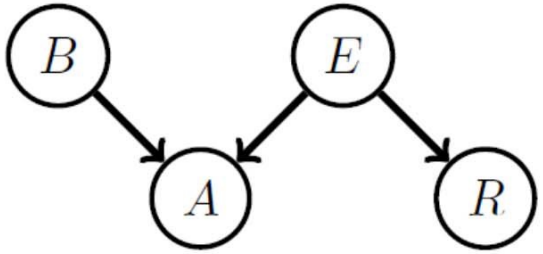
- Without loss of generality, we can write:

$$
\begin{aligned}
& p(A,R,E,B) \\
=& p(A|R,E,B)p(R,E,B) \\
=& p(A|R,E,B)p(R|E,B)p(E,B) \\
=& p(A|R,E,B)p(R|E,B)p(E|B)p(B)
\end{aligned}
$$

# Alarm Example

- Assumptions:
  - The alarm is not directly influenced by any report on the radio
    $p(A|R,E,B) = p(A|E,B)$
  - The radio broadcast is not directly influenced by the burglar variable
    $p(R|E,B) = p(R|E)$
  - Burglaries don't directly `cause' earthquakes
    $p(E|B) = p(E)$
  - Therefore, we have
    $p(A,R,E,B) = p(A|E,B)p(R|E)p(E)p(B)$

# Alarm Example




p(A,R,E,B) = p(A|E,B)p(R|E)p(E)p(B)

$p(A|B,E)$

| Alarm = 1 | Burglar | Earthquake |
|---|---|---|
| 0.9999 | 1 | 1 |
| 0.99 | 1 | 0 |
| 0.99 | 0 | 1 |
| 0.0001 | 0 | 0 |

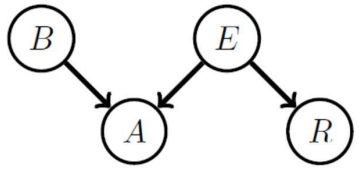
$p(R|E)$

| Radio = 1 | Earthquake |
|---|---|
| 1 | 1 |
| 0 | 0 |

We also have p(B = 1) = 0.01 and p(E = 1) = 0.000001

# Alarm Example

p(A,R,E,B) = p(A|E,B)p(R|E)p(E)p(B)




- Initial evidence:
  - the alarm is sounding

$p(A|B,E)$

| Alarm = 1 | Burglar | Earthquake |
|---|---|---|
| 0.9999 | 1 | 1 |
| 0.99 | 1 | 0 |
| 0.99 | 0 | 1 |
| 0.0001 | 0 | 0 |

$p(R|E)$

| Radio = 1 | Earthquake |
|---|---|
| 1 | 1 |
| 0 | 0 |

We also have p(B = 1) = 0.01 and p(E = 1) = 0.000001

$$p(B=1|A=1) = \frac{\sum_{E,R} p(B=1,E,A=1,R)}{\sum_{B,E,R} p(B,E,A=1,R)}$$

$$= \frac{\sum_{E,R} p(A=1|B=1,E)p(B=1)p(E)p(R|E)}{\sum_{B,E,R} p(A=1|B,E)p(B)p(E)p(R|E)} \approx 0.99$$

# Alarm Example: Inference

- Additional evidence: the radio broadcasts an earthquake warning
- A similar calculation gives
  - $p(B = 1 \mid A = 1, R = 1) \approx 0.01$
- Initially, because the alarm sounds, Sally thinks that she‘s been burgled. However, this probability drops dramatically when she hears that there has been an earthquake.
- The earthquake `explains away‘ to an extent the fact that the alarm is ringing

# Task 2: How to derive the graph (Example 2)

- One morning Tracey leaves her house and realizes that her grass is wet (T).
- Is it due to overnight Rain (R) or did she forget to turn off the sprinkler (S) last night?
- Next she notices that the grass of her neighbor, Jack, is also wet (J). This explains away to some extent the possibility that her sprinkler was left on, and she concludes therefore that it has probably been raining.

- $R \in \{0, 1\}$, R = 1 means that it has been raining, and 0 otherwise
- $S \in \{0, 1\}$, S = 1 means that Tracey has forgotten to turn off the sprinkler, and 0 otherwise
- $J \in \{0, 1\}$, J = 1 means that Jack's grass is wet, and 0 otherwise
- $T \in \{0, 1\}$, T = 1 means that Tracey's Grass is wet, and 0 otherwise

# Wet Grass Example

- The number of values that need to be specified in general scales exponentially with the number of variables in the model
  - This is impractical in general and motivates simplifications

- Conditional independence:
  - $p(T|J,R,S) = p(T|R,S)$
  - $p(J|R,S) = p(J|R)$
  - $p(R|S) = p(R)$

$R \in \{0, 1\}$, R = 1 means that it has been raining
$S \in \{0, 1\}$, S = 1 means that Tracey has forgotten to turn off the sprinkler
$J \in \{0, 1\}$, J = 1 means that Jack‘s grass is wet
$T \in \{0, 1\}$, T = 1 means that Tracey's Grass is wet

# Wet Grass Example

- Original equation

$p(T,J,R,S) = p(T|J,R,S)p(J,R,S) = p(T|J,R,S)p(J|R,S)p(R,S)$

$= p(T|J,R,S)p(J|R,S)p(R|S)p(S)$

- Now becomes

$p(T,J,R,S) = p(T|R,S)p(J|R)p(R)p(S)$

$p(T|J,R,S) = p(T|R,S)$
$p(J|R,S) = p(J|R)$
$p(R|S) = p(R)$

$R \in \{0, 1\}$, R = 1 means that it has been raining
$S \in \{0, 1\}$, S = 1 means that Tracey has forgotten to turn off the sprinkler
$J \in \{0, 1\}$, J = 1 means that Jack‘s grass is wet
$T \in \{0, 1\}$, T = 1 means that Tracey's Grass is wet

# Wet Grass Example

- $p(R = 1) = 0.2$ and $p(S = 1) = 0.1$
- $p(J = 1 | R = 1) = 1$, $p(J = 1 | R = 0) = 0.2$ (sometimes Jack‘s grass is wet due to unknown effects, other than rain)
- $p(T = 1 | R = 1, S = 0) = 1$,
- $p(T = 1 | R = 1, S = 1) = 1$,
- $p(T = 1 | R = 0, S = 1) = 0.9$ (there‘s a small chance that even though the sprinkler was left on, it didn’t wet the grass noticeably)
- $p(T = 1 | R = 0, S = 0) = 0$

# Wet Grass Example

$$
\begin{aligned}
p(S=1|T=1) &= \frac{p(S=1,T=1)}{p(T=1)} = \frac{\sum_{J,R} p(T=1,J,R,S=1)}{\sum_{J,R,S} p(T=1,J,R,S)} \\
&= \frac{\sum_{J,R} p(J|R)p(T=1|R,S=1)p(R)p(S=1)}{\sum_{J,R,S} p(J|R)p(T=1|R,S)p(R)p(S)} \\
&= \frac{\sum_{R} p(T=1|R,S=1)p(R)p(S=1)}{\sum_{R,S} p(T=1|R,S)p(R)p(S)} \\
&= \frac{0.9 \times 0.8 \times 0.1 + 1 \times 0.2 \times 0.1}{0.9 \times 0.8 \times 0.1 + 1 \times 0.2 \times 0.1 + 0 \times 0.8 \times 0.9 + 1 \times 0.2 \times 0.9} = 0.3382
\end{aligned}
$$

p(T,J,R,S) = p(T|R,S)p(J|R)p(R)p(S)
p(T|J,R,S) = p(T|R,S)
p(J|R,S) = p(J|R)
p(R|S) = p(R)

# Wet Grass Example

$$
\begin{aligned}
p(S=1|T=1,J=1) &= \frac{p(S=1,T=1,J=1)}{p(T=1,J=1)} \\
&= \frac{\sum_R p(T=1,J=1,R,S=1)}{\sum_{R,S} p(T=1,J=1,R,S)} \\
&= \frac{\sum_R p(J=1|R)p(T=1|R,S=1)p(R)p(S=1)}{\sum_{R,S} p(J=1|R)p(T=1|R,S)p(R)p(S)} \\
&= \frac{0.0344}{0.2144} = 0.1604
\end{aligned}
$$

# Intro to Linear Algebra

# Vector Space

- A vector space is a collection of objects called ***vectors***, which may be added together and multiplied ("scaled") by numbers, called ***scalars***.
- A vector space over a field $F$ (such real numbers) is a set $V$ together with **two** operations that satisfy the **eight** axioms listed below.
- Two Operations
  - **Vector addition** or simply **addition** $+ : V \times V \rightarrow V$, takes any two vectors **v** and **w** and assigns to them a third vector which is commonly written as **v** + **w**, and called the sum of these two vectors. (Note that the resultant vector is also an element of the set $V$ ).
  - **Scalar multiplication** $\cdot : F \times V \rightarrow V$, takes any scalar $a$ and any vector **v** and gives another vector $a$**v**. (Similarly, the vector $a$**v** is an element of the set $V$ ).

# Vector Space

## The eight axioms

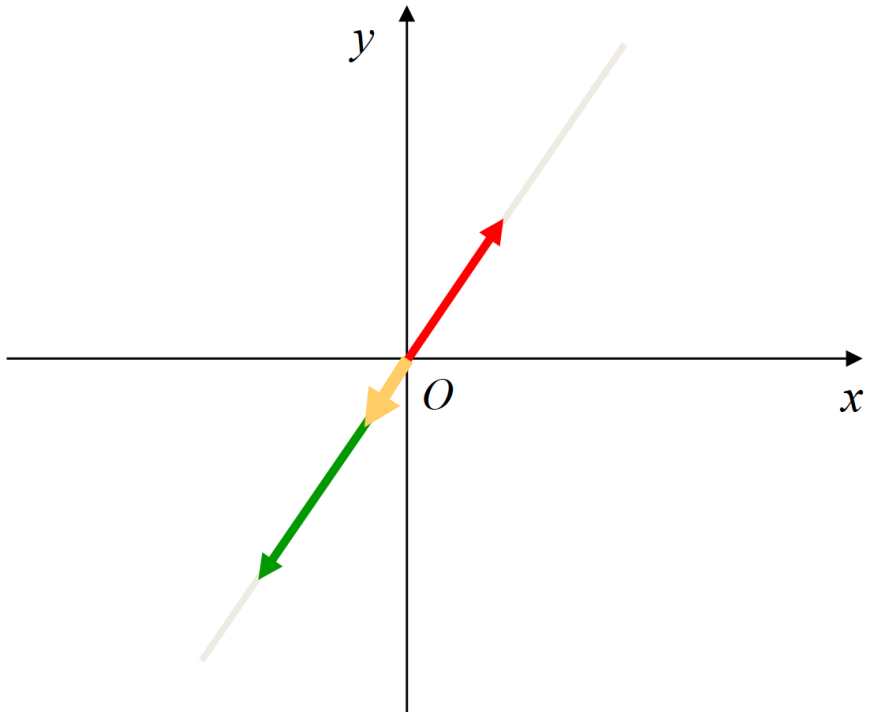
| | |
|---|---|
| Associativity of addition | $u + (v + w) = (u + v) + w$ |
| Commutativity of addition | $u + v = v + u$ |
| Identity element of addition | There exists an element $0 \in V$, called the zero vector, such that $v + 0 = v$ for all $v \in V$. |
| Inverse elements of addition | For every $v \in V$, there exists an element $-v \in V$, called the additive inverse of v, such that $v + (-v) = 0$. |
| Compatibility of scalar multiplication with field multiplication | $a(bv) = (ab)v$ |
| Identity element of scalar multiplication | $1v = v$, where 1 denotes the multiplicative identity in F. |
| Distributivity of scalar multiplication with respect to vector addition | $a(u + v) = au + av$ |
| Distributivity of scalar multiplication with respect to field addition | $(a + b)v = av + bv$ |

# Subspace

- Let *F* be a field, *V* be a vector space over *F*, and let *W* be a subset of *V*. Then *W* is a **subspace** if:
  - The zero vector, **0**, is in *W*.
  - If **u** and **v** are elements of *W*, then the sum **u** + **v** is an element of *W*.
  - If **u** is an element of *W* and *c* is a scalar from *K*, then the scalar product *c***u** is an element of *W*.

# Subspace Example

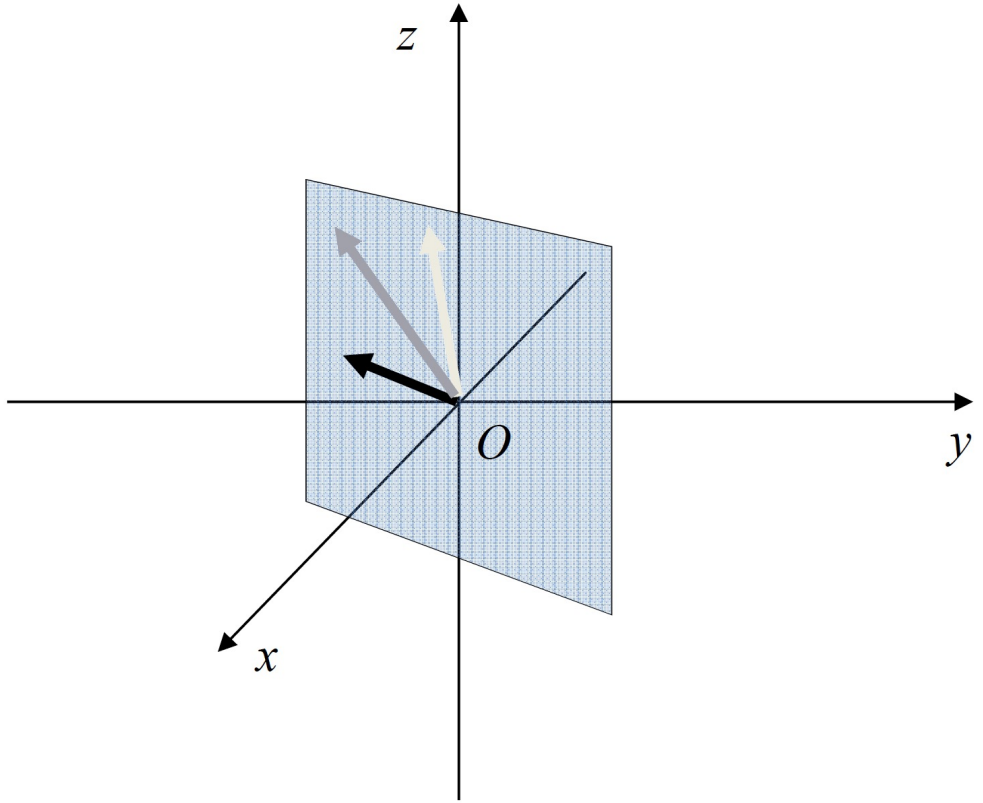
- Let *l* be a 2D line though the origin
- $L = \{\mathbf{p} - O \mid \mathbf{p} \in l\}$ is a linear subspace of $R^2$




O. Sorkine, 2006

# Subspace Example

- Let $\pi$ be a plane through the origin in 3D
- $V = \{\mathbf{p} - O \mid \mathbf{p} \in \pi\}$ is a linear subspace of $R^3$



# Linear Independence

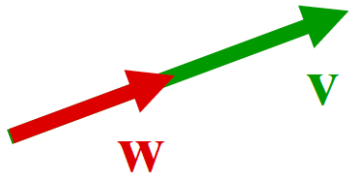
- The vectors $\{\mathbf{v}_1,\ldots, \mathbf{v}_k\}$ are a linearly independent set if:

$$\alpha_1\mathbf{v}_1 + \cdots + \alpha_k\mathbf{v}_k = \mathbf{0} \Leftrightarrow \alpha_i = 0 \; \forall i$$

- It means that none of the vectors can be obtained as a linear combination of the others.

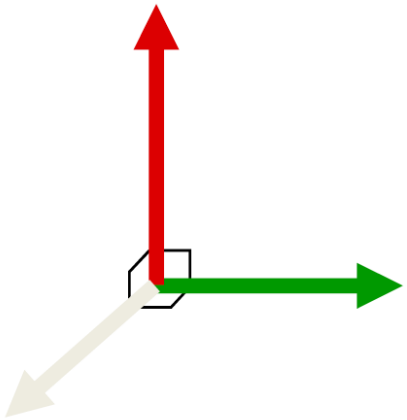
# Linear independence - example

- Parallel vectors are always dependent:




$$\mathbf{v} = 2.4\,\mathbf{w} \quad \Rightarrow \quad \mathbf{v} + (-2.4)\mathbf{w} = 0$$

- Orthogonal vectors are always linearly independent:



# Basis of Vector Space

- $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ are **linear independent**
- $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ **span** the whole vector space $\mathbf{V}$

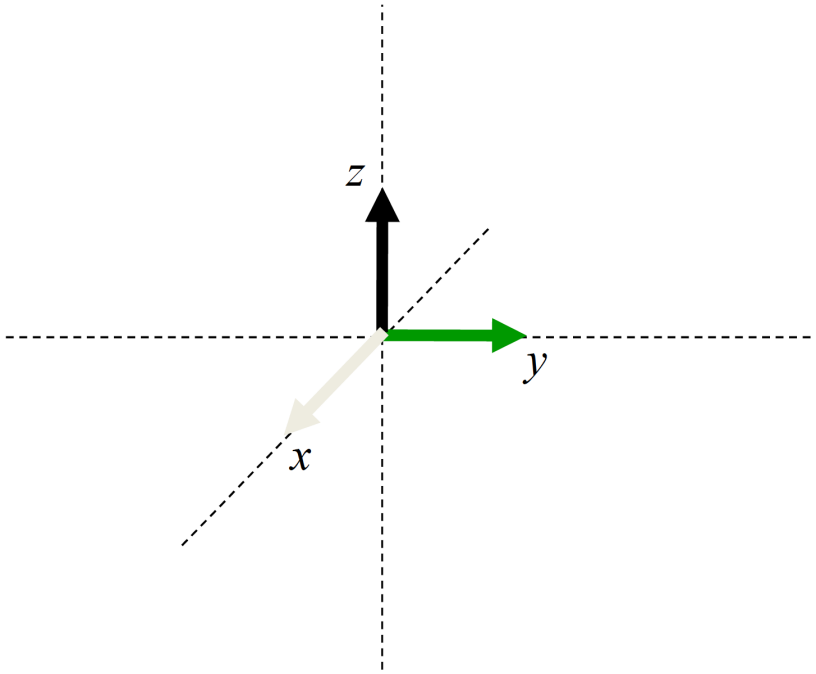
$$\mathbf{V} = \{\alpha_1 \mathbf{v}_1 + \cdots + \alpha_k \mathbf{v}_k | \alpha_1 \in R\}$$

- Any vector in $\mathbf{V}$ is a **unique** linear combination of the basis

- The number of basis vectors is called the **dimension** of $\mathbf{V}$

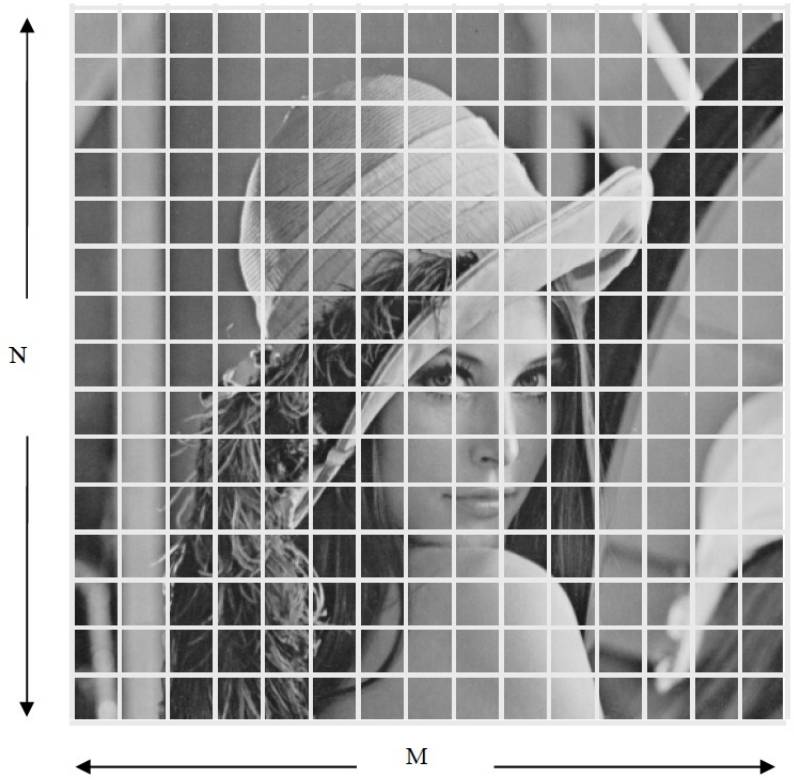
# Basis Example

- The standard basis of $R^3$

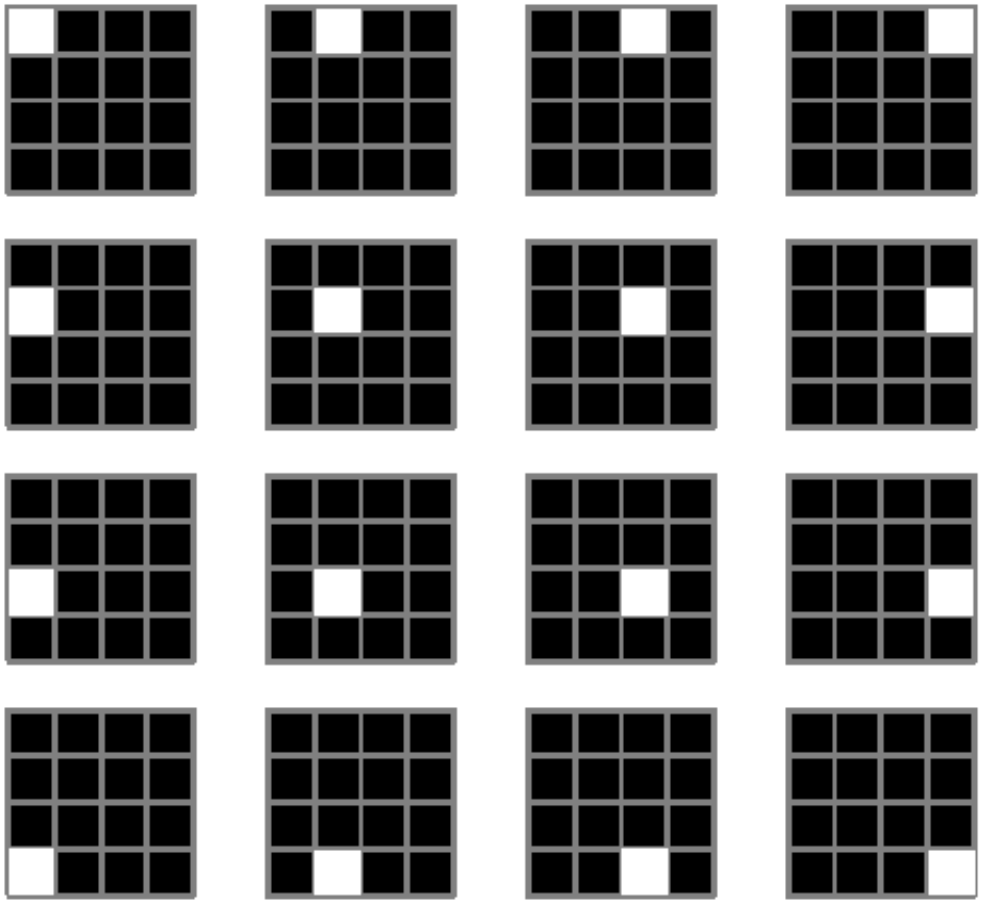


# Basis – Another Example

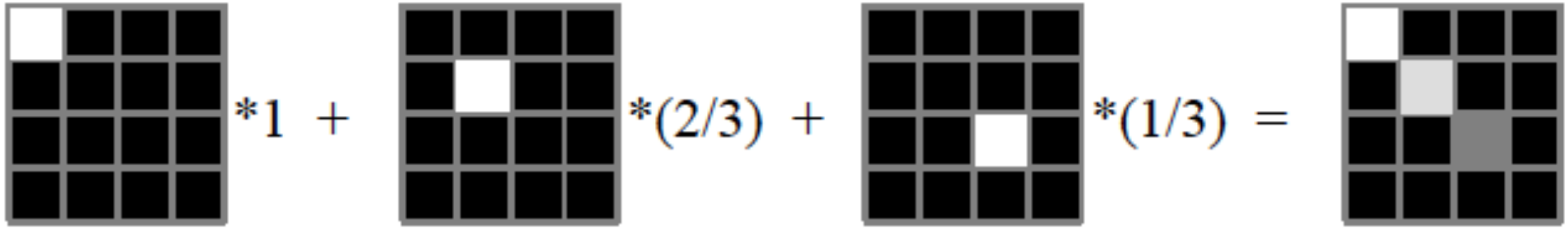
- Grayscale NxM images:
  - Each pixel has value between 0 (black) and 1 (white)
  - The image can be interpreted as a vector



# The “Standard” Basis of a 4x4 Image



# Linear combinations of the basis



# Matrix Representation

- Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ be a basis
- Every $\mathbf{v}$ can be uniquely represented as:

$$\mathbf{v} = \alpha_1 \mathbf{v}_1 + \cdots + \alpha_k \mathbf{v}_k$$

- Denote $\mathbf{v}$ by the column-vector: $\mathbf{v} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_k \end{pmatrix}$
- Denote the basis vectors as: $\begin{pmatrix} \mathbf{1} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \mathbf{1} \\ \vdots \\ \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{1} \end{pmatrix}$

# Linear Operators

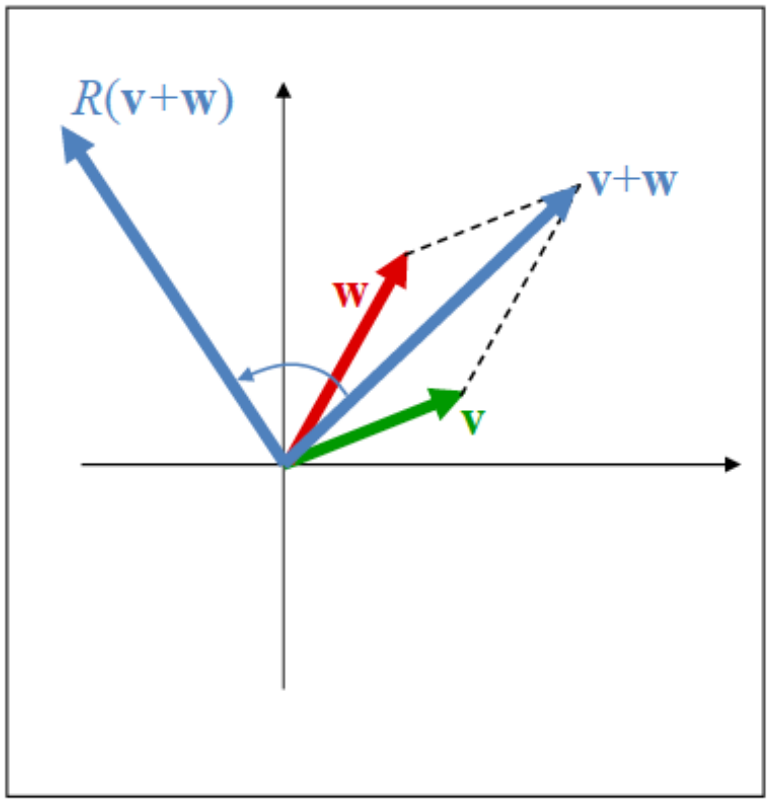
- $A$ : V → W is called linear operator if:
  - $A(\mathbf{v} + \mathbf{w}) = A(\mathbf{v}) + A(\mathbf{w})$
  - $A(\alpha\mathbf{v}) = \alpha A(\mathbf{v})$

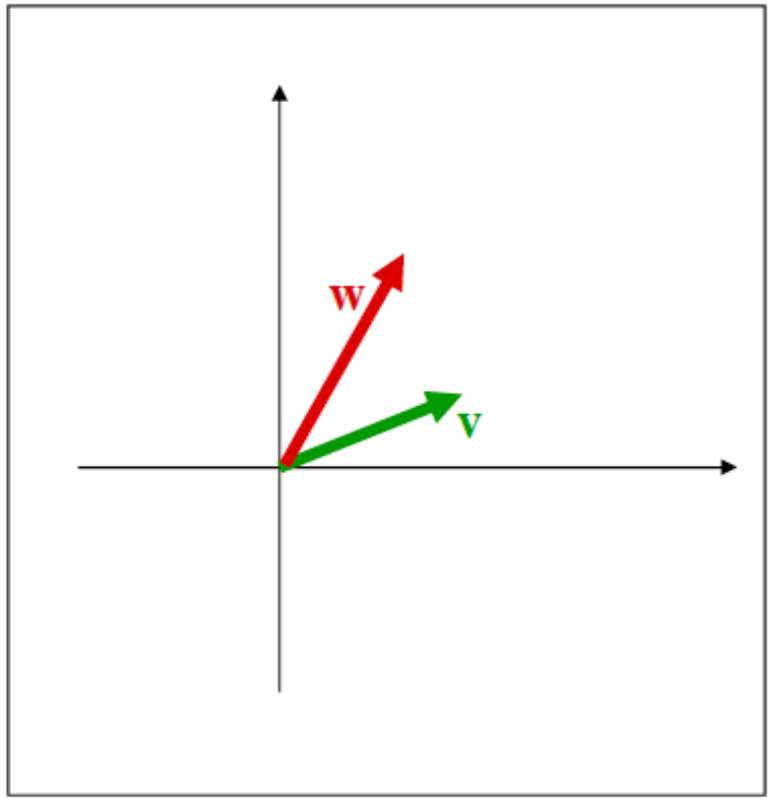
- In particular, $A(\mathbf{0}) = 0$

- Are the following operators linear?
  - Scaling
  - Rotation
  - Translation

# Linear operators - illustration

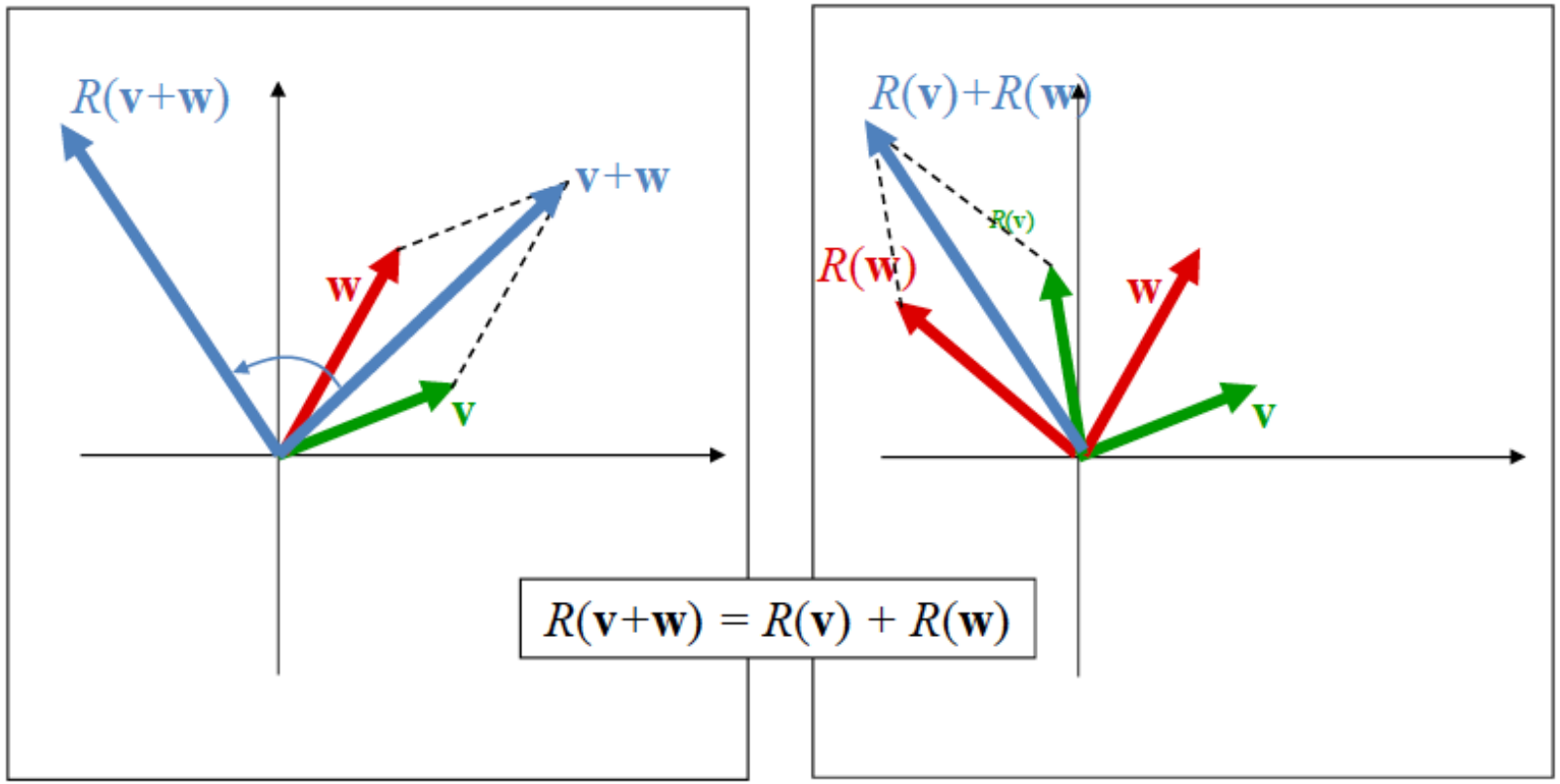
- Rotation is a linear operator:



# Linear operators - illustration

- Rotation is a linear operator:



# Matrix Operations

- Addition, subtraction, scalar multiplication
- Multiplication of matrix by column vector:

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} \sum_i a_{1i} b_i \\ \vdots \\ \sum_i a_{mi} b_i \end{pmatrix} = \begin{pmatrix} < \text{row}_1, \mathbf{b} > \\ \vdots \\ < \text{row}_m, \mathbf{b} > \end{pmatrix}$$

$A$ $\mathbf{b}$

# Matrix by Vector Multiplication

- Sometimes a better way to look at it:
  - *A***b** is a linear combination of *A*'s columns!

$$\begin{pmatrix} | & | & & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \\ | & | & & | \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = b_1 \begin{pmatrix} | \\ \mathbf{a}_1 \\ | \end{pmatrix} + b_2 \begin{pmatrix} | \\ \mathbf{a}_2 \\ | \end{pmatrix} + \ldots + b_n \begin{pmatrix} | \\ \mathbf{a}_n \\ | \end{pmatrix}$$

# Matrix operations

- Transposition: make the rows to be the columns

- $(AB)^T = B^T A^T$

$$A\begin{bmatrix}1 & 2 & 3\\ 4 & 5 & 6\\ 7 & 8 & 9\end{bmatrix} \quad A^T\begin{bmatrix}1 & 4 & 7\\ 2 & 5 & 8\\ 3 & 6 & 9\end{bmatrix}$$

$$A\begin{bmatrix}1 & 4 & 3\\ 8 & 2 & 6\\ 7 & 8 & 3\\ 4 & 9 & 6\\ 7 & 8 & 1\end{bmatrix} \quad A^T\begin{bmatrix}1 & 8 & 7 & 4 & 7\\ 4 & 2 & 8 & 9 & 8\\ 3 & 6 & 3 & 6 & 1\end{bmatrix}$$

# Matrix properties

- Matrix *A* (*nxn*) is **non-singular** if *B, AB = BA = I*
- *B = inv(A)* is called the **inverse** of *A*
- *A* is singular if *det (A) =0*

- If *A* is non-singular then the equation
  - *A***x**=**b** has one unique solution for each **b**
  - the rows of *A* are linearly independent (and so are the columns)

# Orthogonal matrices

- Matrix $A$ $(n \times n)$ is orthogonal if $inv(A) = A^T$
- Follows: $AA^T = A^TA = I$
- The rows of $A$ are orthonormal vectors!

  Proof:

$$I = A^TA = \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \\ \mathbf{v}_n \end{pmatrix} \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & & \mathbf{v}_n \end{pmatrix} = \begin{pmatrix} \mathbf{v}_i{}^T\mathbf{v}_j \end{pmatrix} = \begin{pmatrix} V_{ij} \end{pmatrix}$$

$$<\mathbf{v}_i, \mathbf{v}_i> = 1, \quad \|\mathbf{v}_i\| = 1; \quad <\mathbf{v}_i, \mathbf{v}_j> = 0$$

# Trace

- The trace of a square matrix denoted by tr($A$) is sum of the diagonal elements

$$\mathrm{tr}(A) = \sum_{i=1}^{n} A_{ii}$$

# Determinant

- For a square matrix $A$, the determinant is denoted by $|A|$ or $\det(A)$

$$
\begin{aligned}
|A| &= \sum_{i=1}^{n} (-1)^{i+j} a_{ij} |A_{\setminus i, \setminus j}| \quad (\text{for any } j \in 1, \ldots, n) \\
&= \sum_{j=1}^{n} (-1)^{i+j} a_{ij} |A_{\setminus i, \setminus j}| \quad (\text{for any } i \in 1, \ldots, n)
\end{aligned}
$$

$$
\begin{aligned}
|[a_{11}]| &= a_{11} \\
\left| \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \right| &= a_{11}a_{22} - a_{12}a_{21} \\
\left| \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \right| &= \begin{array}{l} a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ \quad -a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31} \end{array}
\end{aligned}
$$

# Determinant

- $|A| = |A^T|$
- $|AB| = |A|\ |B|$
- $|A| = 0$, if and only if $A$ is singular
  - Else, $|A^{-1}| = 1/|A|$

# The Covariance Matrix

# Covariance

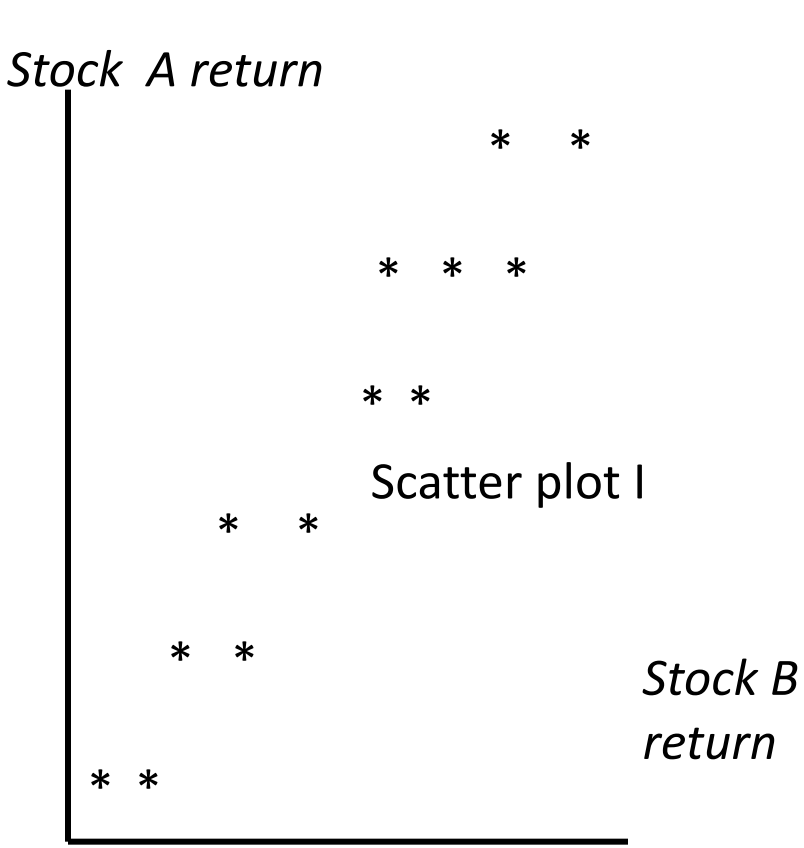
- Covariance is a numerical measure that shows how much two random variables change together

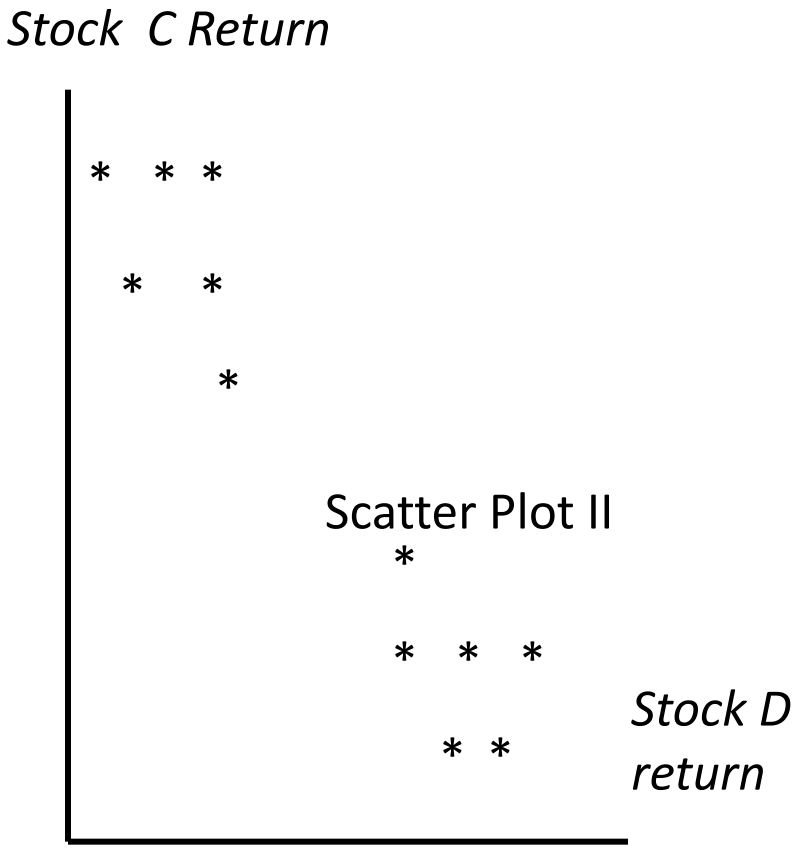
$$\sigma_{jk} = E\left[(Y_{ij} - \mu_j)(Y_{ik} - \mu_k)\right]$$

- More precisely: the covariance is a measure of the *linear* dependence between the two variables

# Covariance example

- Relationships between the returns of different stocks



# Correlation coefficient

- One may be tempted to conclude that if the covariance is larger, the relationship between two variables is stronger (in the sense that they have stronger linear relationship)

- The correlation coefficient is defined as:

$$\rho_{jk} = \frac{E\left[(Y_{ij} - \mu_j)(Y_{ik} - \mu_k)\right]}{\sigma_j \sigma_k}$$

# Correlation coefficient

- The correlation coefficient, unlike covariance, is a measure of dependence that **is free of scales** of measurement of $Y_{ij}$ and $Y_{ik}$.
- By definition, correlation must take values between −1 and 1
- A correlation of 1 or −1 is obtained when there is a perfect linear relationship between the two variables

$$\rho_{jk} = \frac{E\left[(Y_{ij} - \mu_j)(Y_{ik} - \mu_k)\right]}{\sigma_j \sigma_k}$$

# Covariance matrix

- For the vector of repeated measures, $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})$, we define the covariance matrix, $\text{Cov}(Y_i)$:

$$\text{Cov}\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{pmatrix} = \begin{pmatrix} \text{Var}(Y_{i1}) & \text{Cov}(Y_{i1}, Y_{i2}) & \cdots & \text{Cov}(Y_{i1}, Y_{in}) \\ \text{Cov}(Y_{i2}, Y_{i1}) & \text{Var}(Y_{i2}) & \cdots & \text{Cov}(Y_{i2}, Y_{in}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(Y_{in}, Y_{i1}) & \text{Cov}(Y_{in}, Y_{i2}) & \cdots & \text{Var}(Y_{in}) \end{pmatrix}$$

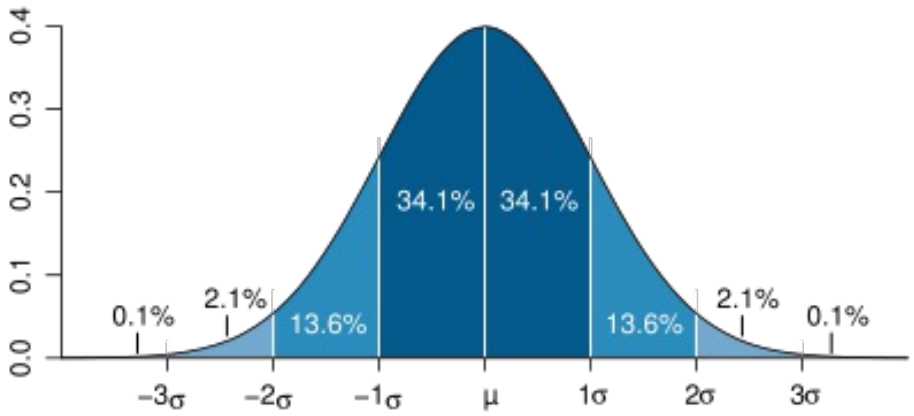
$$= \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix},$$

$$\text{where } \text{Cov}\,(Y_{ij}, Y_{ik}) = \sigma_{jk} = \sigma_{kj} = \text{Cov}(Y_{ik}, Y_{ij}).$$

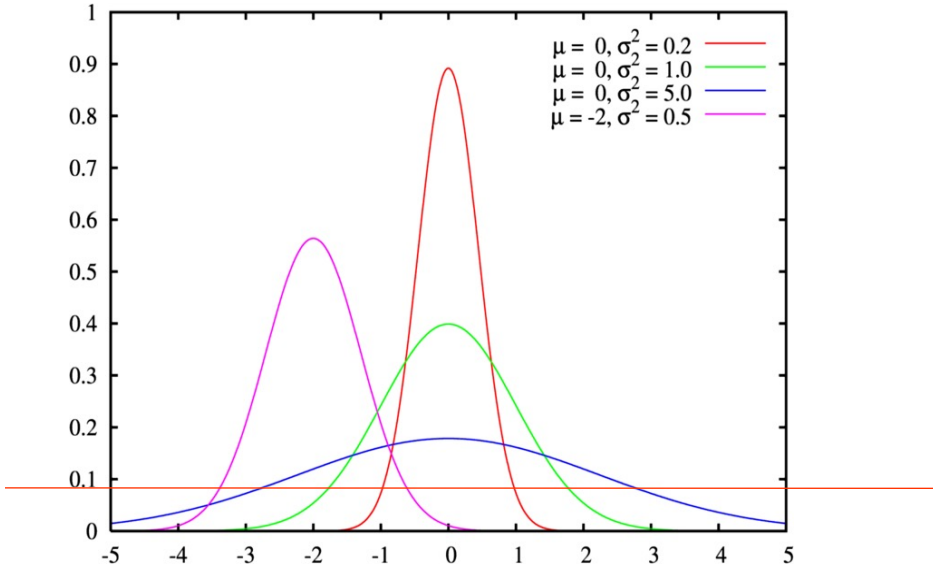
- It is a symmetric, square matrix

# Variance and Confidence Intervals

- Single Gaussian (Normal) Random Variable




$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = N(\mu, \sigma^2)$$



# Multivariate Normal Density

- The multivariate normal density in d dimensions is:

$$P(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp\left[-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)\right]$$

where:

$x = (x_1, x_2, \ldots, x_d)^t$

$\mu = (\mu_1, \mu_2, \ldots, \mu_d)^t$ mean vector

$\Sigma$ = d×d covariance matrix

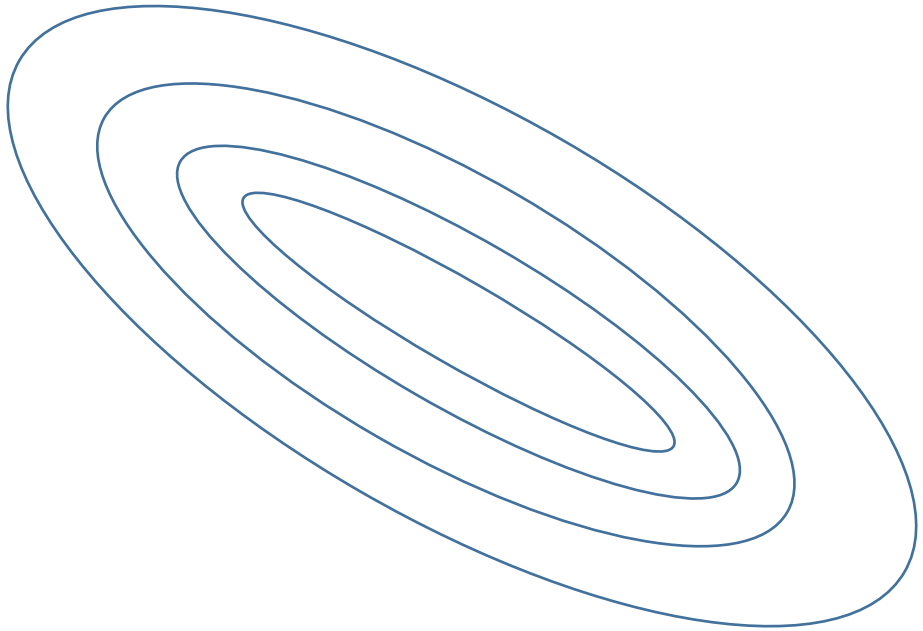
$|\Sigma|$ and $\Sigma^{-1}$ are the determinant and inverse respectively

# Confidence Intervals: Multi-Variate Case

- Same concept: how large is the area that contains X% of samples drawn from the distribution
- Confidence intervals are ellipsoids for normal distribution

# Confidence Intervals: Multi-Variate Case

- Increasing X%, increases the size of the ellipsoids, but not their orientation and aspect ratio.

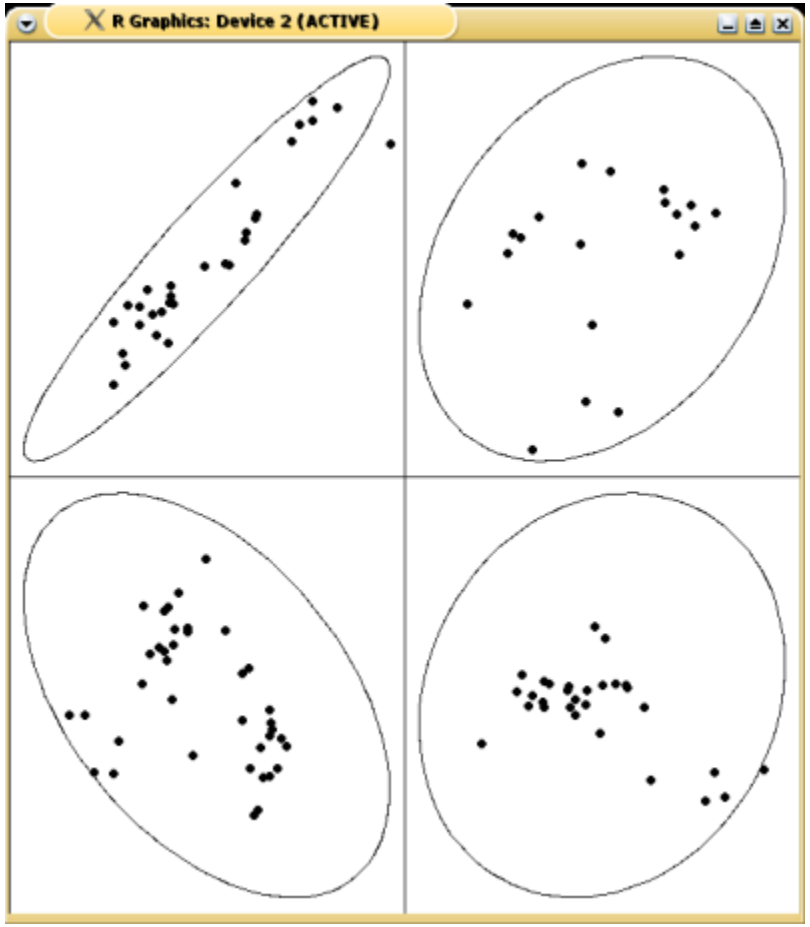


# The Multi-Variate Normal Density

- Σ is positive semi definite (**$x^t \Sigma x >= 0$, for non-zero x)**
  - If **$x^t \Sigma x = 0$ , then det(Σ)=0.** This case is not interesting, **p(x) is not defined**
    - The feature vector is a constant (has zero variance)
    - Two or more features are linearly dependent
- So we will assume Σ is positive definite (**$x^t \Sigma x > 0$)**
- If Σ is positive definite then so is $\Sigma^{-1}$

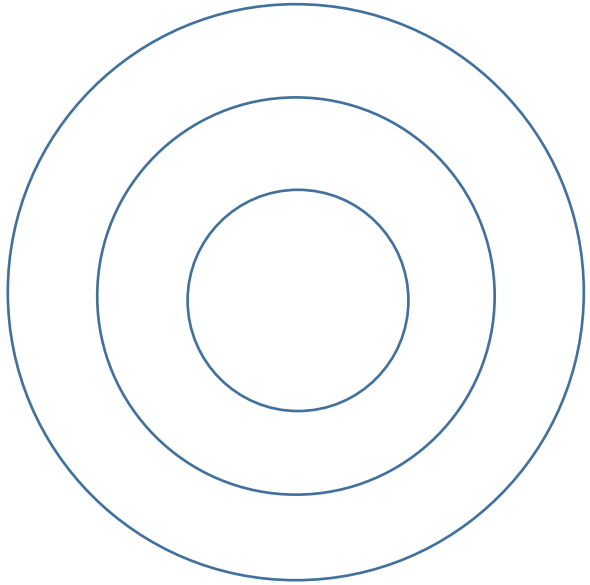
# Confidence Intervals: Multi-Variate Case

- Covariance matrix determines the shape



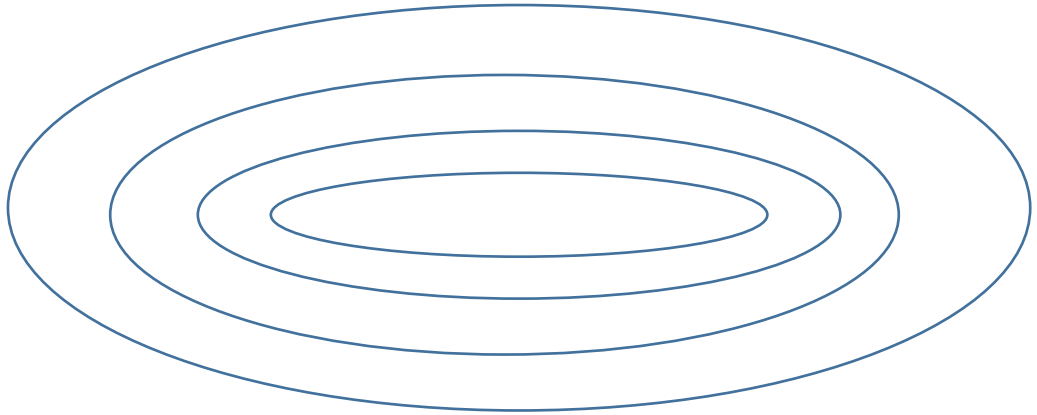
# Confidence Intervals: Multi-Variate Case

- Case I: covariance matrix is an identical matrix
  - All variables are uncorrelated and have equal variance
- Confidence intervals are circles



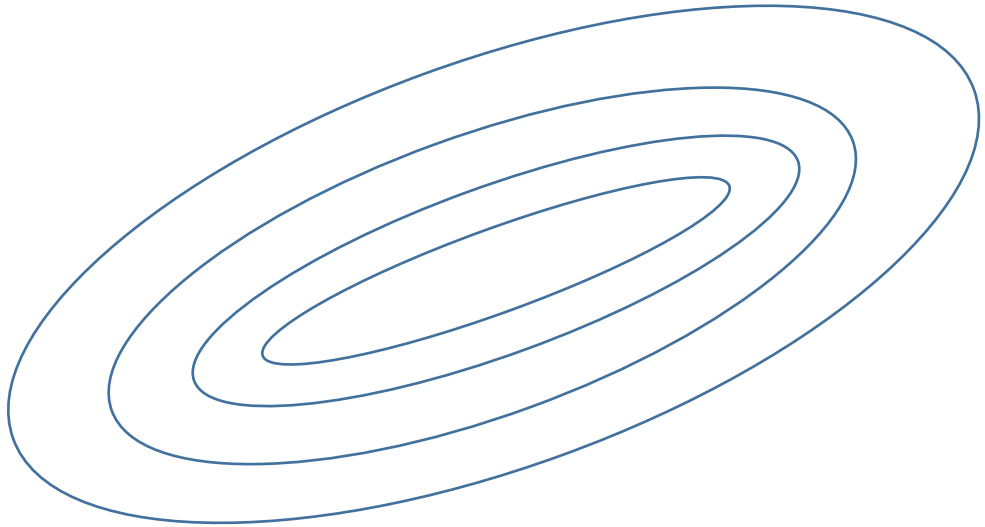
# Confidence Intervals: Multi-Variate Case

- Case II: covariance matrix is diagonal, with unequal elements
  - All variables are uncorrelated but have different variances
- Confidence intervals are axis-aligned ellipsoids



# Confidence Intervals: Multi-Variate Case

- Case III: covariance matrix is arbitrary
  - Variables may be correlated and have different variances
- Confidence intervals are arbitrary ellipsoids



# Eigenvalue and Eigenvector

# Eigenvalues and Eigenvectors

- For an n×n **square** matrix A, e is an eigenvector with eigenvalue λ if

$$Ae = \lambda e$$

- Or

$$(A- \lambda \mathbf{I})e=0$$

- If $(A- \lambda \mathbf{I})$ is invertible, the only solution is e=0 (trivial)

# Eigenvalues and Eigenvectors

$$(A- \lambda \mathbf{I})e=0$$

- For non-trivial solutions:

$$\det(A- \lambda \mathbf{I})=0$$

- Above equation is called the “characteristic polynomial”
- Solutions are not unique
  - If e is an eigenvector $\alpha e$ is also an eigenvector

# Simple Example

- For a 2×2 matrix

$$\det[\mathbf{A}-\lambda\mathbf{I}]=\begin{vmatrix} a_{11}-\lambda & a_{12} \\ a_{21} & a_{22}-\lambda \end{vmatrix}=(a_{11}-\lambda)(a_{22}-\lambda)-a_{12}a_{21}=0$$

$$0=a_{11}a_{22}-a_{12}a_{21}-\lambda(a_{11}+a_{22})+\lambda^2$$

$$\mathbf{A}=\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

# Simple Example

$$0 = a_{11}a_{22} - a_{12}a_{21} - (a_{11} + a_{22})\lambda + \lambda^2$$

$$0 = 1 \cdot 4 - 2 \cdot 2 - (1 + 4)\lambda + \lambda^2$$

$$(1 + 4)\lambda = \lambda^2$$

- The solutions are λ=0 and λ=5
- The eigenvector for the first eigenvalue, λ=0 is:

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}, \quad (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$$

$$\left[\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}\right] \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1x + 2y \\ 2x + 4y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- One solution for both equations is x=2, y=-1

# Simple Example

- The second eigenvalue is λ=5

$$\left[\begin{bmatrix}1 & 2\\ 2 & 4\end{bmatrix}-\begin{bmatrix}5 & 0\\ 0 & 5\end{bmatrix}\right]\cdot\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}-4 & 2\\ 2 & -1\end{bmatrix}\cdot\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}-4x+2y\\ 2x-1y\end{bmatrix}=\begin{bmatrix}0\\ 0\end{bmatrix}$$

$$-4x+2y=0,\ \text{and}\ 2x-y=0,\ \text{so},\ x=1, y=2$$

# Properties

- The product of the eigenvalues = |A|
- The sum of the eigenvalues = trace(A)
- The eigenvectors are pairwise orthogonal