

CS 559: Machine Learning Fundamentals and Applications

2nd Set of Notes

Instructor: Philippos Mordohai

Webpage: www.cs.stevens.edu/~mordohai

E-mail: Philippos.Mordohai@stevens.edu

Office: Lieb 215

Overview

- **Introduction to Graphical Models**
- **Belief Networks**

- **Linear Algebra Review**
 - See links on class webpage
 - Email me if you need additional resources

Example: Disease Testing

- Suppose you have been tested positive for a disease; what is the probability that you actually have the disease?
- It depends on the accuracy and sensitivity of the test, and on the background (prior) probability of the disease

Example: Disease Testing (cont.)

- Let $P(\text{Test}=+ \mid \text{Disease}=\text{true}) = 0.95$
- Then the false negative rate, $P(\text{Test}=- \mid \text{Disease}=\text{true}) = 5\%$.
- Let $P(\text{Test}=+ \mid \text{Disease}=\text{false}) = 0.05$, (the false positive rate is also 5%)
- Suppose the disease is rare: $P(\text{Disease}=\text{true}) = 0.01$

$$\begin{aligned} P(\text{Disease} = \text{true} \mid \text{Test} = +) &= \\ &= \frac{p(\text{Test} = + \mid \text{Disease} = \text{true}) P(\text{Disease} = \text{true})}{p(\text{Test} = + \mid \text{Disease} = \text{true}) P(\text{Disease} = \text{true}) + p(\text{Test} = + \mid \text{Disease} = \text{false}) P(\text{Disease} = \text{false})} = \\ &= \frac{0.95 * 0.01}{0.95 * 0.01 + 0.05 * 0.99} = 0.161 \end{aligned}$$

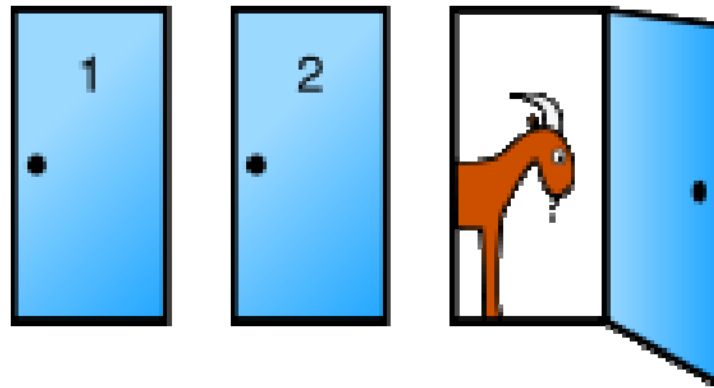
Example: Disease Testing (cont.)

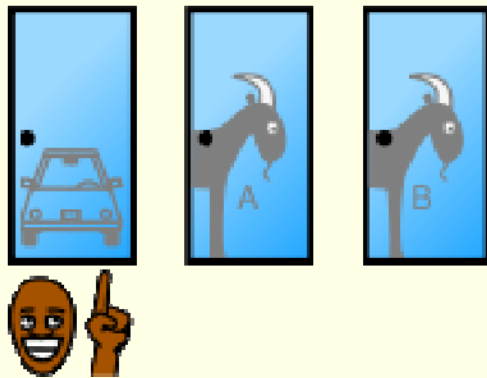
- Probability of having the disease given that you tested positive is just 16%.
 - Seems too low, but ...
- Of 100 people, we expect only 1 to have the disease, and that person will probably test positive.
- But we also expect about 5% of the others (about 5 people in total) to test positive by accident.
- So of the 6 people who test positive, we only expect 1 of them to actually have the disease; and indeed $1/6$ is approximately 0.16.

Monty Hall Problem

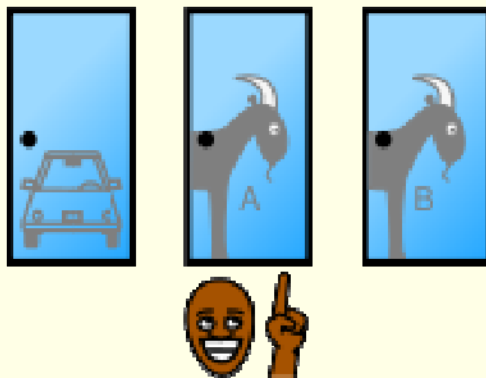
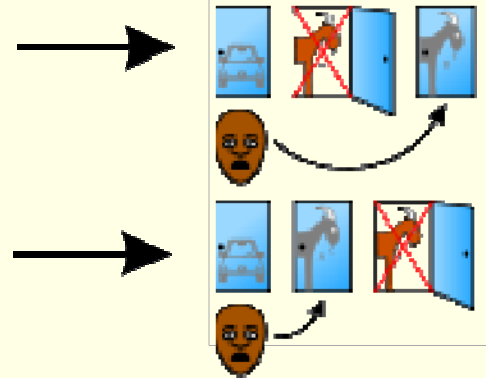
Slides by Jingrui He (CMU), 2007

- You're given the choice of three doors: Behind one door is a car; behind the others, goats.
- You pick a door, say No. 1
- The host, who knows what's behind the doors, opens another door, say No. 3, which has a goat.
- Do you want to pick door No. 2 instead?

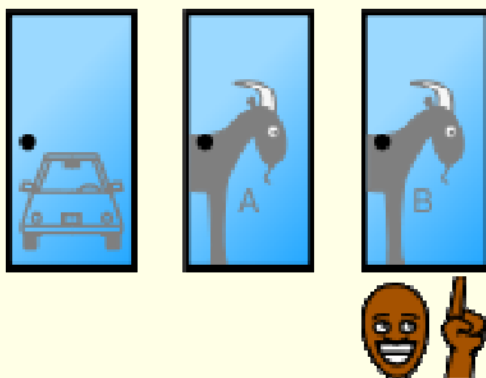




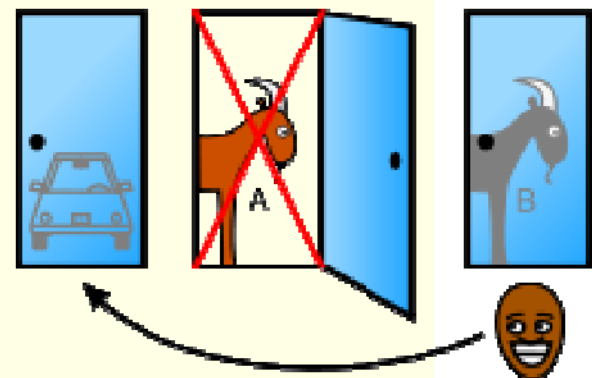
*Host reveals
Goat A
or
Host reveals
Goat B*



*Host must
reveal Goat B*



*Host must
reveal Goat A*



Monty Hall Problem: Bayes Rule

- C_i : the car is behind door i , $i = 1, 2, 3$
- $P(C_i) = 1/3$
- H_{ij} : the host opens door j after you pick door i

$$P(H_{ij} | C_k) = \begin{cases} 0 & i = j \\ 0 & j = k \\ 1/2 & i = k \\ 1 & i \neq k, j \neq k \end{cases}$$

Monty Hall Problem: Bayes Rule cont.

- WLOG, $i=1, j=3$

- $$P(C_1 | H_{13}) = \frac{P(H_{13} | C_1) P(C_1)}{P(H_{13})}$$

- $$P(H_{13} | C_1) P(C_1) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$$

Monty Hall Problem: Bayes Rule cont.

- $$\begin{aligned} P(H_{13}) &= P(H_{13}, C_1) + P(H_{13}, C_2) + P(H_{13}, C_3) \\ &= P(H_{13} | C_1)P(C_1) + P(H_{13} | C_2)P(C_2) \\ &= \frac{1}{6} + 1 \cdot \frac{1}{3} \\ &= \frac{1}{2} \end{aligned}$$
- $$P(C_1 | H_{13}) = \frac{1/6}{1/2} = \frac{1}{3}$$

Monty Hall Problem: Bayes Rule cont.

- $P(C_1 | H_{13}) = \frac{1/6}{1/2} = \frac{1}{3}$
- $P(C_2 | H_{13}) = 1 - \frac{1}{3} = \frac{2}{3} > P(C_1 | H_{13})$
- *You should switch!*

Introduction to Graphical Models

Barber Ch. 2

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
- Directed graphs model conditional distributions (e.g. Belief Networks)
- Undirected graphs represented relationships between variables (e.g. neighboring pixels in an image)

Definition

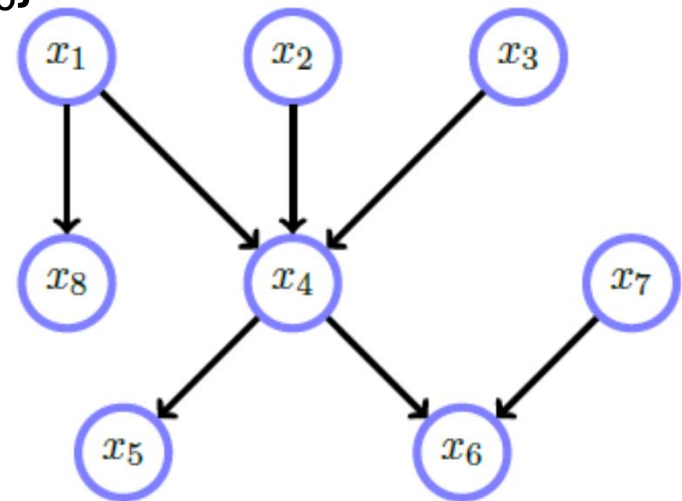
- A graph G consists of nodes (also called vertices) and edges (also called links) between the nodes
- Edges may be directed (they have an arrow in a single direction) or undirected
 - Edges can also have associated weights
- A graph with all edges directed is called a directed graph, and one with all edges undirected is called an undirected graph

More Definitions

- A **path** $A \rightarrow B$ from node A to node B is a sequence of nodes that connects A to B
- 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

- The parents of x_4 are $pa(x_4) = \{x_1, x_2, x_3\}$
- The children of x_4 are $ch(x_4) = \{x_5, x_6\}$
- Graphs can be encoded using the edge list $L = \{(1,8), (1,4), (2,4) \dots\}$ or the adjacency matrix



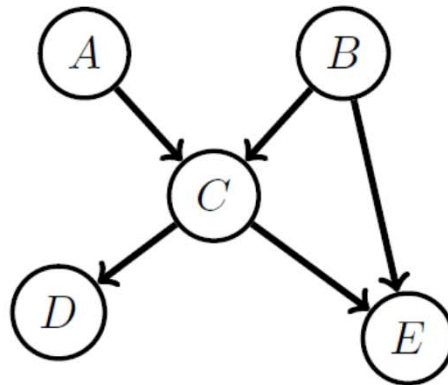
Belief Networks

Barber Ch. 3

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:

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



Alarm Example

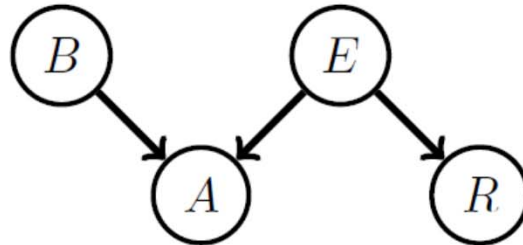
- Sally's burglar Alarm is sounding. Has she been Burgled, or was the alarm triggered by an Earthquake? She turns the car Radio on for news of earthquakes.
- Choosing an ordering
 - Without loss of generality, we can write

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

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

Alarm Example



$$p(A|B, E)$$

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

The remaining data are $p(B = 1) = 0.01$ and $p(E = 1) = 0.000001$

Alarm Example: Inference

- Initial evidence: the alarm is sounding

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

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

Wet Grass Example

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

- Define:

$R \in \{0, 1\}$ $R = 1$ 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:
 $p(T|J,R,S) = p(T|R,S)$
 $p(J|R,S) = p(J|R)$
 $p(R|S) = p(R)$

Wet Grass Example

- Original equation

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

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

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

- Becomes

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

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

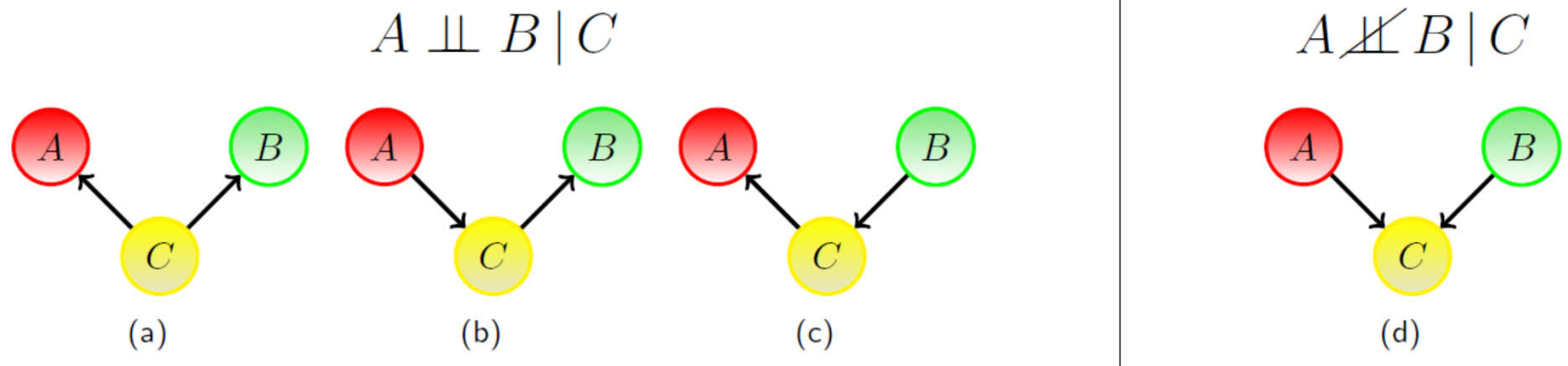
$$\begin{aligned} p(S = 1|T = 1) &= \frac{p(S = 1, T = 1)}{p(T = 1)} = \frac{\sum_{J,R} p(T = 1, J, R, S = 1)}{\sum_{J,R,S} p(T = 1, J, R, S)} \\ &= \frac{\sum_{J,R} p(J|R)p(T = 1|R, S = 1)p(R)p(S = 1)}{\sum_{J,R,S} p(J|R)p(T = 1|R, S)p(R)p(S)} \\ &= \frac{\sum_R p(T = 1|R, S = 1)p(R)p(S = 1)}{\sum_{R,S}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

- Note that $\sum_J p(J|R)p(R) = p(R)$

Wet Grass Example

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

Independence in Belief Networks



- In (a), (b) and (c), A, B are conditionally independent given C

$$(a) \quad p(A, B|C) = \frac{p(A, B, C)}{p(C)} = \frac{p(A|C)p(B|C)p(C)}{p(C)} = p(A|C)p(B|C)$$

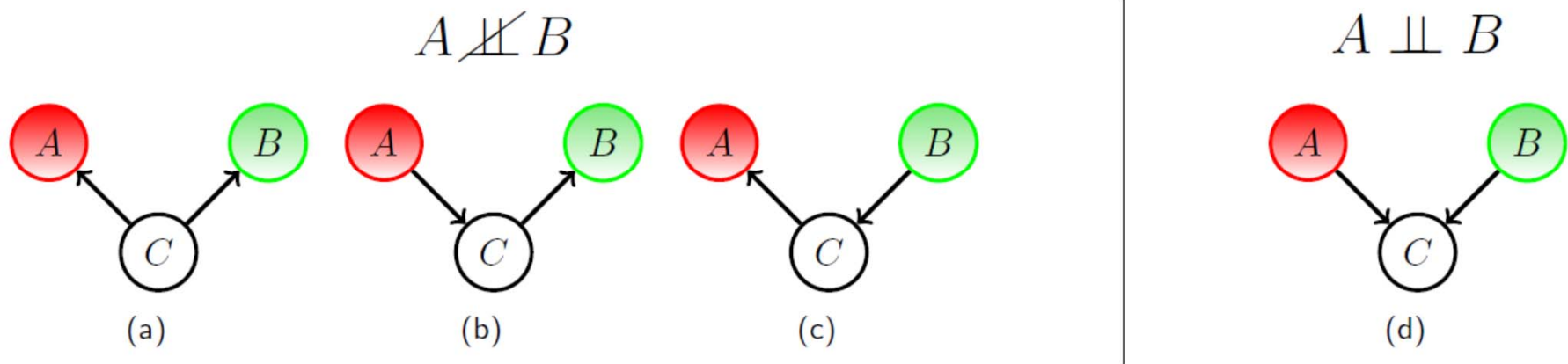
$$(b) \quad p(A, B|C) = \frac{p(A)p(C|A)p(B|C)}{p(C)} = \frac{p(A, C)p(B|C)}{p(C)} = p(A|C)p(B|C)$$

$$(c) \quad p(A, B|C) = \frac{p(A|C)p(C|B)p(B)}{p(C)} = \frac{p(A|C)p(B, C)}{p(C)} = p(A|C)p(B|C)$$

- In (d) the variables A, B are conditionally dependent given C:

$$p(A, B|C) \propto p(C|A, B)p(A)p(B)$$

Independence in Belief Networks



- In (a), (b) and (c), A, B are marginally dependent
- In (d) the variables A, B are marginally independent

$$p(A, B) = \sum_C p(A, B, C) = \sum_C p(A)p(B)p(C|A, B) = p(A)p(B)$$

Intro to Linear Algebra

Slides by Olga Sorkine
(ETH Zurich)

Vector space

- Informal definition:

- $V \neq \emptyset$ (a non-empty set of vectors)

- $\mathbf{v}, \mathbf{w} \in V \Rightarrow \mathbf{v} + \mathbf{w} \in V$ (closed under addition)

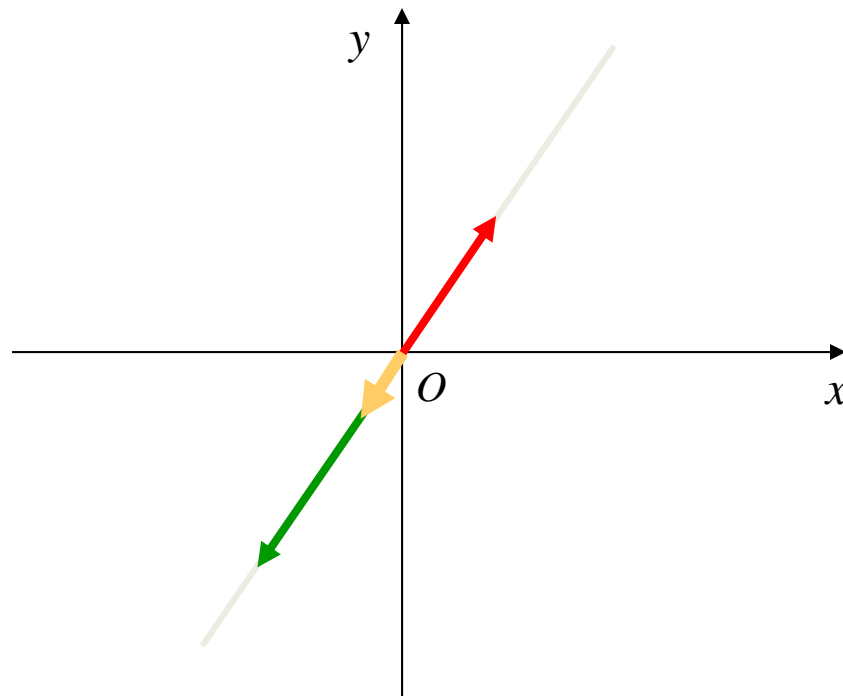
- $\mathbf{v} \in V, \alpha \text{ is scalar} \Rightarrow \alpha\mathbf{v} \in V$ (closed under multiplication by scalar)

- Formal definition includes axioms about associativity and distributivity of the $+$ and \cdot operators.

- $0 \in V$ always!

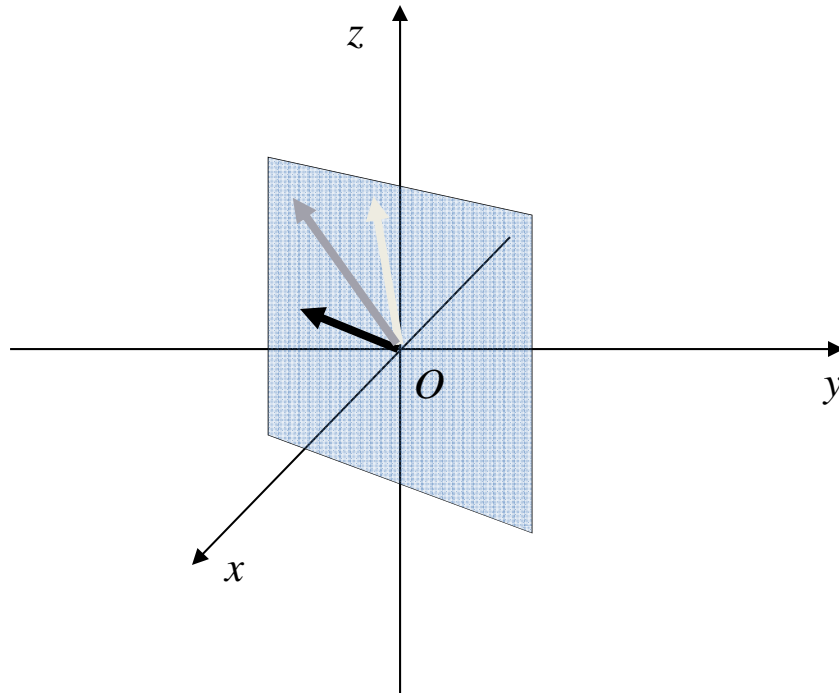
Subspace - example

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



Subspace - example

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



Linear independence

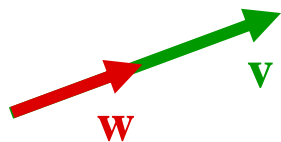
- The vectors $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ are a linearly independent set if:

$$\alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \dots + \alpha_k \mathbf{v}_k = \mathbf{0} \iff \alpha_i = 0 \quad \forall i$$

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

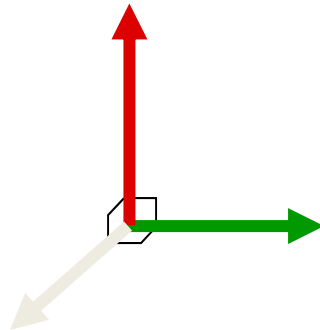
Linear independence - example

- Parallel vectors are always dependent:



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

- Orthogonal vectors are always **linearly** independent



Basis of V

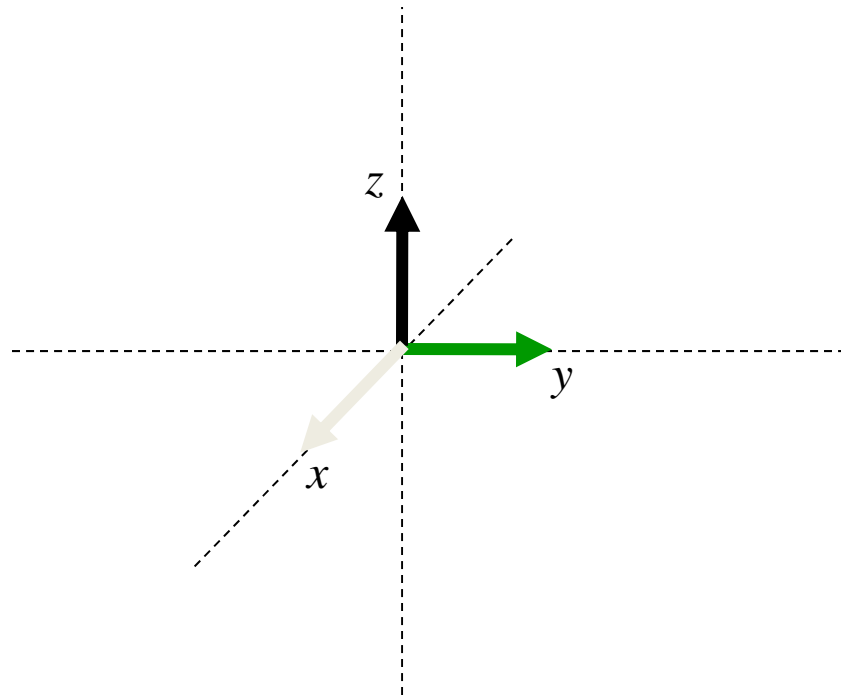
- $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ are **linearly independent**
- $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ **span** the whole vector space V :

$$V = \{ \alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \dots + \alpha_n \mathbf{v}_n \mid \alpha_I \text{ scalars} \}$$

- Any vector in V is a **unique** linear combination of the basis
- The number of basis vectors is called the **dimension** of V

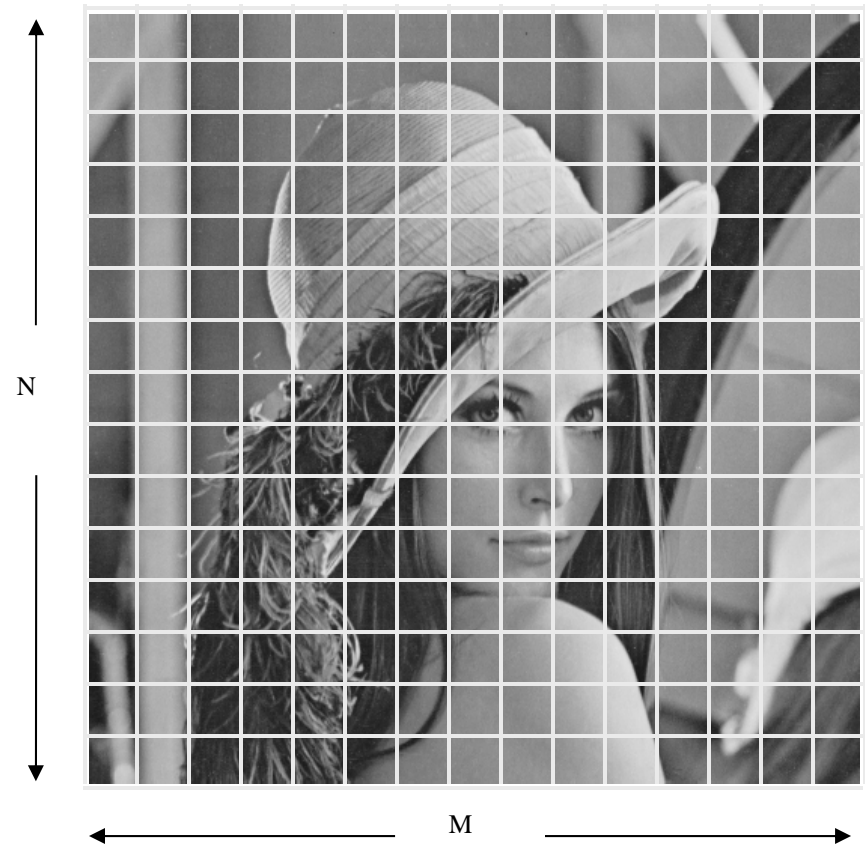
Basis - example

- The standard basis of \mathbb{R}^3 - three unit orthogonal vectors $\hat{x}, \hat{y}, \hat{z}$: (sometimes called i, j, k or $\hat{e}_1, \hat{e}_2, \hat{e}_3$)

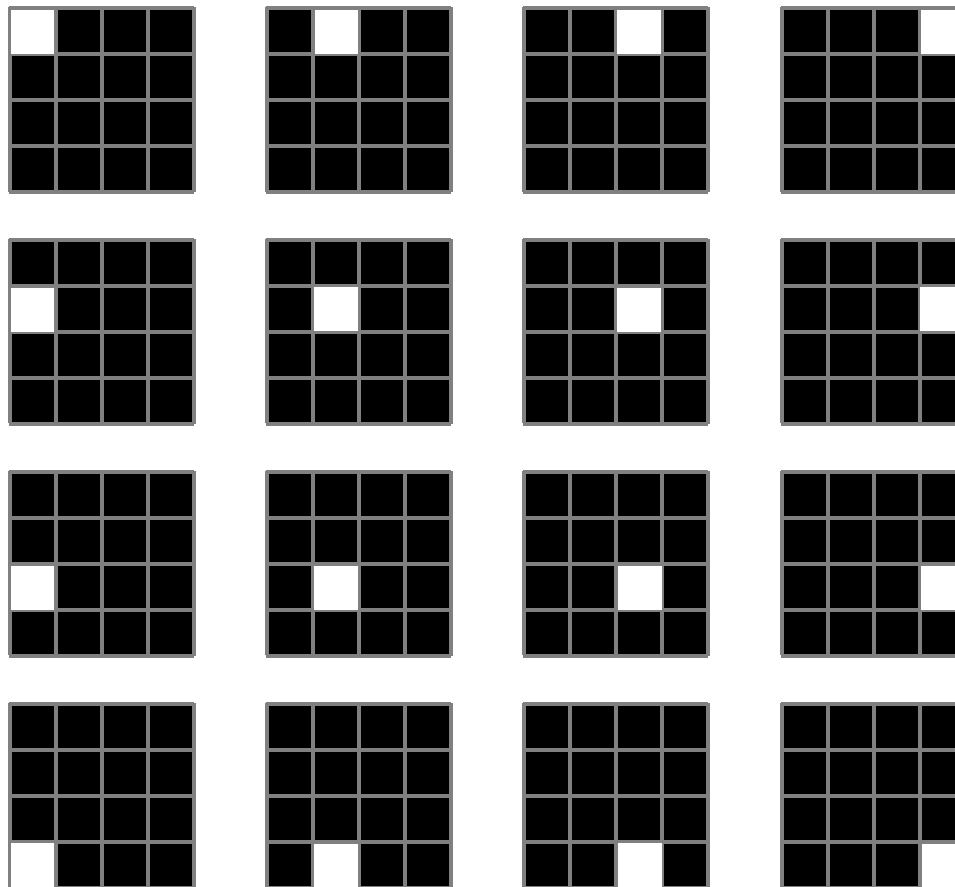


Basis - another example

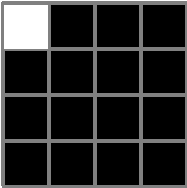
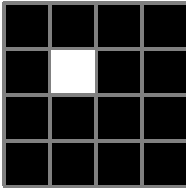
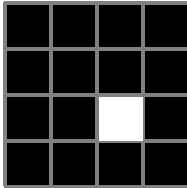
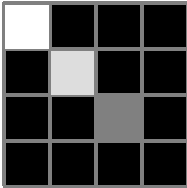
- Grayscale $N \times M$ images:
 - Each pixel has value between 0 (black) and 1 (white)
 - The image can be interpreted as a vector $\in \mathbb{R}^{N \cdot M}$



The “standard” basis (4×4)



Linear combinations of the basis

 $*1$ +  $*(2/3)$ +  $*(1/3) =$ 

Matrix representation

- Let $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ be a basis of V
- Every $\mathbf{v} \in V$ has a unique representation

$$\mathbf{v} = \alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \dots + \alpha_n \mathbf{v}_n$$

- Denote \mathbf{v} by the column-vector:

$$\mathbf{v} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}$$

- The basis vectors are therefore denoted:

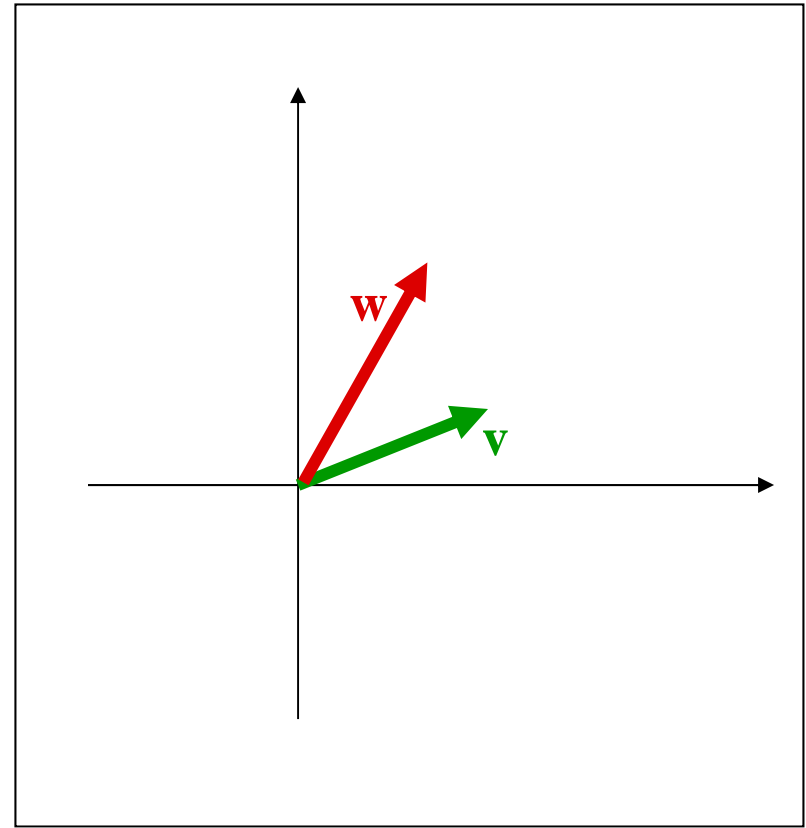
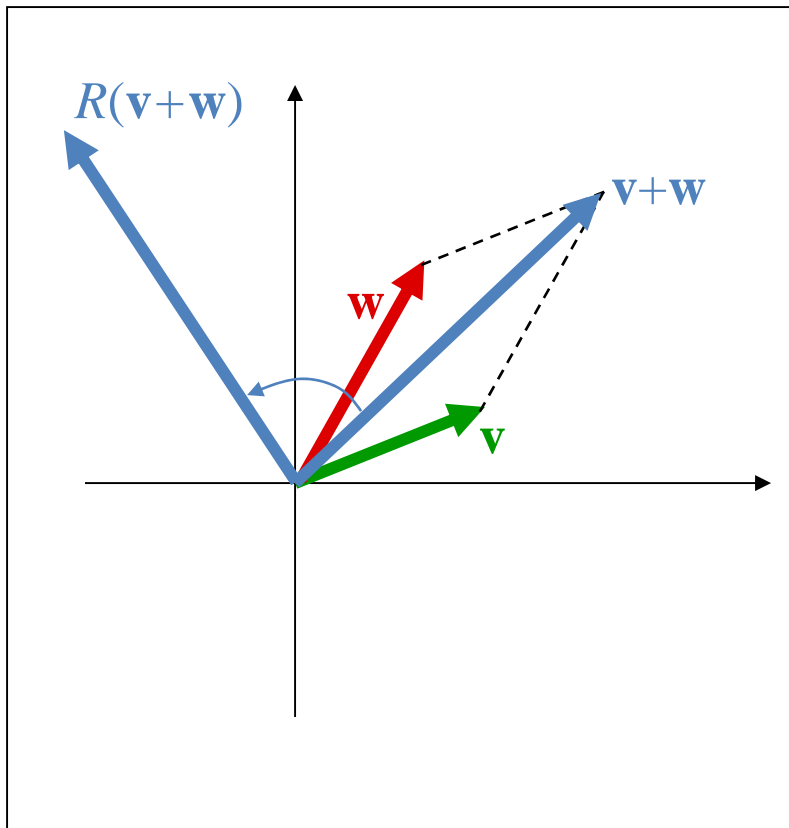
$$\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \dots, \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

Linear operators

- $A : V \rightarrow W$ 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, $A(\mathbf{0}) = \mathbf{0}$
- Linear operators we know:
 - Scaling
 - Rotation, reflection
 - Translation is not linear - moves the origin

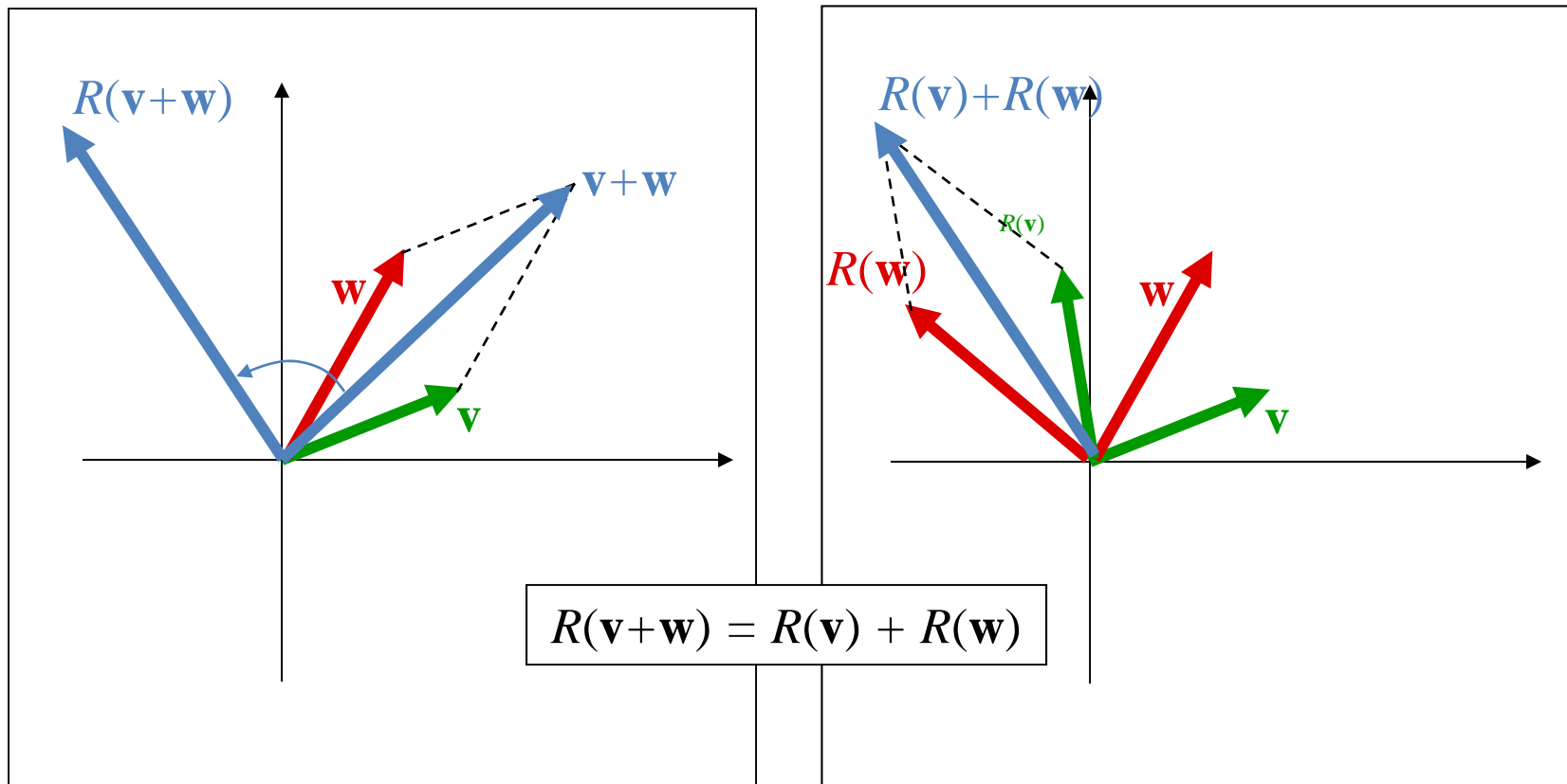
Linear operators - illustration

- Rotation is a linear operator:



Linear operators - illustration

- Rotation is a linear operator:



Matrix operations

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

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

A \mathbf{b}

Matrix by vector multiplication

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

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

Matrix operations

- Transposition: make the rows to be the columns

$$\left(\begin{array}{ccc} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{array} \right)^T = \left(\begin{array}{ccc} a_{11} & \cdots & a_{m1} \\ \vdots & \ddots & \vdots \\ a_{1n} & \cdots & a_{mn} \end{array} \right)$$

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

Matrix properties

- Matrix A ($n \times n$) is **non-singular** if $\exists B, AB = BA = I$
- $B = A^{-1}$ is called the **inverse** of A
- A is non-singular $\Leftrightarrow \det A \neq 0$

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

Orthogonal matrices

- Matrix A ($n \times n$) is **orthogonal** if $A^{-1} = A^T$
- Follows: $AA^T = A^T A = I$
- The rows of A are **orthonormal vectors!**

Proof:

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

$$\Rightarrow \langle \mathbf{v}_i, \mathbf{v}_i \rangle = 1 \Rightarrow \|\mathbf{v}_i\| = 1; \quad \langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0$$

The Trace

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

$$\text{tr}A = \sum_{i=1}^n A_{ii}.$$

The Determinant

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

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

$$= \sum_{j=1}^n (-1)^{i+j} a_{ij} |A_{\setminus i, \setminus j}| \quad (\text{for any } i \in 1, \dots, n)$$

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

The Determinant

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

The Covariance Matrix (Interlude)

Covariance

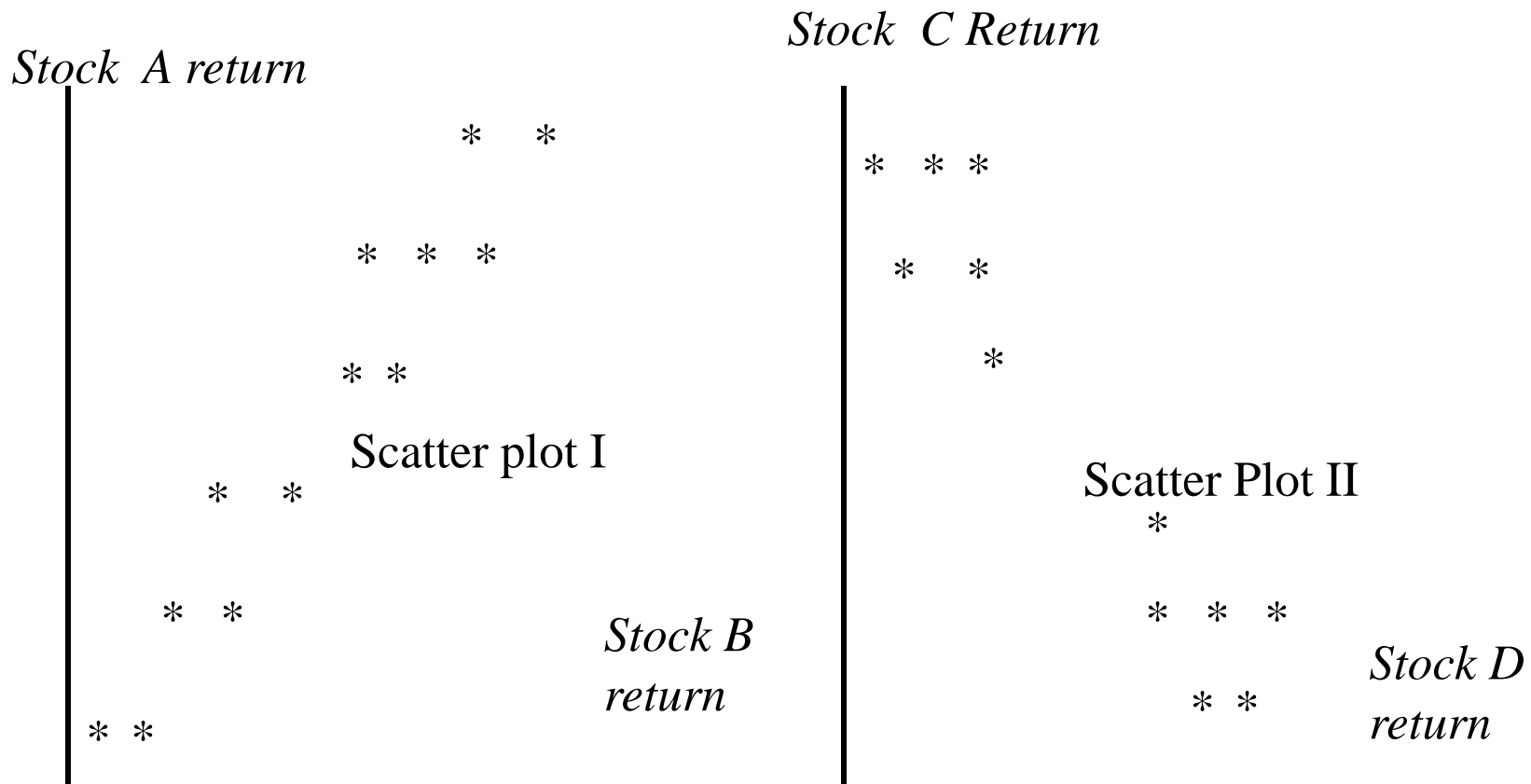
- Covariance is a numerical measure that shows how much two random variables change together

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

- Positive covariance: if one increases, the other is likely to increase
- Negative covariance: ...
- 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:

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

Correlation Coefficient

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

- 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

Covariance Matrix

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

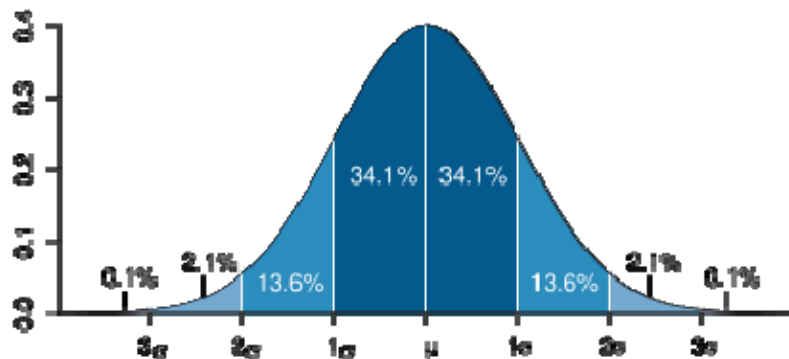
$$\begin{aligned} \text{Cov} \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{pmatrix} &= \begin{pmatrix} \text{Var}(Y_{i1}) & \text{Cov}(Y_{i1}, Y_{i2}) & \cdots & \text{Cov}(Y_{i1}, Y_{in}) \\ \text{Cov}(Y_{i2}, Y_{i1}) & \text{Var}(Y_{i2}) & \cdots & \text{Cov}(Y_{i2}, Y_{in}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(Y_{in}, Y_{i1}) & \text{Cov}(Y_{in}, Y_{i2}) & \cdots & \text{Var}(Y_{in}) \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix}, \end{aligned}$$

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

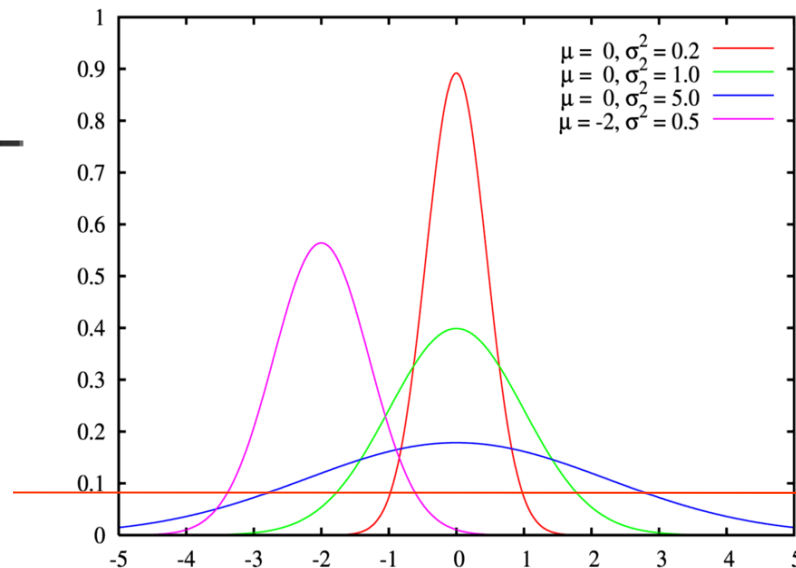
- It is a symmetric, square matrix

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:

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

where:

$\mathbf{x} = (x_1, x_2, \dots, x_d)^t$

$\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_d)^t$ mean vector

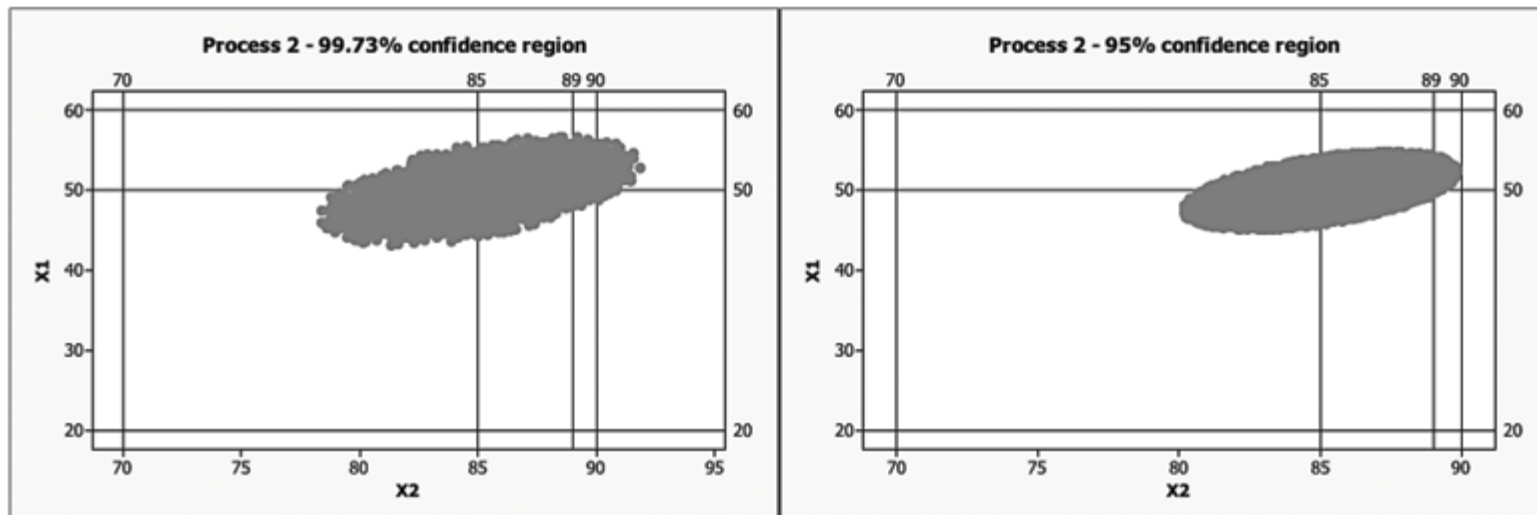
$\Sigma = d \times d$ covariance matrix

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

$P(\mathbf{x})$ is larger for smaller exponents!

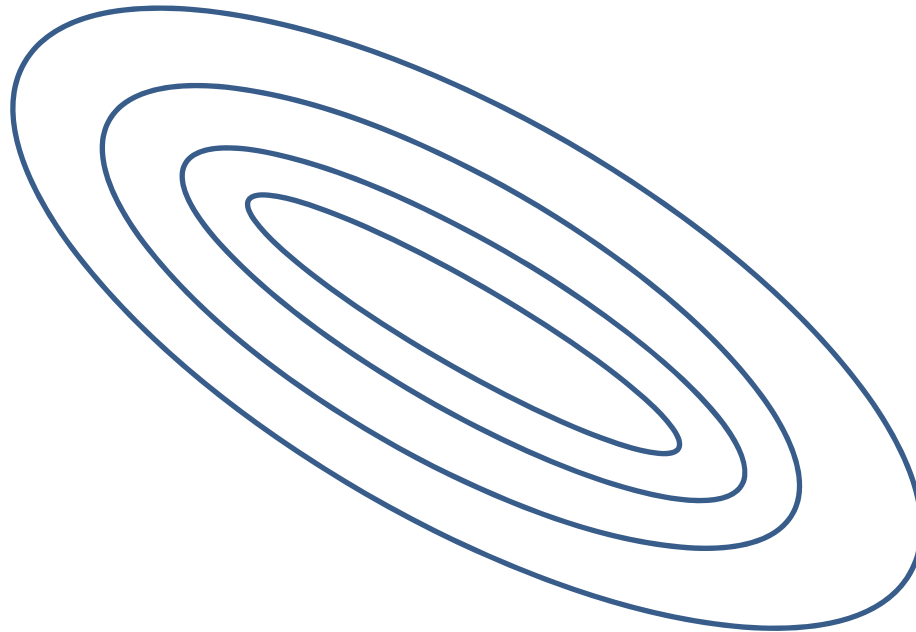
Confidence Intervals: Multi-Variate Case

- 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

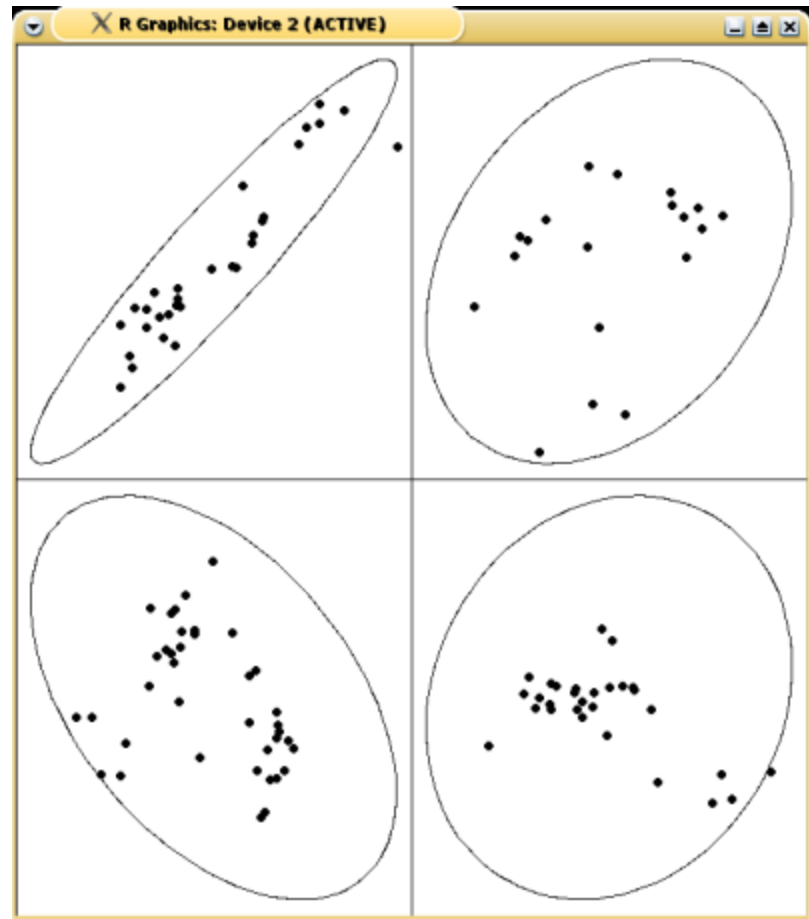


The Multi-Variate Normal Density

- Σ is positive semi definite ($x^t \Sigma x \geq 0$)
 - If $x^t \Sigma x = 0$ for non-zero x then $\det(\Sigma) = 0$. 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 Σ^{-1}

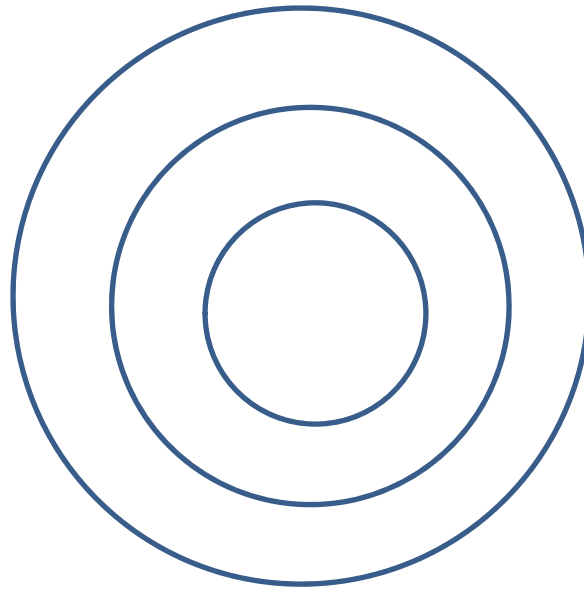
Confidence Intervals: Multi-Variate Case

- Covariance matrix determines the shape



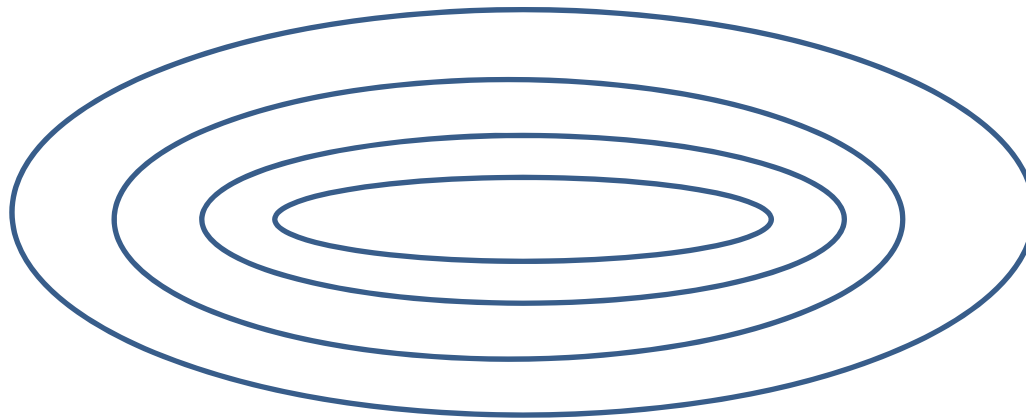
Confidence Intervals: Multi-Variate Case

- Case I: $\Sigma = \sigma^2 I$
 - All variables are uncorrelated and have equal variance
- Confidence intervals are circles



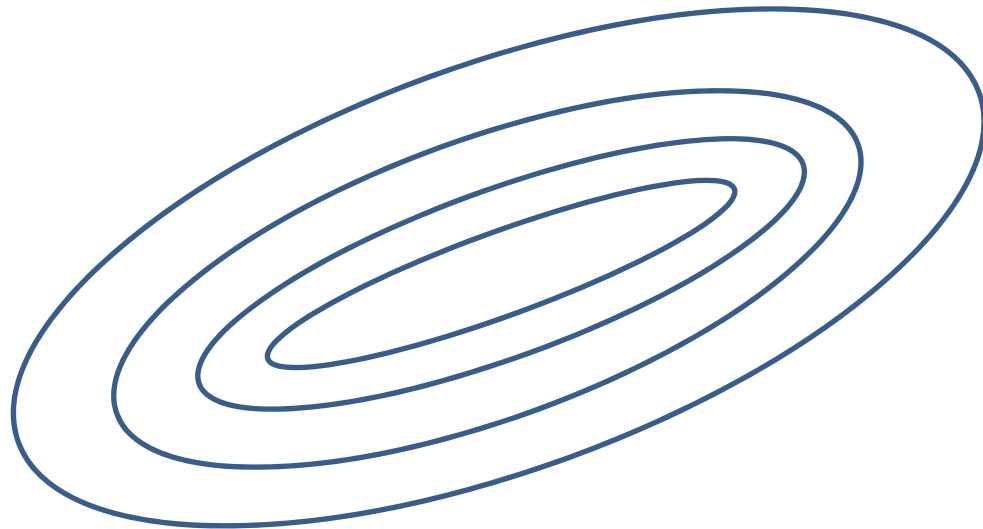
Confidence Intervals: Multi-Variate Case

- Case II: Σ 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: Σ arbitrary
 - Variables may be correlated and have different variances
- Confidence intervals are arbitrary ellipsoids



Eigen-interlude

Based on D. Barber's slides

Eigenvalues and Eigenvectors

- For an $n \times n$ square matrix A , e is an eigenvector with eigenvalue λ if

$$Ae = \lambda e$$

- Or

$$(A - \lambda I)e = 0$$

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

Eigenvalues and Eigenvectors

$$(A - \lambda I)e = 0$$

- For non-trivial solutions:

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

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

Simple Example

- For a 2×2 matrix

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

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

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

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

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

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

The solutions are $\lambda=0$ and $\lambda=5$

The eigenvector for the first eigenvalue, $\lambda=0$ is:

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

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

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

For the other eigenvalue, $\lambda=5$:

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

$-4x + 2y = 0$, and $2x - y = 0$, so, $x = 1, y = 2$

Properties

- The product of the eigenvalues = $|A|$
- The sum of the eigenvalues = $\text{trace}(A)$
- The eigenvectors are pairwise orthogonal

$$(\mathbf{e}^i)^\top \mathbf{e}^j = \delta_{ij} (\mathbf{e}^i)^\top \mathbf{e}^i$$

Spectral Decomposition

- A symmetric matrix has real eigenvalues
- A real symmetric matrix can be written as:

$$\mathbf{A} = \sum_{i=1}^n \lambda_i \mathbf{e}_i \mathbf{e}_i^{\top}$$

$$\mathbf{A} = \mathbf{E} \mathbf{\Lambda} \mathbf{E}^{\top}$$

Back to the Covariance Matrix

Geometric Interpretation

- Start from $N(0, I)$ and construct multivariate distribution with desired covariance matrix

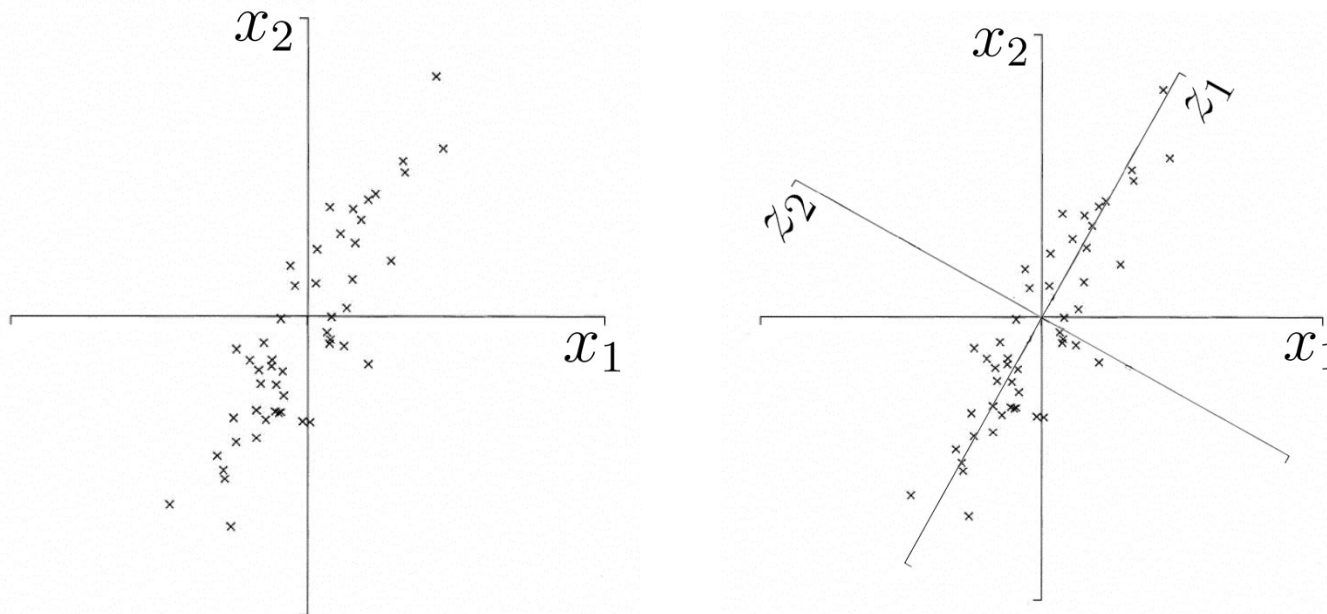
$$X \sim N(\mu, \Sigma) \iff X \sim \mu + U\Lambda^{1/2}N(0, I) \iff X \sim \mu + UN(0, \Lambda).$$

translation rotation anisotropic scaling

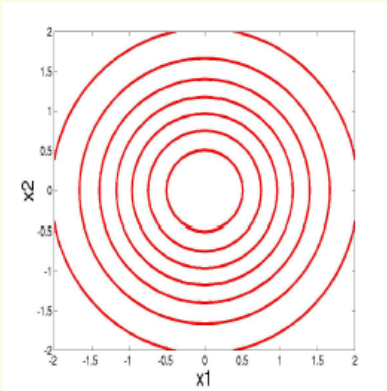
The diagram illustrates the geometric interpretation of a multivariate normal distribution. It shows the equivalence between three representations: $X \sim N(\mu, \Sigma)$, $X \sim \mu + U\Lambda^{1/2}N(0, I)$, and $X \sim \mu + UN(0, \Lambda)$. Blue arrows point from the labels 'translation', 'rotation', and 'anisotropic scaling' to the corresponding terms in the middle expression: μ , U , and $\Lambda^{1/2}$ respectively.

Eigenvectors of the Covariance Matrix

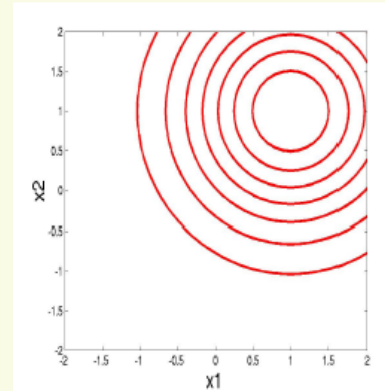
- New basis aligned with ellipsoids
- Major axis \Leftrightarrow eigenvector with max eigenvalue



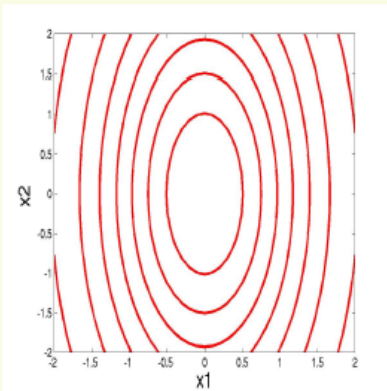
2D Examples



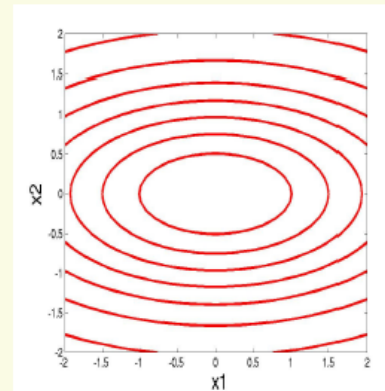
$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$
$$\mu = [0, 0]$$



$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$
$$\mu = [1, 1]$$

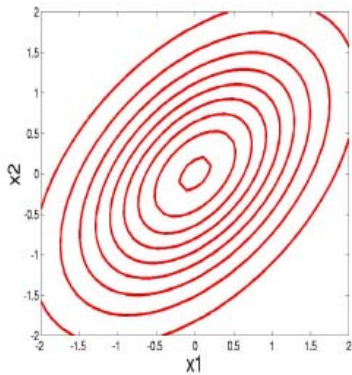


$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$$
$$\mu = [0, 0]$$

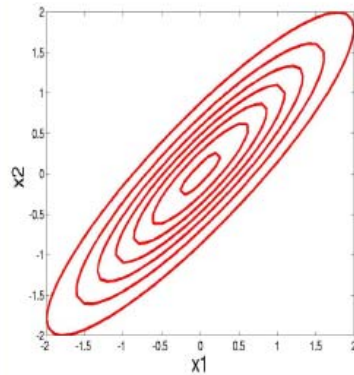


$$\Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$
$$\mu = [0, 0]$$

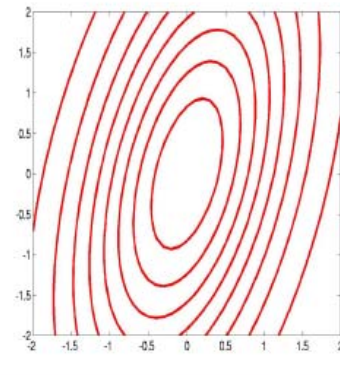
2D Examples



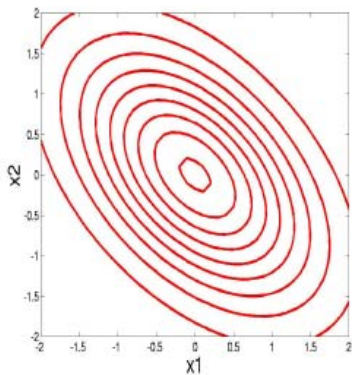
$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$



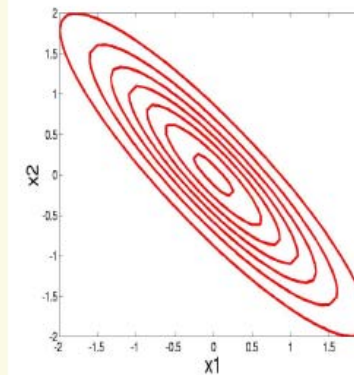
$$\Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$$



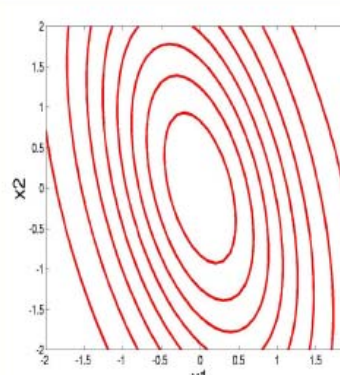
$$\Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 4 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 4 \end{bmatrix}$$