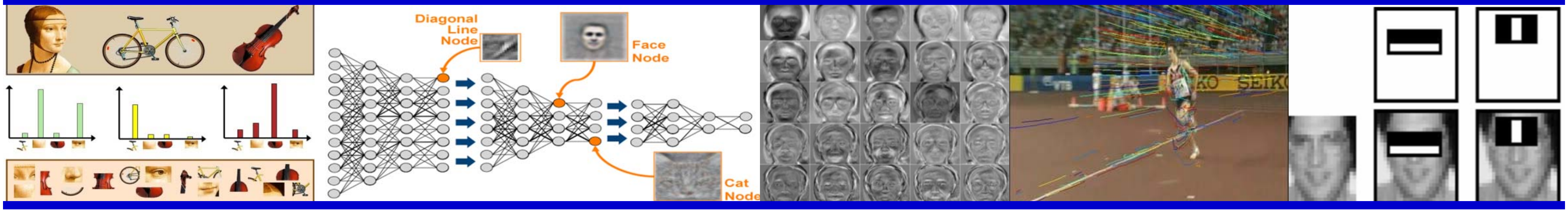


EE5907R: Pattern Recognition

Lecture 0: Review of Probability & Linear Algebra

National University of Singapore

Dr. Tam Nguyen

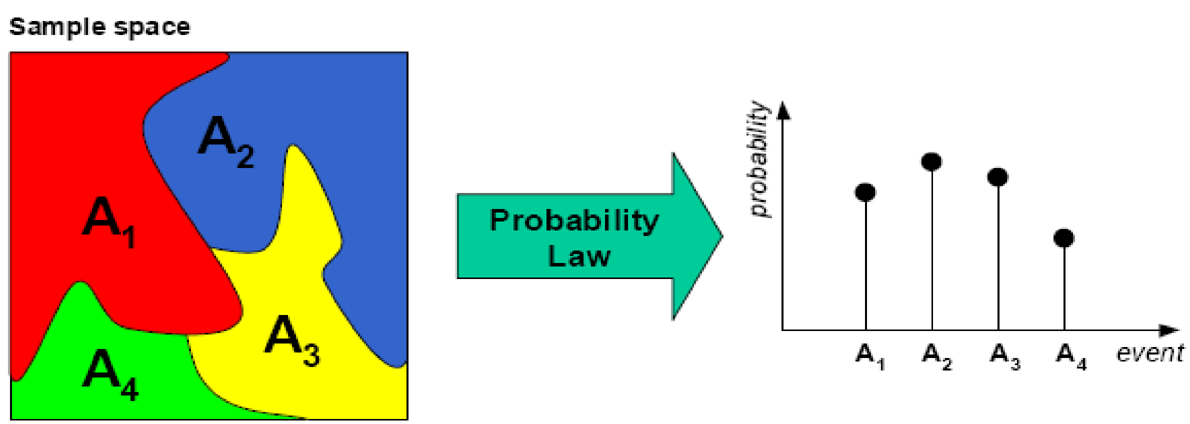
Review of Probability

- Probability
  - Axioms and properties
  - Conditional probability
  - Law of total probability
  - Bayes theorem
- Random Variables
  - Discrete
  - Continuous
- Random Vectors
- Gaussian Random Variables

Some of the following slides are taken from lecture notes by Ricardo Gutierrez-Osuna

Basics of Probability

- **Definitions (informal)**
  - Probabilities are numbers assigned to events that indicate "***how likely***" it is that the event will occur when a random experiment is performed
  - A **probability law** for a random experiment is a rule that assigns probabilities to the events in the experiment
  - The **sample space** S of a random experiment is the set of all possible outcomes



- **Axioms of probability**
  - Axiom I: $0 \le P[A_i]$
  - Axiom II: $P[S] = 1$
  - Axiom III: if $A_i \cap A_j = \varnothing$, then $P[A_i \cup A_j] = P[A_i] + P[A_j]$

Properties of Probability

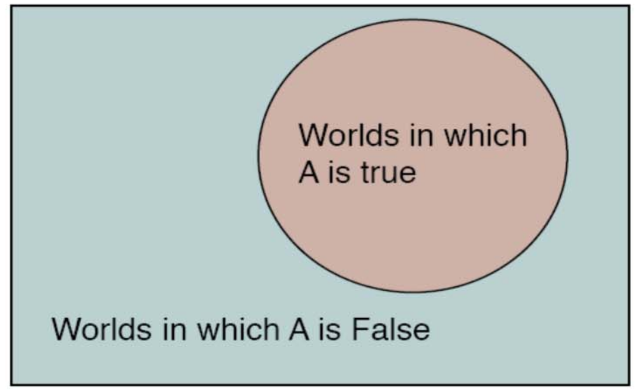
$$1: \quad P[A^C] = 1 - P[A]$$

$$2: \quad P[A] \leq 1$$

$$3: \quad P[\varnothing] = 0$$

Event space of all possible worlds

Its area is 1



P(A) = Area of reddish oval

$$4: \quad \text{given } \{A_1, A_2, \ldots A_N\}, \quad \text{if } \{A_i \cap A_j = \varnothing \quad \forall i,j\} \quad \text{then } P[\bigcup_{k=1}^{N} A_k] = \sum_{k=1}^{N} P[A_k]$$

$$5: \quad P[A_1 \cup A_2] = P[A_1] + P[A_2] - P[A_1 \cap A_2]$$

$$6: \quad P[\bigcup_{k=1}^{N} A_k] = \sum_{k=1}^{N} P[A_k] - \sum_{j<k}^{N} P[A_j \cap A_k] + \ldots + (-1)^{N+1} P[A_1 \cap A_2 \cap \ldots \cap A_N]$$

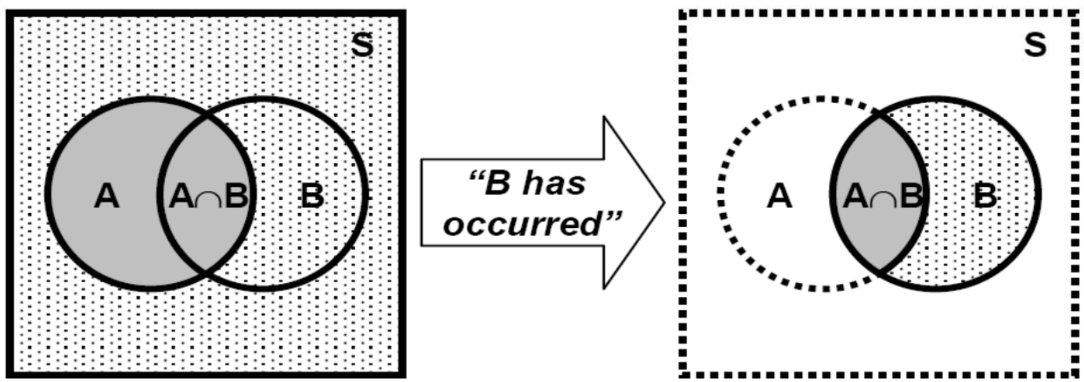
$$7: \quad \text{if } A_1 \subset A_2, \text{ then } P[A_1] \leq P[A_2]$$

Conditional Probability

- **If A and B are two events, the probability of event A when we already know that event B has occurred is defined by the relation**

$$P[A \mid B] = \frac{P[A \cap B]}{P[B]} \quad \text{for} \quad P[B] > 0$$

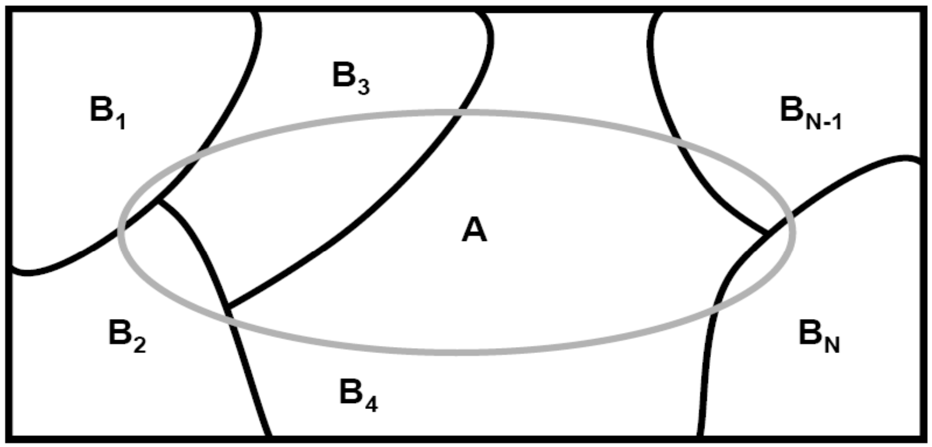
- **This conditional probability P[A|B] is read:**
  - the "conditional probability of A conditioned on B", or simply
  - the "probability of A given B"



Law of Total Probability

- **Let B_1, B_2, …, B_N be mutually exclusive events whose union equals the sample space S. We refer to these sets as a <u>partition</u> of S.**
- **An event A can be represented as:**

$$A = A \cap S = A \cap (B_1 \cup B_2 \cup \ldots \cup B_N) = (A \cap B_1) \cup (A \cap B_2) \cup \ldots (A \cap B_N)$$



- **Since B_1, B_2, …, B_N are mutually exclusive, then**

$$P[A] = P[A \cap B_1] + P[A \cap B_2] + \ldots + P[A \cap B_N]$$

- **and, therefore**

$$P[A] = P[A \mid B_1]P[B_1] + \ldots P[A \mid B_N]P[B_N] = \sum_{k=1}^{N} P[A \mid B_k]P[B_k]$$

Bayes Theorem

- **Given B_1, B_2, …, B_N, a partition of the sample space S. Suppose that event A occurs; what is the probability of event B_j?**
  - Using the definition of conditional probability and the Theorem of total probability we obtain

$$P[B_j \mid A] = \frac{P[A \cap B_j]}{P[A]} = \frac{P[A \mid B_j] \cdot P[B_j]}{\sum_{k=1}^{N} P[A \mid B_k] \cdot P[B_k]}$$

- **This is known as Bayes Theorem or Bayes Rule, and is (one of) the most useful relations in probability and statistics**
  - Bayes Theorem is definitely the fundamental relation in Statistical Pattern Recognition

Rev. Thomas Bayes (1702-1761)

Example

- You can play tennis if there is no rain on either Saturday or Sunday. The probability of it raining on a Saturday is 80% and the chance of it raining on Sunday is 60%.

a) What is the probability of playing tennis in the weekend?

P (Tennis) = 1 - P (Tennis^c) = 1 - (0.8)(0.6)= 0.52

b) Given that you played tennis over the weekend, what is the probability that that it rained on the weekend?

P (Rain|Tennis)= P(Rain and Tennis)/P(Tennis)

 | 

= [P(Rain Sat^c & Rain Sun) + P(Rain Sat and Rain Sun^c)]/P(Tennis)

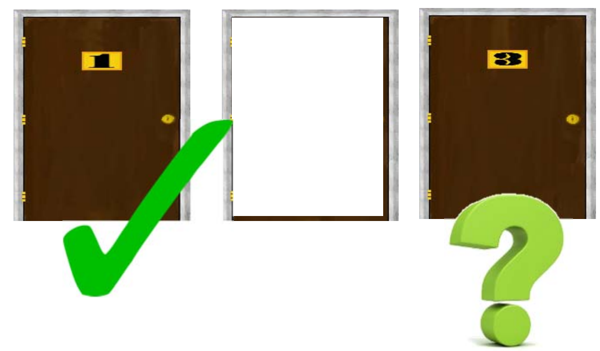
= (0.2*0.6 + 0.8*0.4) / 0.52

≈ 0.846

An Interesting Example

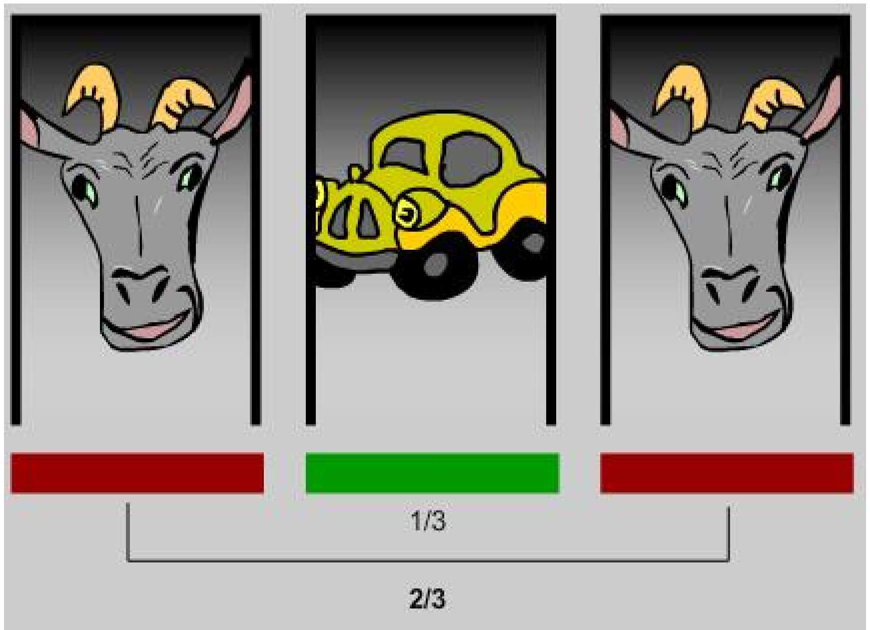
You are a contestant in a game show, and the game show host tells you there is a prize behind one of the three doors you face. You have to guess which door to open.

But when you make your guess, instead of opening the door you picked, the game show host opens a *different* door - one that he *knows* has nothing behind it (the host will *never reveal* the prize at this stage). So now you're down to *two* doors. And the game show host says, "I'll let you change your choice, if you want to."



And the question is, do you change your guess? Or keep your original choice? Does it make any difference?

Monty Hall Problem



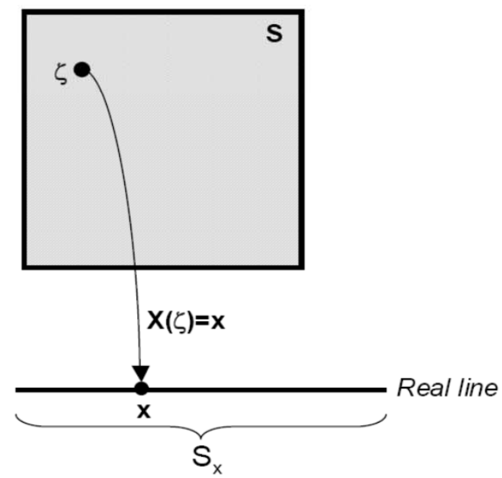
Explanations: http://www.youtube.com/watch?v=mhlc7peGlGg

http://www.youtube.com/watch?v=9vRUxbzJZ9Y

Try it out: http://math.ucsd.edu/~crypto/Monty/monty.html

Random Variables

- **When we perform a random experiment we are usually interested in some measurement or numerical attribute of the outcome**
  - When we sample a population we may be interested in their weights
  - When rating the performance of two computers we may be interested in the execution time of a benchmark
  - When trying to recognize an intruder aircraft, we may want to measure parameters that characterize its shape
- **These examples lead to the concept of *random variable***
  - **A random variable X is a function that assigns a real number X(ζ) to each outcome ζ in the sample space of a random experiment**
    - This function X(ζ) is performing a mapping from all the possible elements in the sample space onto the real line (real numbers)



Two Types of Random Variables

- Discrete Random Variable
  - has countable number of values
  - e.g., the resulting number of rolling a dice (any number from (1, 2, 3, 4, 5, 6)
  - probability distribution defined by **probability mass function**

- Continuous Random Variable
  - has values that are continuous
  - e.g., the weight of an individual (any real number within the range of human weight)
  - defined by **probability density function**

Statistical Characterization of RVs

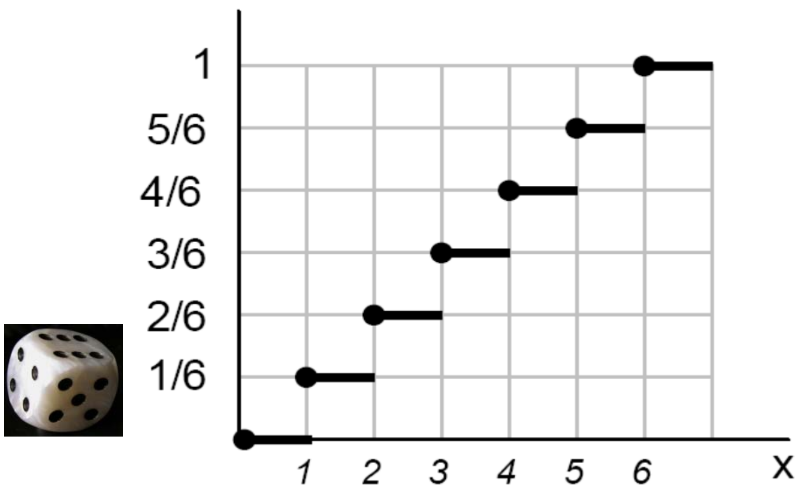
- **The cdf or the pdf are SUFFICIENT to fully characterize a random variable, However, a random variable can be PARTIALLY characterized with other measures**
  - **Expectation** $$E[X] = \mu = \int_{-\infty}^{+\infty} x f_X(x) dx$$
    - The expectation represents the center of mass of a density
  - **Variance** $$VAR[X] = E[(X - E[X])^2] = \int_{-\infty}^{+\infty} (x - \mu)^2 f_X(x) dx$$
    - The variance represents the spread about the mean
  - **Standard deviation** $$STD[X] = VAR[X]^{1/2}$$
    - The square root of the variance. It has the same units as the random variable.

Cumulative Distribution Function

- **The cumulative distribution function $F_X(x)$ of a random variable X is defined as the probability of the event $\{X \le x\}$**

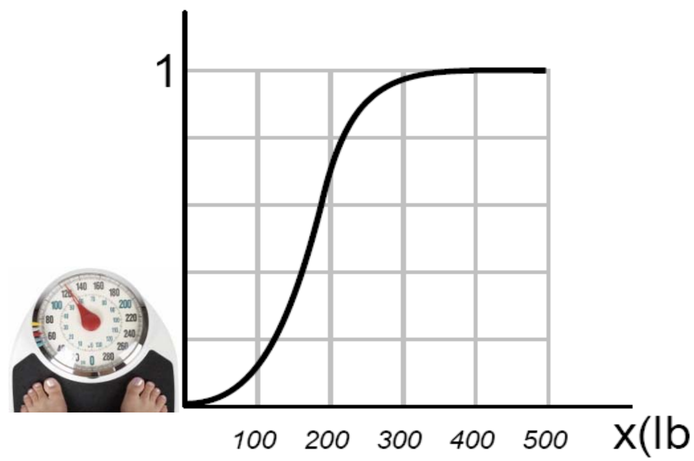
$$F_X(x) = P[X \le x] \quad \text{for} \quad -\infty < x < +\infty$$

CDF for discrete RV



cdf for rolling a dice

CDF for continuous RV



cdf for a person's weight

Properties of CDF

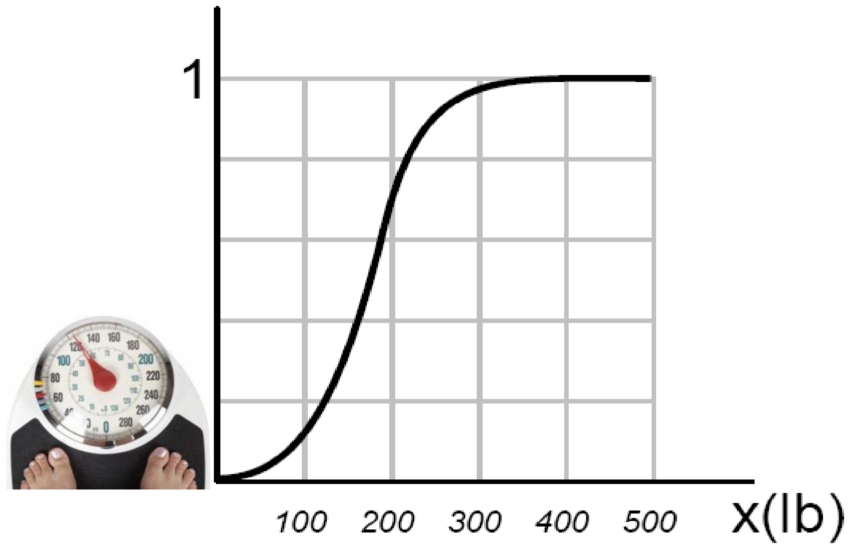
$$0 \le F_X(x) \le 1$$

$$\lim_{x \to \infty} F_X(x) = 1$$

$$\lim_{x \to -\infty} F_X(x) = 0$$

$$F_X(a) \le F_X(b) \text{ if } a \le b$$

$$F_X(b) = \lim_{h \to 0} F_X(b+h) = F_X(b^+)$$



cdf for a person's weight

Questions about ***X*** can be asked in terms of CDF

$$\boxed{P(a < X \le b) = F(b) - F(a)}$$

P(a person's weight between 100 and 200)=F(200)-F(100)

Discrete Random Variable: Probability Mass Function

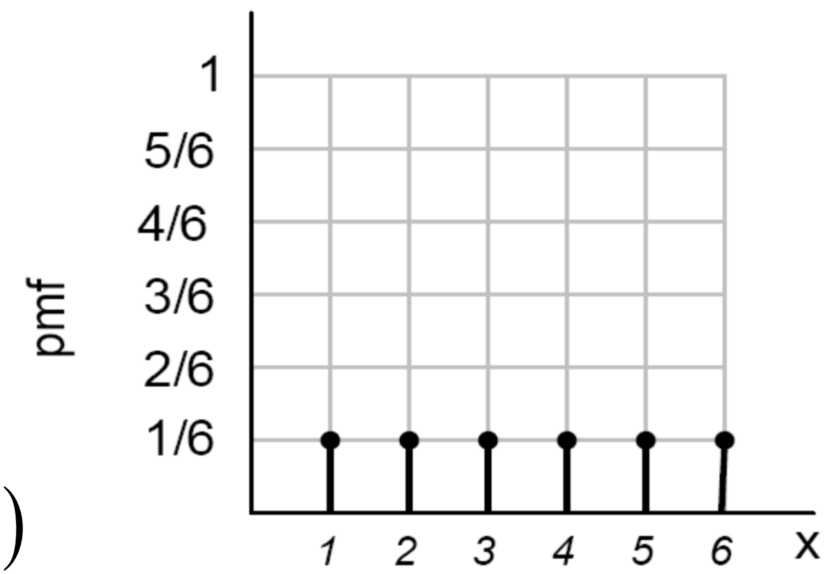
- Given a discrete random variable ***X***, the probability mass function is defined as

$$P(a) = P(X = a)$$

- Satisfies all axioms of probability

- CDF satisfies

$$F(a) = P(X \leq a) = \sum_{k \leq a} P(X = k)$$



pmf for rolling a (fair) dice

Continuous Random Variable: Probability Density Function

- PDF is the derivative of CDF

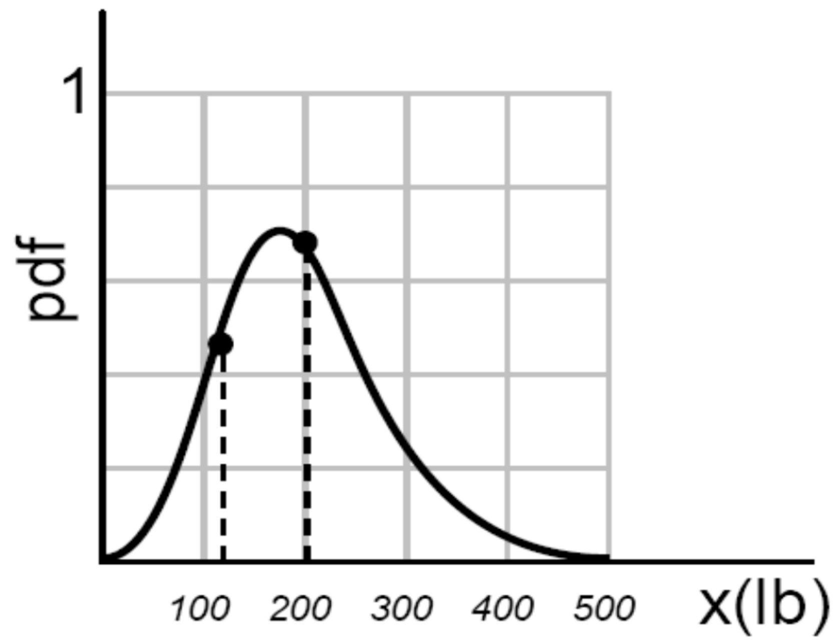
$$f_X(x) = \frac{dF_X(x)}{dx}$$

- CDF satisfies

$$F(a) = P(X \le a) = \int_{-\infty}^{a} f(x)dx$$

- General usage

$$P(a < X \le b) = \int_{a}^{b} f(x)dx$$



pdf for a person's weight

Random Vectors

- **The notion of a random vector is an extension to that of a random variable**
  - A vector random variable X is a function that assigns a vector of real numbers to each outcome ζ in the sample space S
  - We will always denote a random vector by a **column vector**
- **The notions of cdf and pdf are replaced by 'joint cdf' and 'joint pdf'**
  - Given random vector, $\underline{X} = [x_1\, x_2 \dots x_N]^T$ we define
    - **Joint Cumulative Distribution Function** as:

$$F_{\underline{X}}(\underline{x}) = P_{\underline{X}}[\{X_1 \le x_1\} \cap \{X_2 \le x_2\} \cap \dots \cap \{X_N \le x_N\}]$$

    - **Joint Probability Density Function** as:

$$f_{\underline{X}}(\underline{x}) = \frac{\partial^N F_{\underline{X}}(\underline{x})}{\partial x_1 \partial x_2 \dots \partial x_N}$$

–Example: When trying to recognize the aircraft, we may consider its shape, size and color.

Covariance Matrix

- **The covariance matrix indicates the tendency of each pair of features (dimensions in a random vector) to vary together, i.e., to <u>co-vary</u>***
- **The covariance has several important properties**
  - If $\mathbf{x}_i$ and $\mathbf{x}_k$ tend to increase together, then $\mathbf{c}_{ik}>0$
  - If $\mathbf{x}_i$ tends to decrease when $\mathbf{x}_k$ increases, then $\mathbf{c}_{ik}<0$
  - If $\mathbf{x}_i$ and $\mathbf{x}_k$ are **uncorrelated**, then $\mathbf{c}_{ik}=0$
  - $|\mathbf{c}_{ik}| \le \sigma_i \sigma_k$, where σ_i is the standard deviation of $\mathbf{x}_i$
  - $\mathbf{c}_{ii} = \sigma_i^2 = \mathrm{VAR}(\mathbf{x}_i)$
  - symmetric: $c_{ij} = c_{ji}$
  - Positive semi-definite:
    - eigenvalues are nonnegative
    - determinant is nonnegative, $|C| \geq 0$

$$E[X] = \left[E[X_1]E[X_2]\ldots E[X_N]\right]^T = \left[\mu_1 \mu_2 \ldots \mu_N\right] = \mu$$

$$\mathrm{COV}[X] = \Sigma = E[(X-\mu)(X-\mu)^T]$$

$$= \begin{bmatrix} E[(x_1-\mu_1)(x_1-\mu_1)] & \ldots & E[(x_1-\mu_1)(x_N-\mu_N)] \\ \ldots & \ldots & \\ E[(x_N-\mu_N)(x_1-\mu_1)] & \ldots & E[(x_N-\mu_N)(x_N-\mu_N)] \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \ldots & c_{1N} \\ \ldots & \ldots & \\ c_{1N} & \ldots & \sigma_N^2 \end{bmatrix}$$

Covariance Matrix: Quiz

You are given the heights and weights of a certain set of individuals in unknown units. Which one of the following four matrices is the most likely to be the sampled covariance matrix?

ASYMMETRIC

$$(a)\begin{bmatrix} 1.232 & 0.867 \\ -0.867 & 2.791 \end{bmatrix}$$

ARE HEIGHT AND WEIGHT NOT CORRELATED?

$$(b)\begin{bmatrix} 1.232 & -0.867 \\ -0.867 & 2.791 \end{bmatrix}$$

$$(c)\begin{bmatrix} 1.232 & 0.867 \\ 0.867 & 2.791 \end{bmatrix}$$

HEIGHT AND WEIGHT OVER CORRELATED?

$$(d)\begin{bmatrix} 1.232 & 3.307 \\ 3.307 & 2.791 \end{bmatrix}$$

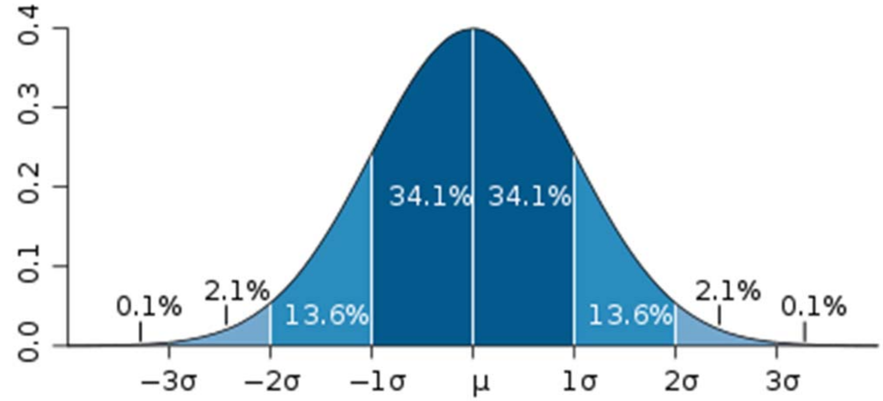
The answer is C. Why?

The Normal or Gaussian Distribution of a Random Variable

- Probability density function:

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

- μ = mean (or expected value) of x
- σ^2 = expected squared deviation or variance

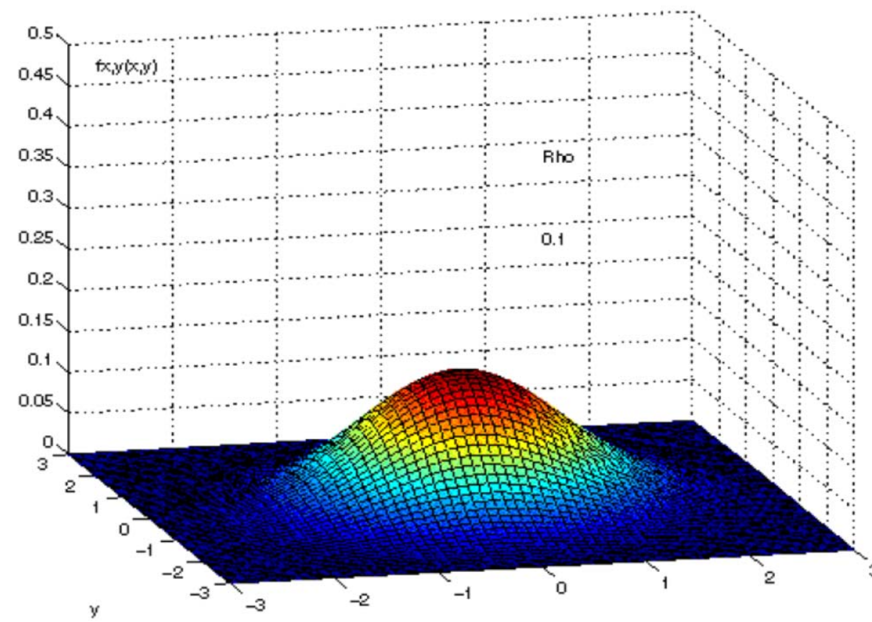


Multivariate Gaussian (Random Vector)

- Probability density function:

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

- mean vector $\boldsymbol{\mu}$
- covariance matrix Σ



Why Gaussian?

- The parameters ($\boldsymbol{\mu}, \boldsymbol{\Sigma}$) are *sufficient* to uniquely characterize the distribution.
- If $\boldsymbol{x_i}$'s are mutually *uncorrelated,* then they are also *independent*.
- Practical – nice to work with
  - The *marginal* and *conditional densities* are also Gaussian.
  - Any *linear transformation* of any *N* jointly Gaussian RV's results in *N* RV's also Gaussian.

Review of Linear Algebra

- Vectors
- Products and norms
- Linear Dependence and Independence
- Vector spaces and basis
- Matrices
- Linear transformations
- Eigenvalues and eigenvectors

Vectors

- An n-dimensional column vector and its transpose (row vector) are represented as

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \quad \text{and} \quad x^T = [x_1 \, x_1 \cdots x_d]$$

- The inner product (dot product or scalar product) of two vectors:

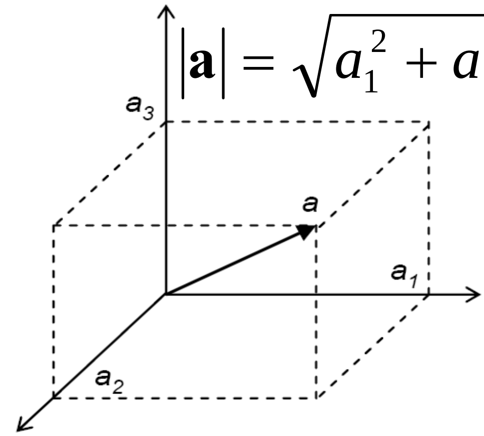
$$\langle x, y \rangle = x^T y = y^T x = \sum_{k=1}^{d} x_k y_k$$

Vectors (Cont'd)

- Euclidean norm or length

$$|\mathbf{x}| = \sqrt{\mathbf{x}^T\mathbf{x}} = \left[\sum_{k=1}^{d} x_k x_k\right]^{1/2}$$

$$|\mathbf{a}| = \sqrt{a_1^2 + a_2^2 + a_3^2}$$



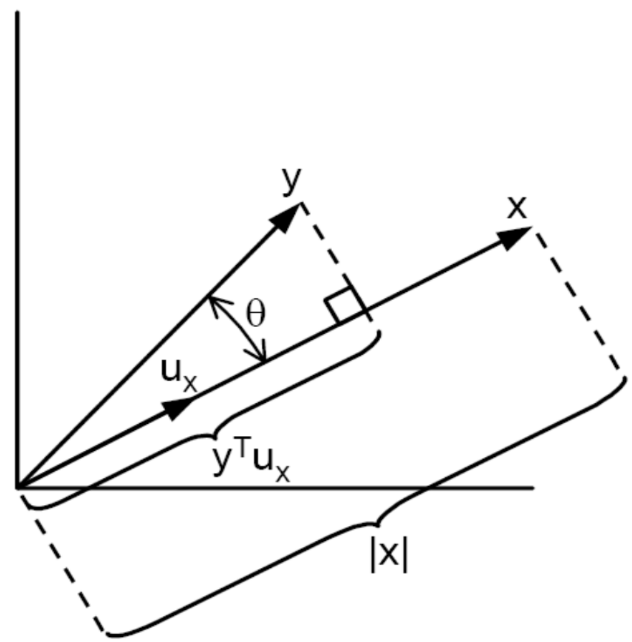
$$\frac{\mathbf{a}}{|\mathbf{a}|}$$

- Normalized (unit) vector

$$|x|=1$$

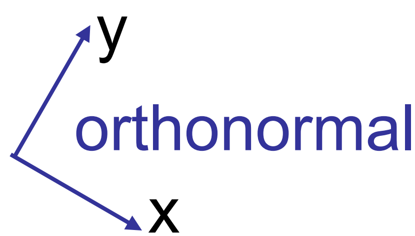
- Angle between vectors x and y

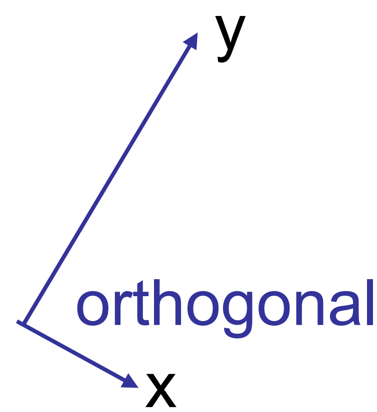
$$\cos\theta = \frac{\langle x, y\rangle}{|x| \cdot |y|}$$



Vectors (Cont'd)

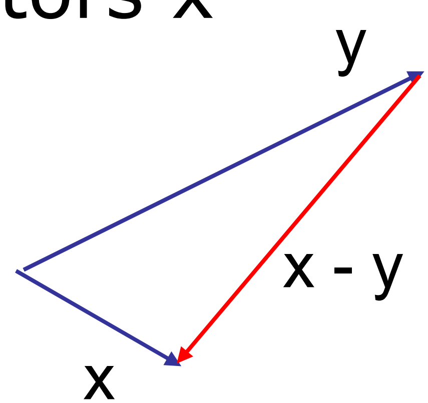
- Two vectors x and y are
  - orthogonal if $\cos\theta=0$ or $\langle x,y\rangle=0$
  - orthonormal if they are orthogonal and $|x|=|y|=1$





- Euclidean distance between vectors x and y

$$|\mathbf{x}-\mathbf{y}| = \sqrt{\sum_{i=1}^{n}\left(x_i - y_i\right)^2}$$



Linear Dependence and Independence

- Vectors $x_1, x_2, \ldots, x_n$ are linearly dependent if there exists a set of coefficients $a_1, a_2, \ldots, a_n$ (at least one $a_i \neq 0$) such that

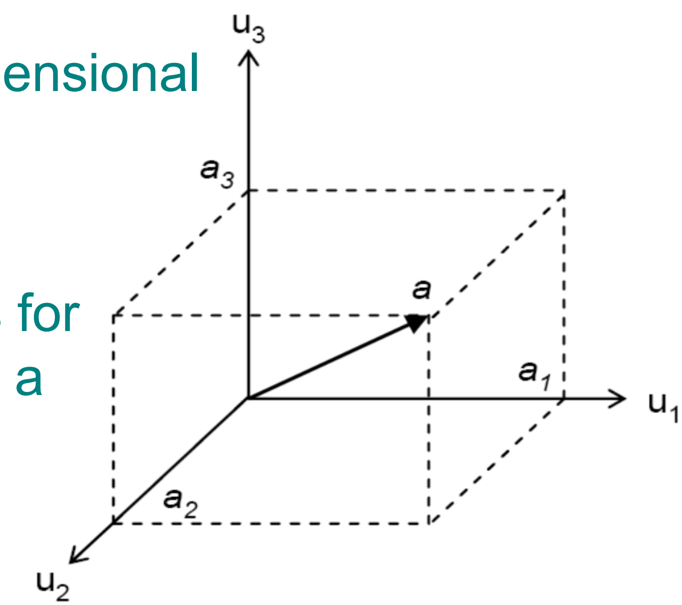
$$a_1 x_1 + a_2 x_2 + \cdots a_n x_n = 0$$

- Vectors $x_1, x_2, \ldots, x_n$ are linearly independent if

$$a_1 x_1 + a_2 x_2 + \cdots a_n x_n = 0 \Rightarrow a_k = 0 \quad \forall k$$

Vector Spaces and Basis

- Vector Space:
  - The n-dimensional space in which all the n-dimensional vectors reside



- Basis:
  - A set of vectors $\{u_1, u_2, \dots u_n\}$ are called a basis for a vector space if any vector x can be written as a linear combination of $\{u_i\}$.

$$x = a_1 u_1 + a_2 u_2 + \cdots a_n u_n$$

  - $u_1, u_2, \dots u_n$ are independent implies they form a basis, and vice versa.
  - A basis $\{u_i\}$ is orthonormal if
    - ✓ Basis vectors are pairwise orthogonal, and have unit length, i.e., $|u_i|=1$.

$$u_i^T u_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

Matrices

- An *n* by *d* matrix A and its transpose A^T

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1d} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & & a_{nd} \end{bmatrix}_{n \times d} \qquad A^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ a_{13} & a_{23} & \cdots & a_{n3} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1d} & a_{2d} & & a_{nd} \end{bmatrix}_{d \times n}$$

- Product of two matrices:

$$AB = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1d} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2d} \\ \vdots & \vdots & \vdots & \ddots & \\ a_{m1} & a_{m2} & a_{m3} & & a_{md} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ b_{31} & b_{32} & \cdots & b_{3n} \\ \vdots & \vdots & \ddots & \\ b_{d1} & b_{d2} & & b_{dn} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} & \cdots & c_{1n} \\ c_{21} & c_{22} & c_{23} & \cdots & c_{2n} \\ c_{31} & c_{32} & c_{33} & \cdots & c_{3n} \\ \vdots & \vdots & \vdots & \ddots & \\ c_{m1} & c_{m2} & c_{m3} & & c_{mn} \end{bmatrix}$$

$$\boxed{c_{ij} = \sum_{k=1}^{d} a_{ik} b_{kj}}$$

Matrices (Cont'd)

- Determinant of square matrix **A**

$$|A| = \sum_{k=1}^{d} a_{ik} |A_{ik}| (-1)^{k+i}$$

– minor matrix $\mathbf{A}_{ik}$ is formed by removing the i^{th} row and the k^{th} column of **A**

$$\mathbf{A} = \begin{pmatrix} 1 & 4 & 7 \\ 3 & 0 & 5 \\ -1 & 9 & 11 \end{pmatrix} \quad \mathbf{A}_{23} = \begin{vmatrix} 1 & 4 & \square \\ \square & \square & \square \\ -1 & 9 & \square \end{vmatrix} = \begin{vmatrix} 1 & 4 \\ -1 & 9 \end{vmatrix}$$

– Its transpose has the same determinant: $|\mathbf{A}| = |\mathbf{A}^T|$

- Trace: sum of diagonal elements

$$tr(A) = \sum_{k=1}^{d} a_{kk}$$

Matrices (Cont'd)

- Rank: the number of linearly independent rows (or columns)
- Singular and Non-singular
  - –A singular matrix has a zero determinant
  - –A non-singular matrix has a non-zero determinant
    - ✓ Rank equals the number of rows (or columns)
- Identity matrix: $\mathbf{I}$

$$\mathbf{IA} = \mathbf{AI} = \mathbf{A}$$

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix} \qquad \mathbf{A} = \begin{bmatrix} 1 & 2 & 8 & 6 \\ 2 & 3 & 4 & 9 \\ 8 & 4 & 5 & 6 \\ 6 & 9 & 6 & 7 \end{bmatrix}$$

- Symmetric: $\mathbf{A} = \mathbf{A}^T$

Matrices (Cont'd)

- For a square matrix **A**
  - Orthonormal:

$$\mathbf{A}\mathbf{A}^T = \mathbf{A}^T\mathbf{A} = \mathbf{I}$$

  - Positive definite:

$$\text{if } x^T A x > 0 \text{ for all } x \neq 0$$

  - Semi-positive definite:

$$\text{if } x^T A x \geq 0 \text{ for all } x \neq 0$$

Matrices (Cont'd)

- Inverse
  - The inverse of a square matrix $\mathbf{A}$ is $\mathbf{A}^{-1}$

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

  - The inverse $\mathbf{A}^{-1}$ exists **if and only if** $\mathbf{A}$ is non-singular

$$|\mathbf{A}|\neq 0$$

- Pseudo-inverse

$$\mathbf{A}^{\dagger}=[\mathbf{A}^{T}\mathbf{A}]^{-1}\mathbf{A}^{T} \text{ with } \mathbf{A}^{\dagger}\mathbf{A}=\mathbf{I}$$

  - Assuming $\mathbf{A}^{T}\mathbf{A}$ is non-singular
  - Used whenever $\mathbf{A}^{-1}$ does not exist, i.e., $\mathbf{A}$ is not square or $\mathbf{A}$ is singular.

Linear Transformations

- Mapping from vector space X^N to vector space Y^M, represented by a matrix

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1N} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2N} \\ \vdots & \vdots & \vdots & \ddots & \\ a_{M1} & a_{M2} & a_{M3} & & a_{MN} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$$

- Note that
  - The dimensionality of the two spaces does not need to be the same.
  - For PR, typically $M<N$, i.e., project onto a lower-dimensionality space.

Eigenvectors and Eigenvalues

- Definition: **v** is an eigenvector of matrix **A** if there exists a scalar λ, such that

$$Av = \lambda v \Leftrightarrow \begin{cases} v \ \text{ is an eigenvector} \\ \lambda \ \text{ is the corresponding eigenvalue} \end{cases}$$

- Computation

$$Av = \lambda v \Rightarrow Av - \lambda v = 0 \Rightarrow (A - \lambda I)v = 0 \Rightarrow \begin{cases} v = 0 & \text{trivial solution} \\ (A - \lambda I) = 0 & \text{non-trivial solution} \end{cases}$$

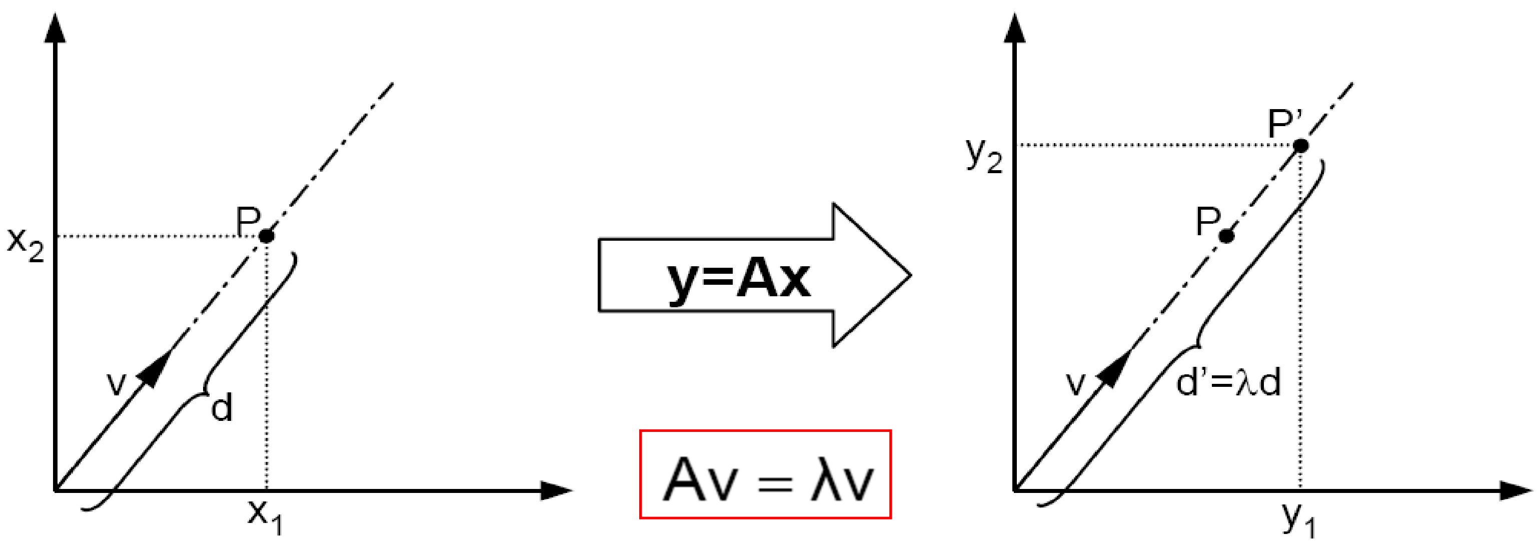
$$(A - \lambda I) = 0 \Rightarrow |A - \lambda I| = 0 \Rightarrow \underbrace{\lambda^N + a_1\lambda^{N-1} + \cdots a_{N-1}\lambda + a_0 = 0}_{\text{Characteristic Equation}}$$

Eigenvectors and Eigenvalues

- Properties
  - If A is non-singular
    - ✓ All eigenvalues are non-zero.
  - If A is real and symmetric
    - ✓ All eigenvalues are real.
    - ✓ The eigenvectors associated with distinct eigenvalues are orthogonal.
  - If A is positive definite
    - ✓ All eigenvalues are positive.

Eigenvectors and Eigenvalues

- Interpretation: an eigenvector represents an invariant direction in the vector space
  - any point lying on the direction defined by **v** remains on that direction
  - its magnitude is multiplied by the corresponding eigenvalue λ

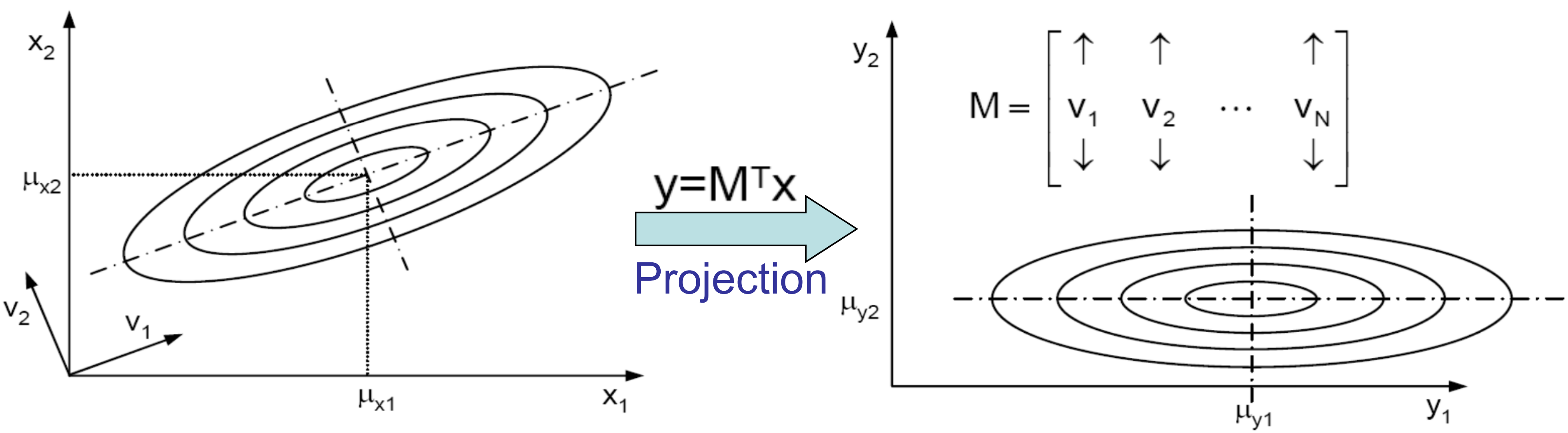


Eigenvectors and Eigenvalues

- For Gaussian distribution
  - The eigenvectors of **Σ** are the principal directions.
  - The eigenvalues are the variances.

$$f_X(x) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu)\right]$$

$$f_Y(y) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left[-\frac{(y_i-\mu_{y_i})^2}{2\lambda_i}\right]$$



Readings

1. Review Appendix A of DHS book, Mathematical foundations