Probability Review

Appendices A.4 & A.5 (Duda et al.)

CS479/679 Pattern Recognition
Dr. George Bebis

Why Bother About Probabilities?

- Probability theory is the proper mechanism to account for uncertainty.
  - e.g., ambiguity in our measurements (features).

- Take into account *a-priori* knowledge.
  - *e.g., "If the fish was caught in the Atlantic ocean, then it is more likely to be salmon than sea-bass"*

Definitions

- Random experiment
  - An experiment whose result is not certain in advance (e.g., throwing a die).

- Outcome
  - The result of a random experiment.
- Sample space
  - The set of all possible outcomes (e.g., {1,2,3,4,5,6}).
- Event
  - A subset of the sample space (e.g., obtain an odd number when throwing a die = {1,3,5}).

Formulation of Probability

- The probability of an event α could be defined as:

$$P(a) = \lim_{n \to \infty} \frac{N(a)}{n}$$

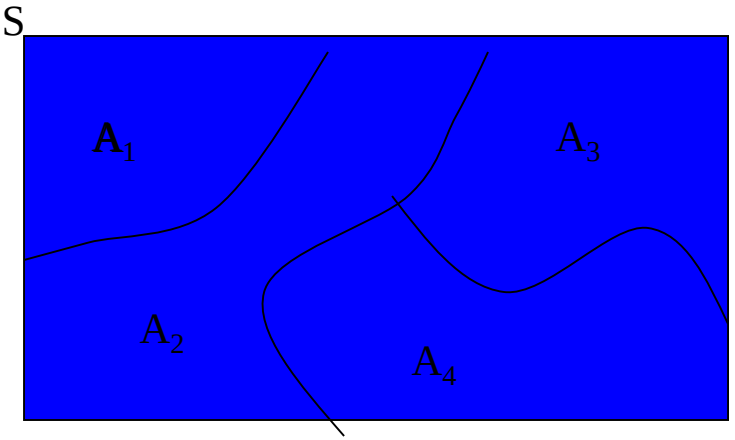
where N(a) is the number of times that event α happens in n trials

- Assumes that all outcomes are *equally likely* (Laplacian definition)

Axioms of Probability

(1) $0 \le P(A) \le 1$

(2) $P(S) = 1$ (S is the sample space)



(3) If $A_1, A_2, \ldots, A_n$ are mutually exclusive events (i.e., $P(A_i \cap A_j) = 0$), then:

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P(A_i)$$

Unconditional (Prior) Probability

- This is the probability of an event in the absence of any evidence.

Example:

P(Cavity)=0.1

"in the absence of any other information, there is a 10% chance that somebody has a cavity".

Conditional (Posterior) Probability

- This is the probability of an event given some evidence (e.g., toothache).

Example:

P(Cavity/Toothache)=0.8

"there is an 80% chance that somebody has a cavity given that he/she is having a toothache"

Conditional (Posterior) Probability (cont'd)

- The conditional probability can be defined as follows:

$$P(A/B)=\frac{P(A\cap B)}{P(B)},\quad P\left(B/A\right)=\frac{P(A\cap B)}{P(A)}$$

Notation:

$$P(A\cap B)\equiv P(A,B)$$

$$P(A/B)=\frac{P(A,B)}{P(B)},\quad P\left(B/A\right)=\frac{P(A,B)}{P(A)}$$

- The above equations lead to the chain rule:

$$P(A,B)=P(A/B)P(B)=P\left(B/A\right)P(A)$$

Law of "Total Probability"

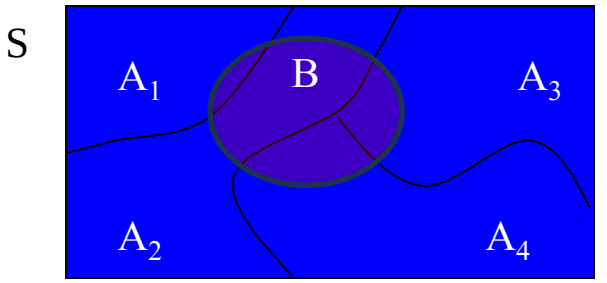
- If $A_1, A_2, \ldots, A_n$ is a partition of mutually exclusive events and B is any event, then:

$$P(B) = P(B, A_1) + P(B, A_2) + \ldots + P(B, A_n) =$$

$$P(B / A_1)P(A_1) + P(B / A_2)P(A_2) + \ldots + P(B / A_n)P(A_n)$$

$$= \sum_{j=1}^{n} P(B / A_j)P(A_j)$$

$$B = (B, A_1) \cup (B, A_2) \cup \ldots \cup (B, A_n)$$



- Special case : $A, \bar{A}$

$$P(B) = P(B, A) + P(B, \bar{A}) = P(B / A)P(A) + P(B / \bar{A})P(\bar{A})$$

Bayes' Rule

- The Bayes' rule can be derived using the definition of conditional probabilities:

$$P(A/B) = \frac{P(A,B)}{P(B)} \quad (i) \qquad P(B/A) = \frac{P(A,B)}{P(A)} \quad (ii)$$

Solve (ii): P(A,B) = P(B/A)P(A)
Substitute in (i)

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

where: $P(B) = P(B/A)P(A) + P(B/\bar{A})P(\bar{A})$

Example

- Consider the probability of *Disease* given *Symptom:*

$$P(Disease / Symptom) = \frac{P(Symptom / Disease)P(Disease)}{P(Symptom)}$$

where:

$$P(Symptom) = P(Symptom / Disease)P(Disease) +$$

$$P(Symptom / \overline{Disease})P(\overline{Disease})$$

Example (cont'd)

- Meningitis causes a stiff neck 50% of the time.

P(S/M)=0.5

- A patient comes in with a stiff neck – what is the probability that he has meningitis?

P(M/S)=?

- Let's use the Bayes rule:

$$P(M/S) = \frac{P(S/M)P(M)}{P(S)}$$

Example (cont'd)

- Need to know:
  - The prior probability of a patient having meningitis:

    e.g., P(M)=1/50,000

  - The prior probability of a patient having a stiff neck:

    e.g., P(S)=1/20

  - What is the probability that he has meningitis?

$$P(M/S) = \frac{P(S/M)P(M)}{P(S)}$$

P(M/S)=0.0002

General Form of Bayes' Rule

- If $A_1, A_2, \ldots, A_n$ is a partition of mutually exclusive events and B is any event, then the Bayes' rule is given by:

$$P(A_i / B) = \frac{P(B / A_i)P(A_i)}{P(B)}$$

where $P(B) = \sum_{j=1}^{n} P(B / A_j)P(A_j)$

Independence

- Two events *A* and *B* are independent iff:

$$P(A,B)=P(A)P(B)$$

- If A and B are independent, then:

$$P(A/B)=P(A) \text{ and } P(B/A)=P(B)$$

- *A* and *B* are conditionally independent given *C* iff:

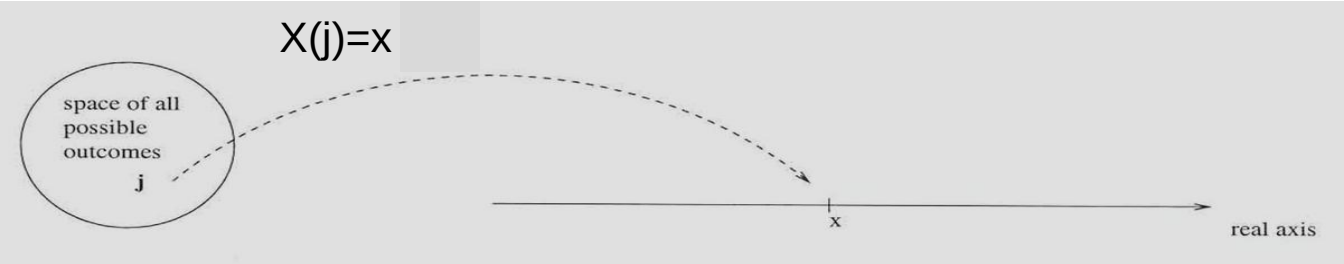
$$P(A,B)\neq P(A)P(B) \quad \underline{but} \quad P(A,B/C)=P(A/C)P(B/C)$$

$$P(A/B,C)=P(A/C)$$

https://www.probabilitycourse.com/chapter1/1_4_4_conditional_independence.php

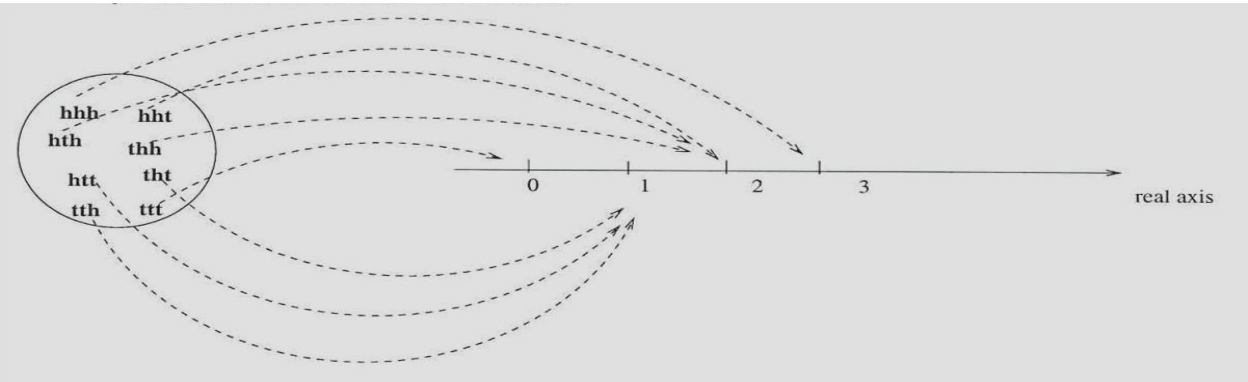
Random Variables

- A random variable (r.v.) is a function that assigns a discrete or continuous value to the outcome of a random experiment.



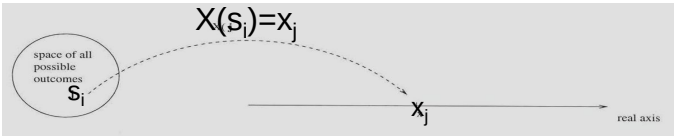
Notation: X: random variable, x: value of random variable

Example: toss a coin 3 times, define X: # of heads



Random Variables (cont'd)

- How do we compute probabilities using random variables?
  - Suppose the sample space is $S=<s_1, s_2, \ldots, s_n>$
  - Let the random variable X assume values in $<x_1,x_2,\ldots,x_m>$
  - What is $P(X = x_j)$?



  - Find all outcomes $s_i \in S$ such that $X(s_i)=x_j$ and take the sum of their probabilities:

$$P(X = x_j) = \sum_{s_i : X(s_i) = x_j} P(s_i)$$

Example

- Consider the random experiment of throwing a pair of dice.
- Let's define X="sum of dice", what is P(X=5)?

e.g., $X = 5$ corresponds to $A_5 = \{(1,4),(4,1),(2,3),(3,2)\}$

$$P(X = x) = P(A_x) = \sum_{s:X(s)=x} P(s) \text{ or}$$

$$P(X = 5) = P((1,4)) + P((4,1)) + P((2,3)) + P((2,3)) = 4/36 = 1/9$$

Random Variables (cont'd)

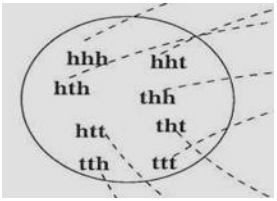
- A random variable (r.v.) can be discrete or continuous:
  - Discrete: assumes only discrete values (countable).
  - Continuous: assumes continuous values (e.g., sensor readings).

- Random variables are associated with a probability distribution.
  - Probability mass function(pmf) P(x) → **discrete** case
  - Probability density function(pdf) p(x) → **continuous** case

Probability mass function (pmf)

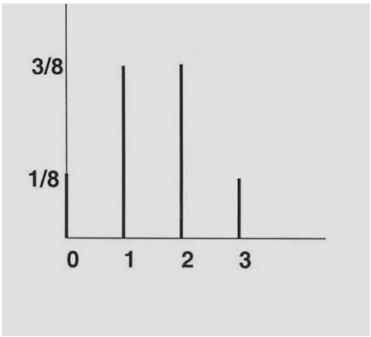
- The *pmf* P(x) of a discrete r.v. *X* assigns a probability to each possible value x of the r.v. *X*.

Example:
toss a coin 3 times
X: # of heads



P(X=0)=P(X=3)=1/8
P(X=1)=P(X=2)=3/8

pmf P(x)



$$\sum_x P(x) = 1$$

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

$P(X = x)$ is denoted as $P(x)$

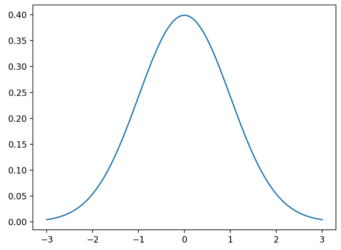
- **Notation:** given two r.v.'s, *X* and *Y*, their *pmfs* are denoted as $P_X(x)$ and $P_Y(y)$; for convenience, we will drop the subscripts and denote them as *P(x)* and *P(y)*. However, keep in mind that these are two different pdfs!

Probability density function (pdf)

- The *pdf* p(x) of a continuous r.v. *X* represents the probability of being close to some value x (i.e., the probability of landing inside an infinitesimal region with area δx is p(x)δx).

Gaussian pdf

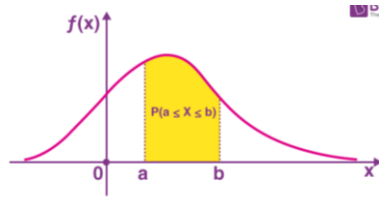
$$p(x) = N(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



$$\int_x p(x)dx = 1$$

$$p(x) \geq 0$$

$$P(a < X < b) = \int_a^b p(x)dx$$



- **Notation:** given two r.v.'s, *X* and *Y*, their *pdf* are denoted as $p_X(x)$ and $p_Y(y)$; for convenience, we will drop the subscripts and denote them as *p(x)* and *p(y)*. However, keep in mind that they are different!

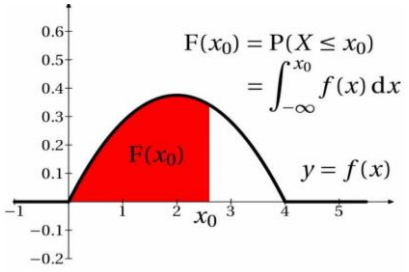
Cumulative Distribution Function (CDF)

- The cumulative distribution function (CDF) is defined as:

$$F(x) = P(X \leq x)$$

- Computing $F(x)$ in the continuous case:

$$F(x) = \int_{-\infty}^{x} p(t)dt$$



- Computing $F(x)$ in the discrete case:

$$F(x) = \sum_{k=0}^{x} P(X = k) = \sum_{k=0}^{x} P(k)$$

F(x) is always non-decreasing

$$0 \leq F(x) \leq 1$$

Example – Discrete Case

$$F(x) = \sum_{k=0}^{x} P(X = k) = \sum_{k=0}^{x} P(k)$$

Example:
toss a coin 3 times
X: # of heads

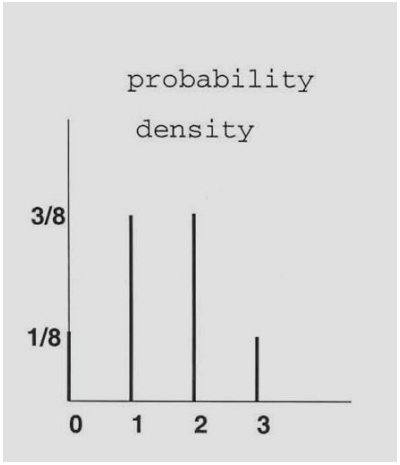
$$F(0) = P(X \le 0) = P(X = 0) = 1/8$$

$$F(1) = P(X \le 1) = P(X = 0) + P(X = 1) = 1/2$$

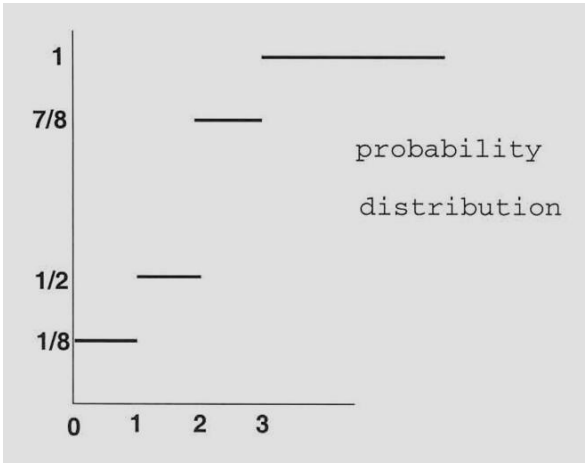
$$F(2) = P(X \le 2) = P(X = 0) + P(X = 1) + P(X = 2) = 7/8$$

$$F(3) = P(X \le 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = 1$$

pmf



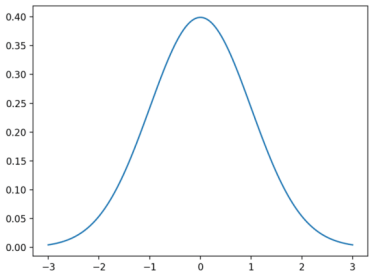
PDF



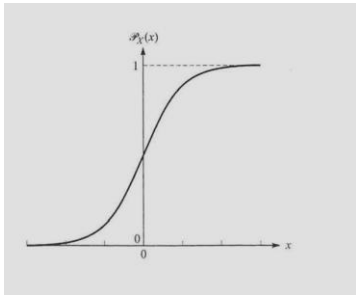
Example – Continuous Case

$$F(x) = \int_{-\infty}^{x} p(t)dt$$

Gaussian pdf



Gaussian PDF

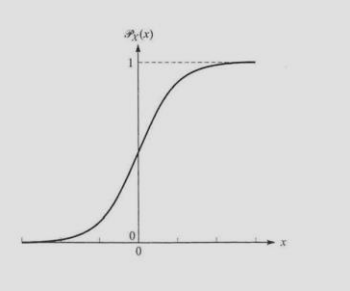


Cumulative Distribution Function (CDF) (cont'd)

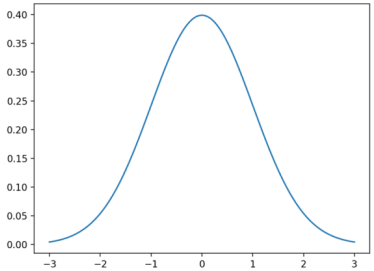
- The pdf can be computed from the PDF using:

$$p(x) = \frac{dF}{dx}(x)$$

Gaussian PDF



Gaussian pdf



Joint *pmf* (discrete case)

- Given *N* random variables ($X_1,X_2,\ldots,X_N$), the joint *pmf* assigns a probability for each set of values ($x_1,x_2,\ldots,x_N$):

$$P(X_1=x_1, X_2=x_2, \ldots, X_N=x_N)=P(x_1,x_2,\ldots,x_N)$$

$$0 \le P(x_1, x_2, \ldots, x_N) \le 1$$

$$\sum_{x_1,x_2,\ldots,x_N} P(x_1, x_2, \ldots, x_N) = 1$$

- **Notation:** the joint *pmf /pdf* of the r.v.'s $X_1, X_2, \ldots, X_N$ and $Y_1, Y_2, \ldots, Y_N$ are denoted as $P_{X1X2\ldots XN}(x_1,x_2,\ldots,x_N)$ and $P_{Y1Y2\ldots YN}(y_1,y_2,\ldots,y_N)$; for convenience, we will drop the subscripts and denote them $P(x_1,x_2,\ldots,x_N)$ and $P(y_1,y_2,\ldots,y_N)$. However, keep in mind that they are <u>different</u> !

Joint *pmf* (discrete case) (cont'd)

- Fully defining the joint *pmf* requires a large number of values!
  - Given N random variables with each variable assuming k values, we need to specify a total of k^N values!

Example:

$$P(X_1, X_2) = P(Cavity, Toothache)$$

Joint Probability

	Toothache -	Toothache +
Cavity -	0.04	0.06
Cavity +	0.01	0.89

(N=2, k=2)

Note: probabilities sum to 1

Joint *pdf* (*continuous* case)

- Given *N* random variables (X_1,X_2,…,X_N), the joint *pdf* shows the probability of being "close" to a set of values (x_1,x_2,…,x_N) (i.e., probabilities are computed over intervals, not single points).

$$p(x_1, x_2, ..., x_N) \geq 0$$

$$\int_{x_1} ... \int_{x_N} p(x_1, x_2, ..., x_N) dx_1 ... dx_N = 1$$

Conditional pdf

- The conditional *pdf* is defined as follows:

$$p(y / x) = \frac{p(x, y)}{p(x)}$$

- The chain rule can be derived from the above definition:

$$p(x, y) = p(y / x)\, p(x)$$

generalizes to n variables

$$p(x_1, x_2, \ldots, x_N) = p(x_1 / x_2, \ldots, x_N)\, p(x_2 / x_3, \ldots, x_N) \ldots p(x_{N-1} / x_N)\, p(x_N)$$

Independence

- We say that *X* and *Y* are independent if:

$$p(x, y) = p(x)\,p(y)$$

2D 1D 1D

- Independence is a very useful property because it can greatly simplify a joint pdf.

Marginalization

- Given p(x,y), what is p(x) or p(y)? (i.e., marginal pdfs)
- Given the joint pmf/pdf of a set of random variables, the computation of the pmf/pdf of **any subset** of the variables is referred to as marginalization.
- Marginalization is simply performed by summing/integrating with respect to the "unwanted" variables.

Examples:

$$\int_{-\infty}^{\infty} p(x,y)dy = p(x)$$

$$\int_{-\infty}^{\infty} p(x_1, x_2, \ldots, x_N)dx_i = p(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N)$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} p(x_1, x_2, \ldots, x_N)dx_3 \ldots dx_N = p(x_1, x_2)$$

Example – Discrete Case

P(X,Y)

	x_1	x_2	x_3
y_1	0.2	0.1	0.1
y_2	0.1	0.2	0.3

Knowledge of the joint pdf/pmf is very powerful since it allows us to compute any other probability!

P(X) → sum with respect to y

P(X)

x_1	x_2	x_3
0.3	0.3	0.4

P(Y) → sum with respect to x

P(Y)

y_1	0.4
y_2	0.6

P(X|Y)

	x_1	x_2	x_3
y_1	0.5	0.25	0.25
y_2	0.167	0.333	0.5

P(Y|X)

	x_1	x_2	x_3
y_1	0.667	0.333	0.25
y_2	0.333	0.667	0.75

Expected Value

- The expected value of a discrete r.v. is given by:

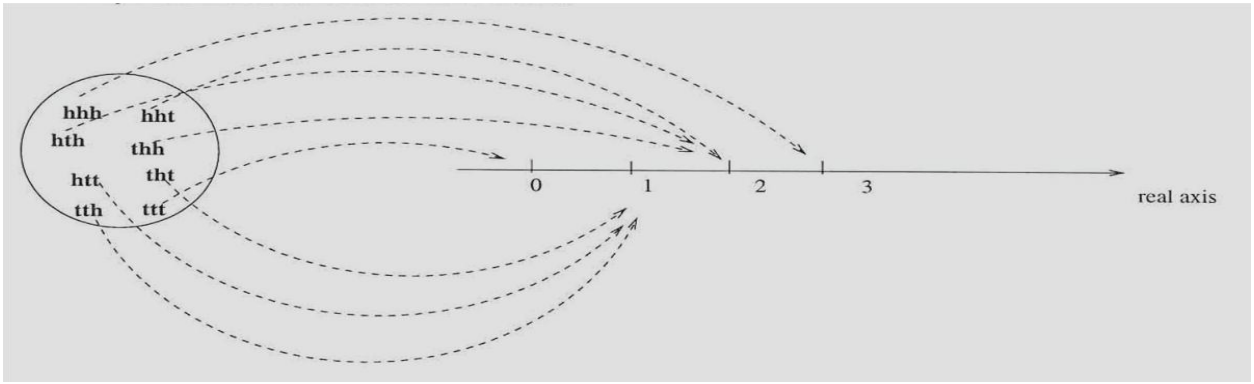
$$\mu_X = E(X) = \sum_x xP(x)$$

- The expected value of a continuous r.v. is given by:

$$\mu_X = E(X) = \int_{-\infty}^{\infty} xp(x)dx$$

Expected Value (cont'd)

Example: toss a coin 3 times, define X: # of heads



$$E(X) = 0P(X=0) + 1P(X=1) + 2P(X=2) + 3P(X=3) =$$

$$0(1/8) + 1(3/8) + 2(3/8) + 3(1/8) = 12/8 = 3/2 = 1.5$$

Expected Value (cont'd)

- In practice, we can approximate E(X) by the sample mean $\hat{\mu}_X$; assuming n data samples x_i:

$$\hat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- The sample mean $\hat{\mu}_X$ and expected value E(X) are related as follows:

$$E(X) = \lim_{n\to\infty} \hat{\mu}_X$$

Variance

- The variance σ_X^2 is defined as:

$$\sigma_X^2 = Var(X) = E((X - \mu_X)^2)$$

- What does σ_X^2 measure?

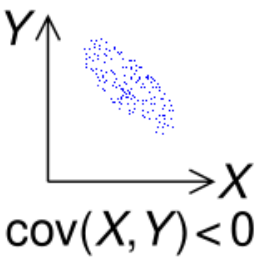
- In practice, we can approximate Var(X) by the sample variance $\hat{\sigma}_X^2$; assuming n data samples x_i :

$$\hat{\sigma}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu}_X)^2$$

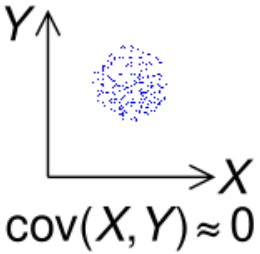
Covariance

- The covariance σ_{XY} of X and Y is defined as:

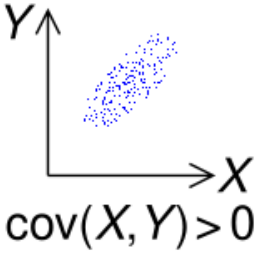
$$\sigma_{XY} = Cov(X,Y) = E((X-\mu_X)(Y-\mu_Y))$$



- What does σ_{XY} measure?



- In practice, we can approximate σ_{XY} by the sample covariance $\hat{\sigma}_{XY}$; assuming n data samples x_i and y_i:



$$\hat{\sigma}_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\hat{\mu}_X)(y_i-\hat{\mu}_Y)$$

Covariance Matrix - 2 variables

- The covariance matrix Σ_{XY} of X and Y is given by:

$$\Sigma_{XY} = \begin{bmatrix} Cov(X,X) & Cov(X,Y) \\ Cov(Y,X) & Cov(Y,Y) \end{bmatrix} = \begin{bmatrix} Var(X) & Cov(X,Y) \\ Cov(Y,X) & Var(Y) \end{bmatrix}$$

where Cov(X,Y)=Cov(Y,X) (i.e., Σ_{XY} is symmetric)

Covariance Matrix – N variables

- The covariance matrix of N variables is given by:

$$\Sigma = \begin{bmatrix} Cov(X_1,X_1) & Cov(X_1,X_2) & \dots & Cov(X_1,X_N) \\ Cov(X_2,X_1) & Cov(X_2,X_2) & \dots & Cov(X_2,X_N) \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ Cov(X_N,X_1) & Cov(X_N,X_2) & \dots & Cov(X_N,X_N) \end{bmatrix} = \begin{bmatrix} Var(X_1) & Cov(X_1,X_2) & \dots & Cov(X_1,X_N) \\ Cov(X_2,X_1) & Var(X_2) & \dots & Cov(X_2,X_N) \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ Cov(X_N,X_1) & Cov(X_N,X_2) & \dots & Var(X_N) \end{bmatrix}$$

$Cov(X_i,X_i)=Var(X_i)$ and $Cov(X_i,X_j)=Cov(X_j,X_i)$ (i.e., symmetric matrix)

- In practice, we approximate Σ by the sample covariance matrix $\hat{\Sigma}$ (i.e., approximate $\sigma_{X_iX_j}$ by $\hat{\sigma}_{X_iX_j}$)

Uncorrelated r.v.'s

- X and Y are uncorrelated iff Cov(X,Y) = 0
- Does uncorrelation imply independence?
- If $X_1, X_2, \ldots, X_N$ are uncorrelated, then Σ is diagonal.

$$\Sigma = \begin{bmatrix} Cov(X_1, X_1) & 0 & \ldots & 0 \\ 0 & Cov(X_2, X_2) & \ldots & 0 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 0 & 0 & \ldots & Cov(X_N, X_N) \end{bmatrix} = \begin{bmatrix} Var(X_1) & 0 & \ldots & 0 \\ 0 & Var(X_2) & \ldots & 0 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 0 & 0 & \ldots & Var(X_N) \end{bmatrix}$$

Properties of Σ

- Σ is always a square and symmetric matrix.
- It has real, non-negative eigenvalues (i.e., due to symmetry).
- Is Σ diagonalizable?

$$\Phi^{-1}\Sigma\Phi = \Lambda$$

- Its eigenvectors form an orthogonal basis.

Decomposition of Σ

- Σ can be decomposed as follows:

$$\Sigma = \Phi \Lambda \Phi^{-1} \quad \text{where} \quad \Phi^{-1} = \Phi^{T}$$

- The columns of Φ are the eigenvectors of Σ

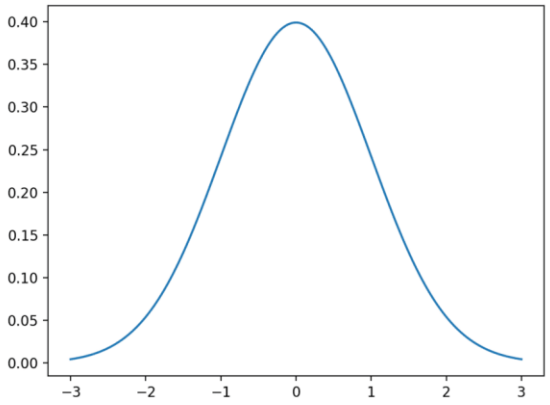
- The diagonal elements of Λ are the eigenvalues of Σ (equal to the variances of Σ)

1D Gaussian Distribution

- The 1D Gaussian pdf is defined as:

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

μ: mean σ: standard deviation



Multivariate Gaussian Distribution

- The multivariate Gaussian (joint) pdf is defined as:

$$\mathbf{x} = (x_1, x_2, ..., x_N) \in R^N$$

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

determinant

μ: mean (N x 1) Σ: covariance matrix (N x N)

- Number of parameters: $N + \frac{N(N+1)}{2} = O(N^2)$

μ Σ

Gaussian Distribution (cont'd)

- Location of distribution is determined by μ
- Shape of distribution is determined by Σ

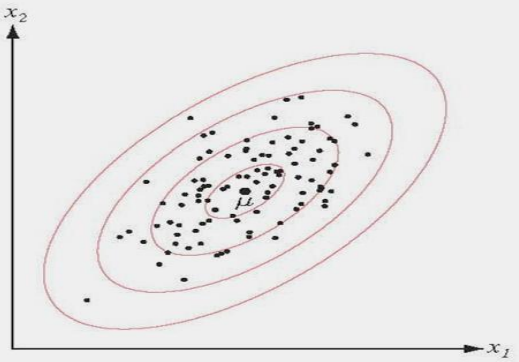


FIGURE 2.9. Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean $\boldsymbol{\mu}$. The ellipses show lines of equal probability density of the Gaussian. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

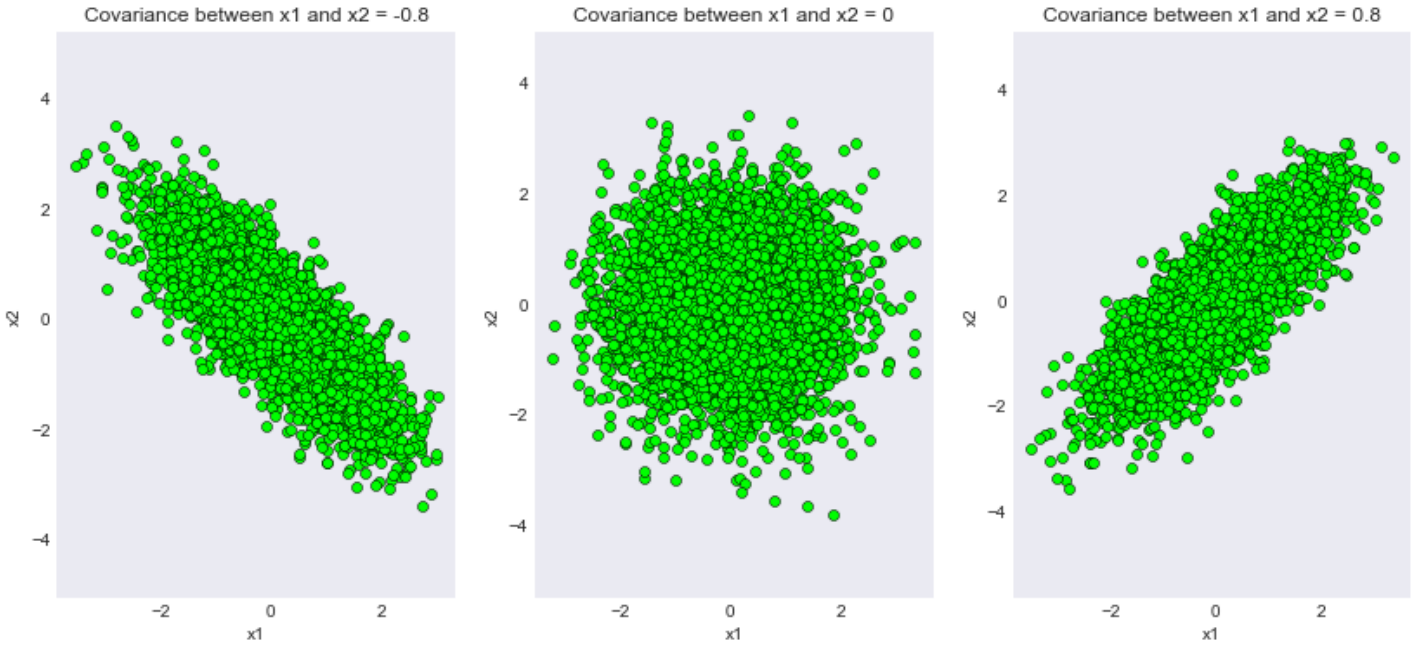
Gaussian Distribution (cont'd)

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$



Gaussian Distribution (cont'd)

- In the case of uncorrelated r.v., the multivariate normal distribution can be greatly simplified:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{N/2} \mid \Sigma \mid^{1/2}} \exp[-\frac{1}{2}(\mathbf{x}-\mu)^t \Sigma^{-1}(\mathbf{x}-\mu)]$$

Σ is diagonal

$$p(\mathbf{x}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp[\frac{(x_i-\mu_i)^2}{2\sigma_i^2}]$$

Linear Transformations

- Let us assume that we obtain a new set of random variables **Y** by applying a linear transformation A on a set of jointly Gaussian distributed variables **X**:

$$p(\mathbf{x}) \sim N(\mu_{\mathbf{X}}, \Sigma_{\mathbf{X}})$$

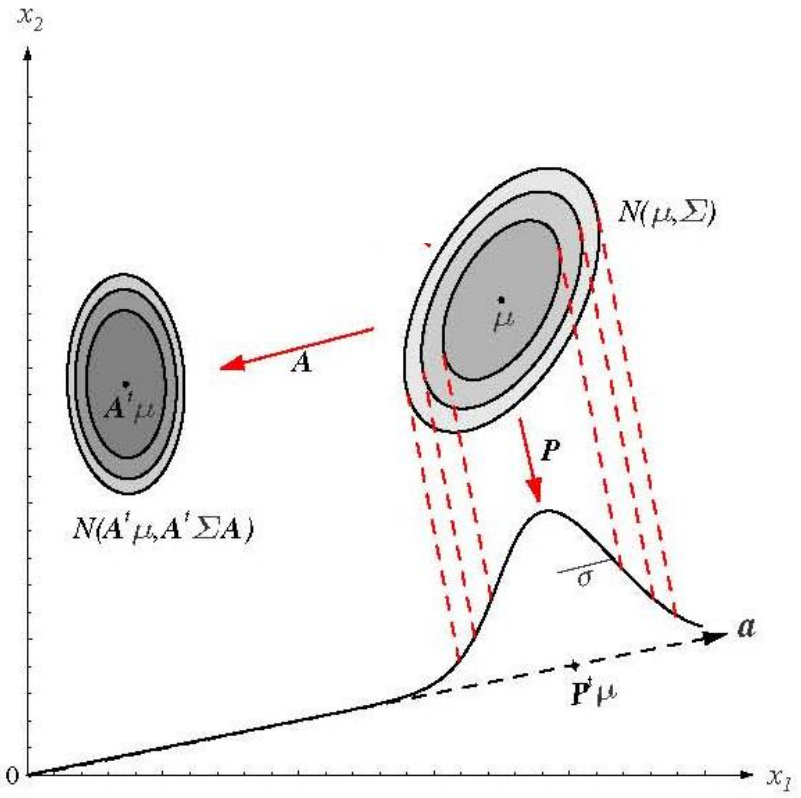
$$\mathbf{Y} = A^t \mathbf{X}$$

- It can be shown that **Y** follows a multivariate Gaussian distribution with the following parameters:

$$p(\mathbf{y}) \sim N(A^t \mu_{\mathbf{X}}, A^t \Sigma_{\mathbf{X}} A)$$

new mean new covariance

Linear Transformations (cont'd)



Whitening Transformation

- Consider the following transformation:

$$A_w = \Phi \Lambda^{-1/2} \qquad \text{where} \qquad \Sigma_{\mathbf{X}} = \Phi \Lambda \Phi^{-1}$$

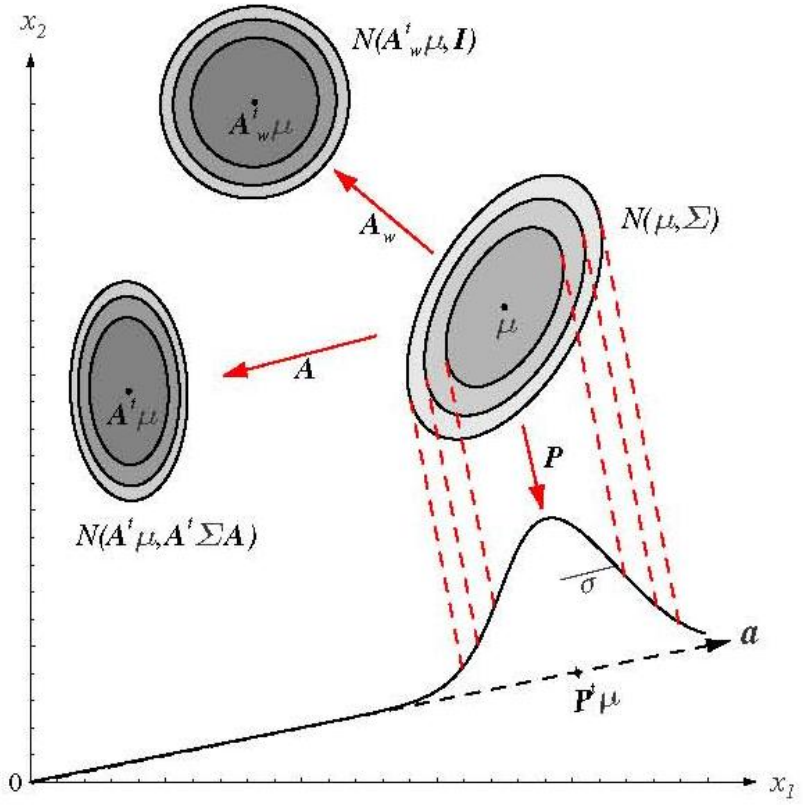
- It can be shown that if $\mathbf{Y} = A_w^t \mathbf{X}$, then:

$$p(\mathbf{x}) \sim N(\mu_{\mathbf{X}}, \Sigma_{\mathbf{X}}) \Rightarrow p(\mathbf{y}) \sim N(A_w^t \mu_{\mathbf{X}}, I)$$

that is: $\mu_{\mathbf{Y}} = A_w^t \mu_{\mathbf{X}}$

$$\Sigma_Y = I \Rightarrow$$ **Y** are uncorrelated, with unit variance

Whitening Transformation (cont'd)



Quiz #2

- When: Wednesday 2/19 @ 2:30pm
  - Closed book/notes

- What: Review of Linear Algebra, Review of Probability

Practice Problems

- Work on the problems shown in the next slides.
- Good practice for the **midterm** and **final** exams!
- Will also help with **quizzes** by enhancing your understanding of the material!

Practice Problem 1

After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease, and that the test is 99% accurate (i.e., the probability of testing positive given that you have the disease is 0.99, as it is the probability of testing negative given that you do not have the disease). The good news is that this is a rare disease, striking only one in 10,000 people.

- Why is it good news that the disease is rare?
- What are the chances that you actually have the disease?

Practice Problem 1 (cont'd)

T^+: test is positive T^-: test is negative

D: has disease $\bar{D}$: does not have disease

Given: $P(T^+/D) = 0.99$ $P(T^-/\bar{D}) = 0.99$

$P(D) = 0.0001$

We can infer: $P(\bar{D}) = 0.9999$, $P(T^+/\bar{D}) = 0.01$

We need to compute:

$$P(D/T^+) = \frac{P(T^+/D)P(D)}{P(T^+)} \text{ where}$$

$$P(T^+) = P(T^+/D)P(D) + P(T^+/\bar{D})P(\bar{D})$$

$$= 0.99 \times 0.0001 + 0.01 \times 0.9999 =$$

$$= 0.010098$$

So, $P(D/T^+) = \frac{0.99 \times 0.0001}{0.010098} = 0.0098$

Practice Problem 2

- The following problem investigates the way in which conditional independence relationships affect the amount of information needed in probabilistic calculations.
- Suppose we wish to calculate $P(H/E_1,E_2)$, and we have no conditional independence information. Which of the following sets of numbers are sufficient for the calculations?

  i) $P(E_1, E_2)$, $P(H)$, $P(E_1/H)$, $P(E_2/H)$

  ii) $P(E_1,E_2)$, $P(H)$, $P(E_1,E_2/H)$

  iii) $P(H)$, $P(E_1/H)$, $P(E_2/H)$

We do not have any independence assumptions:

$$P(H|E_1, E_2) = \Big(P(E_1, E_2|H)P(H)\Big)\Big/P(E1, E2)$$

So we need to know these probabilities: $P(H), P(E_1, E_2|H)$ and $P(E_1, E_2)$, which is (set ii).

Practice Problem 2 (cont'd)

- Suppose we know that $P(E_1/H,E_2)=P(E_1/H)$ for all values of H, E_1, E_2. Now which of the above three sets are sufficient?

If we know that $\boxed{P(E_1|H,E_2) = P(E_1|H)}$ *: (see slide #15)

$$
\begin{aligned}
P(H|E_1,E_2) &= \frac{P(H,E1,E2)}{P(E_1,E_2)} && : P(E_1,H,E_2) = P(E_1|H,E_2)P(H,E_2) \\
&= \frac{P(E_1|H,E_2)P(H,E_2)}{P(E_1,E_2)} && : P(E_1|H,E_2) = P(E_1|H) * \\
&= \frac{P(E_1|H)P(E_2,H))}{P(E_1,E_2)} && : P(E_2,H) = P(E_2|H)P(H) \\
&= \frac{P(E_1|H)P(E_2|H)P(H)}{P(E_1,E_2)}
\end{aligned}
$$

Thus we need to know the following probabilities: $P(E_1|H)$, $P(E_2|H)$, $P(H)$, and $P(E_1,E_2)$; which is (set i).

Practice Problem 3

- The task of finding and removing apples that house worms before they get to the grocery store is a big problem. To combat this problem, Wormfinder Inc. has developed an amazing new non-intrusive test for worms in apples. This test is, called WormScan has the incredible false negative rate of exactly 0 (i.e., if an apple is declared by WormScan to be free of worms, it is guaranteed to have no worms in it). Unfortunately, such a performance comes with a cost; the false positive rate is 3% (i.e., 3% of all good apples are marked as having a worm inside). Statistically, it has been found that 0.2% (1 in 500 apples) have worms.
  - What percentage of the apples will test as having worms?
  - Given that an apple has tested as having worms, what is the probability that there is a worm inside?

Practice Problem 3 (cont'd)

T^+: test is positive, T^-: test is negative

W: apple has worm, $\bar{W}$: apple does not have worm

Given: FN = 0 $P(T^-/W) = 0$ $P(T^+/W) = 1.0$

FP = 3% $P(T^+/\bar{W}) = 0.03$

$P(W) = 0.002$

(a) $P(T^+) = P(T^+/W)P(W) + P(T^+/\bar{W})P(\bar{W})$

$= 1 \times 0.002 + 0.03 \times 0.998 = 0.03194$

(b) $P(W/T^+) = \dfrac{P(T^+/W)\,P(W)}{P(T^+)} = \dfrac{1 \times 0.002}{0.03194} = 0.0626$

Practice Problem 4

23. Consider the three-dimensional normal distribution $p(\mathbf{x}|\omega) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 5 & 2 \\ 0 & 2 & 5 \end{pmatrix}$.

(a) Find the probability density at the point $\mathbf{x}_0 = (.5, 0, 1)^t$.

(b) Construct the whitening transformation $\mathbf{A}_w$. Show your $\boldsymbol{\Lambda}$ and $\boldsymbol{\Phi}$ matrices. Next, convert the distribution to one centered on the origin with covariance matrix equal to the identity matrix, $p(\mathbf{x}|\omega) \sim N(\mathbf{0}, \mathbf{I})$.

(c) Apply the same overall transformation to $\mathbf{x}_0$ to yield a transformed point $\mathbf{x}_w$.

(d) By explicit calculation, confirm that the Mahalanobis distance from $\mathbf{x}_0$ to the mean $\boldsymbol{\mu}$ in the original distribution is the same as for $\mathbf{x}_w$ to $\mathbf{0}$ in the transformed distribution.

(e) Does the probability density remain unchanged under a general linear transformation? In other words, is $p(\mathbf{x}_0|N(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = p(\mathbf{T}^t\mathbf{x}_0|N(\mathbf{T}^t\boldsymbol{\mu}, \mathbf{T}^t\boldsymbol{\Sigma}\mathbf{T}))$ for some linear transform $\mathbf{T}$? Explain.

(f) Prove that a general whitening transform $\mathbf{A}_w = \boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2}$ when applied to a Gaussian distribution insures that the final distribution has covariance proportional to the identity matrix $\mathbf{I}$. Check whether normalization is preserved by the transformation.

Practice Problem 4 (cont'd)

Consider the 3-dimensional normal distribution $p(\mathbf{x}|\omega) \sim N(\mu, \Sigma)$ where:

$$\mu = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 5 & 2 \\ 0 & 2 & 5 \end{pmatrix}$$

a)

Find the probability density at the point $\mathbf{x}_0 = (0.5, 0, 1)^T$:

The probability density at point $\mathbf{x}_0$ is:

$$\begin{aligned}
p(\mathbf{x}_0|\omega) &= \frac{1}{(2\pi)^{3/2}|\Sigma|^{1/2}} \exp\left\{\frac{-1}{2}(\mathbf{x}_0 - \mu)^T \Sigma^{-1}(\mathbf{x}_0 - \mu)\right\} \\
&= \frac{1}{(2\pi)^{3/2}\sqrt{21}} \exp\left\{\frac{-1}{2}\left(\begin{pmatrix} 0.5 \\ 0 \\ 1 \end{pmatrix} - \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}\right)^T \frac{1}{21}\begin{pmatrix} 21 & 0 & 0 \\ 0 & 5 & -2 \\ 0 & -2 & 5 \end{pmatrix}\left(\begin{pmatrix} 0.5 \\ 0 \\ 1 \end{pmatrix} - \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}\right)\right\} \\
&= \frac{1}{(2\pi)^{3/2}\sqrt{21}} \exp\left\{\frac{-1}{2}\begin{pmatrix} -0.5 \\ -2 \\ -1 \end{pmatrix}^T \frac{1}{21}\begin{pmatrix} 21 & 0 & 0 \\ 0 & 5 & -2 \\ 0 & -2 & 5 \end{pmatrix}\begin{pmatrix} -0.5 \\ -2 \\ -1 \end{pmatrix}\right\} \\
&= \frac{1}{(2\pi)^{3/2}\sqrt{21}} \exp\left\{\frac{-1}{2} \times 1.0595\right\} \\
&= \frac{1}{(2\pi)^{3/2}\sqrt{21}} \exp\left\{-0.5298\right\} \\
&= \frac{0.5887}{72.1738} \\
&= 0.0082 \Rightarrow \boxed{p(\mathbf{x}_0|\omega) = 0.0082}
\end{aligned}$$

b)

To find the whitening transform A_ω we need to compute the eigen values and eigen vectors of the covariance matrix Σ:

$$\det(\Sigma - \lambda \mathbf{I}) = \mathbf{0} \Rightarrow \det\begin{bmatrix} 1-\lambda & 0 & 0 \\ 0 & 5-\lambda & 2 \\ 0 & 2 & 5-\lambda \end{bmatrix} = 0 \Rightarrow \lambda_1 = 7, \lambda_2 = 3, \lambda_3 = 1$$

So the eigen vector matrix Φ and eigenvalue matrix Λ are as the following:

$$\Phi = \begin{bmatrix} 0 & 0 & 1 \\ -0.7071 & -0.7071 & 0 \\ -0.7071 & 0.7071 & 0 \end{bmatrix} \text{ and } \Lambda = \begin{bmatrix} 7 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Thus the whitening transformation A_ω is computed as:

$$\mathbf{A}_\omega = \begin{bmatrix} 0 & 0 & 1 \\ -0.7071 & -0.7071 & 0 \\ -0.7071 & 0.7071 & 0 \end{bmatrix} \times \begin{bmatrix} 7 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}^{-\frac{1}{2}} = \begin{bmatrix} 0 & 0 & 1 \\ -0.2673 & -0.4082 & 0 \\ -0.2673 & 0.4082 & 0 \end{bmatrix}$$

$A_\omega^T \mu$

To transform the probability distribution to one centered at the origin with covariance matrix equal to the Identity matrix we have to perform this transformation : $\mathbf{y} = \mathbf{x} - \mathbf{A}_\omega^T \mu$. This is because the transformation $\mathbf{T} = \mathbf{A}_\omega^T \mu$ transforms the probability function to a distribution centered at $\mathbf{A}_\omega^T \mu$ with identity covariance matrix : $p(\mathbf{x}) \sim N(\mathbf{A}_\omega^T \mu, \mathbf{A}_\omega^T \Sigma \mathbf{A}_\omega) = N(\mathbf{A}_\omega^T \mu, \mathbf{I})$.

Practice Problem 4 (cont'd)

c)

To apply the same transformation as in b) on the points in space including $\mathbf{x}_0$, we have:

$$\mathbf{x}_\omega = \mathbf{A}_\omega^T \mathbf{x}_0 - \mathbf{A}_\omega^T \mu = [0.8018, 0.4082, -0.5]^T$$

d)

The Mahalanobis distance of vector $\mathbf{x}_0$ from μ is:

$$r_{x_0} = (\mathbf{x}_0 - \mu)^T \Sigma^{-1} (\mathbf{x}_0 - \mu) = \begin{pmatrix} -0.5 \\ -2 \\ -1 \end{pmatrix}^T \frac{1}{21} \begin{pmatrix} 21 & 0 & 0 \\ 0 & 5 & -2 \\ 0 & -2 & 5 \end{pmatrix} \begin{pmatrix} -0.5 \\ -2 \\ -1 \end{pmatrix} = 1.0595$$

Mahalanobis distance of the transformed vector $\mathbf{x}_\omega$ from the origin under the transformed distribution is:

$$r_{x_\omega} = (\mathbf{x}_\omega - 0)^T \mathbf{I}^{-1} (\mathbf{x}_\omega - 0) = \begin{pmatrix} 0.8018 \\ 0.4082 \\ -0.5 \end{pmatrix}^T \begin{pmatrix} 0.8018 \\ 0.4082 \\ -0.5 \end{pmatrix} = 1.0595$$

This confirms that the Mahalanobis distance from a vector $\mathbf{x}$ from the mean of the original distribution is equal to the distance of its transformation under whitening transformation from the origin.

e)

In this part of the problem we want to see whether the probability remains unchanged under a general linear transformation. That is if the following equation holds for all $\mathbf{x}_0$ values; $p(\mathbf{x}_0|N(\mu, \Sigma)) = p(\mathbf{T}^t\mathbf{x}_0|N(\mathbf{T}^t\mu, \mathbf{T}^t\Sigma\mathbf{T}))$. We have:

$$\begin{aligned} p(\mathbf{T}^t\mathbf{x}_0|N(\mathbf{T}^t\mu, \mathbf{T}^t\Sigma\mathbf{T})) &= \frac{1}{(2\pi)^{d/2}|\mathbf{T}^t\Sigma\mathbf{T}|^{1/2}} \exp\left[-(\mathbf{T}^t\mathbf{x}_0 - \mathbf{T}^t\mu)^t(\mathbf{T}^t\Sigma\mathbf{T})^{-1}(\mathbf{T}^t\mathbf{x}_0 - \mathbf{T}^t\mu)\right] \\ &= \frac{1}{(2\pi)^{d/2}|\mathbf{T}^t|^{1/2}|\Sigma|^{1/2}|\mathbf{T}|^{1/2}} \exp\left[-(\mathbf{x}_0 - \mu)\mathbf{T}\mathbf{T}^{-1}\Sigma^{-1}\mathbf{T}^{t^{-1}}\mathbf{T}^t(\mathbf{x}_0 - \mu)\right] \\ &= \frac{1}{(2\pi)^{d/2}|\mathbf{T}^t|^{1/2}\mathbf{T}|^{1/2}|\Sigma|^{1/2}|} \exp\left[-(\mathbf{x}_0 - \mu)\mathbf{I}\Sigma^{-1}\mathbf{I}(\mathbf{x}_0 - \mu)\right] \\ &= \frac{1}{(|\mathbf{T}^t|^{1/2}|\mathbf{T}|^{1/2})((2\pi)^{d/2}|\Sigma|^{1/2})} \exp\left[-(\mathbf{x}_0 - \mu)\Sigma^{-1}(\mathbf{x}_0 - \mu)\right] \\ &= \frac{1}{|\mathbf{T}^t|^{1/2}|\mathbf{T}|^{1/2}} \times p(\mathbf{x}_0|N(\mu, \Sigma)) \end{aligned}$$

Special case: if $T^t = T^{-1}$ (orthogonal)

From the above equation sets we can see that the transformed probability distribution is proportional to the original one with a factor of $\frac{1}{|T^t|^{1/2}|T|^{1/2}}$. If this factor for a linear transformation T is equal to one, then the transformed probability distribution remains unchanged, but in general cases it may vary.

Note: The Mahalanobis distance is the exponent, we will discuss this later in the course!

Practice Problem 4 (cont'd)

f)

In this part of the problem we want to prove that the whitening transformation ensures that the covariance matrix of the transformed distribution is equal to the identity matrix I. From the definition of whitening transform we have:

$$\mathbf{A}_\omega = \Phi\Lambda^{-1/2}$$

Also we know that Φ is an orthogonal matrix, and Λ is a diagonal matrix whose diagonal elements are eigen values of the covariance matrix Σ. We have:

$$\begin{aligned}
\Sigma_\omega &= \mathbf{A}_\omega^T \Sigma \mathbf{A}_\omega \\
&= (\Phi\Lambda^{-1/2})^T \Sigma (\Phi\Lambda^{-1/2}) \\
&= \Lambda^{T^{-1/2}} \Phi^T \Sigma \Phi \Lambda^{-1/2} \Rightarrow && \text{From the properties of } \Phi\text{: } \Phi^T\Sigma\Phi = \Lambda\text{: Singular Value Decomposition} \\
&= \Lambda^{T^{-1/2}} \Lambda\Lambda^{-1/2} \Rightarrow && \Lambda \text{ is diagonal } \Rightarrow \Lambda^T = \Lambda \\
&= \Lambda^{-1/2}\Lambda\Lambda^{-1/2} \\
&= \mathbf{I}
\end{aligned}$$

So we have : $\Sigma_\omega = \mathbf{I}$, which means that the covariance matrix of the transformed distribution is the identity matrix.