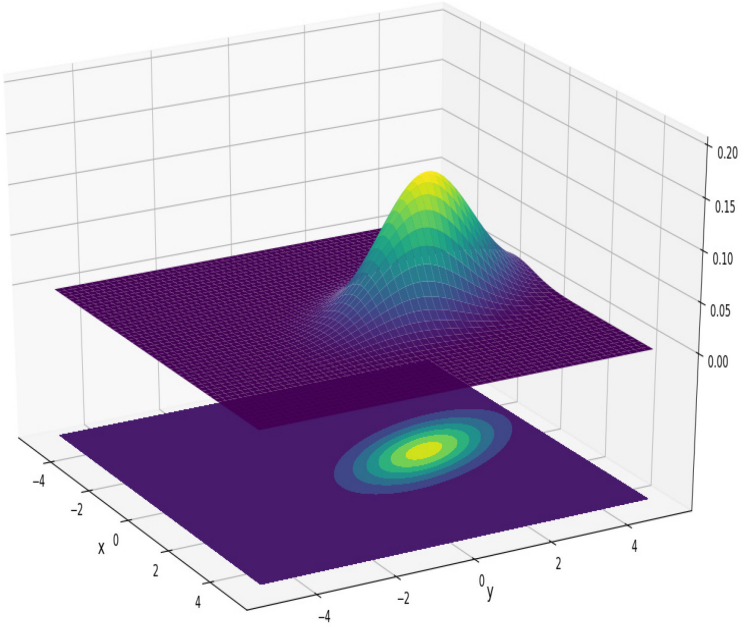
NTB
Interstaatliche Hochschule für Technik Buchs
FHO Fachhochschule Ostschweiz



MSE MachLe
Bayes Theorem

Christoph Würsch
Institute for Computational Enineering ICE
Interstaatliche Hochschule für Technik Buchs, FHO

Bayes Theorem and the Gaussian Distribution

Educational objectives

- Apply Bayesian learning, especially **Bayes' theorem** and the Bayes classifier
- Explain how a **Gaussian Mixture Model (GMM)** is trained and evaluated, given the respective equations and the EM algorithm
- Apply GMMs for **pattern recognition tasks** on audio data

Based on material by
- Stuart Russell, UC Berkeley
- T. Stadelmann, R. Ewerth & B. Freisleben, U Marburg

I. Probability Theory

(a short repetition)

$$p(X,Y) = p(Y|X) \cdot p(X)$$
$$= p(X|Y) \cdot p(Y)$$

Educational Objectives

- you have repeated the basic rules of probability theory.
- you know the difference between a **joint** and a **conditional probability** distribution.
- you know how to apply **Bayes Theorem** to calculate the posterior probability distribution for simple discrete examples.
- You can name the **prior probability** distribution, the **likelihood function**, the **evidence**, and you know how to **marginalize** over a joint probability distribution.
- you know the basic properties of a **multivariate Gaussian probability distribution**. You can plot a 2D Gaussian probability distribution given the mean vector μ and the covariance matrix Σ
- you can estimate the parameters of a multivariate Gaussian distribution from data points using a **kernel density estimation method.**

Basic Probability Theory (recap)

- **Probability** and random variable:

$$p(x = \mathsf{x}) : \text{the probability of variable } x \text{ being in state } \mathsf{x}.$$

$$p(x = \mathsf{x}) = \begin{cases} 1 & \text{we are certain } x \text{ is in state } \mathsf{x} \\ 0 & \text{we are certain } x \text{ is not in state } \mathsf{x} \end{cases}$$

Values between 0 and 1 represent the degree of certainty of state occupancy.

- **Domain**: dom( )

- dom( ) denotes the states x can take. For example:

$$\text{dom}(coin) = \{\mathsf{heads}, \mathsf{tails}\}.$$

- When summing over a variable x, the interpretation is that all states of x are included

$$\sum_x f(x) \equiv \sum_{\mathsf{s} \in \text{dom}(x)} f(x = \mathsf{s}).$$

Basic Rules of Probability

- **Sum Rule:**

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

Marginal

- **Product Rule:**

$$p(X, Y) = p(Y|X) \cdot p(X) = p(X|Y) \cdot p(Y)$$

Conditional

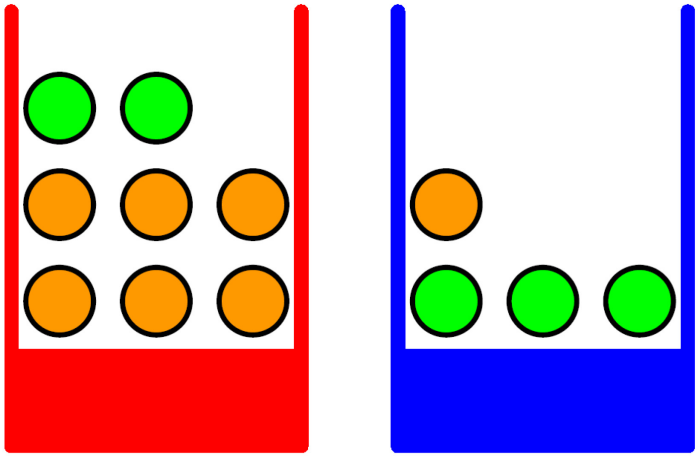
- **Bayes Theorem:**

$$p(Y|X) = \frac{p(X|Y) \cdot p(Y)}{p(X)} = \frac{p(X|Y) \cdot p(Y)}{\sum_{y \in Y} p(X, y)} = \frac{p(X|Y) \cdot p(Y)}{\sum_{y} p(X|y) \cdot p(y)}$$

Bayes Rule

Probability Tables – Joint Probability

- Two colored boxes (X=r, X=b) with apples (Y=a) and oranges (Y=o)
- We may ask: What is the overall probability that the selection procedure will pick an apple? $p(Y = a) = ?$
- Or: "given that we have chosen an orange, what is the probability that the box we chose was the blue one?".



$$p(X = b|Y = o) = ?$$

- Using the concept of the **joint probability (density),** the marginal probability and the sum and product rule, we can answer all those questions. We define:

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

Most simple non-trivial example

- If we count the frequencies, we get the following table. In total, there are 12 fruits, 8 are in a red box, 4 in a blue box. We have 5 apples and 7 oranges in total.

$n_{ij}(X,Y)$	$Y=a$	$Y=o$	$n(X)$
$X=r$	2	6	8
$X=b$	3	1	4
$n(Y)$	5	7	$N=12$

- We can convert this into a probability table by normalization:

$p(X,Y)$	$Y=a$	$Y=o$	$p(X)$
$X=r$	$\frac{2}{12}$	$\frac{6}{12}$	$\frac{8}{12}$
$X=b$	$\frac{3}{12}$	$\frac{1}{12}$	$\frac{4}{12}$
$p(Y)$	$\frac{5}{12}$	$\frac{7}{12}$	1

Conditional Probabilities

- This table is the joint probability of this problem, the colums at the borders are the marginals. Now we can easily calculate the **conditional probabilities using the joint and the marginals.**

$$p(Y = a|X = r) = \frac{p(Y = a, X = r)}{p(X = r)} = \frac{\frac{2}{12}}{\frac{8}{12}} = \frac{1}{4}$$

$$p(Y = o|X = r) = \frac{p(Y = o, X = r)}{p(X = r)} = \frac{\frac{6}{12}}{\frac{8}{12}} = \frac{3}{4}$$

$$p(Y = a|X = b) = \frac{p(Y = a, X = b)}{p(X = b)} = \frac{\frac{3}{12}}{\frac{4}{12}} = \frac{3}{4}$$

$$p(Y = o|X = b) = \frac{p(Y = o, X = b)}{p(X = b)} = \frac{\frac{1}{12}}{\frac{4}{12}} = \frac{1}{4}$$

Bayes Theorem

- Using **Bayes Theorem**, we can calculate the probability to draw an apple (5 from 12) resp. Orange (7 from 12):

$$p(Y=a) = p(Y=a|X=r) \cdot p(X=r) + p(Y=a)|X=b) \cdot p(X=b)$$
$$= \frac{1}{4} \cdot \frac{8}{12} + \frac{3}{4} \cdot \frac{4}{12} = \frac{5}{12}$$

- The conditionals are not symmetric:

$$p(Y=a|X=r) = \frac{p(Y=a, X=r)}{p(X=r)} = \frac{\frac{2}{12}}{\frac{8}{12}} = \frac{1}{4}$$

$$p(X=r|Y=a) = \frac{p(Y=a, X=r)}{p(Y=a)} = \frac{\frac{2}{12}}{\frac{5}{12}} = \frac{2}{5}$$

Bayes Theorem

- We can test Bayes rule:

$$\frac{1}{4} = p(Y = a|X = r) = \frac{p(X = r|Y = a) \cdot p(Y = a)}{p(X = r)} = \frac{\frac{2}{5} \cdot \frac{5}{12}}{\frac{8}{12}}$$

Inspector Clouseau

Inspector Clouseau arrives at the scene of a crime. The Butler (B) and Maid (M) are his main suspects. The inspector has a prior belief of 0.6 that the Butler is the murderer, and a prior belief of 0.2 that the Maid is the murderer. These probabilities are independent in the sense that $p(B, M) = p(B)p(M)$. (It is possible that both the Butler and the Maid murdered the victim or neither). The inspector's *prior* criminal knowledge can be formulated mathematically as follows:

$$\mathrm{dom}(B) = \mathrm{dom}(M) = \{\text{murderer}, \text{not murderer}\}$$

$$\mathrm{dom}(K) = \{\text{knife used}, \text{knife not used}\}$$

$$p(B = \text{murderer}) = 0.6, \qquad p(M = \text{murderer}) = 0.2$$

$$\begin{aligned}
p(\text{knife used}|B = \text{not murderer}, \quad & M = \text{not murderer}) && = 0.3 \\
p(\text{knife used}|B = \text{not murderer}, \quad & M = \text{murderer}) && = 0.2 \\
p(\text{knife used}|B = \text{murderer}, \quad & M = \text{not murderer}) && = 0.6 \\
p(\text{knife used}|B = \text{murderer}, \quad & M = \text{murderer}) && = 0.1
\end{aligned}$$

The victim lies dead in the room and the inspector quickly finds the murder weapon, a Knife (K). What is the probability that the Butler is the murderer? (Remember that it might be that neither is the murderer).

Inspector Clouseau II

Using b for the two states of B and m for the two states of M,

$$p(B|K) = \sum_m p(B, m|K) = \sum_m \frac{p(B, m, K)}{p(K)} = \frac{p(B) \sum_m p(K|B, m)p(m)}{\sum_b p(b) \sum_m p(K|b, m)p(m)}$$

Plugging in the values we have

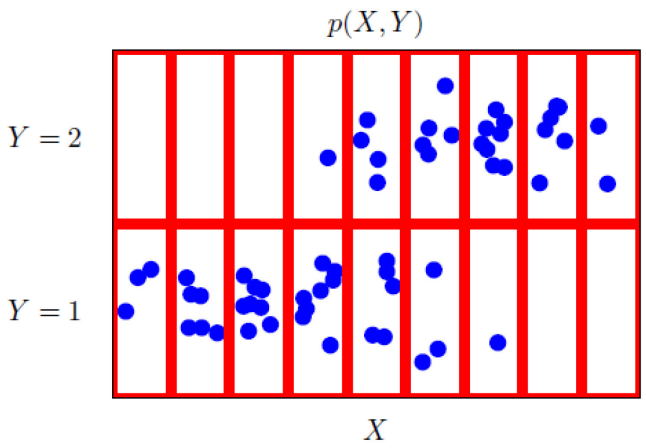
$$p(B = \text{murderer}|\text{knife used}) = \frac{\frac{6}{10}\left(\frac{2}{10} \times \frac{1}{10} + \frac{8}{10} \times \frac{6}{10}\right)}{\frac{6}{10}\left(\frac{2}{10} \times \frac{1}{10} + \frac{8}{10} \times \frac{6}{10}\right) + \frac{4}{10}\left(\frac{2}{10} \times \frac{2}{10} + \frac{8}{10} \times \frac{3}{10}\right)}$$
$$= \frac{300}{412} \approx 0.73$$

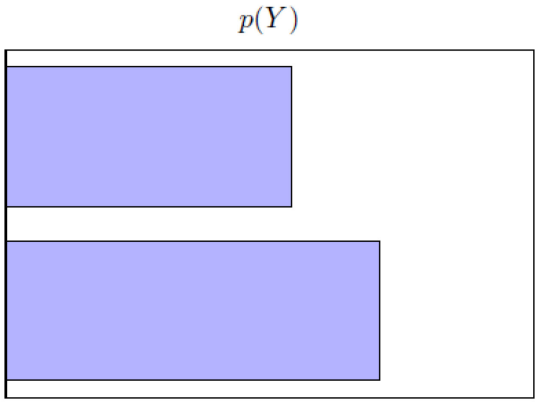
Hence knowing that the knife was the murder weapon strengthens our belief that the butler did it.

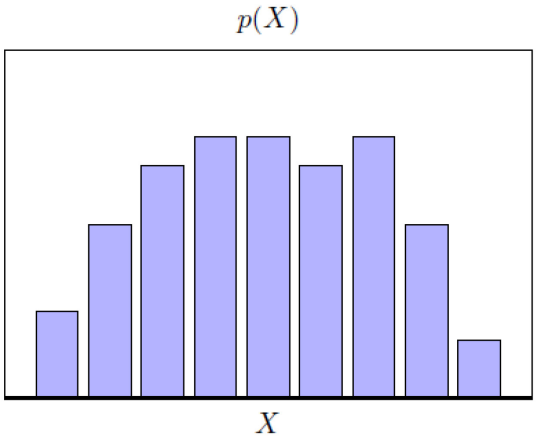
- **Exercise**: compute the probability that the Butler and not the Maid is the murderer.

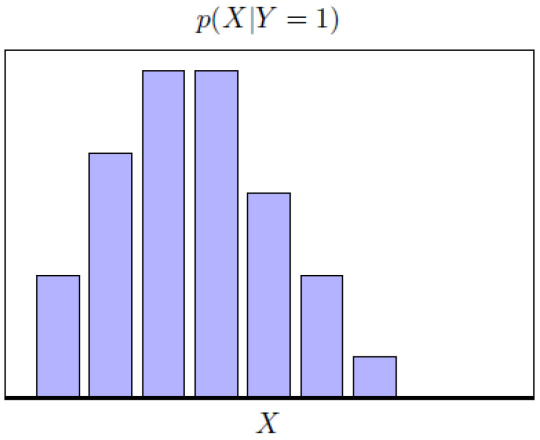
Joint Probability | Joint Probability Density

NTB
Interstaatliche Hochschule für Technik Buchs
FHO Fachhochschule Ostschweiz









Gaussian distribution (1D)

NTB
Interstaatliche Hochschule für Technik Buchs
FHO Fachhochschule Ostschweiz

Normal distribution with mean μ and variance σ^2:

$$p(x) = \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$\mathbf{E}[x] = \int x \cdot p(x) dx = \mu$$

$$\mathbf{E}[x^2] = \int x^2 \cdot p(x) dx = \mu^2 + \sigma^2$$

$$\mathbf{VAR}[x] = \mathbf{E}[x^2] - \mathbf{E}[x]^2 = \sigma^2$$

Normal Distribution

Multivariate Gaussian Distribution

NTB Interstaatliche Hochschule für Technik Buchs
FHO Fachhochschule Ostschweiz

Gaussian defined over a vector x of continuous variables in a D-dimensional space **with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$**, where $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$.

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right\}$$

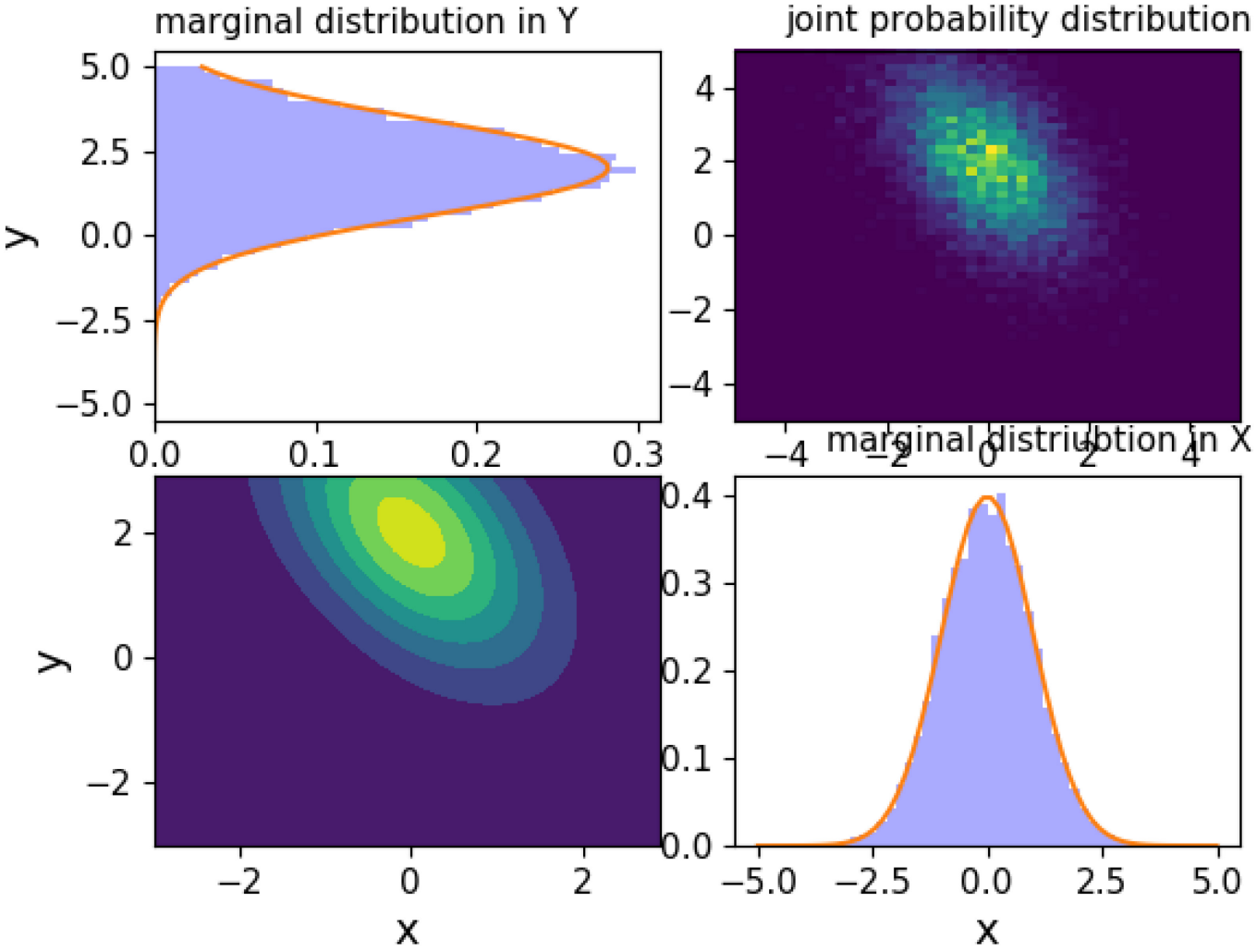
Multivariate Gaussian Distribution

The **quadratic form** in the argument of the exponential is called *Mahalanobis distance*:

$$\Delta = (\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)$$

Without loss of generality, we can assume that Σ is **symmetric** with **real eigenvalues** and an orthonormal set of eigenvectors $\boldsymbol{u_i}$.

A bivariate Gaussian distribution



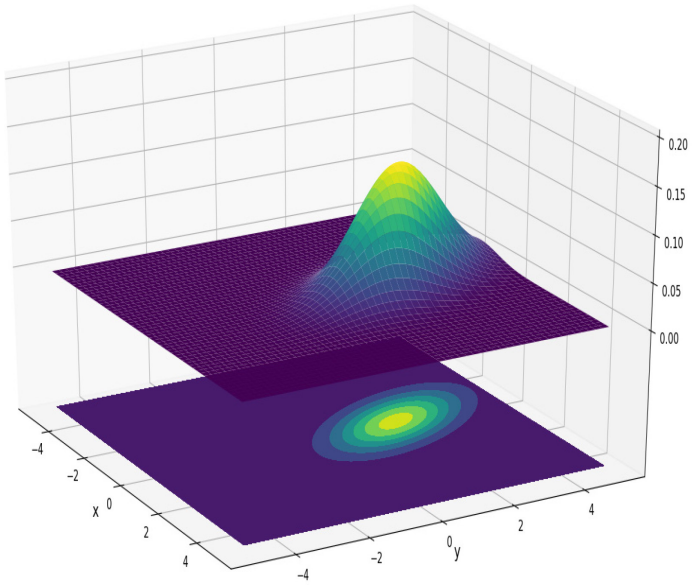
Multivariate Normal Distribution

NTB
Interstaatliche Hochschule für Technik Buchs
FHO Fachhochschule Ostschweiz

```
from scipy.stats import multivariate_normal
import numpy as np import matplotlib.pyplot as plt

F = multivariate_normal(mu, Sigma)
#draw random samples from the multivariate distribution #and try to
reconstruct the gaussian distribution
NSamples=10000
x, y = np.mgrid[-3:3:.1, -3:3:.1]
pos = np.dstack((x, y))
MVGauss = multivariate_normal(mu, Sigma)
MVGSamples = MVGauss.rvs(size=NSamples)
XS = MVGSamples[:,0]
YS = MVGSamples[:,1]

fig2 = plt.figure('Using Scikit Learn')
ax2 = fig2.add_subplot(111)
ax2.contourf(x, y, F.pdf(pos))
ax2.set_xlabel("x")
ax2.set_ylabel("y")
```




```
fig3 = plt.figure('Using Scikit Learn to draw random samples')
ax3 = fig3.add_subplot(111) ax3.scatter(XS,YS) ax3.set_xlabel("x")
ax3.set_ylabel("y")
```

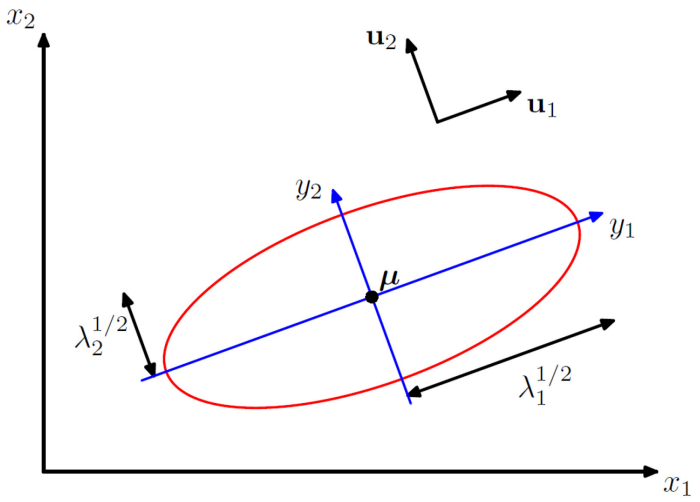
Multivariate Gaussian Distribution

- Since Σ ist real and symmetic, the eigenvectors λ_i are real and the eigenvectors can be chosen from an orthonormal set, so that:

$$\mathbf{u}_i^T \mathbf{u}_j = \delta_{ij}$$

$$\mathbf{\Sigma} = \sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^T$$

$$\mathbf{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T$$



- The quadratic form becomes

$$\Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i} \qquad y_j = \mathbf{u_j}^T (\mathbf{x} - \mu)$$

- This describes D-dimensional ellipsoids with main axes $\sqrt{\lambda_i}$.

- The deterimant is given by the product of the eigenvalues:

$$|\boldsymbol{\Sigma}|^{1/2} = \prod_{j=1}^{D} \lambda_j^{1/2}$$

- In the orthonormal coordinate of the eigenvectors, the multivariate Gaussian distribution takes the form:

$$p(\mathbf{y}) = p(\mathbf{x}) \cdot |J| = \prod_{j=1}^{D} \frac{1}{(2\pi\lambda_j)^{1/2}} \exp\left\{-\frac{y_j^2}{2\lambda_j}\right\}$$

- One can show:

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu} \qquad \textbf{Expectation value}$$

$$\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathbf{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma} \qquad \textbf{Correlation}$$

$$\mathrm{cov}[\mathbf{x}] = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^{\mathbf{T}}] = \boldsymbol{\Sigma} \qquad \textbf{Covariance}$$

Conditional Gaussian distributions

- we partition x into two disjoint subsets x_a and x_b. Without loss of generality, we can take x_a to form the first M components of x, with x_b comprising the remaining $D - M$ components:

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

- In many situations, it will be convenient to work with the **inverse of the covariance matrix, the precision matrix:**

$$\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$$

$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

Conditional Gaussian distributions II

- To calculate the conditional $p(x_a|x_b)$, we have a look at the exponent:

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) =$$

$$-\frac{1}{2}(\mathbf{x_a}-\boldsymbol{\mu}_a)^T\boldsymbol{\Lambda}_{aa}(\mathbf{x_a}-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x_a}-\boldsymbol{\mu}_a)^T\boldsymbol{\Lambda}_{ab}(\mathbf{x_b}-\boldsymbol{\mu}_b)$$

$$-\frac{1}{2}(\mathbf{x_b}-\boldsymbol{\mu}_b)^T\boldsymbol{\Lambda}_{ba}(\mathbf{x_a}-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x_b}-\boldsymbol{\mu}_b)^T\boldsymbol{\Lambda}_{bb}(\mathbf{x_b}-\boldsymbol{\mu}_b)$$

- We see that as a function of x_a, this is again a quadratic form, and hence the corresponding conditional distribution $p(x_a|x_b)$, will be Gaussian.
- Because this distribution is completely characterized by its **mean** and its **covariance**, our goal will be to identify expressions for the mean and covariance of $p(x_a|x_b)$, by inspection of the above equation:
- *Completing the square:*

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \text{const}$$

Conditional Gaussian distributions III

- **Quadratic** term: $-\frac{1}{2}\mathbf{x}_a^T \boldsymbol{\Lambda}_{aa}\mathbf{x}_a$
- The covariance of $p(x_a|x_b)$ is therefore given by: $\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1}$
- **Linear** term:

$$\mathbf{x}_a^T \left\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\right\} = \mathbf{x}_a^T \left\{\boldsymbol{\Sigma}_{a|b}^{-1}\boldsymbol{\mu}_{a|b}\right\}$$

- The term in curly brackets must be equal to: $\boldsymbol{\Sigma}_{a|b}^{-1}\boldsymbol{\mu}_{a|b}$
  Therefore we have:

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\Sigma}_{a|b}\left\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\right\} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)$$

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)$$

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1}$$

Summary

NTB
Interstaatliche Hochschule für Technik Buchs
FHO Fachhochschule Ostschweiz

Covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

Precision matrix

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

- **Conditional** probability distribution

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}\left(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1}\right)$$

Conditional Probability Distribution of a Gaussian

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \qquad \boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b)$$

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1} \qquad \boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$$

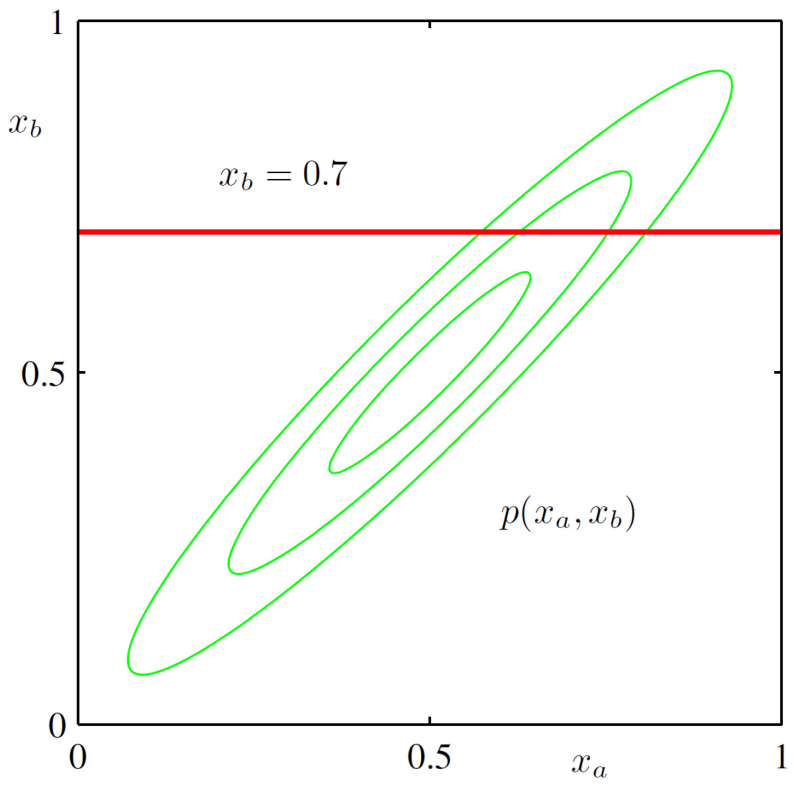
- **Marginal** probability distribution:

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})$$

Example: Squirrel on a Tree (Mitchell)

Joint probability density – marginal – conditional

NTB
Interstaatliche Hochschule für Technik Buchs
FHO Fachhochschule Ostschweiz



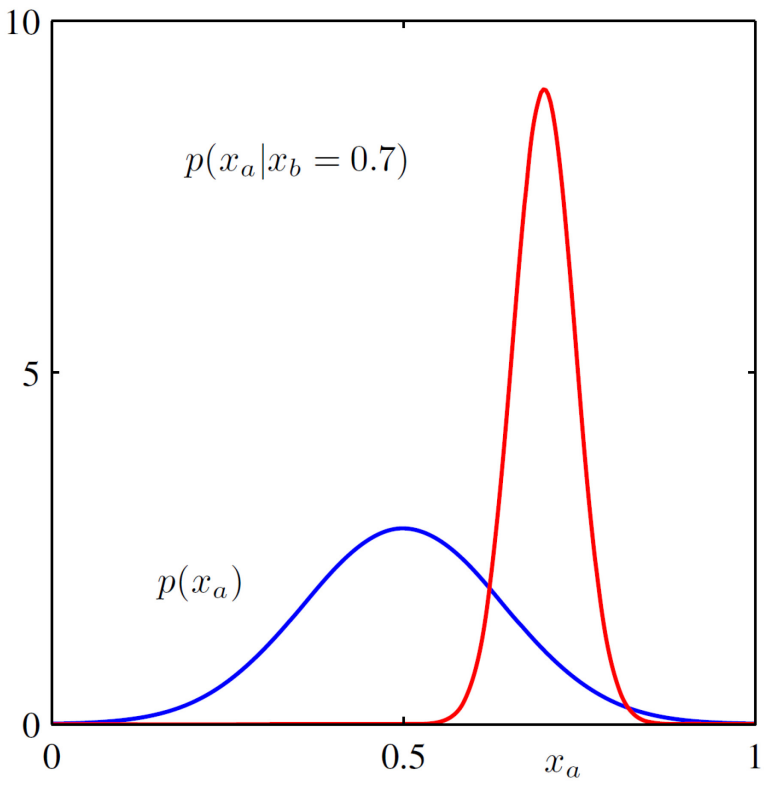
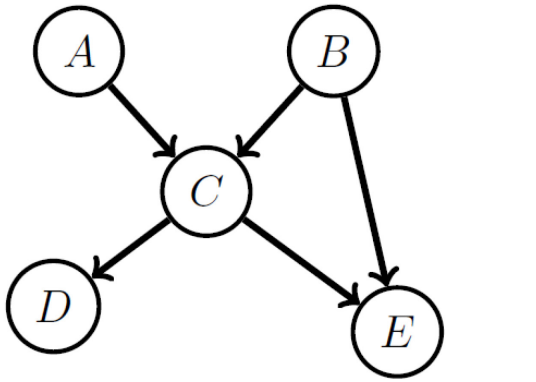


Figure 2.9 The plot on the left shows the contours of a Gaussian distribution $p(x_a, x_b)$ over two variables, and the plot on the right shows the marginal distribution $p(x_a)$ (blue curve) and the conditional distribution $p(x_a|x_b)$ for $x_b = 0.7$ (red curve).

Belief Networks (Bayesian Networks)

- Definition: A **belief network** is a directed acyclic graph (DAG) 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:

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



$p(E|B, C)$

Example: Burglar Alarm

- Sally's burglar **A**larm is sounding. Has she been **B**urgled, or was the alarm triggered by an **E**arthquake? She turns the car **R**adio on for news of earthquakes.

- General case:

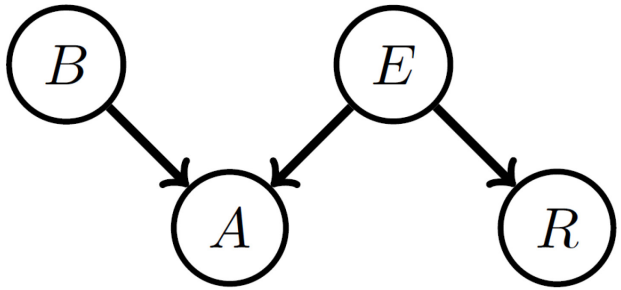
$$\begin{aligned} p(A,R,E,B) &= p(A|R,E,B)p(R,E,B) \\ &= p(A|R,E,B)p(R|E,B)p(E,B) \\ &= p(A|R,E,B)p(R|E,B)p(E|B)p(B) \end{aligned}$$

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:

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

Directed acyclic graph



$p(A|B,E)$

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 tables are $p(B = 1) = 0.01$ and $p(E = 1) = 10^{-6}$
- The tables and graphical structure fully specify the distribution.

Inference

- **Initial evidence:** The alarm is sounding $A = 1$:

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

- **Additional Evidence**: $(R = 1)$, 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.

Examples of Belief Networks in Machine Learning

NTB
Interstaatliche Hochschule für Technik Buchs
FHO Fachhochschule Ostschweiz

- **Prediction** (discriminative): $p(class \mid input)$
- **Prediction** (generative): $p(class \mid input) \propto p(input \mid class) \cdot p(class)$
- **Time-series:** Markov chains, Hidden Markov Models.
- **Unsupervised learning:** $p(data) = \sum_{latent} p(data|latent) \cdot p(latent)$
- **And many more**: Personally I find the framework very useful for understanding and rationalising the many different approaches in machine learning and related areas.

Conditional Independence I

$$X \perp\!\!\!\perp Y \mid Z$$

- denotes that the two sets of variables X and Y are **independent** of each other **given** the state of the set of variables **Z**. This means that

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

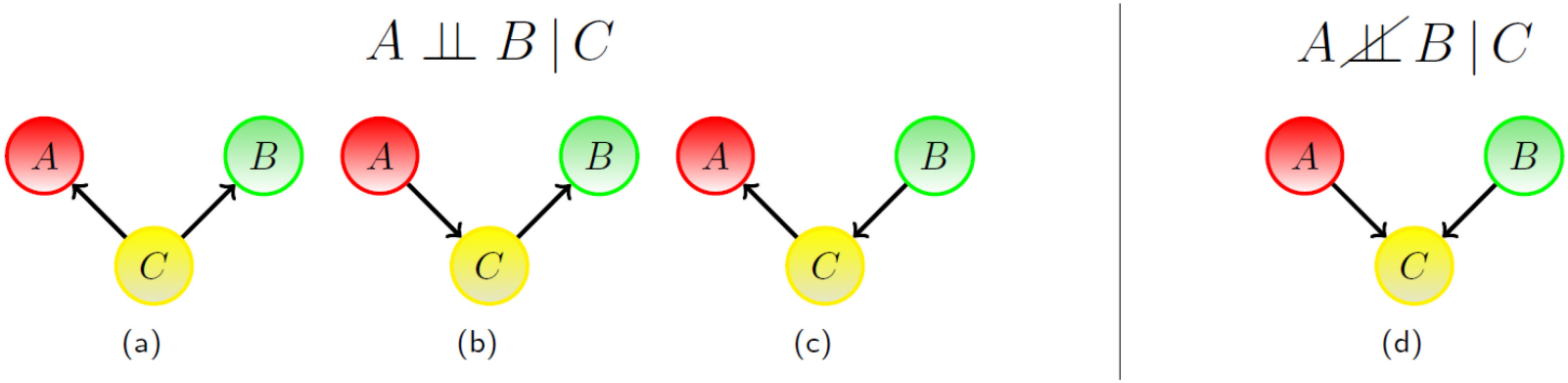
- for all states of X, Y, Z. In case the conditioning set is empty we may also write

$$X \perp\!\!\!\perp Y$$

in which case X is **(unconditionally) independent** of Y.

Conditional Independence II

NTB Interstaatliche Hochschule für Technik Buchs
FHO Fachhochschule Ostschweiz



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

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

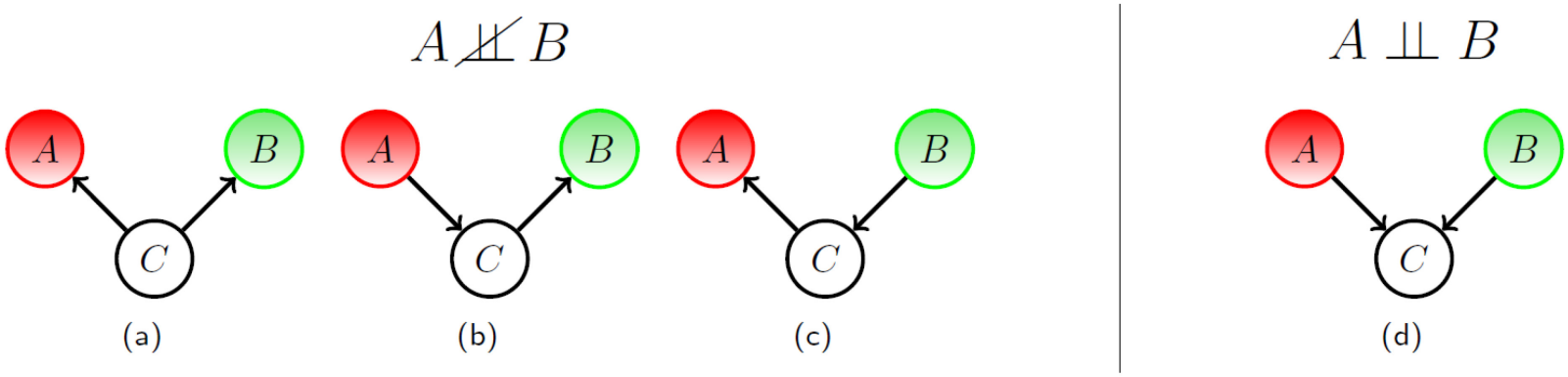
$$\text{(b)}\quad p(A,B|C) = \frac{p(A)p(C|A)p(B|C)}{p(C)} = \frac{p(A,C)p(B|C)}{p(C)} = p(A|C)p(B|C)$$

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

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

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

Conditional Independence III

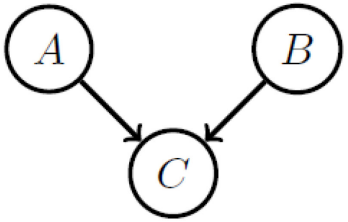


- In (a), (b) and (c), the variables A;B are **marginally dependent**.
- In (d) the variables A, B are **marginally independent**.

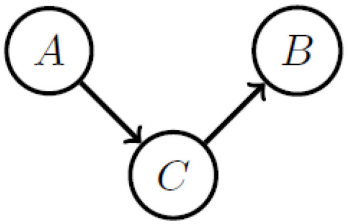
$$p(A,B) = \sum_C p(A,B,C) = \sum_C p(A)p(B)p(C|A,B) = p(A)p(B)$$

Collider

- A collider contains two or more incoming arrows along a chosen path.
- Summary of two previous slides:



If C has more than one incoming link, then $A \perp\!\!\!\perp B$ and $A \not\perp\!\!\!\perp B \mid C$. In this case C is called collider.



If C has at most one incoming link, then $A \perp\!\!\!\perp B \mid C$ and $A \not\perp\!\!\!\perp B$. In this case C is called non-collider.

Summary

- We can reason with certain or uncertain evidence using repeated application of **Bayes' rule**.
- A belief network represents a **factorisation of a distribution into conditional probabilities** of variables dependent on parental variables.
- Belief networks correspond to **directed acyclic graphs (DAG).**
- Variables are **conditionally independent** $x \perp\!\!\!\perp y|z$ if $p(x, y|z) = p(x|z) \cdot p(y|z)$; the absence of a link in a belief network corresponds to a conditional independence statement.
- If in the graph representing the belief network, two variables are independent, then they are independent *in any distribution consistent with the belief network* structure.
- Belief networks are natural for representing `causal' influences.

Recap: Probability distributions and PDFs

Terminology: its probability density function (pdf) is one way to describe a distribution.

What does a pdf tell about a set of data?

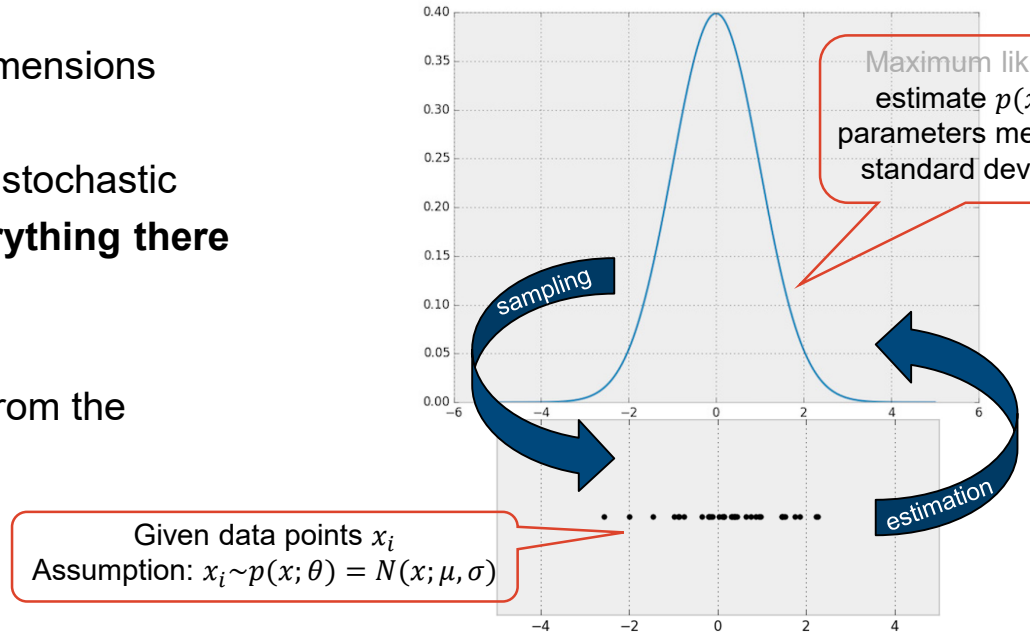
- Where to expect samples
  …with which probability
- Correlation/covariance of dimensions

➔ For data coming from some stochastic processes, the pdf tells **everything there is to know** about the data

➔ **Allows for sampling** data from the underlying distribution
(generative modeling)

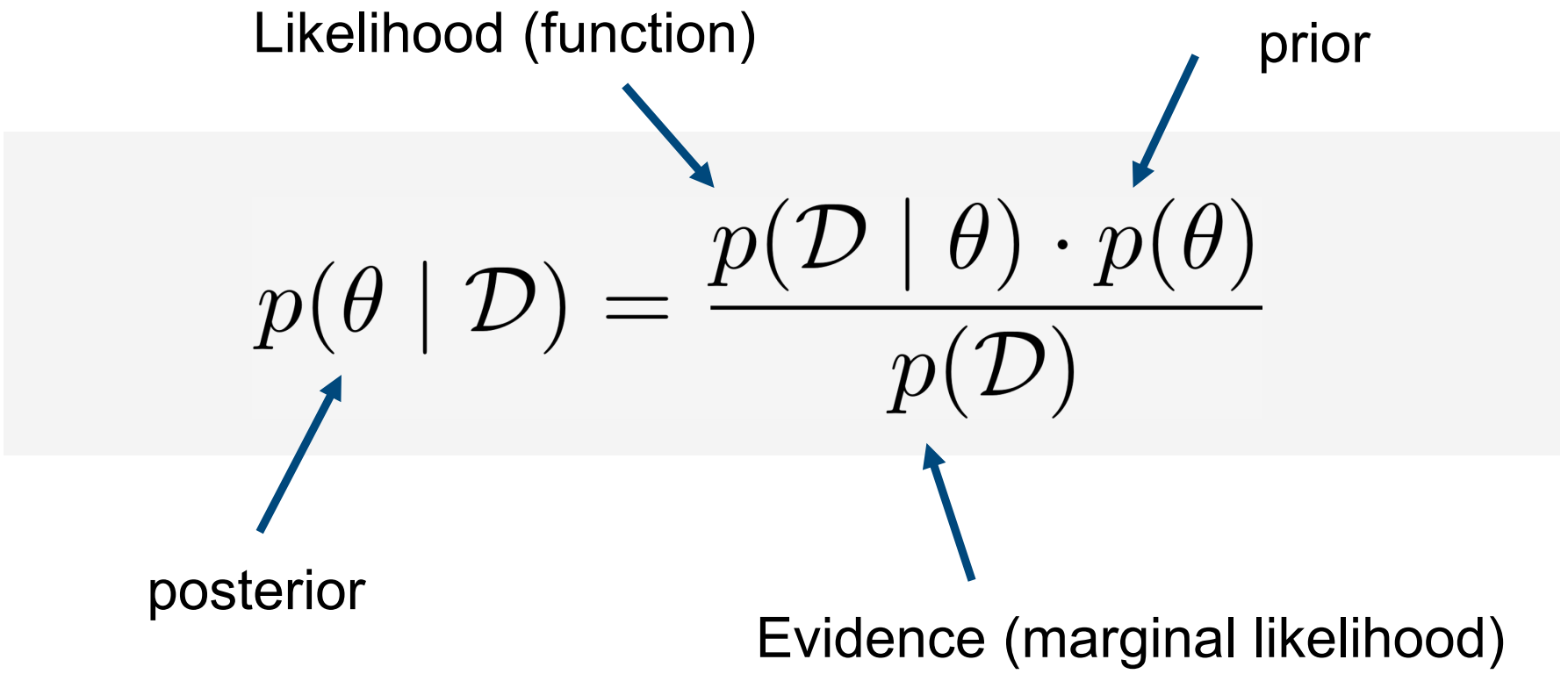
An example generative model

- The univariate Gaussian
  A parametric pdf, recoverable from data (Gaussianity given)



Source: Brandon Amos, *«Image Completion with Deep Learning in TensorFlow»*, 2016, https://bamos.github.io/2016/08/09/deep-completion/

II. Bayes' Theorem



Bayes' theorem

One of the cornerstones of modern data analysis

NTB
Interstaatliche Hochschule für Technik Buchs
FHO Fachhochschule Ostschweiz

$$p(h|X) = \frac{p(X|h) \cdot p(h)}{p(X)}$$

Bayes Theorem

with (in a machine learning context with training data X and model h)

$p(X|h)$ the **likelihood** of the data, given the model → called the **evidence** for h

$p(X)$ the a priori probability of the training data X → this normalization factor is **rarely needed**/used

$p(h)$ is the **a priori** probability of hypothesis h → **often neglected** in practice due to dominance of evidence

Use cases

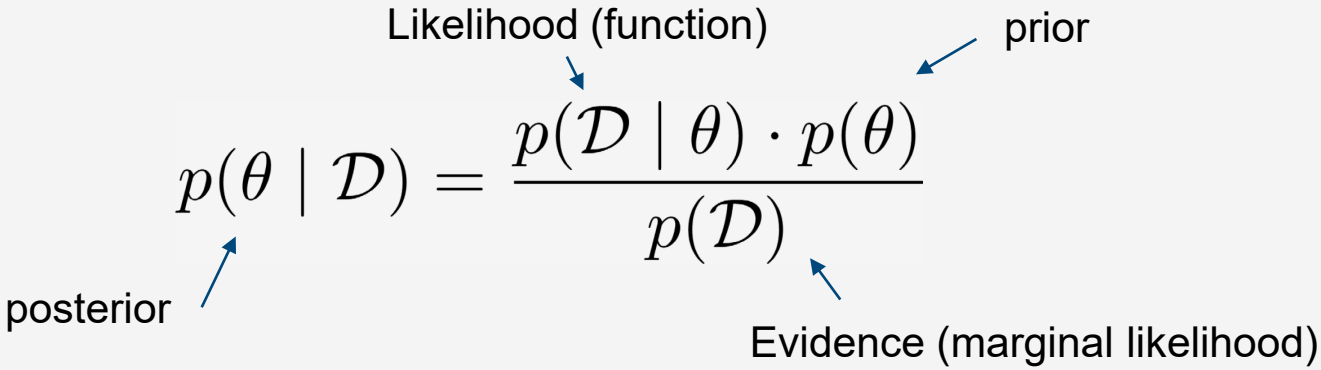
- Generally: **Convert** between prior and posterior probabilities
- Specific example: **Model selection**

  ➔ Given competing $h_i \in \mathcal{H}$, one can calculate the likelihood $p(X|h_i)$, then select best $\hat{h} = \max_{h_i} p(h_i|X) \approx \max_{h_i} p(X|h_i)$

Rev. Thomas Bayes, 1701-1761

Prior, Likelihood and Posterior

- For data D and variable θ, Bayes' rule tells us how to update our prior beliefs about the variable θ in light of the data to a posterior belief:



- The *evidence* is also called the *marginal likelihood*.
- The term *likelihood* is used for the probability that a model generates observed data.

Prior, Likelihood and Posterior

- if we condition on the model M, we have

$$p(\theta \mid \mathcal{D}, M) = \frac{p(\mathcal{D} \mid \theta, M) \cdot p(\theta \mid M)}{p(\mathcal{D} \mid M)}$$

- **The MAP assignment**: The **M**ost probable **A P**osteriori (MAP) setting is that which maximises the posterior,

$$\hat{\theta}_{\mathrm{MAP}} = \underset{\theta}{\operatorname{argmax}} \{p(\theta \mid \mathcal{D}, M)\} = \underset{\theta}{\operatorname{argmax}} \{p(\theta, \mathcal{D} \mid M)\}$$

- The **Max. Likelihood assignment**: when $p(\theta|M) = \mathrm{const}$

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta}{\operatorname{argmax}} \{p(\theta|\mathcal{D}, M)\} = \underset{\theta}{\operatorname{argmax}} \{p(\mathcal{D}|\theta, M)\}$$

The Bayes optimal classifier

Classification's «gold standard»

NTB Interstaatliche Hochschule für Technik Buchs
FHO Fachhochschule Ostschweiz

- Theoretically optimal (=most probable) classification
- Combine predictions of all hypotheses, weighted by their posterior probabilities: (where y_i is a label from the set Y of classes)

$$P(Y = y_k | X_1, \ldots, X_n) = \frac{P(Y = y_k) P(X_1, \ldots, X_n | Y = y_k)}{\sum_j P(Y = y_j) P(X_1, \ldots, X_n | Y = y_j)}$$

$$= \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

$$Y \longleftarrow \underset{y_k}{\operatorname{argmax}} P(Y = y_k) \prod_i P(X_i | Y = y_k)$$

Naive Bayes Classifier

- No other method using the same $\mathcal{H}$ and X can do better on average
- In particular outperforms simply taking the classification of the **MAP hypothesis,** enforces the idea of **ensemble learning**
- Computationally intractable (linear in $|\mathcal{H}|$ → see Reader (also on Moodle): http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf)

Maximum Likelihood revisited

NTB Interstaatliche Hochschule für Technik Buchs
FHO Fachhochschule Ostschweiz

A quick example of a maximum likelihood estimate

- You flip a coin 10 times and observe the following sequence (H, T, T, H, T, T, T, T, H, T)¶
- What's the MLE of observing 3 heads in 10 trials?
- simple answer:

The frequentist MLE is (# of successes) / (# of trials) or 3/10

Maximum Likelihood Approach: What is the expected probability for head given the observed data? Solving first derivative of binomial distribution :

$$L(\theta|\, X = 3) = \binom{10}{3} \cdot \theta^3 \cdot (1-\theta)^7$$

$$\log\{L(\theta)\} = \log\binom{10}{3} + 3\log\theta \; + 7\log(1-\theta)$$

$$\frac{\partial \log\{L(\theta)\}}{\partial\theta} = \frac{3}{\theta} - \frac{7}{1-\theta} = 0$$

$$\theta_{ML} = \frac{3}{10}$$

Maximum Likelihood for a Gaussian

- Given a data set $\mathbf{X} = (x_1, x_2, \ldots, x_n)^T$ in which the observations $\{x_n\}$ are assumed to be drawn independently from a multivariate Gaussian distribution, we can estimate the parameters of the distribution by **maximum likelihood.**

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{N \cdot D}{2} \ln(2\pi) - \frac{N}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{n=1}^{N} (x_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (x_n - \boldsymbol{\mu})$$

- By simple rearrangement, we see that the **log likelihood** function depends on the data set only through the two quantities (sufficient statistics for a Gaussian)

$$\sum_{n=1}^{N} x_n \qquad\qquad \sum_{n=1}^{N} x_n x_n^T$$

- Derivative with respect to $\boldsymbol{\mu}$ and setting it to zero yields the ML estimate for $\boldsymbol{\mu}$

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1} (x_n - \boldsymbol{\mu}) \Longrightarrow \boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

sample mean

ML estimate for the covariance is biased

- If we evaluate the expectations of the maximum likelihood solutions under the true distribution, we obtain the following results:

$$\mathbb{E}[\mu_{\mathbf{ML}}] = \int \boldsymbol{\mu}_{\mathrm{ML}} \cdot p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathbf{dX} = \boldsymbol{\mu}$$

$$\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathbf{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^T + \delta_{mn}\boldsymbol{\Sigma}$$

$$\mathbb{E}[\boldsymbol{\Sigma}_{\mathbf{ML}}] = \frac{N-1}{N}\boldsymbol{\Sigma}$$

- We see that the expectation of the maximum likelihood estimate for the mean is equal to the true mean. However, the maximum likelihood estimate for the covariance has an expectation that is less than the true value, and hence it is biased.

$$\tilde{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \boldsymbol{\mu}_{\mathrm{ML}})(x_n - \boldsymbol{\mu}_{\mathrm{ML}})^T$$

Unbiased estimator of the covariance

Bayesian Inference for a Gaussian

- The maximum likelihood framework gave point estimates for the parameters μ and Σ. Now we develop a Bayesian treatment by introducing prior distributions over these parameters.

$$p(\mathbf{X}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

- NB: the **likelihood function** $p(\mathbf{X}|\mu)$ is not a probability distribution over μ and is not normalized. if we choose a prior $p(\mu)$ given by a Gaussian, it will be a **conjugate distribution** for this likelihood function because the corresponding posterior will be a product of two exponentials of quadratic functions of μ and hence will also be Gaussian.

$$p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2) \qquad p(\mu|\mathbf{X}) \propto p(\mathbf{X}|\mu) \cdot p(\mu)$$

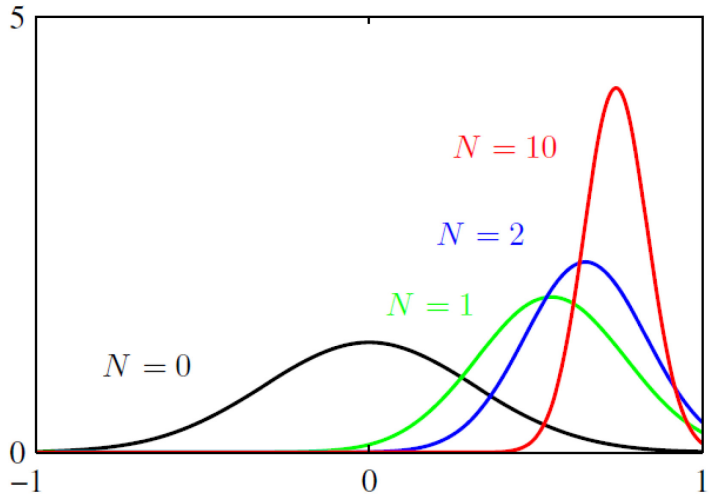
- In the exercices using completing the square in the exponent, you will be able to show that the posterior distribution is given by:

$$p(\mu|\mathbf{X}) = \mathcal{N}(\mu|\mu_N, \sigma_N^2)$$

Bayesian Inference for a Gaussian II

- The **posterior distribution** is a compromise between the prior mean μ_0 and the maximum likelihood solution μ_{ML}. If the number of observed data points $N = 0$, then μ_N reduces to the prior mean μ_0 as expected. For $N \to \infty$, the posterior mean μ_N is given by the maximum likelihood solution μ_{ML}.

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \cdot \mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \cdot \mu_{\mathrm{ML}}$$

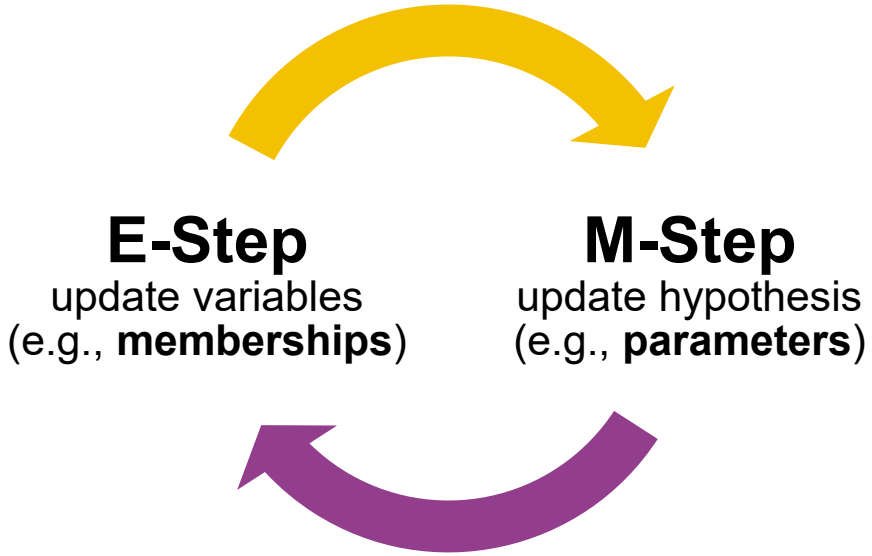


$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

the precisions (inverse variance) are additive, so that the precision of the posterior is given by the precision of the prior plus one contribution of the data precision from each of the observed data points.
As we increase the number of observed data points, the precision steadily increases, the posterior gets infinitly peaked around the ML solution.

III. Gaussian Mixture Models

(& the EM algorithm)



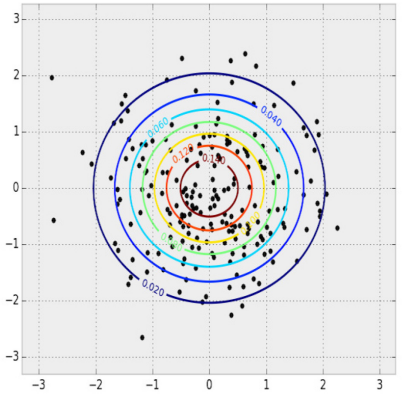
Recap: Probabilistic mixture models

Generative models for unknown, multivariate distributions

NTB Interstaatliche Hochschule für Technik Buchs
FHO Fachhochschule Ostschweiz

Mixture Models

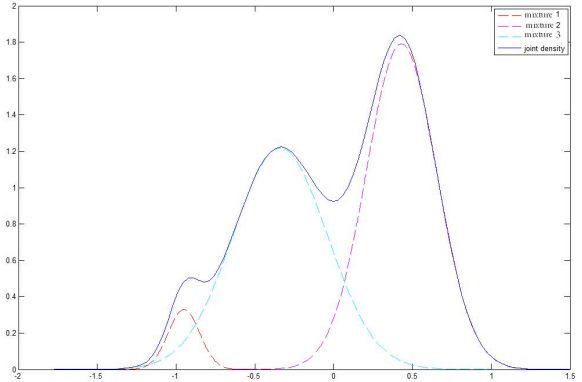
- **Approximate** an **arbitrary** distribution **by** a **linear combination of** a **simpler**, "well-behaved" distribution
  ➔ Mathematically **tractable**, **compact** formulation, allows **sampling** & **inference**



Example of a multivariate (2D) Gaussian distribution: samples and contour plot.

The Gaussian Mixture Model (GMM)

- **Modeled by** a weighted sum of N **multivariate Gaussians** (N being sufficiently large)
- Often used because of and "**nice**" mathematical **properties** of Gaussian pdf and central limit theorem
  (~ data from natural phenomena tend to be Gaussian distributed)
- The Gaussians' **parameters** can be estimated efficiently **using** the **EM** algorithm



Example of a multimodal (but univariate) distribution, approximated by a GMM with 3 mixtures.

The EM algorithm

A general-purpose, *unsupervised* learning algorithm

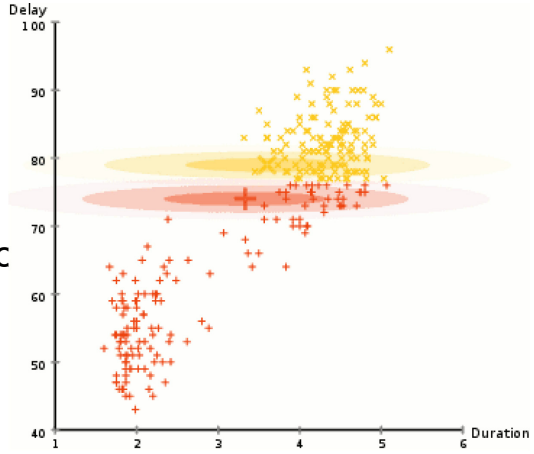
NTB Interstaatliche Hochschule für Technik Buchs
FHO Fachhochschule Ostschweiz

EM (expectation maximization)

- Iterative method to **learn in the presence of unobserved variables**
  ➔ A **typical hidden variable** is some sort of **group**/cluster **membership**
- Good convergence guarantees (finds local maximum)

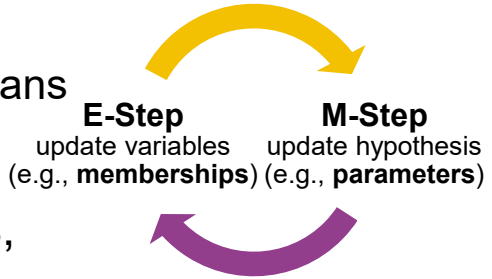
Example

- A given dataset is known to be generated by either of 2 Gaussians (with equal probability)
- Only the data is observed
  ➔ **Which** Gaussian **generated** a certain point **is unobserved (Z)**
  ➔ The Gaussians' **parameters** θ are unknown
- The means & variances of these Gaussians shall be learned
  ➔ Needs an estimation of the membership probability of each point to either Gaussian



EM algorithm used to iteratively optimize the parameters of 2 Gaussians (animated)
Source: https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm)

The EM algorithm (contd.)



1. Start with a random initial hypothesis
   Example: **Pretend to know the parameters** μ, σ^2 of the 2 Gaussians
   (e.g., pick random values)

2. **E-Step**: Estimate expected values of unobserved variables,
   assuming the current hypothesis holds
   Example: **Compute probabilities p_{ti}** that feature vector x_t was produced by Gaussian i
   (i.e., $p_{ti} = p(G = i|x_t) = \frac{p(x_t|G=i)p(G=i)}{p(x_t)} \approx p(x_t|G = i) = g_i(x_t, \mu_i, \sigma_i)$ with g_i being the Gaussian pdf
   and G the unobserved random variable indicating membership to one of the Gaussians)

3. **M-Step**: Calculate new **Maximum Likelihood** (ML) estimate of hypothesis,
   assuming the expected values from (2) hold
   Example: **Calculate the μ_i, ${\sigma_i}^2$**, given the currently assigned membership
   (i.e., using standard ML estimation: $\mu_i = \frac{1}{T}\sum_{t=1}^{T} p_{ti} \cdot x_t$, $\sigma_i{}^2 = \frac{1}{T}\sum_{t=1}^{T} p_{ti} \cdot (x_t - \mu_i)^2$)

4. Repeat with step 2 until convergence
   Always replacing old estimates with new ones

Example: EM for GMM

NTB Interstaatliche Hochschule für Technik Buchs
FHO Fachhochschule Ostschweiz

This suggests an **iterative algorithm**, in the case where both $\boldsymbol{Z}$ and $\boldsymbol{\theta}$ are unknkown,e.g. for two unknown classes (membership $\boldsymbol{Z} = \{1,2\}$, e.g. male, female) is unknown. Input data is $\boldsymbol{X} = \{x_1, x_2, \dots, x_N\}$, weights and heights of N male and female persons.

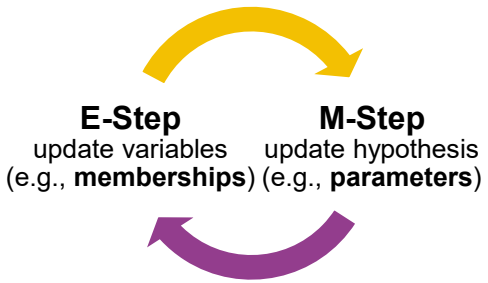
1. First, initialize the model parameters $\boldsymbol{\theta} = \{\mu_k, \sigma_k^2\}$ to some random values, here a Gaussian for each gender $k = 1,2$.
2. **E-Step:** Compute the probability of each possible value of $Z = \{1,2\}$, given θ = probability that x_i is member of class z_j

$$z_{ij} = p(Z = j \mid x_i) = \frac{p(x_i|Z = j\,) \cdot p(Z = j)}{\sum_{k=1}^{2} p(x_i|Z = k)} \approx \frac{p(x = x_i|\,\mu_j\,)}{\sum_{k=1}^{2} p(x = x_i|\,\mu_k\,)} = \frac{e^{-\frac{1}{2\sigma_j^2}(x_i-\mu_j)^2}}{\sum_{k=1}^{2} e^{-\frac{1}{2\sigma_k^2}(x_i-\mu_k)^2}}$$

3. **M-Step:** Then, use the just-computed values of $\boldsymbol{Z}$ to compute a better estimate for the parameters $\boldsymbol{\theta} = \{\mu_n, \sigma^2\}$

$$\hat{\mu}_j = \mathbb{E}(\mu_j) = \frac{\sum_{i=1}^{N} p(Z = j \mid x_i) \cdot x_i}{\sum_{i=1}^{N} p(Z = j \mid x_i)}$$

$$\widehat{\sigma^2}_j = \mathbb{E}(\sigma_j^2) = \frac{1}{N}\sum_{i=1}^{N} z_{ij} \cdot (x_i - \hat{\mu}_j)^2$$



4. Repeat with step 2 until convergence
   Always replacing old estimates with new ones

General Description: EM Algorithm

NTB Interstaatliche Hochschule für Technik Buchs
FHO Fachhochschule Ostschweiz

Given the statistical model which generates a set $\boldsymbol{X}$ of observed data, a set of unobserved latent data $\boldsymbol{Z}$ (e.g membership), and a vector of unknown parameters $\boldsymbol{\theta}$, along with a *likelihood function* $L(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Z}) = p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ the maximum likelihood estimate (MLE) of the unknown parameters θ is determined by **maximizing** the (log) *marginal likelihood* of the observed data

$$L(\theta; X) = p(X|\theta) = \int_Z p(X, Z|\theta)\, dZ \quad \Rightarrow \quad \hat{\theta}_{ML} = \arg\max_\theta \{\log[p(X|\theta)]\}$$

However, the exact calculation of the sum is extremely difficult: The EM algorithm seeks to find the MLE of the marginal likelihood by **iteratively applying these two steps**:

1. **Expectation step (E-step)**

   Calculate the expected value of the log likelihood function, with respect to the conditional distribution of Z given X under the current estimate of the parameters θ_t

$$Q(\theta|\theta_t) = \mathbb{E}_{Z|X,\theta_t}[\log\, p(X, Z \mid \theta)]$$

2. **Maximization step (M-step):** Find the parameters that maximize this quantity:

$$\theta_{t+1} = \arg\max_\theta Q(\theta|\theta_t)$$

Motivation: EM for GMM

- If the value of the parameters $\boldsymbol{\theta}$ is known, usually the value of the latent variables $\boldsymbol{Z}$ can be found by maximizing the **log-likelihood** over all possible values of $\boldsymbol{Z}$, either simply by iterating over $\boldsymbol{Z}$ or using a **Viterbi algorithm for hidden Markov** models.

- Conversely, if we know the value of the latent variables $\boldsymbol{Z}$, we can find an estimate of the parameters $\boldsymbol{\theta}$ fairly easily, typically by simply grouping the observed data points according to the value of the associated latent variable and averaging the values.

Applications of the EM algorithm

Applications:

- **Unsupervised learning of clusters**
- Filling in missing data from a sample set
- Discovering values of latent (i.e. hidden) variables (Z)
- Estimating parameters of HMMs (hidden Markov models)
- Estimating parameters of finite mixtures [mixture models]

3. APPLICATION TO VOICE RECOGNITION

GMMs as generative models for voice modeling

NTB
Interstaatliche Hochschule für Technik Buchs
FHO Fachhochschule Ostschweiz

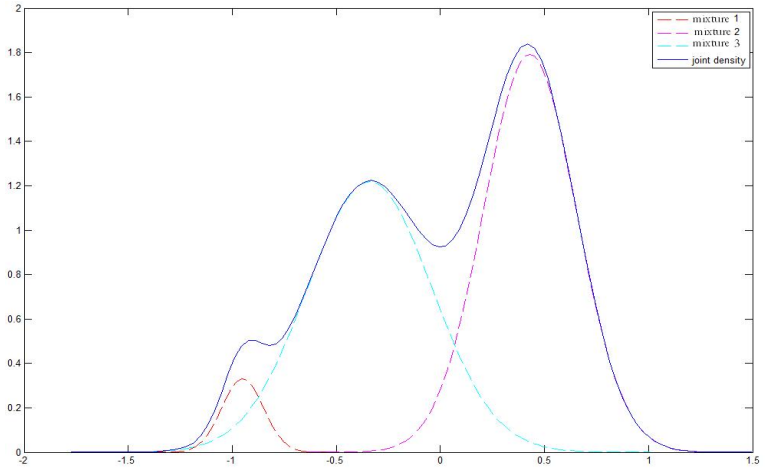
Reference

- Reynolds, Rose, *«Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models»*, 1995

Key ideas

- **Take** the estimated probability density function (**pdf**) $p(x|h)$ of a speaker's D-dim. training vectors x **as** a **model of** his **voice**

- Model the **pdf as** a **weighted sum of** M D-dimensional **Gaussians** (e.g., $M = 32$, $D = 16$)



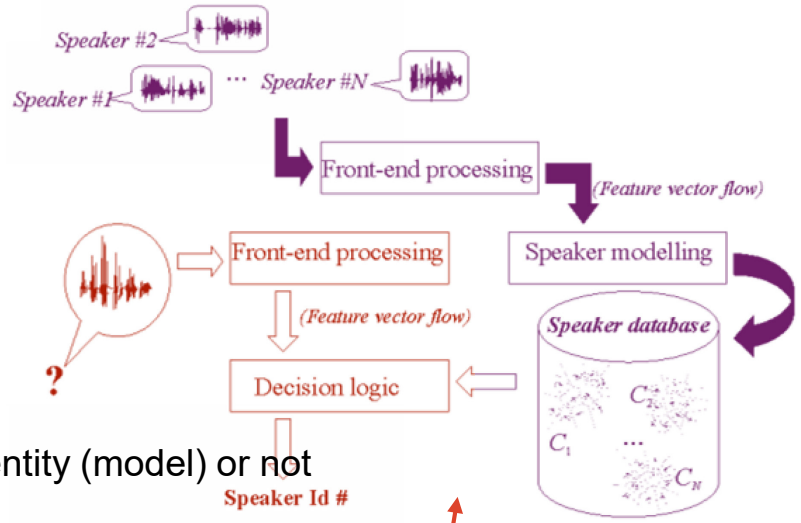
GMM with 3 mixtures in 1 dimension. Solid line shows GMM density, dashed lines show constituting Gaussian densities.

The task of speaker recognition

(→ see appendix for an introduction to speech processing)

Speaker recognition

- **Tell identity** of an **utterance**s' speaker
- Typical: score feature-sequence against a speaker model



Three subsequently more complex settings

Verification: Verify that a **given utterance fits** a **claim**ed identity (model) or not

Identification: **Find** the actual **speaker among** a list of **prebuilt models**
(or declare as unknown: open set identification)

Diarization (a.k.a. tracking, clustering): **Segment** an audio-**stream by voice** identity
(who spoke when, no prior knowledge of any kind)

GMM rationale

Hybrid solution between non-parametric clusters (vector quantization) and compact smoothing (single Gaussian):

Smooth approximation of arbitrary densities

Implicit clustering into broad phonetic classes

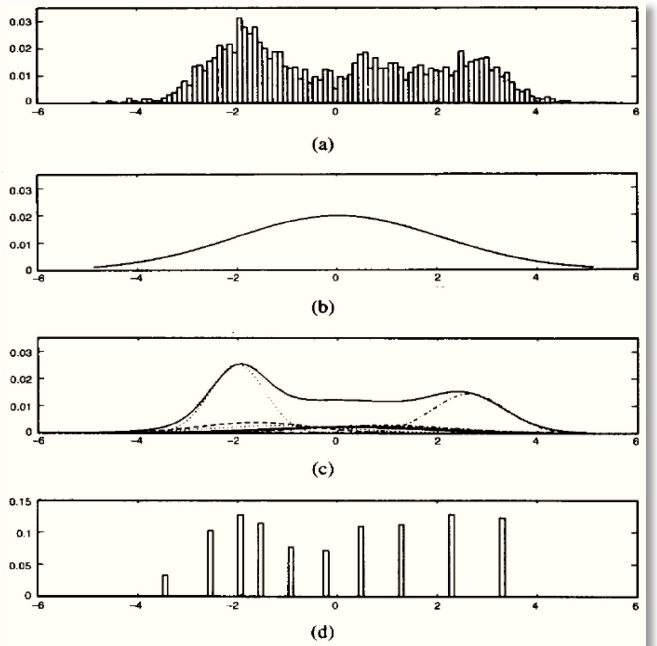


Fig. 3. Comparison of distribution modeling: (a) Histogram of a single cepstral coefficient from a 25 second utterance by a male speaker; (b) maximum likelihood unimodal Gaussian model; (c) GMM and its 10 underlying component densities; (d) histogram of the data assigned to the VQ centroid locations of a 10-element codebook.

GMM comparison with other techniques; from [Reynolds and Rose, 1995].

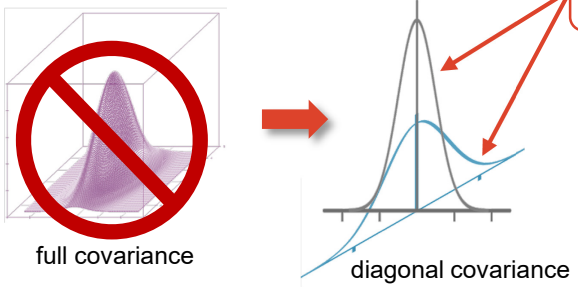
Mathematical formulation of the GMM

Using diagonal covariance (→ see appendix for reasons)

NTB Interstaatliche Hochschule für Technik Buchs
FHO Fachhochschule Ostschweiz

Notation

- h: **model** (GMM)
- w: **weight** (scalar)
- μ: **mean** vector
- σ^2: **variance** vector (the diagonal of the covariance matrix)
- g_i: **Gaussian pdf** of i^{th} (out of M) mixtures
- x: **feature vector**
- D: **dimensionality** of x, μ, σ^2
- p: density/**likelihood** of a feature vector given the model



Formulae

- **Model** consists of: $h = \{w_i, \mu_i, {\sigma_i}^2\}$
  → subject to $i = 1..M$ and $\sum_{i=1}^{M} w_i = 1$

Condition on weights to sum up to 1

- The **multimodal Gaussian** with **diagonal covariance** computes as

$$g_i(x, \mu_i, {\sigma_i}^2) = \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi {\sigma_{id}}^2}} \cdot e^{-\frac{(x-\mu_{id})^2}{2{\sigma_{id}}^2}}$$

Just the product over the assumedly independent marginals (dimensions)

The univariate Gaussian pdf

- Model **evaluation**:

$$p(x|h) = \sum_{i=1}^{M} w_i \cdot g_i(x, \mu_i, \Sigma_i)$$

GMM training via the EM algorithm

Maximum likelihood training

→ see appendix

Initialize model $h = \{w_i, \mu_i, {\sigma_i}^2\}$ using data $X = \{x_1 .. x_T\}$

➔ Instead of pure random initialization, find better values via subsequent clustering (e.g., with k-means)

E-Step:

$$p_{ti}(i|x_t, h) = \frac{w_i \cdot g_i(x_t, \mu_i, I_D \cdot {\sigma_i}^2)}{\sum_{i=1}^{M} w_i \cdot g_i(x_t, \mu_i, I_D \cdot {\sigma_i}^2)}$$

The (properly normalized) probability of x_t being issued by mixture i

M-Step:

$$w_i = \frac{1}{T} \sum_{t=1}^{T} p_{ti}(i|x_t, h)$$

Mixture i's weight is just the mean probability of all training vectors being assigned to it

$$\mu_i = \frac{1}{T \cdot w_i} \sum_{t=1}^{T} p_{ti}(i|x_t, h) \cdot x_t$$

$${\sigma_i}^2 = \left(\frac{1}{T \cdot w_i} \sum_{t=1}^{T} p_{ti}(i|x_t, h) \cdot {x_t}^2 \right) - {\mu_i}^2$$

Alternative: Training via maximum a posteriori (MAP) adaptation (i.e. uses a priori knowledge)

➔ see Reynolds, Quatieri, Dunn, *«Speaker Verification Using Adapted Gaussian Mixture Models»*, 2000

Doing speaker identification

Finding the speaker s of a new utterance, given a set of trained speaker models

- Utterance represented by its feature vector sequence $X = \{x_1 .. x_T\}$
- Speakers models given by $\{h_1 .. h_S\}$

$$s = \arg\max_s p(X|h_s)$$

The prob. of a set of feature vectors is the product of the individual probs (independence assumed)

$$= \arg\max_s \prod_{t=1}^{T} p(x_t|h_s)$$

Using the log turns the product into a sum ➔ makes the computation numerically stable

$$= \arg\max_s \sum_{t=1}^{T} \log p(x_t|h_s)$$

Model comparison via generalized likelihood ration (GLR)

Absolute likelihood values are not meaningful, but their ratios are

➔ To decide if given models h_1, h_2 trained on utterances X_1, X_2 are actually of the same speaker, threshold GLR **distance measure**:

$$GLR(h_1, h_2) = \log\left(\frac{p(X_1|h_1) \cdot p(X_2|h_2)}{p(X_1 \cup X_2|h_{1\cup 2})}\right)$$

What GMMs do not capture

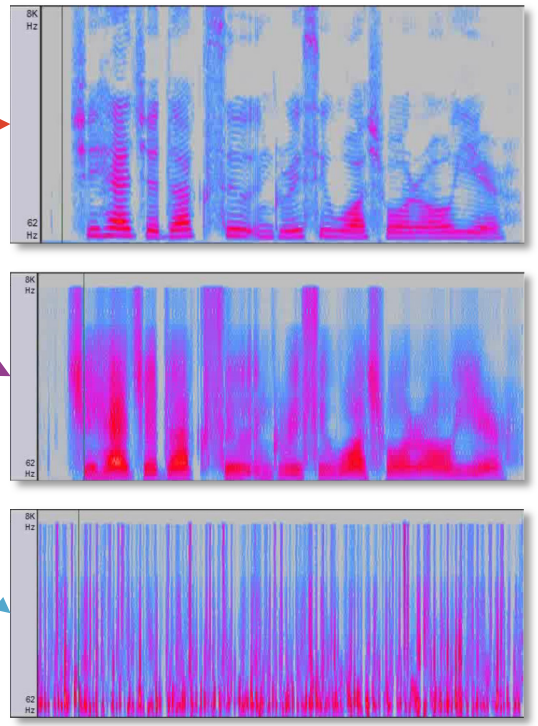
NTB
Interstaatliche Hochschule für Technik Buchs
FHO Fachhochschule Ostschweiz

Re-synthesizing speech from intermediate stages of the speaker modeling pipeline

Original utterance

Resynthesized feature vectors (MFCCs)

Resynthesized MFCCs from GMM



Implication

Temporal context isn't modeled by GMMs

More on temporal context modeling:

Friedland, Vinyals, Huang, Müller, *«Prosodic and other Long-Term Features for Speaker Diarization»*, 2009

Stadelmann, Freisleben, *«Unfolding Speaker Clustering Potential – A Biomimetic Approach»*, 2009

Lukic, Vogt, Dürr, Stadelmann, *«Speaker Identification and Clustering using Convolutional Neural Networks»*, 2016

Review

NTB Interstaatliche Hochschule für Technik Buchs
FHO Fachhochschule Ostschweiz

- Understanding uncertain events as random variables gives us a potent arsenal of tools for modeling: E.g., **probability density function** (pdf) of a random variable **tells us everything** there is to know about this function
- Thus, **estimating** the **pdf** is a rewarding **target for** (unsupervised) **learning**

- **Bayes' theorem** is used to turn priors (i.e., prior knowledge) into posteriors (i.e., taking all evidence & priors into account)

- **Speaker recognition** comes in the flavors of **verification**, **identification** or **diarization**
- The classic approach is **MFCC** features and **GMM** models
- Optimal parameters are best found using **best practices** (→ see appendix)

- **EM** training **iterates between** estimating updates values of hidden variables (based on assumed parameters of the sought distribution – **E-step**), and updating these parameters (based on these new estimates – **M-step**)

GMM best practices

NTB
Interstaatliche Hochschule für Technik Buchs
FHO Fachhochschule Ostschweiz

- Use **log-likelihoods** instead of likelihoods
  ➔ Likelihoods become so small that one ends up with numerical instabilities otherwise
- Use a **diagonal covariance** matrix
  ➔ Simpler/faster training, same/better results due to more compact model (with more mixtures)
- Use a **variance limit** and beware of **curse of dimensionality**
  ➔ Prohibit artifacts through underestimation of components

- Use **16-32 mixtures** and a minimum of **30s of speech** (ML)
- Adapt only means from 512-1024 mixtures per gender (MAP)

Score only with top-scoring mixtures

- **Find optimal number of mixtures** for data via brute force and BIC
- **Compare** models via

Score-wise (more precise): Generalized Likelihood Ratio (GLR)

Parameter-wise (faster): Earth Mover's Distance (EMD) or this paper:

Beigi, Maes, Sorensen, *«A distance measure between collections of distributions and its application to speaker recognition»*, 1998