

Basic Mathematical Tools for Imaging and Visualization

Some Basic Tools from Statistics, Probability and Information Theory

Nassir Navab

24 November 2006

chair for computer aided medical procedures and augmented reality

department of computer science | technische universität münchen

Today's Menu

- Covariance, Correlation and PCA revisited

- Inferential Statistics (Parameter Estimation Basics)
 - Bayesian Prediction
 - Maximum a Posteriori
 - Maximum Likelihood Estimation (MLE)

- Information Theory Tools
 - Entropy
 - Mutual Information

Covariance, Correlation and PCA revisited

Covariance

- Covariance is the generalization of the Variance
- Let x and y be random variables with respective expectations $E[x]$ and $E[y]$

$$\begin{aligned} \text{Var}(x) &= E[(x_i - E[x])^2] \\ &= E[x^2] - E[x]^2 \\ &= E[xx] - E[x]E[x] \end{aligned}$$

$$\begin{aligned} \text{Cov}(x, y) &= E[(x - E[x])(y - E[y])] \\ &= E[xy] - E[x]E[y] \end{aligned}$$

Covariance Matrix

- Let $X=(X_1, X_2, \dots, X_n)$ be a multivariate random variable
- Then the Covariance Matrix is defined as

$$COV(X) = \begin{bmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) & \dots & Cov(X_1, X_n) \\ Cov(X_2, X_1) & Cov(X_2, X_2) & \dots & Cov(X_2, X_n) \\ \vdots & \vdots & & \vdots \\ Cov(X_n, X_1) & Cov(X_n, X_2) & \dots & Cov(X_n, X_n) \end{bmatrix}$$

Covariance Matrix: Measurement of a 3D Point

- PDF for the measurements is not known beforehand
→ assume that every measurement has the same probability (Uniform PDF)
- $P = (Px_i, Py_i, Pz_i)$ are the n measurements, $i=1:n$
- The measurements are modelled as a multivariate random variable $X=(x,y,z)$
- The function X is discretized by using the measurements, such that $x_i = Px_i$, $y_i = Py_i$ and $z_i = Pz_i$

$$COV(X) = \frac{1}{n} \sum_{k=1}^n \begin{pmatrix} (x_k - \bar{x})(x_k - \bar{x}) & (x_k - \bar{x})(y_k - \bar{y}) & (x_k - \bar{x})(z_k - \bar{z}) \\ (y_k - \bar{y})(x_k - \bar{x}) & (y_k - \bar{y})(y_k - \bar{y}) & (y_k - \bar{y})(z_k - \bar{z}) \\ (z_k - \bar{z})(x_k - \bar{x}) & (z_k - \bar{z})(y_k - \bar{y}) & (z_k - \bar{z})(z_k - \bar{z}) \end{pmatrix}$$

Correlation

If we normalize the covariance with the product of the standard deviations of x and y $\sigma_x\sigma_y$ we get the so called *correlation*:

$$\text{cor}(x, y) = \frac{\text{cov}(x, y)}{\sigma_x\sigma_y}$$

- Normalized version of the Covariance
- Takes values from $[-1,1]$
- Expresses Linear Dependency of x and y
 - $\text{Cor}(x,y)=0 \rightarrow x$ and y *linearly independent*
(non-linear dependency is however still possible)

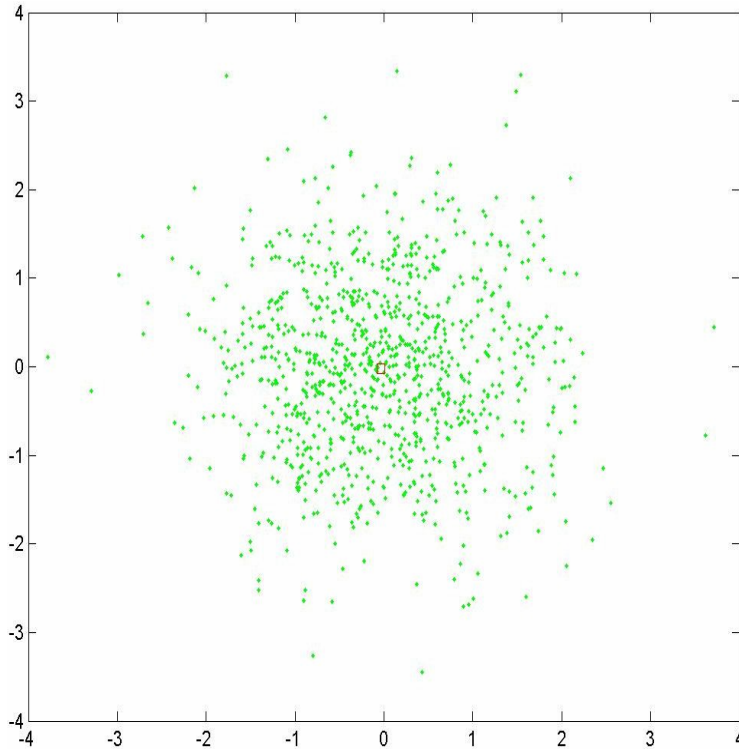
Correlation Matrix

- Let $X=(X_1, X_2, \dots, X_n)$ be a multivariate random variable
- Then the Correlation Matrix is defined as

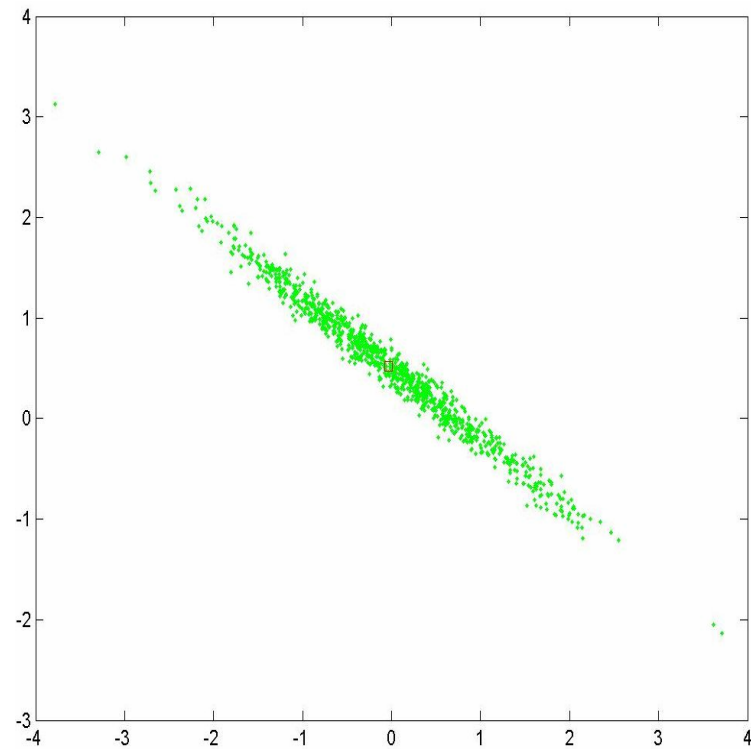
$$COR(X) = \begin{bmatrix} Cor(X_1, X_1) & Cor(X_1, X_2) & \dots & Cor(X_1, X_n) \\ Cor(X_2, X_1) & Cor(X_2, X_2) & \dots & Cor(X_2, X_n) \\ \vdots & \vdots & \dots & \vdots \\ Cor(X_n, X_1) & Cor(X_n, X_2) & \dots & Cor(X_n, X_n) \end{bmatrix}$$

- Properties:
 - Entries give the respective correlation coefficients
 - Diagonal equals to 1

Covariance and Correlation tell you about Certainty of Measurements

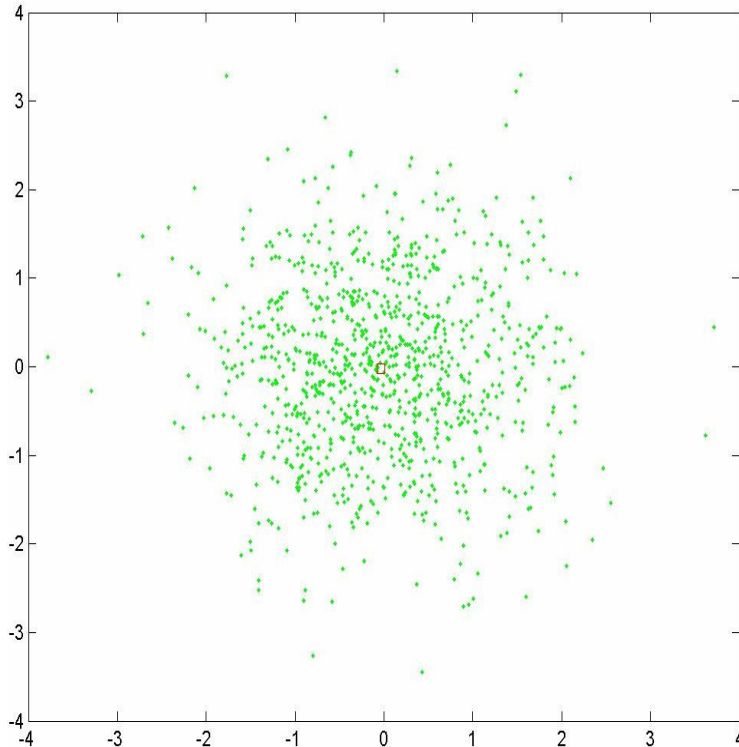


$$C = \begin{bmatrix} 0.9717 & -0.0069 \\ -0.0069 & 0.9966 \end{bmatrix}$$

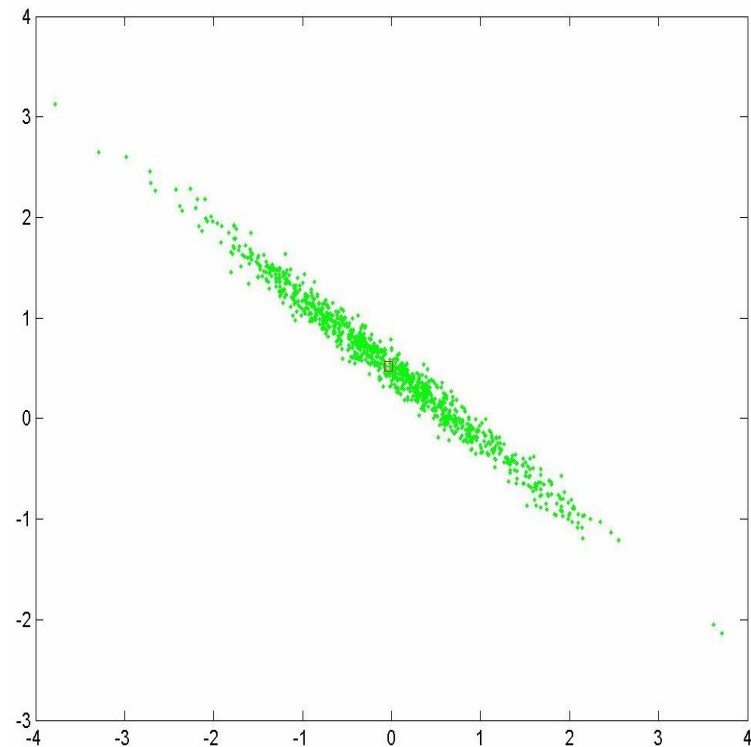


$$C = \begin{bmatrix} 0.9717 & -0.6781 \\ -0.6781 & 0.4834 \end{bmatrix}$$

Covariance and Correlation tell you about Independence of Variables



$$C = \begin{bmatrix} 0.9717 & -0.0069 \\ -0.0069 & 0.9966 \end{bmatrix}$$



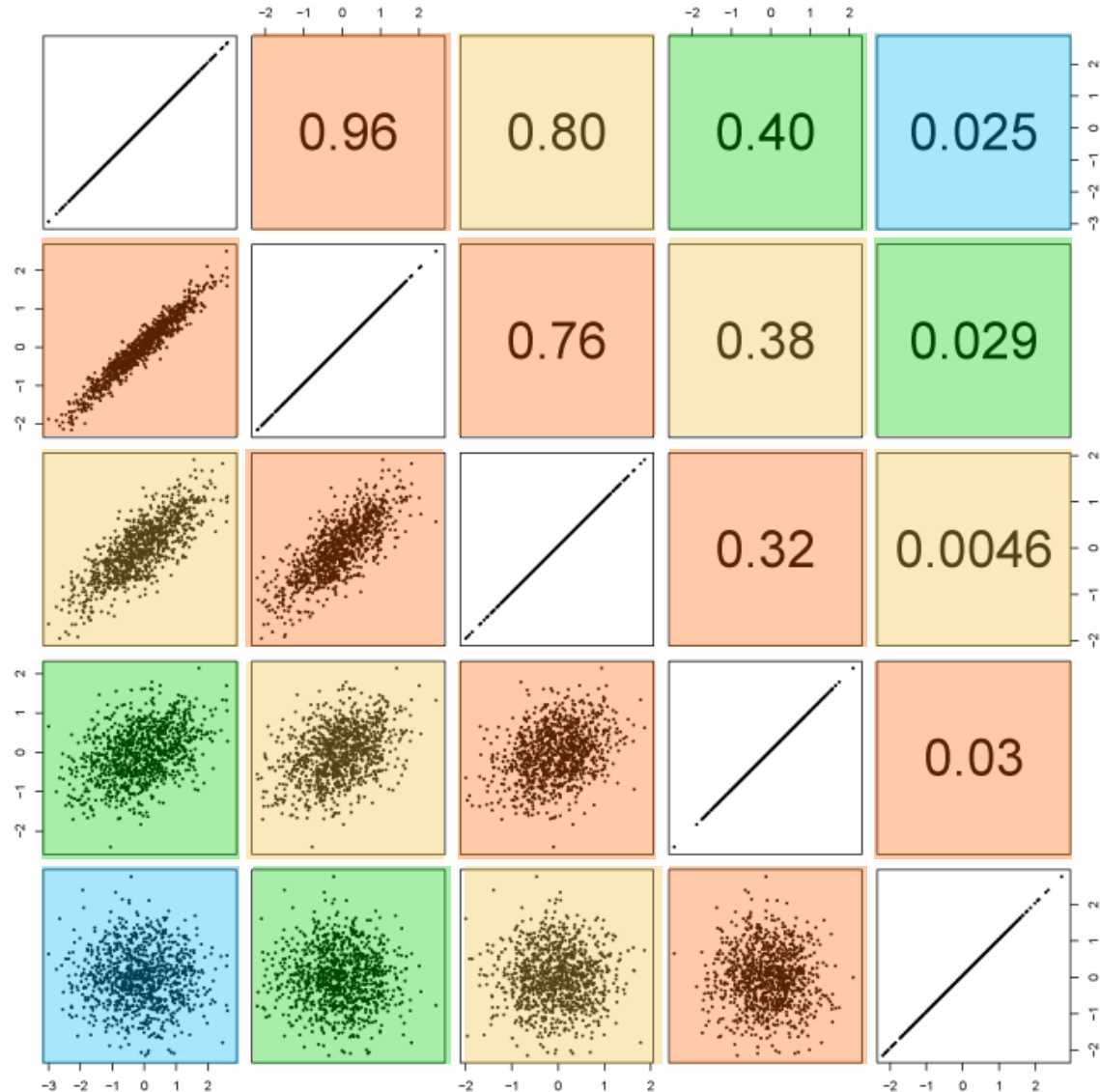
$$C = \begin{bmatrix} 0.9717 & -0.6781 \\ -0.6781 & 0.4834 \end{bmatrix}$$

Examples for Linear Correlations between two Variables X and Y

(1000 Measurements)

The data are graphed on the lower left and their correlation coefficients listed on the upper right. Each square in the upper right corresponds to its mirror-image square in the lower left, the "mirror" being the diagonal of the whole array. Each set of points correlates maximally with itself, as shown on the diagonal (all correlations = +1).

Example adapted from:
<http://en.wikipedia.org/wiki/Correlation>



Application of the most Stuff so far: Principal Component Analysis (PCA)

PCA

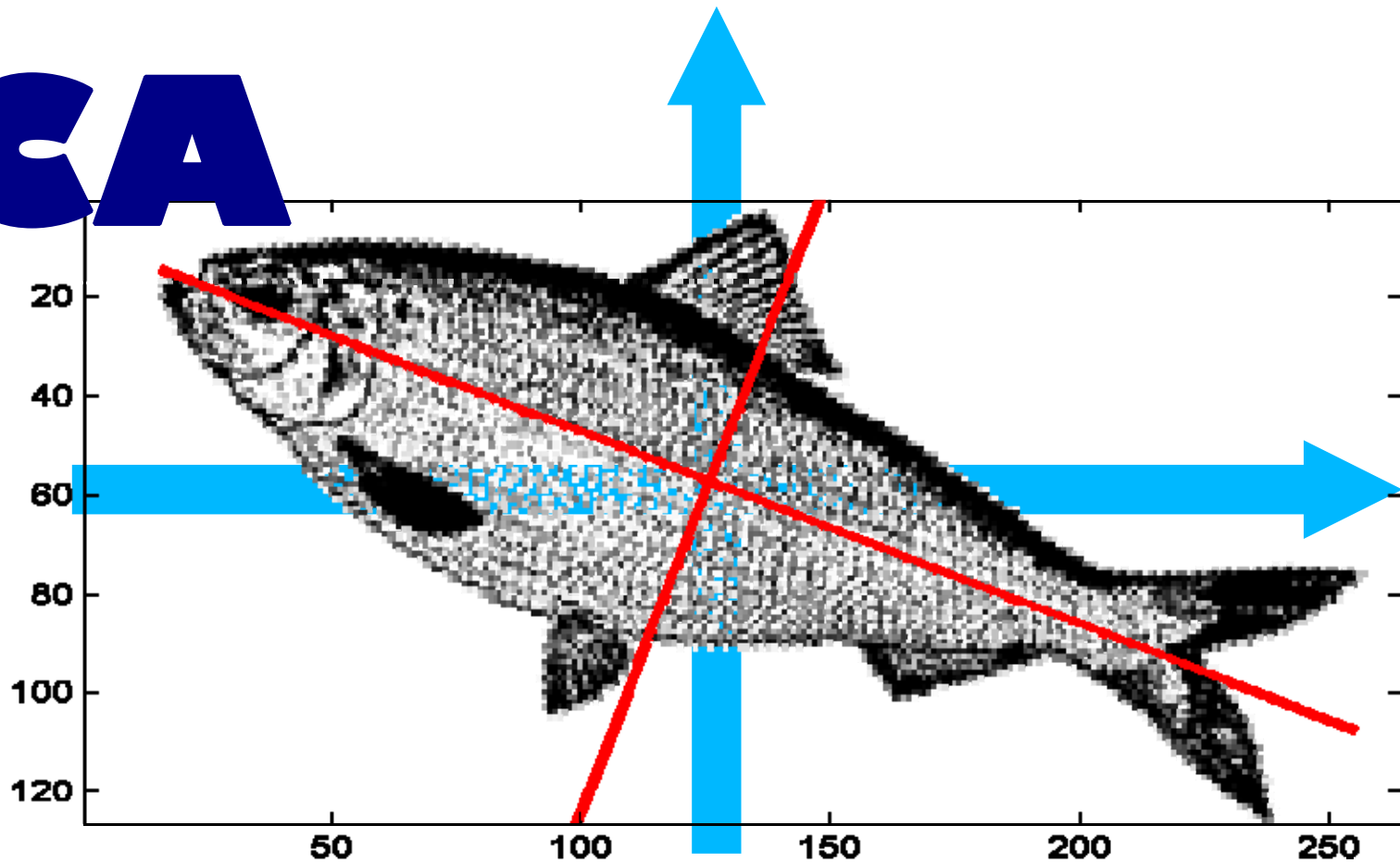


Image from: <http://de.wikipedia.org/wiki/Hauptkomponentenanalyse>

Principal Component Analysis (PCA)

- What is it good for?
 - Finding patterns in data
 - Reducing number of dimensions → Compression

PCA Steps

1. Get Data
2. Subtract the mean
3. Compute the Covariance Matrix C
4. Compute Eigenvectors and Eigenvalues of C
 - Eigenvalue denote Principal Components
 - Largest Eigenvalue is the most important Principal Component
 - Eigenvectors give the intrinsic coordinate system for the data
 $B=[ev_1, \dots, ev_n]$ (ordered eigenvectors)
5. Applications:
 - Transforming to standard coordinate system $x = B^T x$
 - Dimension Reduction: use only dimensions corresponding to large principal components

Inferential Statistics Basics (Parameter Estimation Approaches)

Reference:

Artificial Intelligence: A Modern Approach (Chapter 20)

by Stuart Russell and Peter Norvig

General Problem of Parameter Estimation

- Fact: We observe Data
- Question: How to find a correct model which explains our observations?
- Or easier: Given a certain parametrized model, how to find the right parameters for this model?



- Notation:
 - Data d
 - Hypothesis h_i : Set of parameters for a certain model

Candy Example

from Russel and Norvig: *Artificial Intelligence*

Image from: <http://de.wikipedia.org/wiki/Bonbon>



- You have a bag of candy
- These bags have different mixtures of lime (ugh...) and cherry (yum...)
 - h_1 : 100% cherry
 - h_2 : 75% cherry 25% lime
 - h_3 : 50% cherry 50% lime
 - h_4 : 25% cherry 75% lime
 - h_5 : 100% lime
- The bags are not distinguishable
- Given a bag, you start eating the candy
- How can you predict what will be the next candy?

Overview of Different Approaches

Bayesian Estimation

Bayesian Estimation
optimal, however slow

Maximum a Posteriori (MAP)

Maximum a Posteriori (MAP)
Approximation to Bayesian Estimation

Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation (MLE)
Approximation to the MAP

Bayesian Prediction

- Fact:
 - Given the Observations, it is not so easy to give the probability that a certain model is true

$$P(h_i|d)$$

- Given the Model, it is easy to compute the probability of a certain Observation

$$P(d|h_i)$$

- Solution ? → Bayes' Rule

Bayesian Prediction : Bayes' Rule

$$P(h_i|d) = \frac{P(d|h_i) P(h_i)}{P(d)} = \alpha P(d|h_i) P(h_i)$$

- Estimates the right parameters given the data by using the probability for the data given the parameters
- Notation:
 - Hypothesis Prior : $P(h_i)$
 - Data Likelihood : $P(d|h_i)$

Bayesian Prediction

- Make a prediction based on a mixture of all predictions, weighted by their respective probabilities

$$\begin{aligned} P(X|d) &= \sum_i P(X|h_i) P(h_i|d) \\ &= \sum_i P(X|h_i) P(d|h_i) P(h_i) \end{aligned}$$

- Properties:
 - We can expect that with sufficient measurements, the true hypothesis will dominate the mixture of predictions
 - Bayesian Prediction is optimal:
Given the hypothesis prior, it will be more correct than any other prediction

Candy Example

from Russel and Norvig: Artificial Intelligence

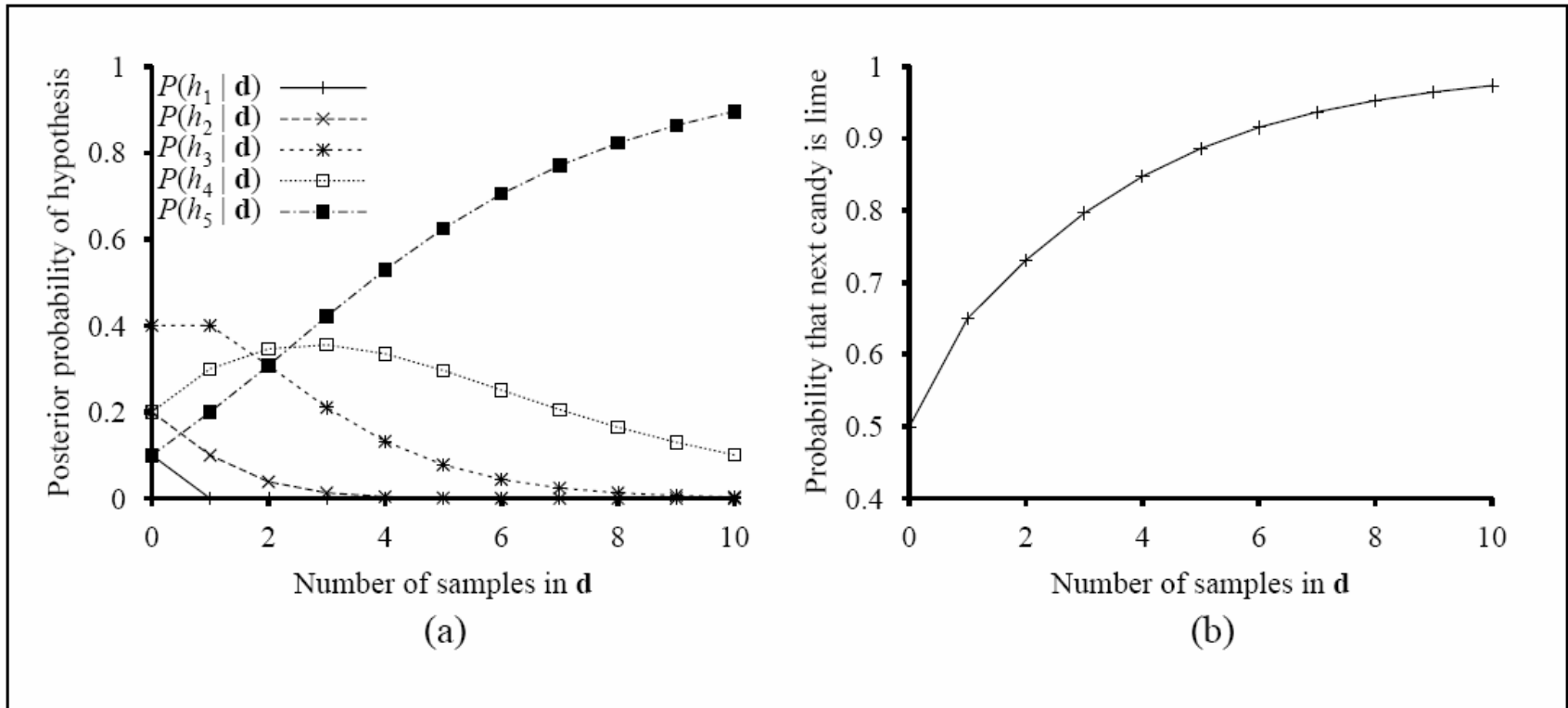


Figure 20.1 (a) Posterior probabilities $P(h_i | d_1, \dots, d_N)$ from Equation (20.1). The number of observations N ranges from 1 to 10, and each observation is of a lime candy. (b) Bayesian prediction $P(d_{N+1} = lime | d_1, \dots, d_N)$ from Equation (20.2).

Maximum a Posteriori (MAP)

- Use only one hypothesis for prediction instead of a mixture
- Replaces a large summation (Bayesian Estimation) by optimization

$$\begin{aligned}h_{MAP} &= \arg \max_{h_i} P(h_i|d) \\ &= \arg \max_{h_i} \alpha P(d|h_i) P(h_i) \\ &= \arg \max_{h_i} P(d|h_i) P(h_i)\end{aligned}$$

- Properties:
 - Gives the most simple hypothesis which is consistent with the given data
(natural embodiment of Ockham's Razor)

Maximum Likelihood Estimation (MLE)

- Approximation to / Simplification of the MAP
- Assume that each hypothesis has the same probability ($P(h_i)=P(h_j)$)
- Seems reasonable if we know nothing about the hypotheses a priori

$$\begin{aligned}h_{ML} &= \arg \max_{h_i} P(h_i|d) \\ &= \arg \max_{h_i} \alpha P(d|h_i) P(h_i) \\ &= \arg \max_{h_i} P(d|h_i) P(h_i) \\ &= \arg \max_{h_i} P(d|h_i)\end{aligned}$$

Summary of Different Estimation Approaches

Bayesian Estimation

Bayesian Estimation

- Mixture of all predictions
- Optimal
- Slow

Maximum a Posteriori (MAP)

Maximum a Posteriori (MAP)

- Prediction based on one hypothesis which is best supported by data
- Faster than Bayesian Estimation

Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation (MLE)

- Simplification of MAP
- Assumption: All hypotheses are equally probable a priori

Some Basic Information Theory Tools

Reference:

Pattern Recognition and Machine Learning

Christopher M. Bishop

Overview

Entropy

measures the amount of information in a random variable

Mutual Information

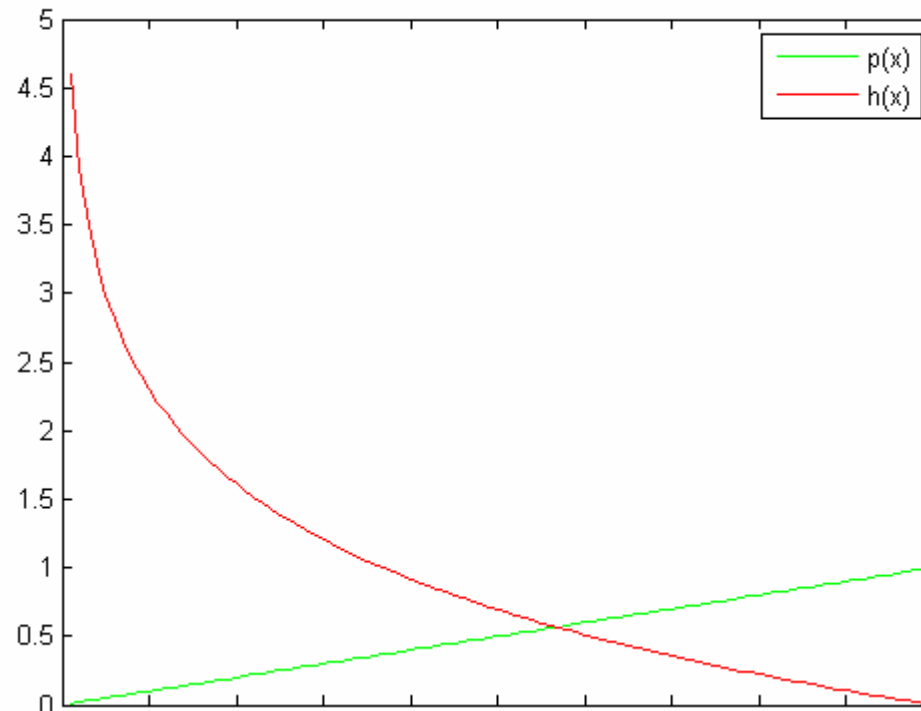
measures the amount of information shared by two random variables

Modelling of Information

- Given a random variable X , let x denote a single event
- How much information do we receive, when we observe a specific value of this variable? ('degree of surprise')
 - event x very probable \rightarrow little amount of information
 - event x not very probable \rightarrow large amount of information
- Amount of information $h(x)$
- $h(x)$ should monotonically depend on $p(x)$
- Assumptions for unrelated events x, y
 - $h(x, y) = h(x) + h(y)$
 - $p(x, y) = p(x)p(y)$
- $\rightarrow h(x) = -\log(p(x))$

Modelling of Information

- Example of relation between h and p for $p(x)=x$



Entropy

- Average amount of information in the random variable X ?

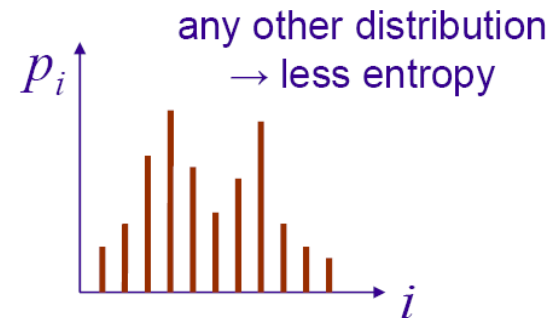
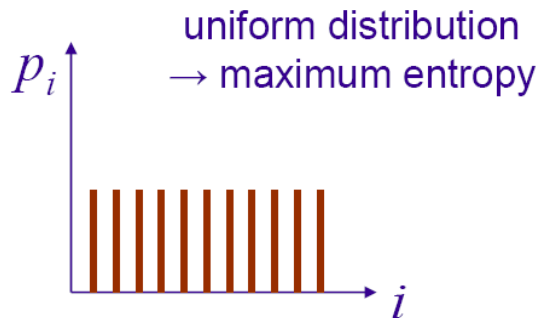
(Amount of information in single event x)
multiplied by
(Probability for the event x)

$$H[X] = - \sum_x p(x) \log(p(x))$$

Entropy

$$H[X] = - \sum_x p(x) \log(p(x))$$

- Properties:
 - Maximal for uniform distribution of X
 - Expresses the minimal length for encoding X
(*Noiseless Coding Theorem, Shannon 1948*)



Mutual Information

★ The joint entropy is defined as

$$H[X, Y] = - \sum_x \sum_y p(x, y) \log(p(x, y))$$

★ Remember that for independent X, Y it holds that

$$H(X, Y) = H(X) + H(Y)$$

★ If X and Y share information, then

$$H(X, Y) < H(X) + H(Y)$$

★ We will denote the amount of this shared information by S

$$H(X, Y) = H(X) + H(Y) - S(X, Y)$$

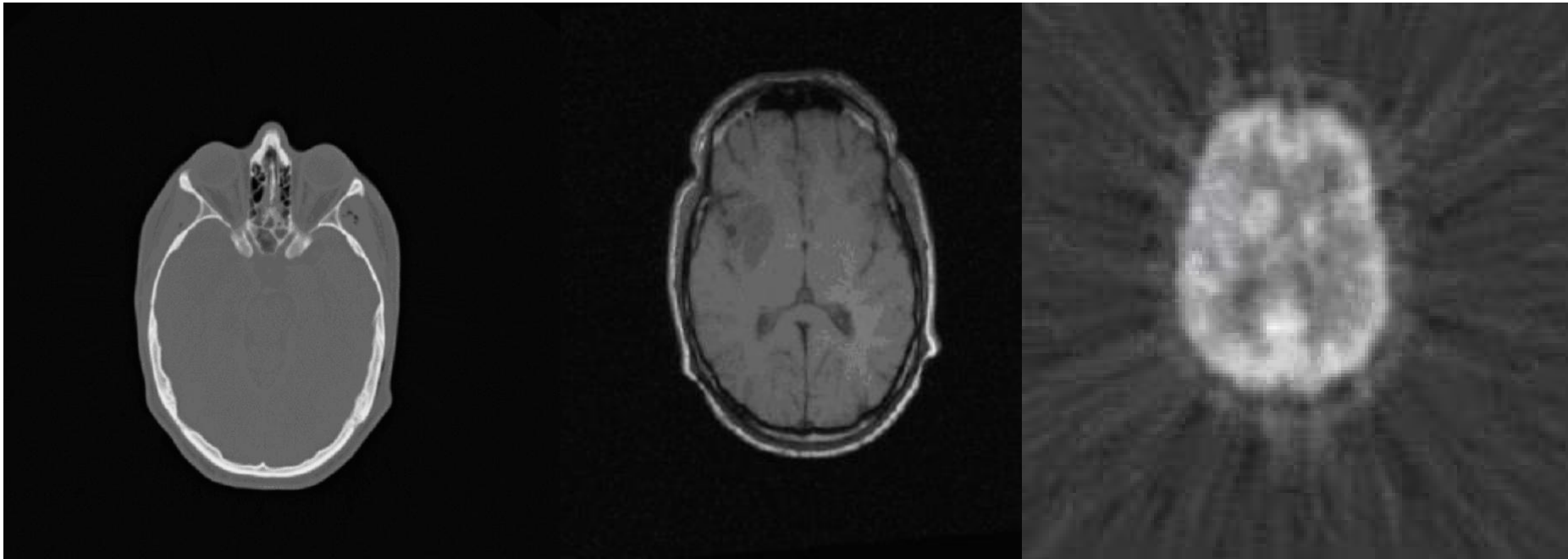
Mutual Information

$$S(X, Y) = H(X) + H(Y) - H(X, Y)$$

$$\begin{aligned} S(X, Y) &= H(X) + H(Y) - H(X, Y) \\ &= - \sum_x p(x) \log(p(x)) - \sum_y p(y) \log(p(y)) + \sum_x \sum_y p(x, y) \log(p(x, y)) \\ &= \sum_x \sum_y p(x, y) \log \left(\frac{p(x, y)}{p(x) \cdot p(y)} \right) \end{aligned}$$

Why bother?

- For example:
Measurement of similarity between different images → Later in the lecture



Images from: <http://www.itk.org/HTML/MutualInfo.htm>

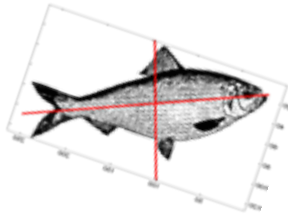
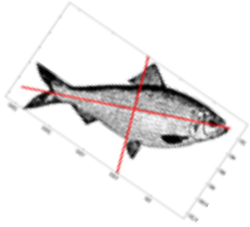
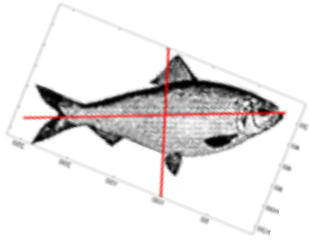
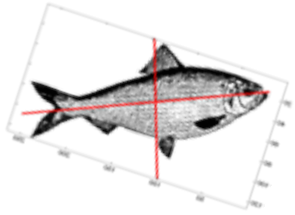
Summary

Entropy

measures the amount of information in a random variable

Mutual Information

measures the amount of information shared by two random variables



End

