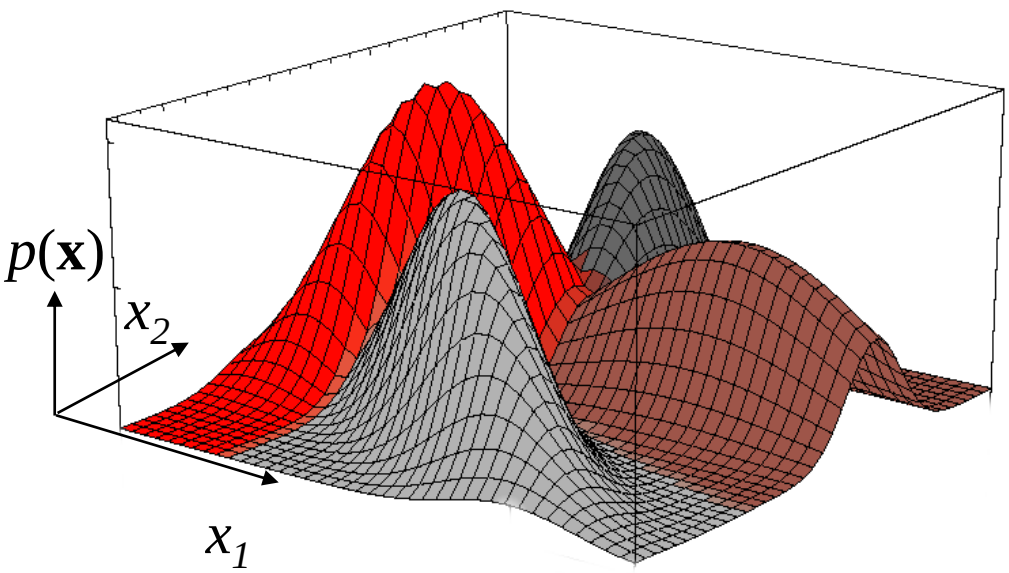
Statistical Estimation Theory

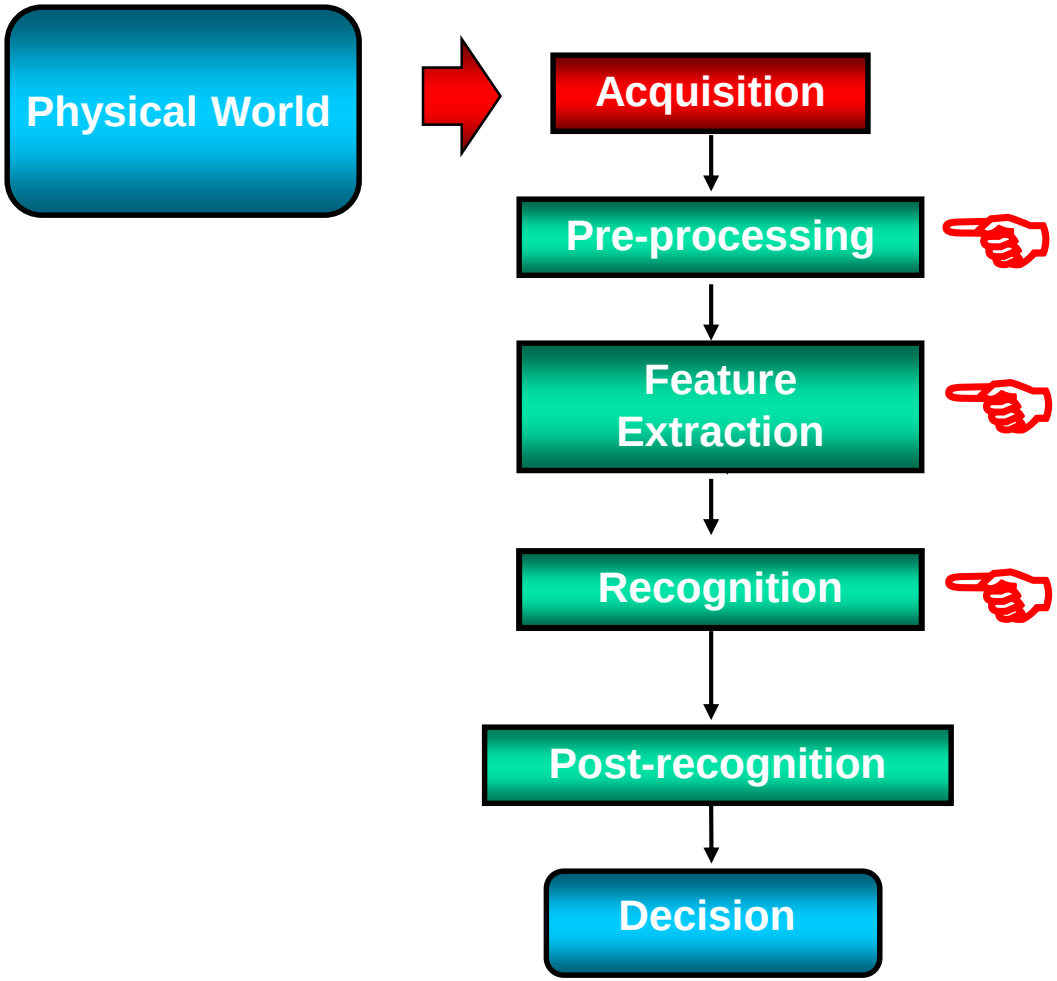
- Introduction
- PDF Estimation
- Maximum Likelihood Estimation
- Bayesian Estimation
- K-NN Estimation
- Parzen Windows Estimation
- Expectation-Maximization Algorithm

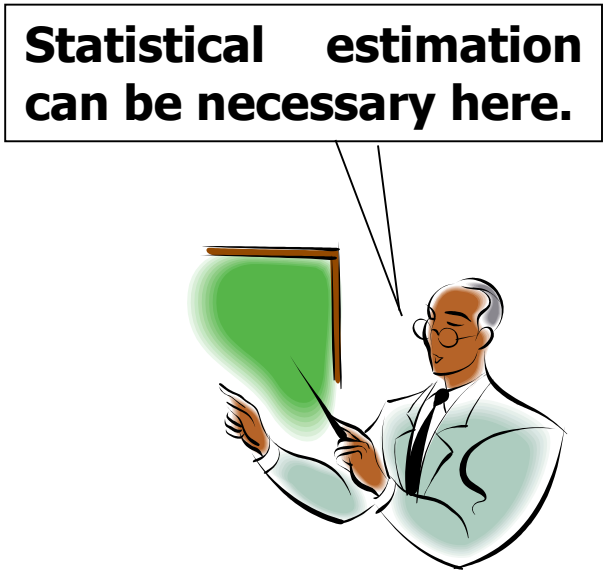
Introduction

- Statistical estimation is rather important as it may intervene in different parts of a system based on machine learning.
- The objective of this chapter is to study methods and algorithms for estimating the statistical distribution of a stochastic signal from a set of available observations (samples).

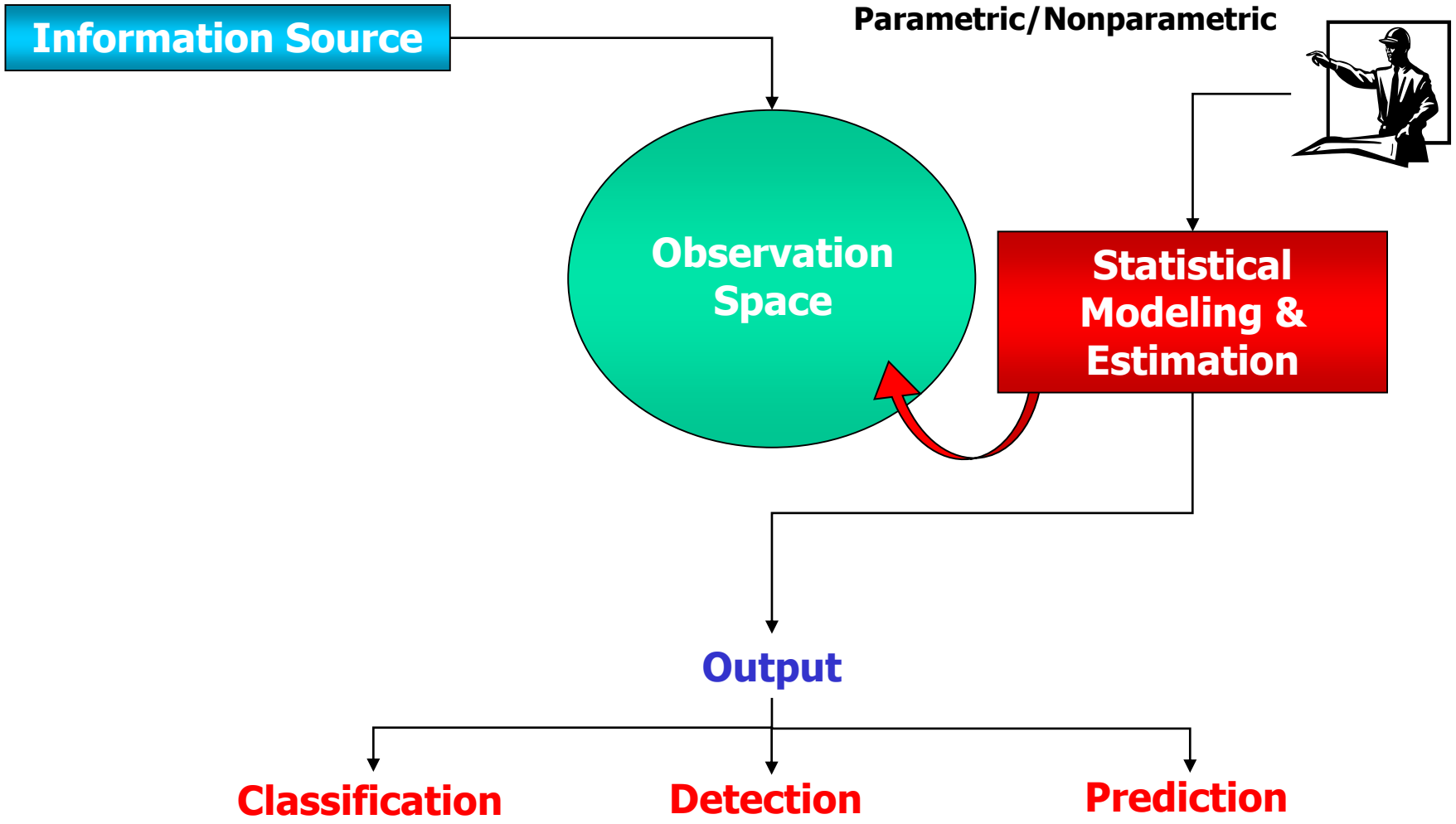


Introduction





Introduction



Introduction

- In the estimation theory, stochastic signals can be subdivided into three categories:
  - Noisy deterministic signals

    The information source is completely known. The noise interference intervenes during the transmission and/or the acquisition phases (e.g., transmission of signals in telecommunication systems based on PAM).
  - Noisy parametric signals

    The information source is only partially known. The observations allow estimating the random parameters controlling the behavior of the signal (e.g., target speed velocity estimation in sonar systems).
  - Noisy random signals

    The signal is completely unknown. In this case, the estimation should rely completely on the available observations (e.g., buried object detection with ground penetrating radar).

Introduction

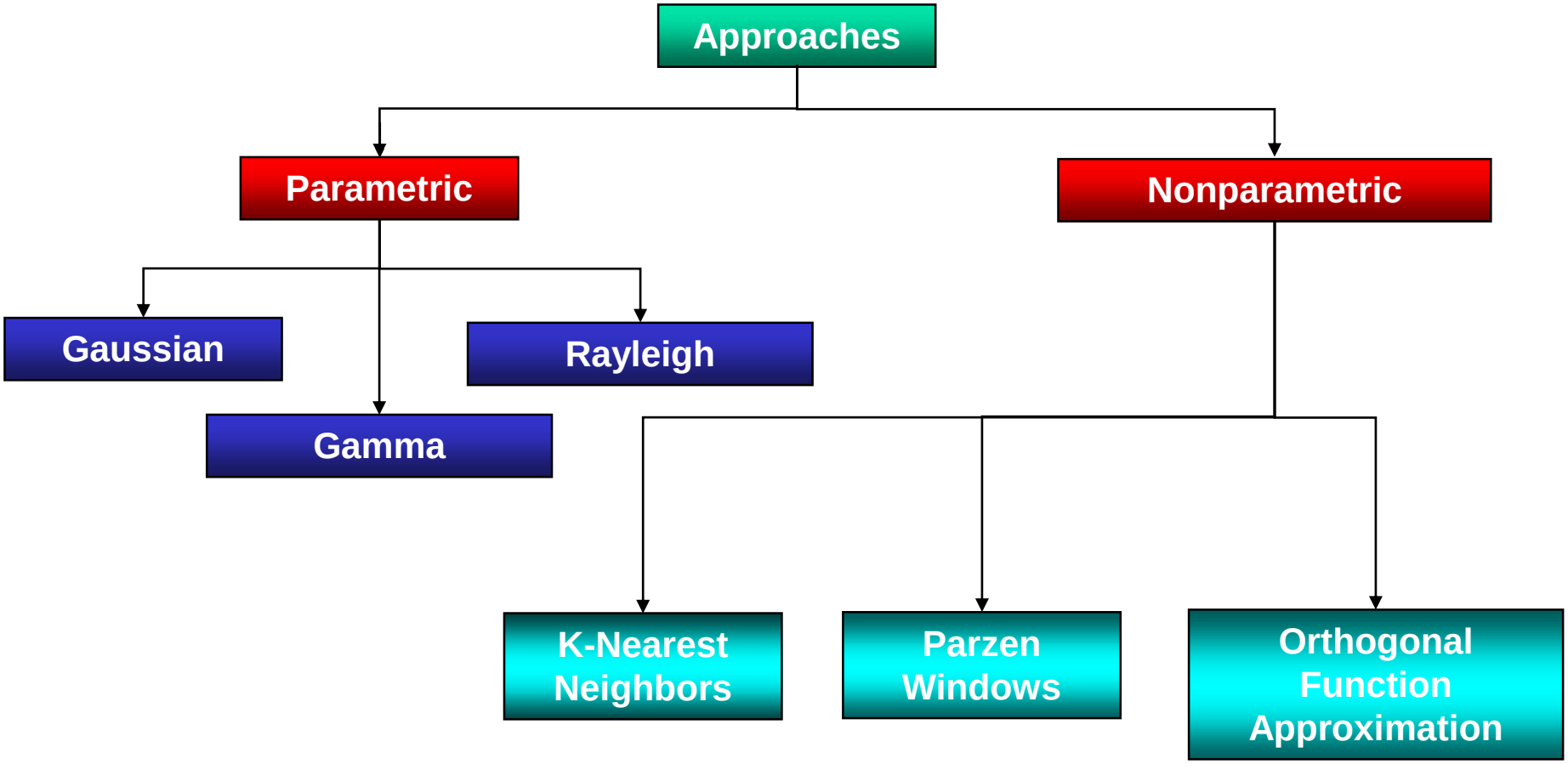
- In the following, focus will be given to the second and third categories, which are the most common in real applications.
- In particular, we will face the problem of the estimation of the probability density function (pdf), which is important as we have understood before.

PDF Estimation

- Let $\mathbf{x}=(x_1, x_2, ..., x_n)$ be a vector of n features with unknown pdf $p(\mathbf{x})$.
- Let $\boldsymbol{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ be a finite set of N independent and identically distributed (iid) samples (observations) drawn from the considered pdf.
- Let us term these samples as training samples.

Objective: to determine an estimate $\hat{p}(\mathbf{x})$ on the basis of the available samples $\boldsymbol{X}$, which is as close as possible to the true pdf $p(\mathbf{x})$.

PDF Estimation



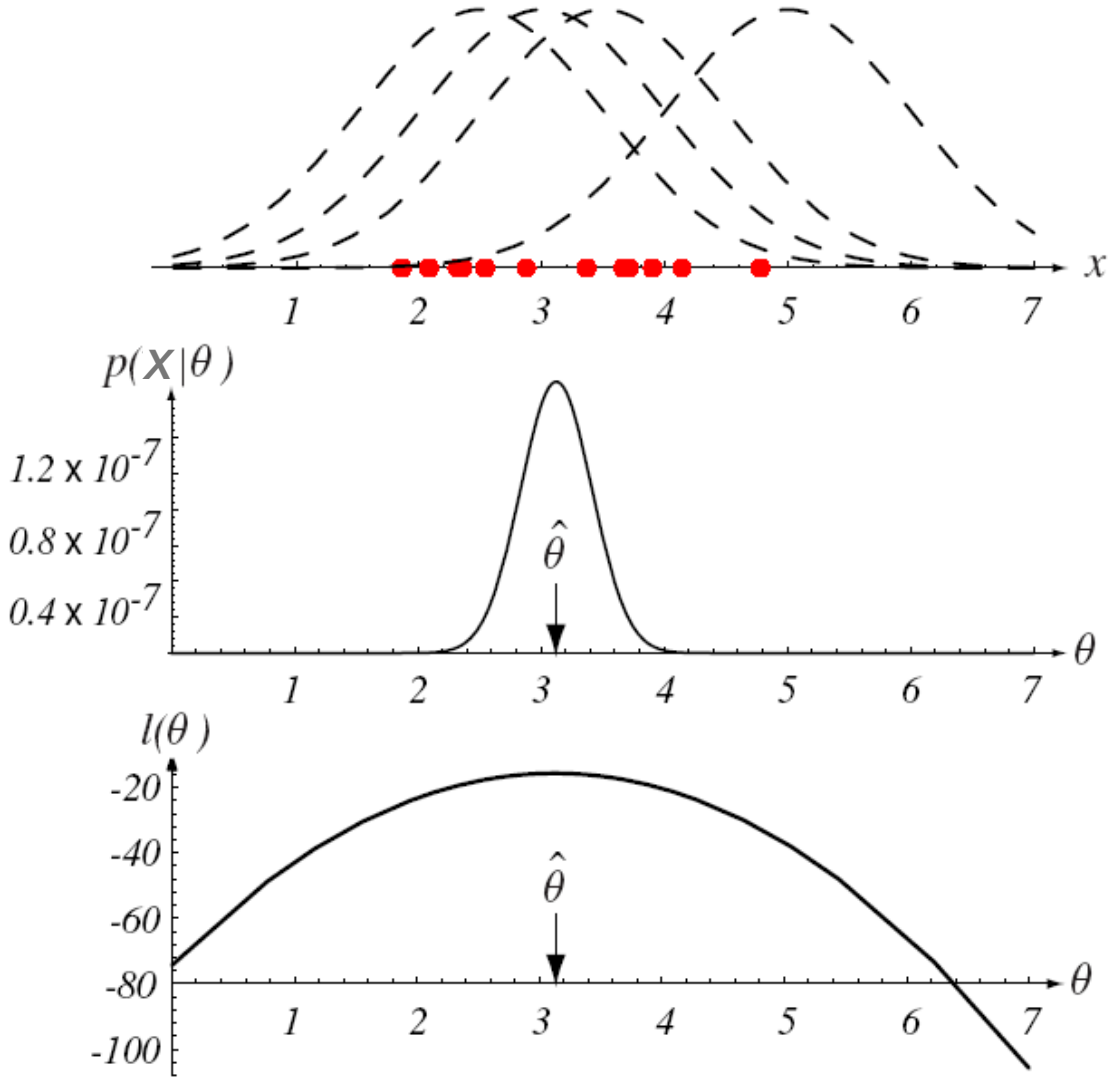
Parametric Estimation

- Let us assume that the model of $p(\mathbf{x})$ is characterized by r parameters which define a parameter vector $\boldsymbol{\theta}=(\theta_1, \theta_2 \ldots\ldots \theta_r.)$.
- The dependency of the model on $\boldsymbol{\theta}$ is underlined by the following notation $p(\mathbf{x}|\boldsymbol{\theta})$.
- Since $\boldsymbol{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ is a vector of iid random variables, we can define a related likelihood function as follows:

$$p(\mathbf{X} \mid \boldsymbol{\theta}) = \prod_{k=1}^{N} p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

- This function defines the likelihood of θ with respect to the considered set of samples ($\boldsymbol{X}$).
- In other words, it provides a useful measure of compatibility/agreement between θ and $\boldsymbol{X}$.

Likelihood Function: Example



Estimation Procedures

Main Procedures

Maximum Likelihood Estimation

Θ is viewed as a vector of quantities whose values are fixed but unknown. The best estimate of their value is defined to be the one that maximizes the probability of obtaining the samples actually observed.

Bayesian Estimation

The parameters are viewed as random variables having some known prior distribution. Observation of the samples converts this to a posterior density.

Estimation Goodness

- The estimate of the vector of parameters depends on the observation vector ***X***:

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{X})$$

- Therefore, the estimate is a random vector.
- Let us define the estimation error $\boldsymbol{\varepsilon}$ as:

$$\boldsymbol{\varepsilon} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \boldsymbol{\varepsilon}(\mathrm{X}, \boldsymbol{\theta}) = [\hat{\theta}_i - \theta_i : i = 1,2,...,r]$$

- In order to get an ideal estimate for each parameter θ_i (i = 1, 2, ..., r), it is necessary that each corresponding estimation error ε_i :
  - is unbiased
  - has no variance

Estimation Goodness

- The bias of an estimation error is its mean value.
- An estimate is said unbiased if:

$$E\{\boldsymbol{\varepsilon}\} = 0 \quad \Longrightarrow \quad E\{\hat{\boldsymbol{\theta}}\} = \boldsymbol{\theta}$$

- Its variance is defined as:

$$\mathrm{var}\{\varepsilon_i\} = E\{(\hat{\theta}_i - \bar{\hat{\theta}}_i)^2\} \qquad (i = 1,2,\ldots, r)$$

- In order to assess the goodness of the variance of our estimate, it is necessary to refer it to the so-called Cramér-Rao bound.

Cramér-Rao Bound

It expresses a lower bound on the variance of an unbiased statistical estimator, based on the Fisher information.

$$\mathrm{var}\{\varepsilon_i\} \geq [\mathrm{I}^{-1}(\boldsymbol{\theta})]_{ii}, \qquad i = 1, 2, \ldots, r$$

where $\mathrm{I}(\boldsymbol{\theta}) = E\{\nabla_{\boldsymbol{\theta}} ln[p(\mathbf{X}|\ \boldsymbol{\theta})]\nabla_{\boldsymbol{\theta}} ln[p(\mathbf{X}|\ \boldsymbol{\theta})]^t\}$ is the Fisher information matrix which is defined as:

$$[\mathrm{I}(\theta)]_{ij} = E\left\{\frac{\partial \ln[p(\mathbf{X}\,|\,\boldsymbol{\theta})]}{\partial \theta_i} \cdot \frac{\partial \ln[p(\mathbf{X}\,|\,\boldsymbol{\theta})]}{\partial \theta_j}\right\}$$

Nota: When the bound is reached, the estimator is said **efficient**.

Asymptotic Properties

- Often, the estimates used in real problems are obtained by biased and inefficient estimators.
- In order to judge better the goodness of the considered estimator, one analyzes its behavior for large sets of observations.
- In other words, an estimator is said to be good if it has good asymptotic properties.
- An estimate is said to be asymptotically unbiased if:

$$\lim_{N\to+\infty} E\{\boldsymbol{\varepsilon}\} = \mathbf{0} \quad \Longrightarrow \quad \lim_{N\to+\infty} E\{\hat{\boldsymbol{\theta}}\} = \boldsymbol{\theta}$$

- It is asymptotically efficient if:

$$\lim_{N\to+\infty} \frac{\mathrm{var}\{\varepsilon_i\}}{[\mathrm{I}^{-1}(\boldsymbol{\theta})]_{ii}} = 1, \qquad i = 1,2,\ldots,r$$

Asymptotic Properties

- An estimate is said to be consistent if it converges to the true value when $N \to +\infty$:

$$\lim_{N \to +\infty} P\{\|\boldsymbol{\varepsilon}\| < \delta\} = 1 \qquad \forall \delta > 0$$

- The necessary condition for consistency is that the estimate is asymptotically unbiased and with variance converging to zero when $N \to +\infty$.

Maximum Likelihood Estimation

Definition

- The maximum likelihood (ML) estimate of $\boldsymbol{\theta}$ is:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{X} \mid \boldsymbol{\theta})$$

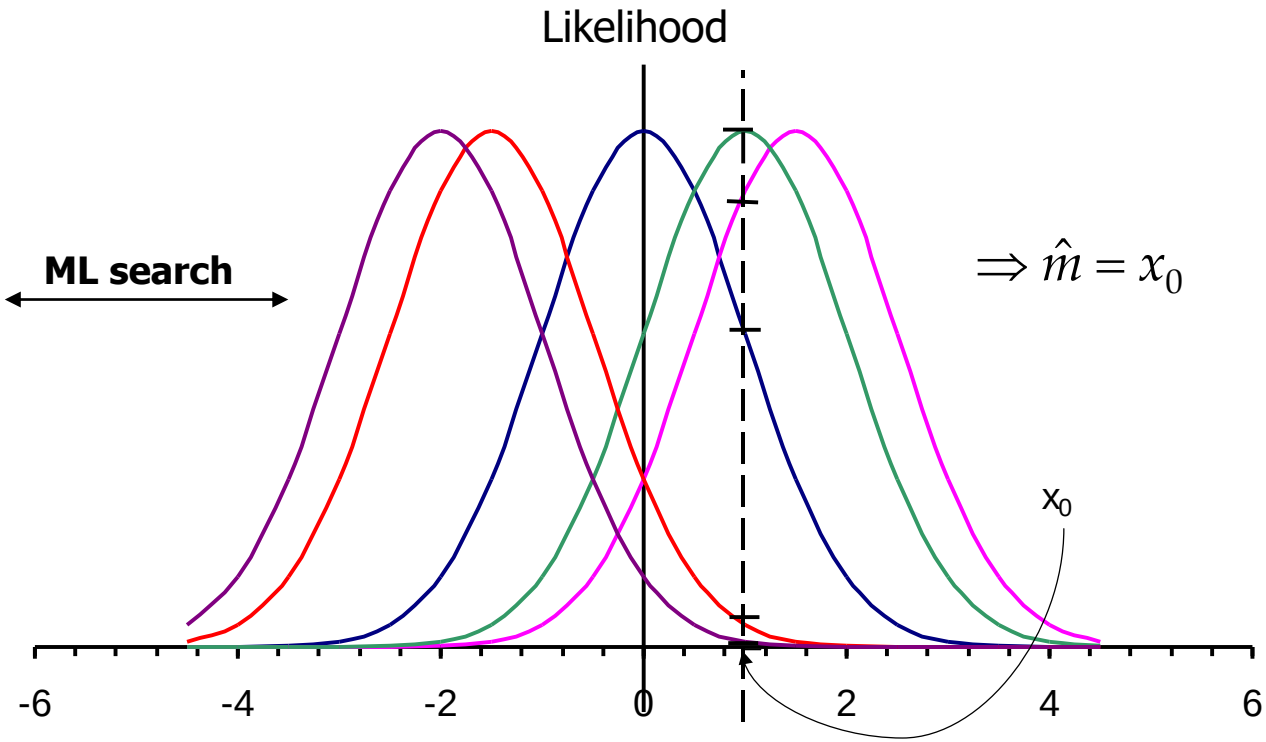
Observations

- By varying $\boldsymbol{\theta}$, one obtains different pdfs, each yielding a likelihood value.
- Intuitively, the ML estimate corresponds to the value of $\boldsymbol{\theta}$ that in some sense best agrees with the actually observed training samples.
- For analytical purposes, it is usually easier to work with the logarithm of the likelihood than with the likelihood itself.
- Since the logarithm is monotonically increasing, the $\boldsymbol{\theta}$ that maximizes the log-likelihood also maximizes the likelihood.

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ln p(\mathbf{X} \mid \boldsymbol{\theta})$$

ML Estimation: Example

ML estimation of the mean of a mono-dimensional Gaussian pdf ($\hat{m}$) with known variance basing on a single observation x_0.



ML Estimation: Properties

- In certain hypotheses regarding $p(\mathbf{x}|\boldsymbol{\theta})$, it can be shown that, if it exists an efficient estimate and if the ML estimate is unbiased, then the efficient estimate is that provided by ML estimation.
- Even if it does not exist an efficient estimate, the ML estimate exhibits good asymptotic properties since it is:
  - asymptotically unbiased
  - asymptotically efficient
  - consistent
- Such properties are behind the common use of ML estimation in numerous real problems.

Statistical Model Selection

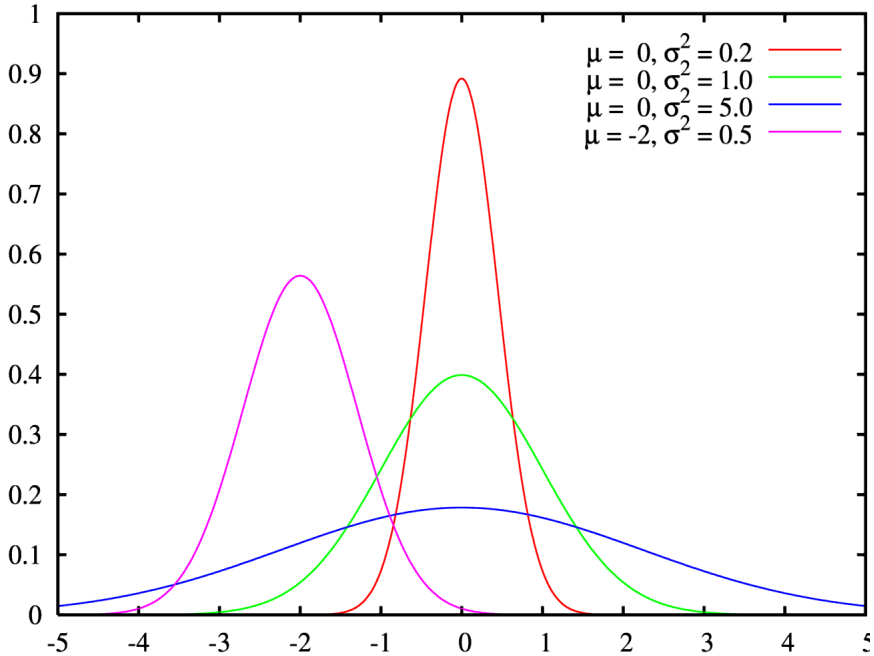
- In the selection of a statistical model for $p(\mathbf{x})$, three main aspects should be considered:
  - The intrinsic statistical nature of the physical phenomenon under analysis;
  - The noise introduced during the (passive/active) signal transmission and acquisition phases by the propagation medium and the sensor, respectively;
  - The pre-processing and feature extraction steps adopted before feeding the recognition system.

Statistical Model Selection

- Among the popular statistical models, one can find:
  - Gaussian model
  - Generalized Gaussian model
  - Gamma model
  - Rayleigh model
  - Chi-square model
  - Log-Normal model

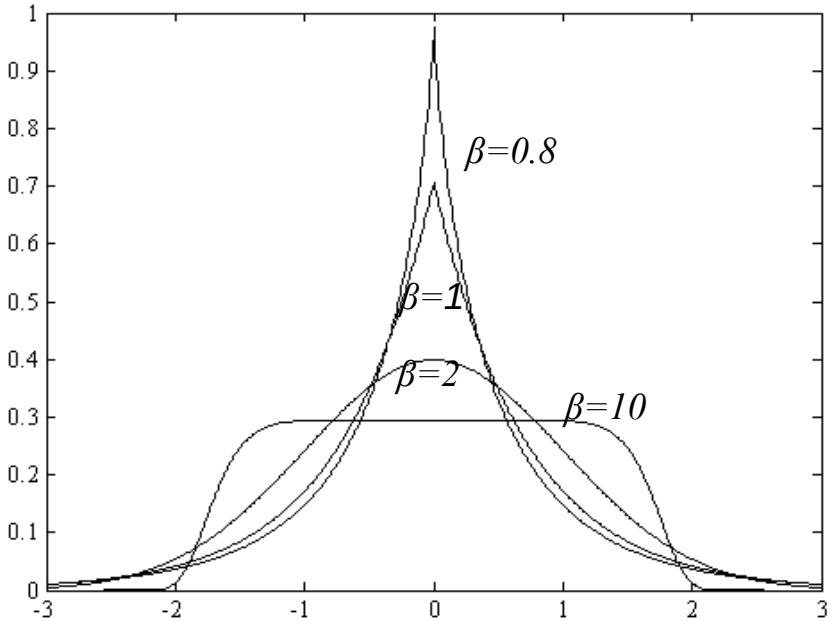
PDF Examples

Gaussian



$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

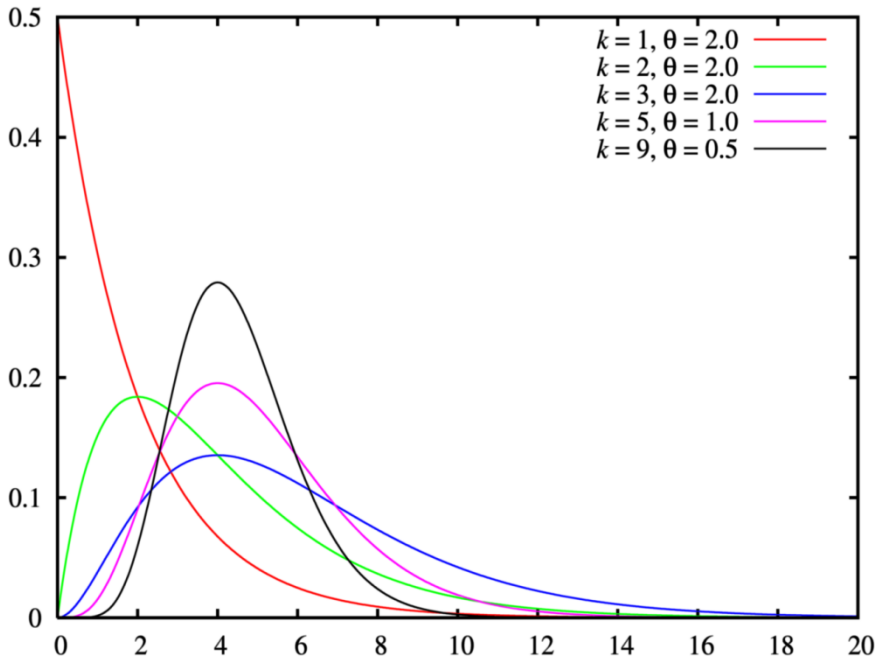
Generalized Gaussian



$$f(x;m,b,\beta) = \frac{b\beta}{2\Gamma(1/\beta)} e^{-\left[b|x-m|\right]^{\beta}}$$

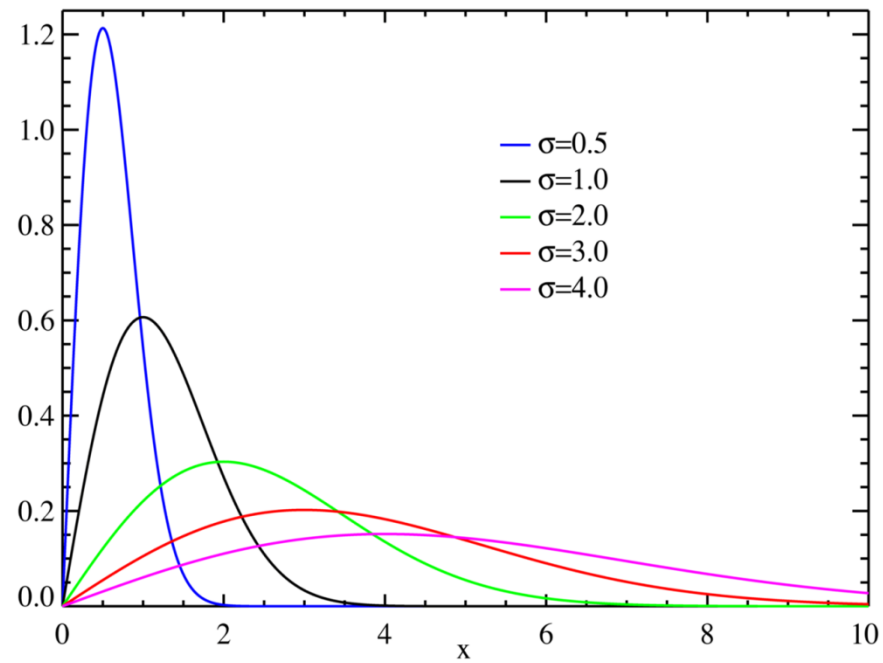
PDF Examples

Gamma



$$f(x;k,\theta) = x^{k-1}\frac{e^{-x/\theta}}{\theta^k\,\Gamma(k)} \text{ for } x > 0$$

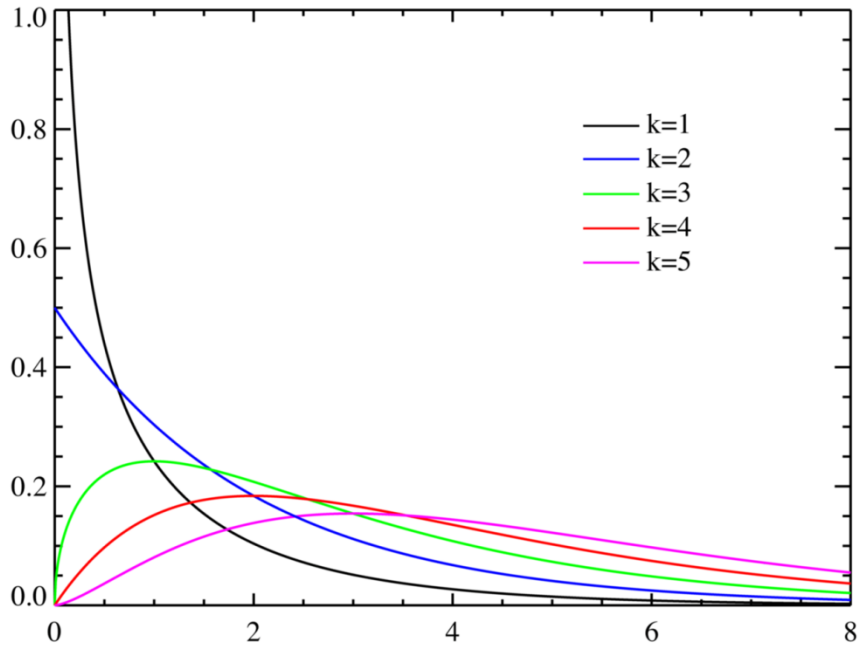
Rayleigh



$$f(x|\sigma) = \frac{x\exp\left(\frac{-x^2}{2\sigma^2}\right)}{\sigma^2}$$

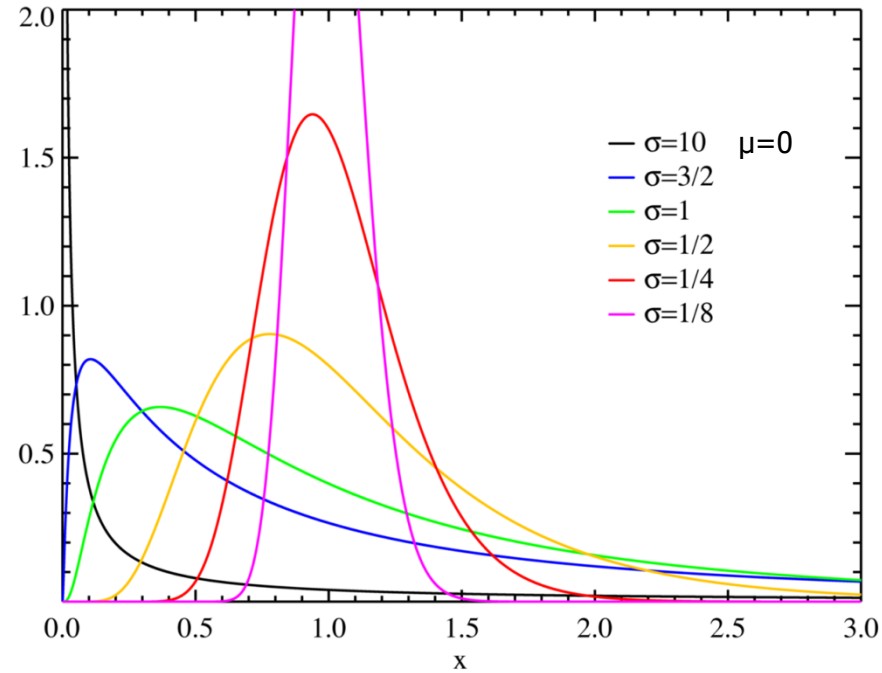
PDF Examples

Chi-Square



$$f(x;k) = \frac{(1/2)^{k/2}}{\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

Log-Normal



$$f(x;\mu,\sigma) = \frac{e^{-(\ln x-\mu)^2/(2\sigma^2)}}{x\sigma\sqrt{2\pi}}$$

Gaussian Model

- The Gaussian model is a parametric model widely used in numerous applications.

- The motivation behind this success can be found in the central limit theorem which states:

> *"if the sum of many iid variables has a finite variance, then it will be approximately normally distributed."*

- Since many real processes take origin from many independent causes and yield distributions with finite variance, this explains the ubiquity of the Gaussian pdf.

Gaussian Model

- The multivariate Gaussian pdf takes the following analytical expression:

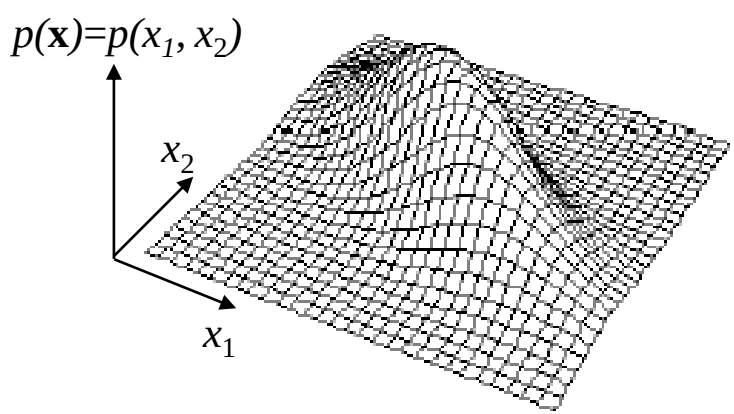
$$p(\mathbf{x}|\theta) = p(\mathbf{x}|\mathbf{m},\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{m})^t \Sigma^{-1}(\mathbf{x}-\mathbf{m})\right]$$

Covariance matrix

Mean vector

N(m, Σ)

Example of a bivariate Gaussian pdf.



Gaussian Model

- The parameters defining completely the multivariate Gaussian pdf are the mean vector and the covariance matrix, which are defined as:

$$\mathbf{m} = E\{\mathbf{x}\}$$

$$\Sigma = Cov\{\mathbf{x}\} = E\{(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^t\} = E\{\mathbf{x}\mathbf{x}^\mathbf{t}\} - \mathbf{m}\mathbf{m}^\mathbf{t}$$

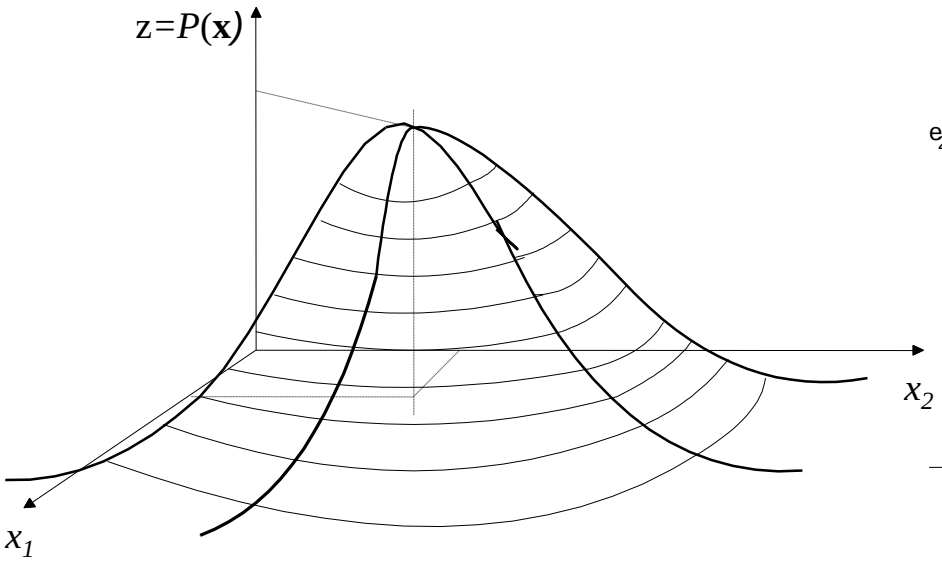
- **Properties of the covariance matrix**
  - Σ is symmetric: $\Sigma = \Sigma^t$.
  - Σ is positive semidefinite.
  - For independent features:

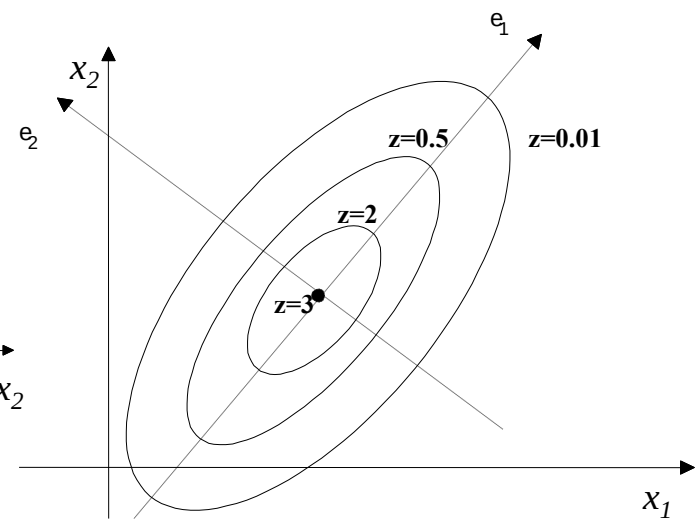
$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix} \Rightarrow \quad p(\mathbf{x}) = p(x_1)\, p(x_2) \dots p(x_n)$$

Gaussian Model

Shape of a 2-D Gaussian PDF

- $p(\mathbf{x})$ can be seen as a bell of unitary volume.
- The horizontal slices of the bell corresponding to isolevels are ellipses whose axes are directed by the eigenvectors Σ.
- The eigenvector corresponding to the largest eigenvalue defines the main axis of the ellipse.





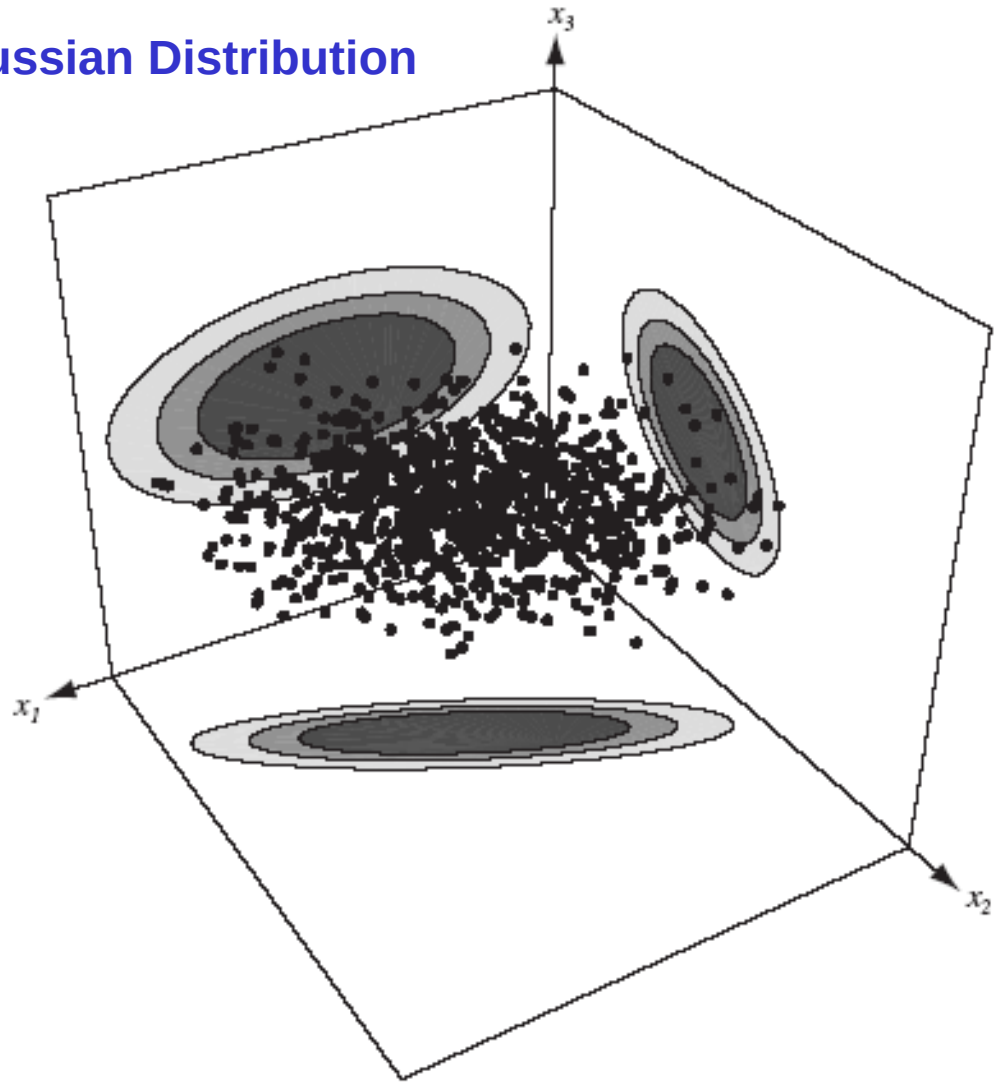
Gaussian Model

Shape of a n-D Gaussian PDF

- The previous observations done for the 2-D case can be generalized to the n-D one:
  - Let λ_1, λ_2, … and λ_n be the eigenvalues of Σ and $\mathbf{e}_1$, $\mathbf{e}_2$, …, $\mathbf{e}_n$ be the corresponding eigenvectors.
  - By convention, the eigenvalues and eigenvectors are ordered so that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$.
  - Since Σ is symmetric and positive semidefinite, the eigenvalues will take positive values.
  - The eigenvectors will form an orthogonal basis.
  - The isolevels of $p(\mathbf{x})$ are hyperellipses in $\Re^n$, whose axis directions are governed by the eigenvectors.
  - The first eigenvector will define the principal axis while the last one will determine the smallest axis.

Gaussian Model

Example of a 3-D Gaussian Distribution



Nota: In this example, the features x_1 and x_3 are independent!

Gaussian Model: ML Estimation

- It can be shown that the ML estimation of the mean vector and the covariance matrix of a multivariate Gaussian pdf, given a set of N iid training samples $\boldsymbol{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, leads to:

$$\hat{\mathbf{m}} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_k, \quad \hat{\Sigma} = \frac{1}{N}\sum_{k=1}^{N}(\mathbf{x}_k - \hat{\mathbf{m}})(\mathbf{x}_k - \hat{\mathbf{m}})^t = \frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_k\mathbf{x}_k^t - \hat{\mathbf{m}}\hat{\mathbf{m}}^t$$

- Such estimates are asymptotically unbiased and efficient, and consistent.
- In addition, we have: $E\{\hat{\mathbf{m}}\} = \mathbf{m}, \quad E\{\hat{\Sigma}\} = \frac{N-1}{N}\Sigma$
- An elementary unbiased estimator for Σ is:

$$\hat{\Sigma} = \frac{1}{N-1}\sum_{k=1}^{N}(\mathbf{x}_k - \hat{\mathbf{m}})(\mathbf{x}_k - \hat{\mathbf{m}})^t$$

Bayesian Estimation

- Bayesian estimation (termed also as Bayesian learning) is conceptually different with respect to ML estimation.

  - **In ML methods, we view the true parameter vector (θ) we seek to be fixed.**

  - **In Bayesian learning, we consider θ to be a random variable, and training data allows us to convert a distribution on this variable into a posterior probability density.**

Bayesian Estimation

- Bayesian estimation can be applied to any situation in which the unknown density can be parameterized.
- Its basic assumptions are as follows:

The form of the density $p(\mathbf{x}|\theta)$ is assumed to be known, but the value of the parameter vector θ is not known exactly.

Our initial knowledge about θ is assumed to be contained in a known a priori density $p(\theta)$.

The rest of our knowledge about θ is contained in a set X of N samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ drawn independently according to the unknown probability density $p(\mathbf{x})$.

Bayesian Estimation

- The third assumption can be converted into a posterior density $p(\boldsymbol{\theta}|\boldsymbol{X})$, which we hope, is sharply peaked about the true value of $\boldsymbol{\theta}$.
- The goal is to compute $p(\mathbf{x}|\boldsymbol{X})$, which is as close as possible to $p(\mathbf{x})$.
- This can be done by:

$$p(\mathbf{x} \mid \boldsymbol{X}) = \int p(\mathbf{x}, \theta \mid \boldsymbol{X}) \mathrm{d}\theta$$

The integration extends over the entire parameter space.

- By using Bayes theorem:

$$p(\mathbf{x} \mid \boldsymbol{X}) = \int p(\mathbf{x} \mid \theta, \boldsymbol{X}) p(\theta \mid \boldsymbol{X}) \mathrm{d}\theta$$

- Since the distribution of $\mathbf{x}$ is completely known once we know the value of the parameter vector $\boldsymbol{\theta}$:

$$p(\mathbf{x} \mid \boldsymbol{X}) = \int p(\mathbf{x} \mid \theta) p(\theta \mid \boldsymbol{X}) \mathrm{d}\theta$$

Bayesian Estimation

Since we do not know the exact value of θ, Bayesian estimation directs us to average $p(\mathbf{x}|\theta)$ over all possible values of θ.

The available observations X exert their influence on $p(\mathbf{x}|X)$ through the posterior probability density $p(\theta|X)$.

- Thus, the basic problem in Bayesian learning is to compute the posterior density $p(\boldsymbol{\theta}|\boldsymbol{X})$.
- For such purpose, one can make use of the following two relationships:
  - From Bayes formula:

$$p(\theta \mid \boldsymbol{X}) = \frac{p(\boldsymbol{X} \mid \theta)\, p(\theta)}{\int p(\boldsymbol{X} \mid \theta)\, p(\theta)\, \mathrm{d}\theta}$$

  - From independence assumption:

$$p(\boldsymbol{X} \mid \theta) = \prod_{k=1}^{N} p(\mathbf{x}_k \mid \theta)$$

Bayesian Estimation: Univariate Gaussian Case

- In the following, let us consider a simple example of computation of the posterior density $p(\boldsymbol{\theta}|\boldsymbol{X})$ and the desired probability density $p(\mathbf{x}|\boldsymbol{X})$ by assuming that:
  - $p(\mathbf{x}|\boldsymbol{\theta})$ is a univariate Gaussian distribution;
  - In $\boldsymbol{\theta}=[\mu,\sigma^2]$, the only unknown parameter is μ.
- Accordingly, $p(\mathbf{x}|\mu) \sim \mathrm{N}(\mu,\sigma^2)$
- We shall make the further assumption that $p(\mu) \sim \mathrm{N}(\mu_0, \sigma_0^2)$ whose parameters are a priori known.
- Using formula previously seen,

$$p(\mu \mid \boldsymbol{X}) = \alpha \prod_{k=1}^{N} p(\mathbf{x}_k \mid \mu)\, p(\mu)$$

Normalization factor

Bayesian Estimation: Univariate Gaussian Case

- Since $p(\mathbf{x}_k|\mu) \sim N(\mu,\sigma^2)$ and $p(\mu) \sim N(\mu_0,\sigma_0^2)$, we can obtain:

$$p(\mu \mid X) = \alpha'' \exp\left\{-\frac{1}{2}\left[\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{N}\mathbf{x}_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right\}$$

Constant

$p(\mu|X)$ is again a normal density: $p(\mu|X) \sim N(\mu_N,\sigma_N^2)$.

- After some calculations, one can get:

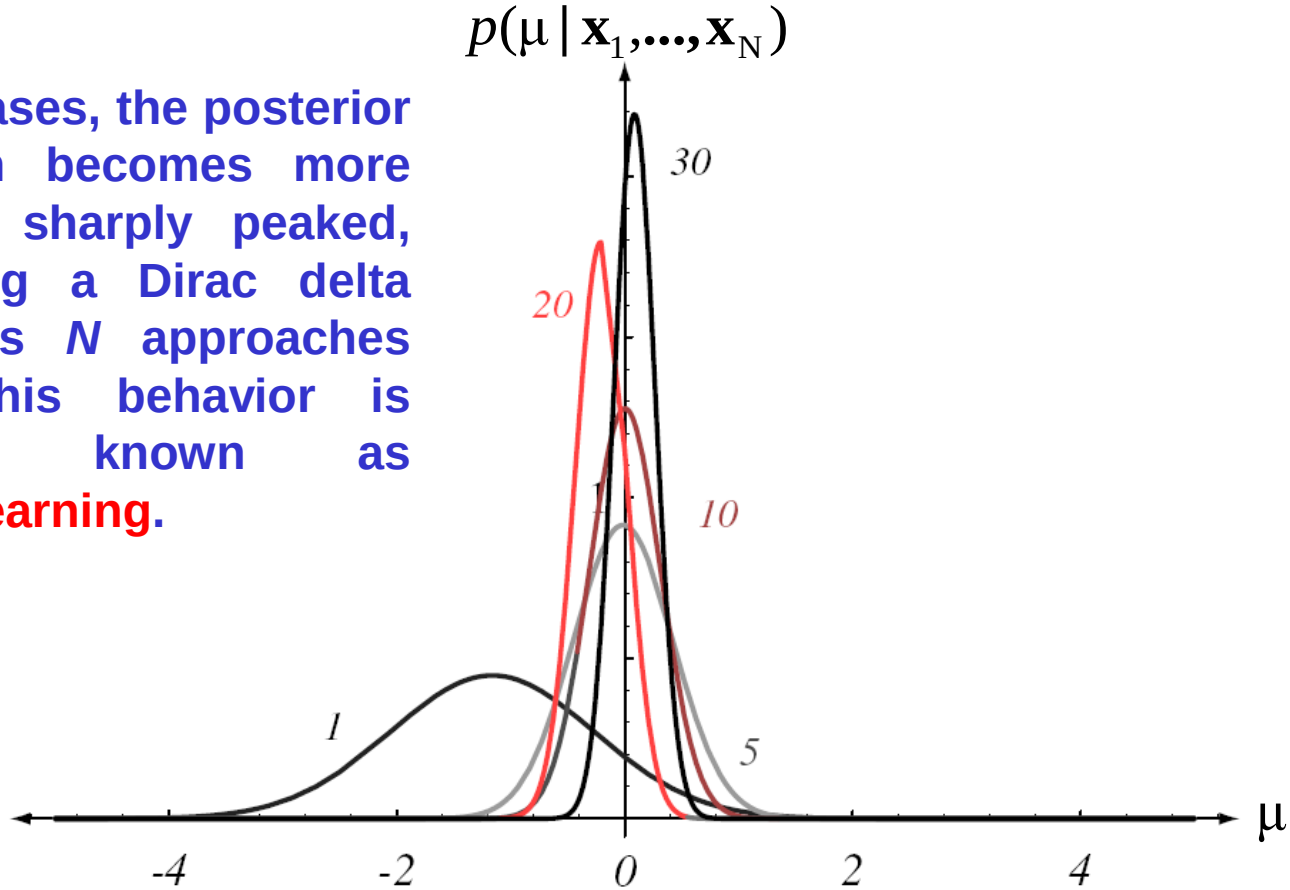
$$\begin{cases} \mu_N = \left(\dfrac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\right)\overline{X} + \left(\dfrac{\sigma^2}{N\sigma_0^2+\sigma^2}\right)\mu_0 \\ \sigma_N^2 = \dfrac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2} \end{cases}$$

Sample mean

Nota: These equations show how the prior information is combined with the empirical information in the samples to obtain the a posteriori density.

Bayesian Estimation: Univariate Gaussian Case

As *N* increases, the posterior distribution becomes more and more sharply peaked, approaching a Dirac delta function as *N* approaches infinity. This behavior is commonly known as **Bayesian learning**.



Bayesian Estimation: Univariate Gaussian Case

- Having obtained the a posteriori density for the mean $p(\mu|\boldsymbol{X})$, all that remains is to obtain the desired density $p(\mathbf{x}|\boldsymbol{X})$:

$$p(\mathbf{x}\,|\,\boldsymbol{X}) = \int p(\mathbf{x}\,|\,\mu)\,p(\mu\,|\,\boldsymbol{X})\mathrm{d}\mu$$

$$= \int \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{\mathbf{x}-\mu}{\sigma}\right)^2\right]\frac{1}{\sqrt{2\pi}\sigma_N}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_N}{\sigma_N}\right)^2\right]\mathrm{d}\mu$$

$$= \frac{1}{2\pi\,\sigma\sigma_N}\exp\left[-\frac{1}{2}\frac{(\mathbf{x}-\mu_N)^2}{\sigma^2+\sigma_N^2}\right]f(\sigma,\sigma_N)$$

This means that $p(x|X) \sim N(\mu_N, \sigma^2 + \sigma_N^2)$ and thus $\mu = \mu_N$ since $p(x|\mu) \sim N(\mu, \sigma^2)$.

ML versus Bayesian Estimation

- The ML approach estimates a point in $\boldsymbol{\theta}$ space while the Bayesian approach estimates a distribution $p(\mathbf{x}|\boldsymbol{X})$ from which $\boldsymbol{\theta}$ is inferred.
- ML and Bayes solutions are equivalent:
  - **in the asymptotic limit of infinite training samples**
  - **or if the prior $p(\boldsymbol{\theta})$ is uniform**
- In practice, ML is often preferred over Bayesian estimation because of:
  - **Lower computational complexity**
  - **Easier interpretability**

Nonparametric Estimation

- Nonparametric estimation becomes necessary when:
  - There is no prior knowledge about the functional form of the pdf characterizing the observed phenomenon;
  - Parametric models do not offer a good approximation of the considered pdf.
- Nonparametric estimation approaches are thus applied directly on the available observations (training samples).

Nonparametric Estimation

- In the following, we will focus our attention on two different popular nonparametric estimation methods:
  - K-nearest neighbor (K-NN) method
  - Parzen windows method

Basic Concepts

- Let $\mathbf{x}^*$ be a generic sample and R a predefined region of the feature space such that $\mathbf{x}^* \in R$.
- Assuming the true pdf $p(\mathbf{x})$ is a continuous function and R is sufficiently small so that $p(\mathbf{x})$ does not vary significantly within it, we can write:

$$P_R = P\{\mathbf{x} \in R\} = \int_R p(\mathbf{x})d\mathbf{x} = p(\mathbf{x}^*)V$$

Volume of R

- Let K be the number of training samples belonging to R (among a total of N training samples).
- A consistent estimate of P_R can be achieved through the computation of the relative frequency:

$$\hat{P}_R = \frac{K}{N}, \quad \lim_{N \to +\infty} P\{|\hat{P}_R - P_R| < \delta\} = 1 \quad \forall \delta > 0$$

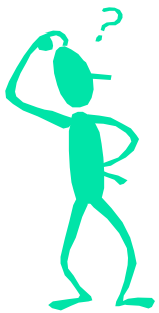
Law of large numbers

Basic Concepts

- PDF Estimation
  - From the estimate of the probability P_R that a sample belongs to R, it can be derived an estimate of the pdf at $\mathbf{x}^*$:

$$\hat{p}(\mathbf{x}^*) = \frac{\hat{P}_R}{V} = \frac{K}{NV}$$

- Observations
  - R should be enough large to contain a number of training samples that is sufficient for applying the law of large numbers.



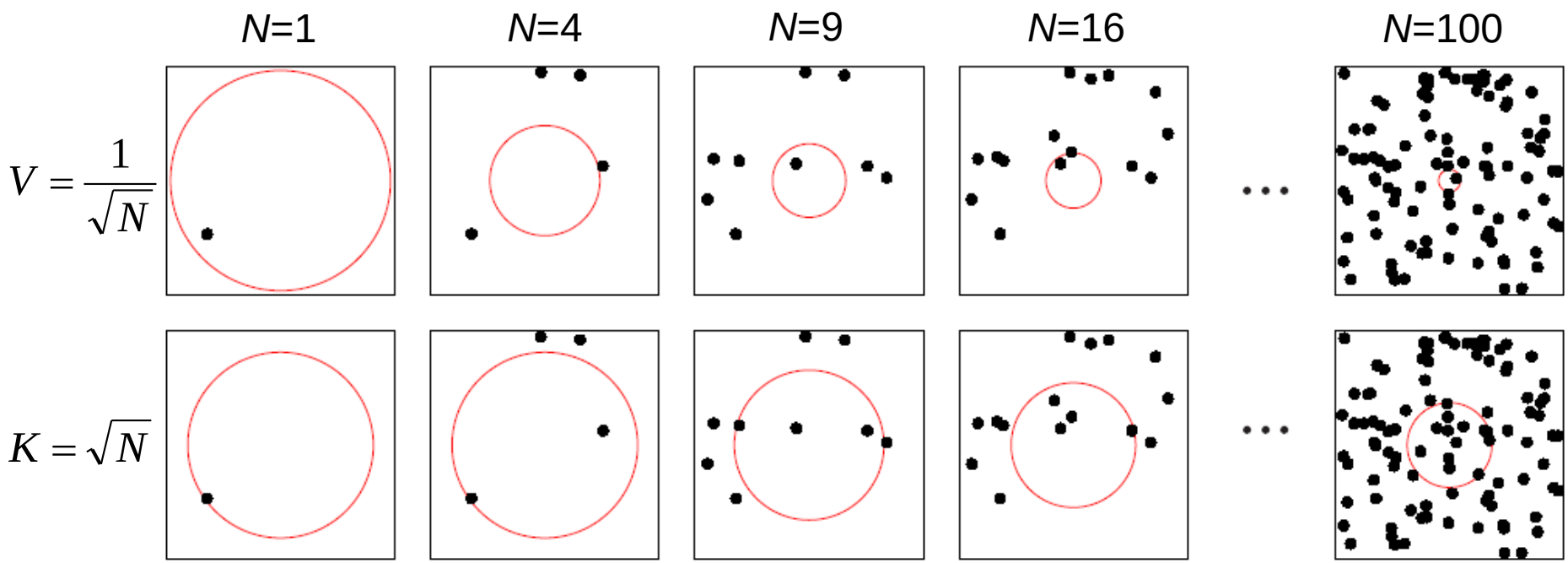
  - At the same time, it should be enough small to limit the variability of $p(\mathbf{x})$ within it.

Basic Concepts

- Depending on the role taken by each of the two parameters K e V, one may lead to two different nonparametric estimation methods:
  - K-nearest neighbor method: K is fixed and the hypervolume V of R is computed on the basis of the training set to deduce the estimate of the pdf;
  - Parzen method: R (and thus V) is fixed and K is calculated (from the training samples) to determine the pdf estimate.
- Both of these methods do in fact converge, although it is difficult to make meaningful statements about their finite-sample behavior.

Basic Concepts

Illustration of Parzen versus K-NN estimation concepts



K-NN Estimation

- **Hypotheses**
  - The number K is set a priori.
  - A shape for the cell of volume centered on $\mathbf{x}^*$ is chosen a priori (e.g., hypersphere).

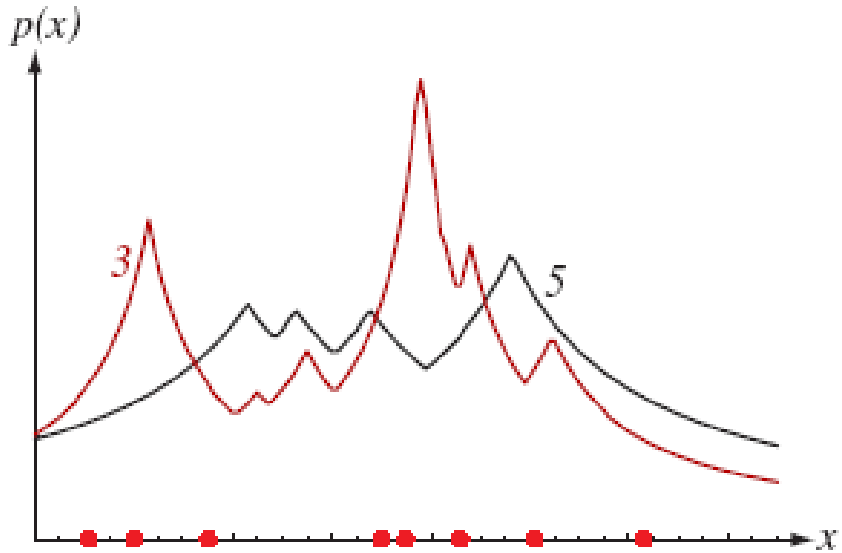
- **Methodology**
  - The K-NN method consists to expand the cell up to spanning K training samples.
  - Let $V_K(\mathbf{x}^*)$ be the resulting volume.
  - The pdf value at $\mathbf{x}^*$ is given by: $\hat{p}(\mathbf{x}^*) = \dfrac{K}{NV_K(\mathbf{x}^*)}$
  - It can be shown that, if K is chosen as a function of N, a necessary and sufficient condition to get a consistent estimate in all points, where $p(\mathbf{x})$ is continuous, is given by:

$$\lim_{N\to+\infty} K_N = +\infty, \qquad \lim_{N\to+\infty} \frac{K_N}{N} = 0$$

$$\text{Example}: \quad K(N) = \sqrt{N}$$

K-NN Estimation

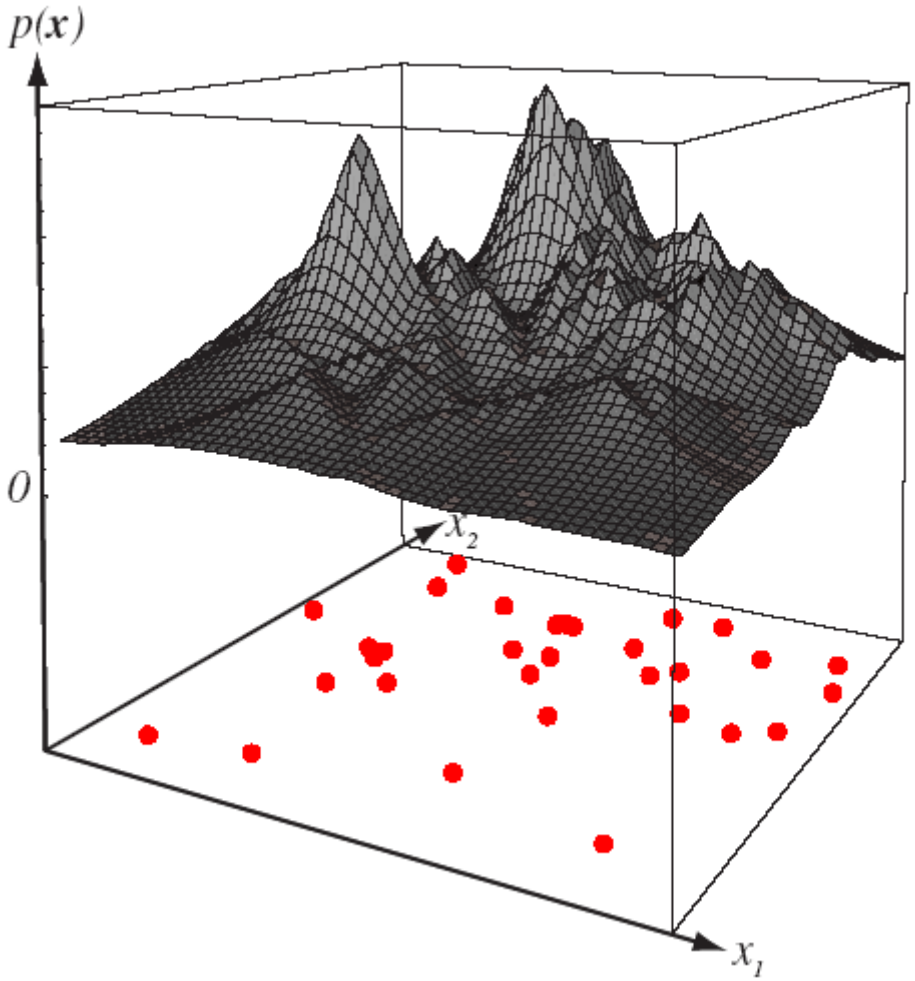
Example of K-NN estimation in a 1-D feature space (K=3 and 5)



Nota: The discontinuities in the slopes in the estimates generally occur away from the positions of the points themselves.

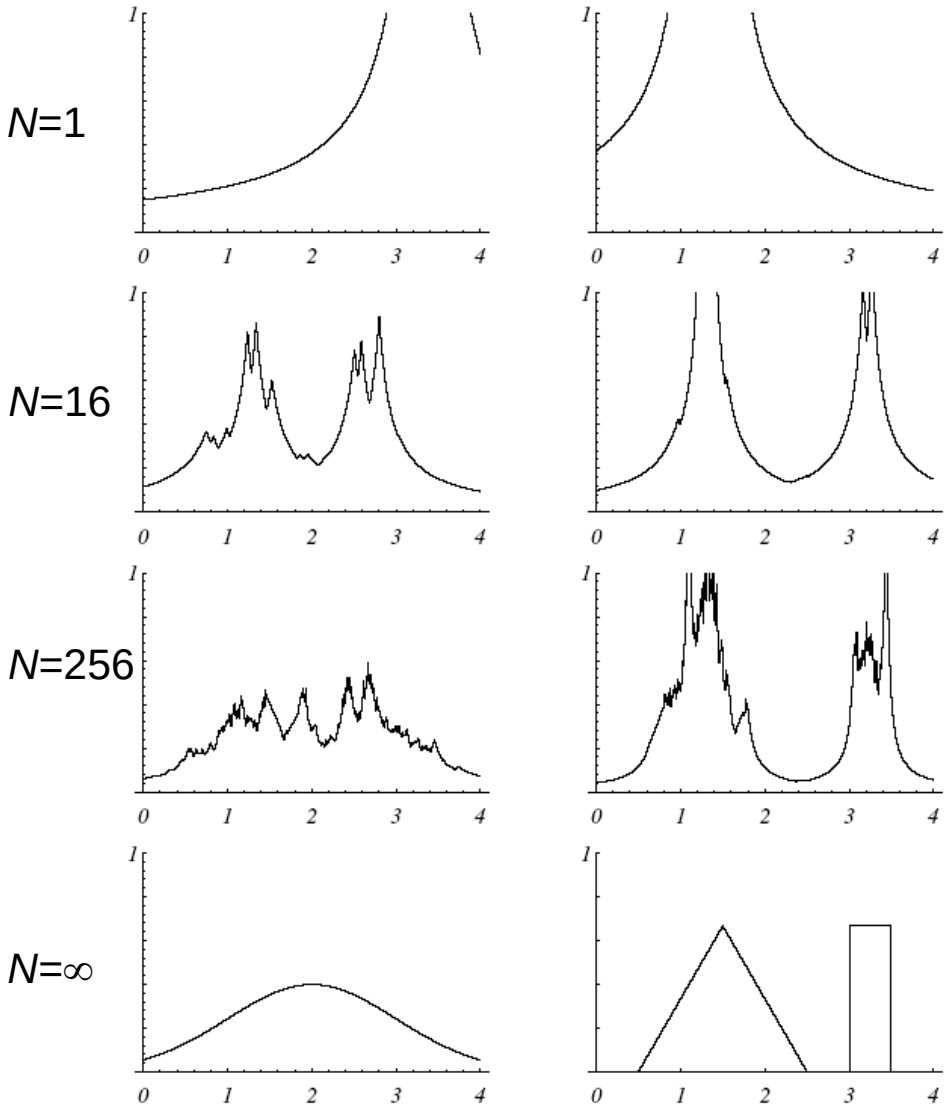
K-NN Estimation

Example of K-NN estimation in a 2-D feature space (K=5)



K-NN Estimation

Several K-NN estimates of two 1-D densities: a Gaussian and a bimodal distribution.



Parzen Windows Estimation

- **Hypotheses**
  - The Parzen window approach to estimating densities can be introduced by temporarily assuming that R is a n-dimensional hypercube.
  - If h is the length of an edge of that hypercube centered on $\mathbf{x}^*$, then its volume is $V = h^n$.

- **Methodology**
  - We can obtain an analytic expression for K, the number of samples falling in the hypercube, by defining the following window function:

$$\gamma(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} \text{ belongs to the hypercube} \\ 0, & \text{otherwise} \end{cases}$$

  - The training sample $\mathbf{x}_k$ belongs to R (with center $\mathbf{x}^*$ and edge length h) if $\gamma[(\mathbf{x}_k - \mathbf{x}^*)/h] = 1$. Otherwise, $\gamma[(\mathbf{x}_k - \mathbf{x}^*)/h] = 0$.

Parzen Windows Estimation

- The number of samples in this hypercube is therefore given by:

$$K = \sum_{k=1}^{N} \gamma\left(\frac{\mathbf{x}_k - \mathbf{x}^*}{h}\right)$$

which leads to the following estimate:

$$\hat{p}(\mathbf{x}^*) = \frac{K}{NV} = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{h^n}\gamma\left(\frac{\mathbf{x}_k - \mathbf{x}^*}{h}\right) = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{h^n}\gamma\left(\frac{\mathbf{x}^* - \mathbf{x}_k}{h}\right)$$

- Such estimate is thus as a collection of contributions coming from rectangular functions, each associated with a single training sample.
- It is also equivalent to a simple count of the number of training samples falling in the predefined hypercube.
- This equation suggests a more general approach to estimating density functions.

Rather than limiting ourselves to the hypercube window function, suppose we allow a more general class of window functions.

Parzen Windows Estimation

- In such a case, our estimate for $p(\mathbf{x})$ becomes an average of functions of $\mathbf{x}$ and the samples $\mathbf{x}_i$:

$$\hat{p}(\mathbf{x}) = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{V(h)}\gamma\left(\frac{\mathbf{x}-\mathbf{x}_k}{h}\right)$$

- In essence, the window function is being used for interpolation - each sample contributing to the estimate in accordance with its distance from $\mathbf{x}$.
- The function $\gamma(\cdot)$ is called Parzen window or kernel and the h parameter is the width of the window (kernel).
- It is natural to ask that the estimate $p(\mathbf{x})$ be a legitimate density function, i.e., that it be nonnegative and integrate to one.

Parzen Windows Estimation

- This can be assured by requiring the window function itself be a density function. To be more precise, if we require that

$$\gamma(x) \geq 0 \quad \forall \mathrm{x} \in \Re^{\mathrm{n}}, \qquad \int_{\Re^n} \gamma(x)dx = 1$$

  and if we maintain the relation $V = h^n$, then it follows at once that $p(\mathbf{x})$ also satisfies these conditions.

- Additional conditions to get a "good" estimate are:
  - $\gamma(\cdot)$ takes a maximal value at the origin;
  - $\gamma(\cdot)$ is continuous;
  - $\gamma(\mathbf{x}) \rightarrow 0$ as $\mathbf{x} \rightarrow +\infty$.

Window Width Effect

- Let us examine the effect that the window width h has on $p(\mathbf{x})$.
- If we define the function $\delta(\mathbf{x})$ by:

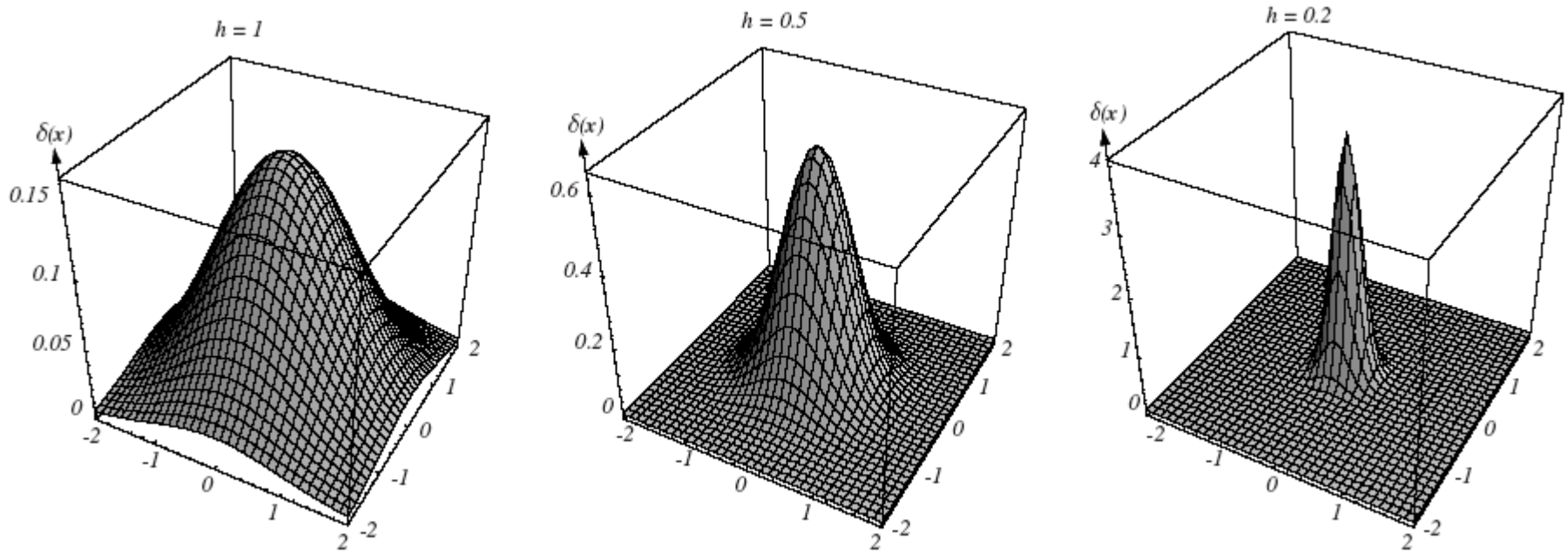
$$\delta(\mathbf{x}) = \frac{1}{V(h)} \gamma\left(\frac{\mathbf{x}}{h}\right)$$

- Then, we can write the estimate for $p(\mathbf{x})$ as the average:

$$\hat{p}(\mathbf{x}) = \frac{1}{N} \sum_{k=1}^{N} \delta(\mathbf{x} - \mathbf{x}_k)$$

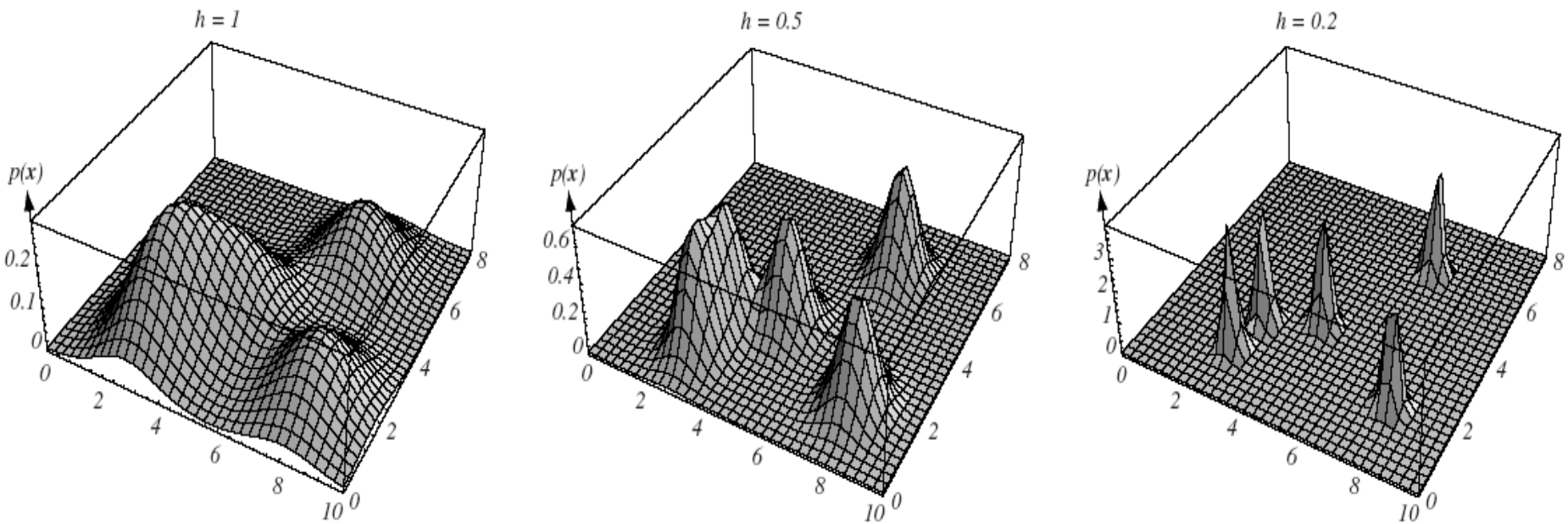
Window Width Effect

Examples of 2-D circularly symmetric normal Parzen windows for three different values of h.



Window Width Effect

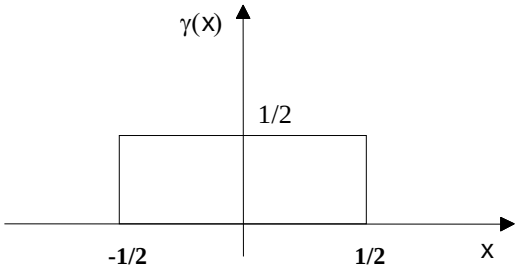
Three Parzen window density estimates based on the same set of five samples.



Kernel Examples

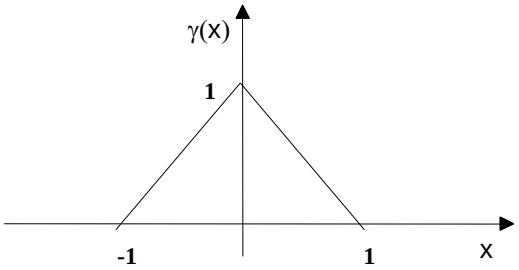
- ***Rectangular kernel***:

$$\gamma(x) = \Pi(x)$$



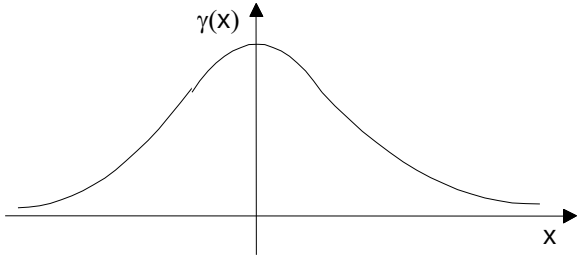
- ***Triangular kernel***:

$$\gamma(x) = \Lambda(x)$$



- ***Gaussian kernel***:

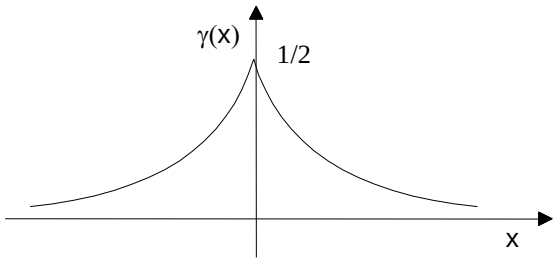
$$\gamma(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$



Kernel Examples

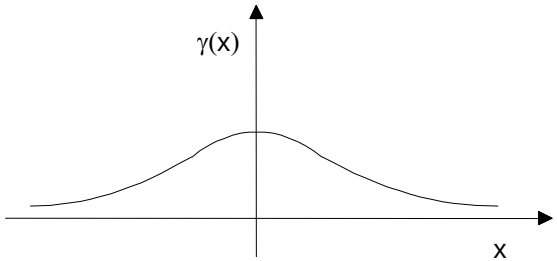
- ***Exponential kernel***:

$$\gamma(x) = \tfrac{1}{2}\exp(-|x|)$$



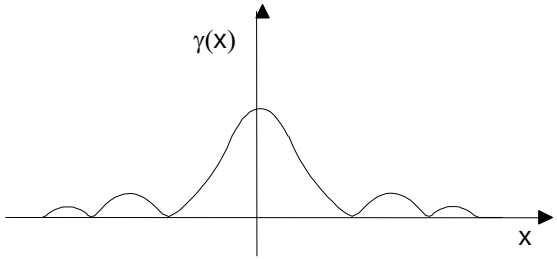
- ***Cauchy kernel***:

$$\gamma(x) = \frac{1}{\pi}\cdot\frac{1}{1+x^2}$$



- ***"sinc²(·)" kernel*** :

$$\gamma(x) = \frac{1}{2\pi}\left(\frac{\sin(x/2)}{x/2}\right)^2$$



Parzen Windows Estimation: Properties

- **Bias**
  - Since the samples $\mathbf{x}_i$ are i.i.d. according to the (unknown) density $p(\mathbf{x})$, we have:

$$E\{\hat{p}(\mathbf{x})\} = p(\mathbf{x}) * \frac{1}{h^n}\gamma\left(\frac{\mathbf{x}}{h}\right)$$

The estimate mean is a blurred version of the true $p(\mathbf{x})$!

  - Parzen estimate is thus biased. However, it can be shown that it is asymptotically unbiased if the kernel width is chosen opportunely, i.e., $h_N \to 0$ as $N \to +\infty$:

$$\lim_{N\to+\infty}\left\{\frac{1}{h_N^n}\gamma\left(\frac{\mathbf{x}}{h_N}\right)\right\} = \delta(\mathbf{x}) \quad \Rightarrow \quad \lim_{N\to+\infty} E\{\hat{p}(\mathbf{x})\} = p(\mathbf{x})$$

Parzen Windows Estimation: Properties

- **Variance**
  - Since the observed distribution is the sum of functions of statistically independent random variables, it can be shown that the estimate variance is bounded by:

$$E\left\{(\hat{p}(\mathbf{x})-\bar{\hat{p}}(\mathbf{x}))^2\right\} \leq \frac{\sup(\gamma(\cdot))\bar{\hat{p}}(\mathbf{x})}{Nh_N^n}$$

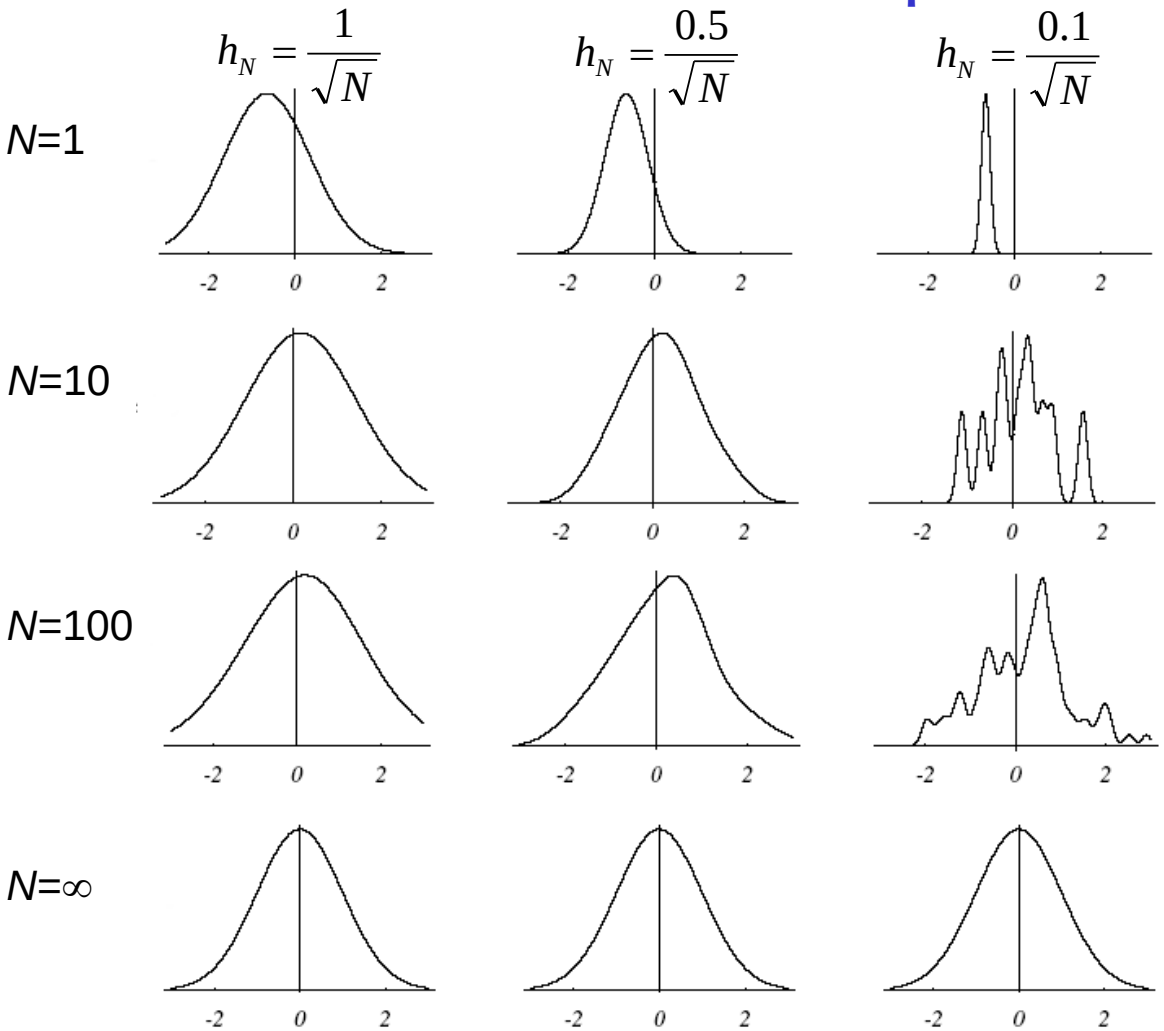
  - It can be demonstrated that the Parzen estimate is consistent if:

$$\lim_{N\to+\infty} h_N = 0, \quad \lim_{N\to+\infty} Nh_N^n = +\infty$$

Example: $h_N = \frac{1}{\sqrt[2n]{N}}$

Parzen Windows Estimation: Examples

1-D normal density using different window widths and numbers of samples.



Parzen Windows Estimation: Examples

1-D bimodal distribution using different window widths and numbers of samples.

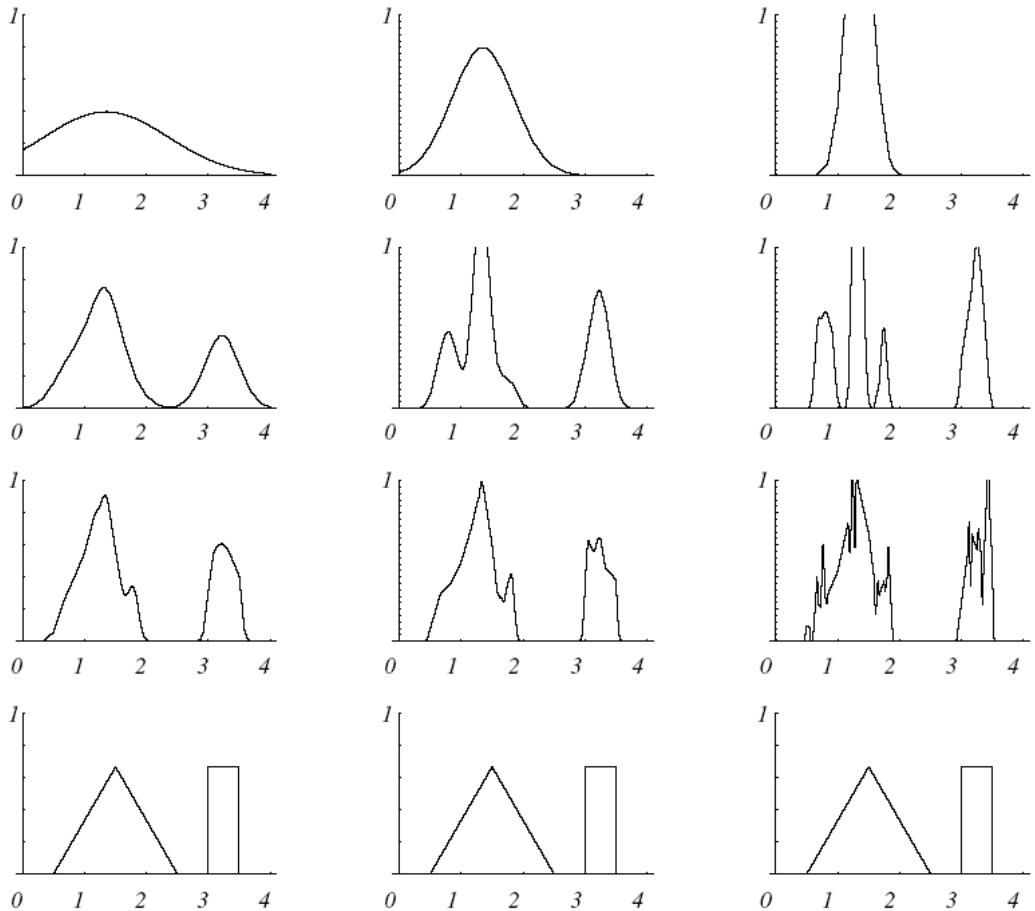
$h_N = \frac{1}{\sqrt{N}}$ $\quad$ $h_N = \frac{0.5}{\sqrt{N}}$ $\quad$ $h_N = \frac{0.1}{\sqrt{N}}$

N=1

N=16

N=256

N=∞



Parzen Windows Estimation: Examples

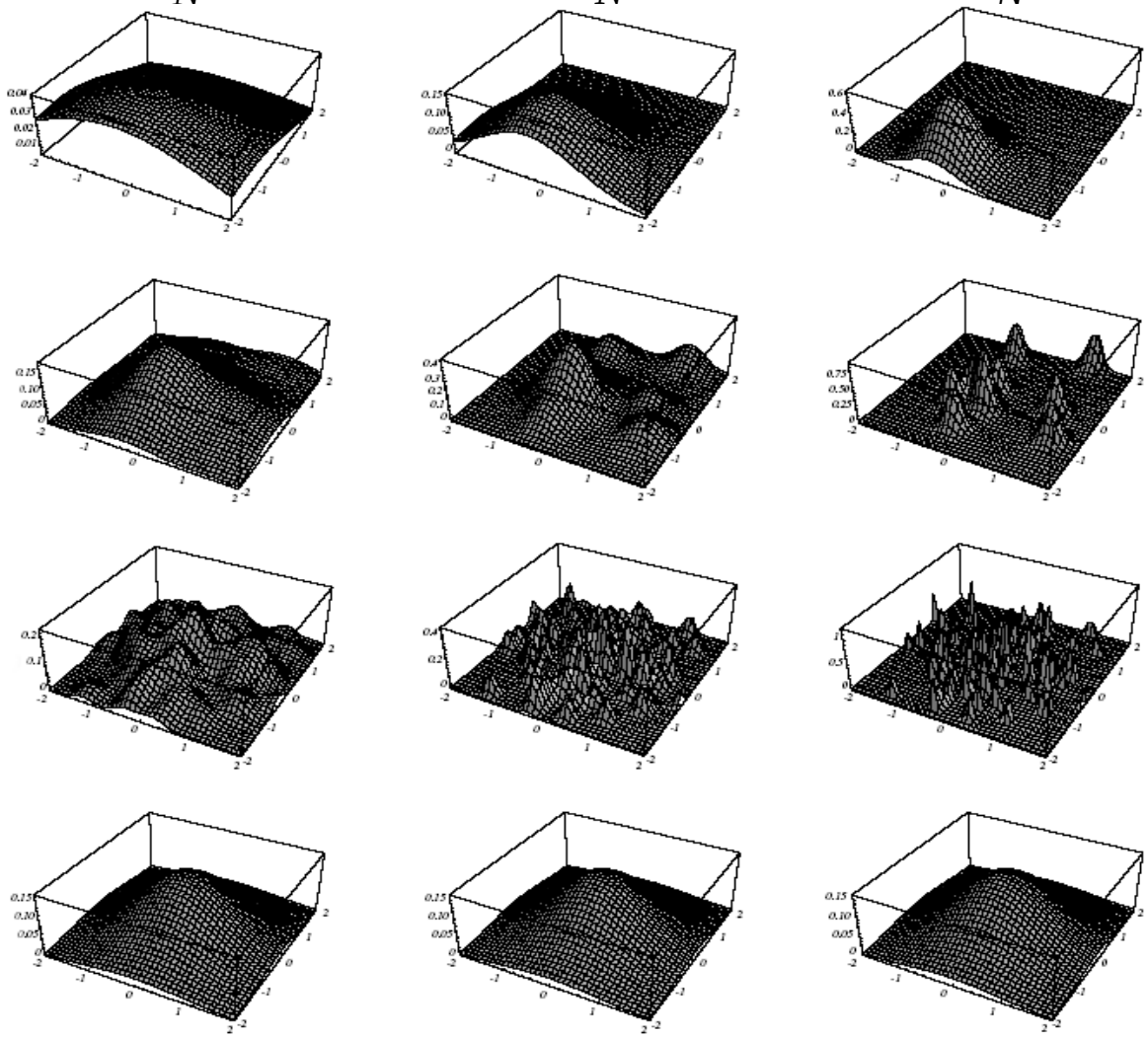
$h_N = \frac{1}{N}$ $\quad$ $h_N = \frac{0.5}{N}$ $\quad$ $h_N = \frac{0.1}{N}$

N=1

N=10

N=100

N=∞



2-D normal distribution using different window widths and numbers of samples.

Gaussian Kernel

- A commonly used kernel in the Parzen windows estimation is the Gaussian kernel with spherical symmetry (Specht method).
- In this case, the pdf can be expressed as:

$$\hat{p}(\mathbf{x}) = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_k\|^2}{2\sigma^2}\right)$$

- The standard deviation σ of the kernel represents thus a smoothing parameter whose setting should be made carefully to avoid both over- and under-fitting.
- Since σ is the same along all features, it is important to normalize them before undertaking the pdf estimation process.
- Furthermore, one may think to compute σ in an adaptive way, i.e. by associating a σ_k to each training sample $\mathbf{x}_k$ ($k = 1, 2, ..., N$).

Gaussian Kernel

- σ_k could be defined as the mean Euclidean distance between $\mathbf{x}_k$ and the related L nearest training samples $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_L$:

$$\sigma_k = \frac{1}{L}\sum_{l=1}^{L}\|\mathbf{x}_k - \mathbf{y}_l\|$$

- The Specht method can be implemented by means of Probabilistic Neural Networks (PNN).

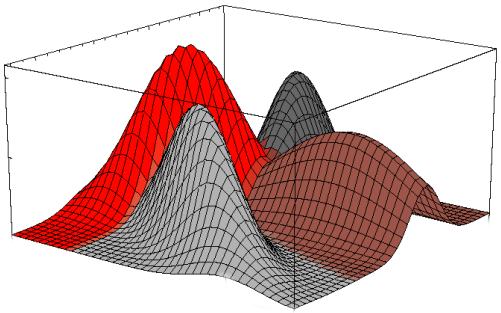
Estimation with Incomplete Data

- Up to now, we have considered estimation problems where the model depends completely on observable variables.
- In some real cases, the model depends on unobserved latent variables.
- Typically, for problems with incomplete data, the estimation approach that is used is the parametric one.
- A solution to this class of estimation problems is the Expectation-Maximization (EM) algorithm.
- In the following, we will illustrate a typical example of estimation with incomplete data.

Estimation with Incomplete Data

- Let us assume to have N i.i.d. observations $\mathbf{x_i}$ ($i = 1, 2, \ldots, N$) defined in a n-dimensional space X.
- Let us assume that these observations are drawn from a distribution $p(\mathbf{x})$ defined as a mixture of M Gaussian modes:

$$p(\mathbf{x}) = \sum_{i=1}^{M} P_i \, p(\mathbf{x} \mid \mathbf{m}_i, \Sigma_i)$$

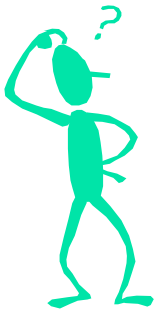


where P_i, $\mathbf{m_i}$ and Σ_i are the prior probability, the mean vector and the covariance matrix of the i-th mixture component, respectively.

- Let $\boldsymbol{\theta} = [P_1, P_2, \ldots, P_M; \mathbf{m_1}, \mathbf{m_2}, \ldots, \mathbf{m_M}; \Sigma_1, \Sigma_2, \ldots, \Sigma_M]$ be the vector of parameters to be estimated.

Estimation with Incomplete Data

- **Objective**: to estimate the distribution of each Gaussian mode of $p(\mathbf{x})$.
- Our estimation problem is thus with incomplete data since:



  - we have just the observations associated with the whole $p(\mathbf{x})$ and not those of the single modes;
  - we do not know to which mode is associated each sample (i.e., which are the samples to use for the estimation of each mode?);
  - we just know that there is a relationship of the summation type between the different (hidden) modes.

Problem Formulation

- We have seen that the ML estimation problem is given by:

$$\boldsymbol{\theta}^* = \arg\max\ l(\boldsymbol{\theta}|\mathbf{X})$$

where

$$l(\boldsymbol{\theta}|\mathbf{X}) = p(\mathbf{X} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathbf{x}_i \mid \boldsymbol{\theta})$$

- Depending on the form of $p(\mathbf{X}|\theta)$, the ML estimation problem can be easy or hard.
- For many problems, it is not possible to find an analytical solution and one must resort to more elaborate techniques.
- The EM algorithm is one of them.

Problem Formulation

- Let's make the following assumptions.
- **X** is observed and generated by some distribution (incomplete dataset).
- Complete dataset exists: **Z** = (**X**, **Y**) where **Y** represents missing data
- The joint density function is given by:

$$p(\mathbf{Z}\,|\boldsymbol{\theta}) = p(\mathbf{X},\mathbf{Y}\,|\boldsymbol{\theta}) = p(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta})\cdot p(\mathbf{X}\,|\boldsymbol{\theta})$$

- The complete-data likelihood function becomes:

$$l(\boldsymbol{\theta}|\mathbf{Z}) = p(\mathbf{Z}\mid\boldsymbol{\theta}) = \mathrm{p}(\mathbf{X},\mathbf{Y}\mid\boldsymbol{\theta})$$

- While the incomplete-data likelihood is:

$$l(\boldsymbol{\theta}|\mathbf{X}) = p(\mathbf{X}\mid\boldsymbol{\theta})$$

EM Algorithm

- The EM algorithm finds the expected value of *log p*(**X**,**Y**|θ)
  - w.r.t. **y**
  - given **X** and a current estimate of θ
- That is, we define the so-called conditional expectation:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = E\{ log(p(\mathbf{X}, \mathbf{Y} \mid \boldsymbol{\theta})) \mid \mathbf{X}, \boldsymbol{\theta}^{(k)} \}$$

$$\Downarrow$$

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = \int log(p(\mathbf{X}, \mathbf{y} \mid \boldsymbol{\theta})) p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}^{(k)}) d\mathbf{y}$$

EM Algorithm

EM alternates between performing an expectation (E) step, which computes an expectation of the likelihood by including the latent variables as if they were observed, and a maximization (M) step, which computes the maximum likelihood estimates of the parameters by maximizing the expected likelihood found on the E step.

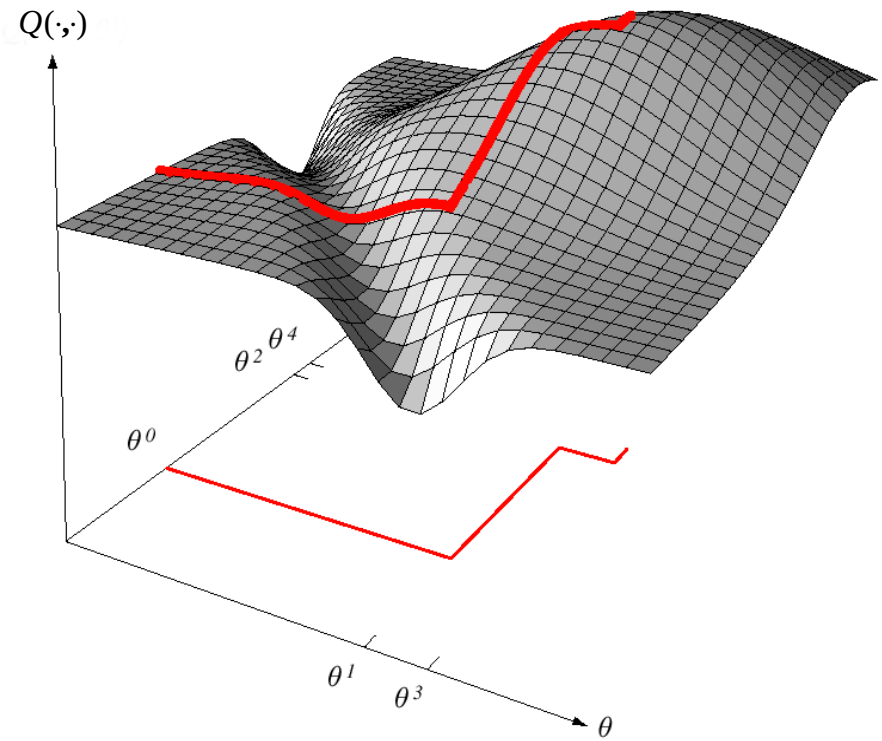
The parameters found on the M step are then used to begin another E step, and the process is repeated.

- EM can thus be formalized into the following two steps:
  - E-step: $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = E\{log(p(\mathbf{X}, \mathbf{Y} \mid \boldsymbol{\theta})) \mid \mathbf{X}, \boldsymbol{\theta}^{(k)}\}$
  - M-step: $\boldsymbol{\theta}^{(k+1)} = \arg\max Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$

Convergence of EM Algorithm

- It can be shown that the sequence of estimates $\{\theta^{[k]}\}$ generated by the EM algorithm allows increasing at each iteration the value of the log-likelihood function $l(\theta)$, i.e.:

$$l(\boldsymbol{\theta}^{(i+1)}) \geq l(\boldsymbol{\theta}^{(i)})$$



Convergence of EM Algorithm

- It is not guaranteed that the EM converges to a global maximum of $L(\theta)$ (i.e., the ML estimate of θ).
- Even though it may suffer from local minima and saddle points, the EM algorithm has proved to be particularly effective in the estimation of mixture components parameters.

EM Algorithm: Gaussian Mixture

- Turning back to the case of a mixture of Gaussian modes, it can be shown that the EM equations for estimating the parameters of each mode are:

$$P_i^{[k+1]} = \frac{1}{N}\sum_{j=1}^{N}\frac{P_j^{[k]}\, p(\mathbf{x}_j \mid \mathbf{m}_\mathbf{i}^{[k]}, \Sigma_i^{[k]})}{p(\mathbf{x}_j \mid \boldsymbol{\theta}^{[k]})}$$

$$\mathbf{m}_\mathbf{i}^{[k+1]} = \frac{\sum_{j=1}^{N}\frac{P_i^{[k]}\, p(\mathbf{x_j} \mid \mathbf{m}_\mathbf{i}^{[k]}, \Sigma_i^{[k]})}{p(\mathbf{x_j} \mid \boldsymbol{\theta}^{[k]})}\cdot \mathbf{x_j}}{\sum_{j=1}^{N}\frac{P_i^{[k]}\, p(\mathbf{x_j} \mid \mathbf{m}_\mathbf{i}^{[k]}, \Sigma_i^{[k]})}{p(\mathbf{x_j} \mid \boldsymbol{\theta}^{[k]})}}$$

$$\Sigma_i^{[k+1]} = \frac{\sum_{j=1}^{N}\frac{P_i^{[k]}\, p(\mathbf{x_j} \mid \mathbf{m}_\mathbf{i}^{[k+1]}, \Sigma_i^{[k]})}{p(\mathbf{x_j} \mid \boldsymbol{\theta}^{[k]})}\cdot(\mathbf{x_j} - \mathbf{m}_\mathbf{i}^{[k+1]})\cdot(\mathbf{x_j} - \mathbf{m}_\mathbf{i}^{[k+1]})^t}{\sum_{j=1}^{N}\frac{P_i^{[k]}\, p(\mathbf{x_j} \mid \mathbf{m}_\mathbf{i}^{[k+1]}, \Sigma_i^{[k]})}{p(\mathbf{x_j} \mid \boldsymbol{\theta}^{[k]})}}$$

EM Algorithm: Example

Example with four data points, one of which is missing the value of the x_1 component.