

01. select the features.
 2. feature selection and extraction
 3. select the classifier to be applied.
-

Pattern Classification

05. Density Estimation

after this lecture, we can build a bayes classifier.
8aleban da el lab el gay.

AbdElMoniem Bayoumi, PhD

Fall 2021

Recap: Gaussian Densities

- Assume a multi dimensional Gaussian density for each $P(\underline{X}|C_i)$
- Features may be independent (or conditionally independent), i.e., independent Gaussians
- Features may be dependent in other cases

Recap: Applying Bayes Rule

- One way on how to apply Bayes rule in practical situations:
 - Obtain the training set $\underline{X}(1), \underline{X}(2) \dots \underline{X}(M)$
 - Assume a multi-dimensional Gaussian density for each class, i.e., $P(\underline{X}|C_i)$
 - To obtain the form of each density we need $\underline{\mu}_i$ and Σ_i for each class $i \rightarrow$ estimate from training set
 - Estimate the a priori probabilities $P(C_i)$ from the training set, i.e., according to the frequencies of each class
 - Using the obtained estimates, plug in Bayes rule to obtain the classification rule

Density Estimation

- In Bayes rule, the probability densities have to be estimated
- One way is to assume that they are multivariate Gaussian and estimate μ & Σ of these distributions
- Estimate the densities from data

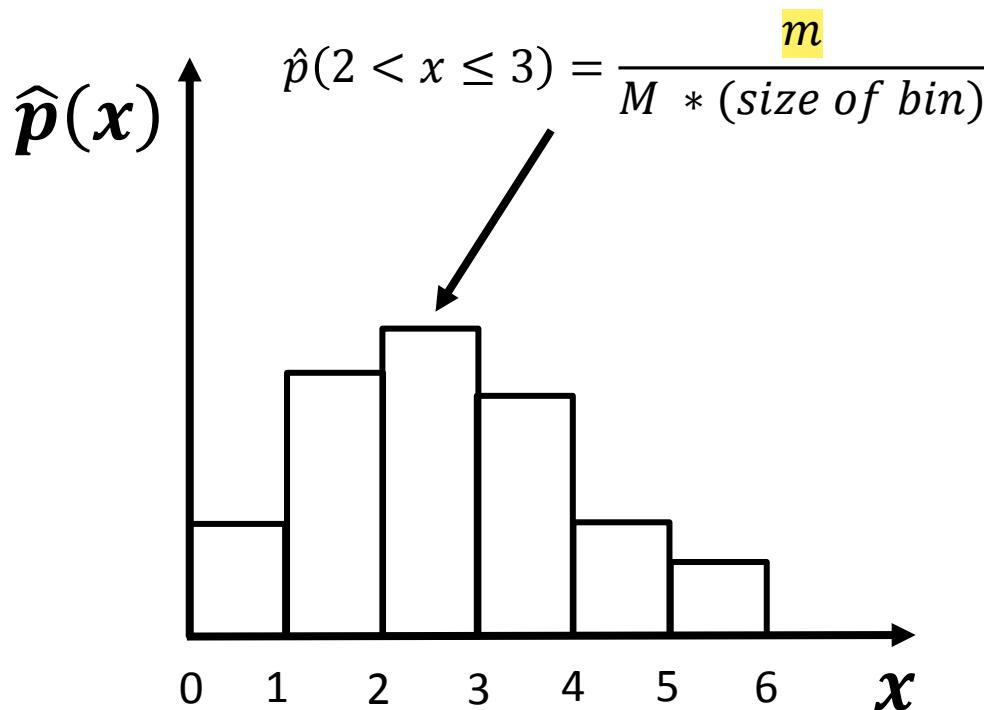
Histogram Analysis

$$\hat{p}(x) = \frac{m}{M * (\text{size of bin})}$$

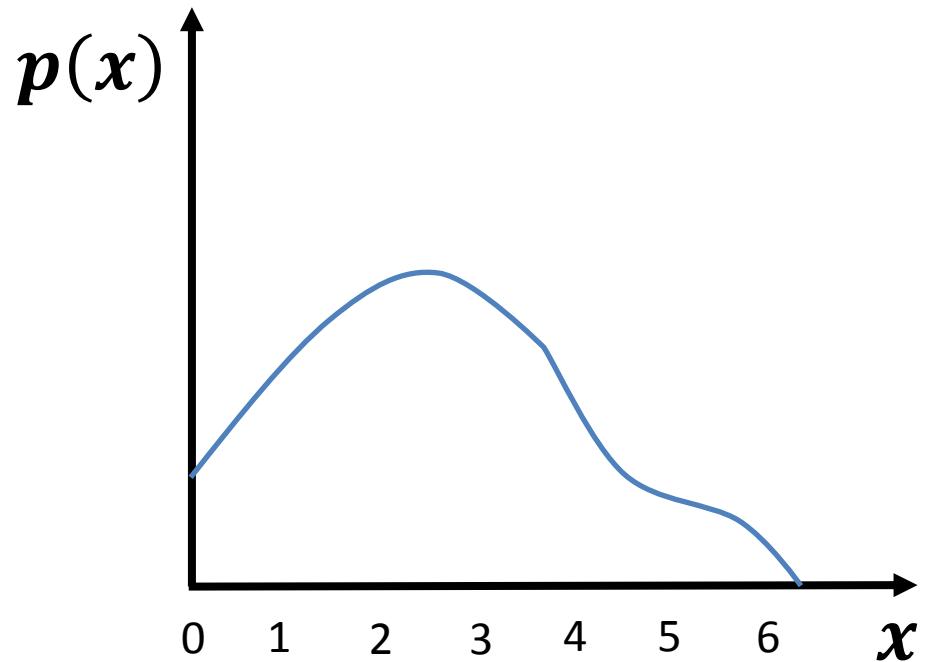
- m is the number of data points falling within a given range, i.e., histogram bin
- M is the total number of points (that belongs to the same class)
- Size of bin: size of the histogram bin

Histogram Analysis

- Consider 1-D example:
 - m is number of data points within the given range, e.g., $2 < x \leq 3$



$$\hat{p}(2 < x \leq 3) = \frac{m}{M * (\text{size of bin})}$$



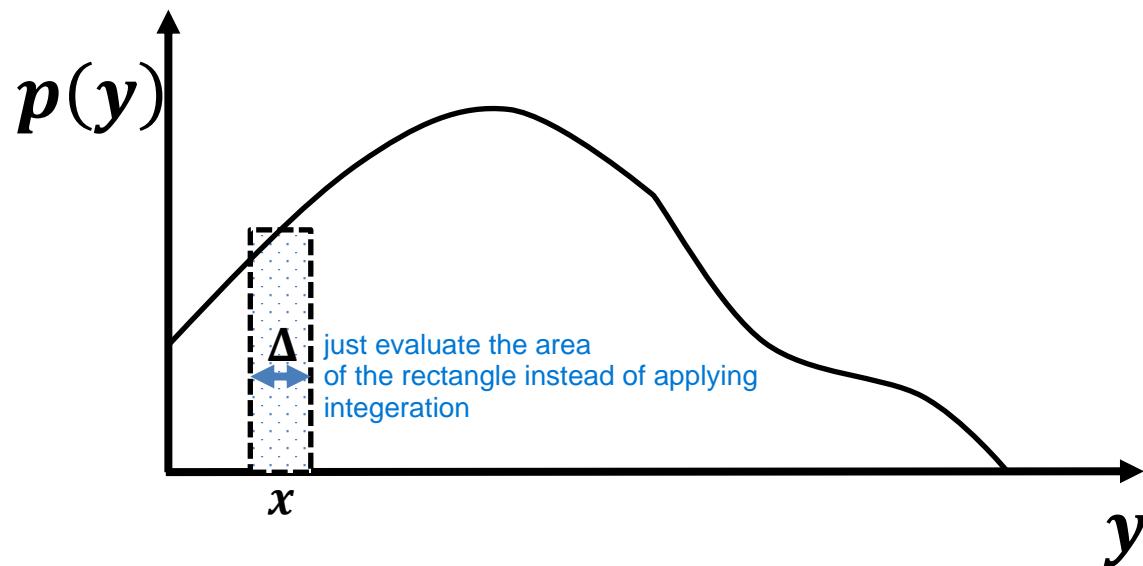
Data was originally generated
from this density

Histogram Analysis

proof for the equation

$$\int_{x-\frac{\Delta}{2}}^{x+\frac{\Delta}{2}} p(x)dx \approx \Delta \cdot p(x)$$

- Probability (generated point $\in [x - \frac{\Delta}{2}, x + \frac{\Delta}{2}]$) $\approx \Delta \cdot p(x) \equiv z$



Bin size $\equiv \Delta$

Histogram Analysis

$$\int_{x-\frac{\Delta}{2}}^{x+\frac{\Delta}{2}} p(x)dx \approx \Delta \cdot p(x)$$

- Probability (generated point $\epsilon [X - \frac{\Delta}{2}, X + \frac{\Delta}{2}]$) $\approx \Delta \cdot p(x) \equiv z$
- Assume we draw a number M of points according to $p(x)$
→ binomial distribution
- Binomial distribution with probability z for number of points falling in BIN

Histogram Analysis

$$P(k \text{ points falling in BIN out of } M \text{ points}) = \binom{M}{k} z^k (1 - z)^{M-k}$$

$$\begin{aligned} E(\# \text{ points in BIN}) &= M \cdot z \\ &= M \cdot p(x) \cdot \Delta \end{aligned}$$

- Example: flip a coin 10 times

$$P(8 \text{ Heads}) = \binom{10}{8} p^8 (1 - p)^{10-8}$$

$$E(\# \text{ Heads}) = p \cdot M = 0.5 * 10 = 5$$

$p \equiv \text{probability of head}$

Histogram Analysis

$$\begin{aligned} E(\# \text{ points in BIN}) &= M.z \\ &= M.p(x).\Delta \end{aligned}$$

- If k points fall in the histogram range then assuming:

$$k \approx M.p(x).\Delta$$

- Then, estimate of $p(x)$ is:

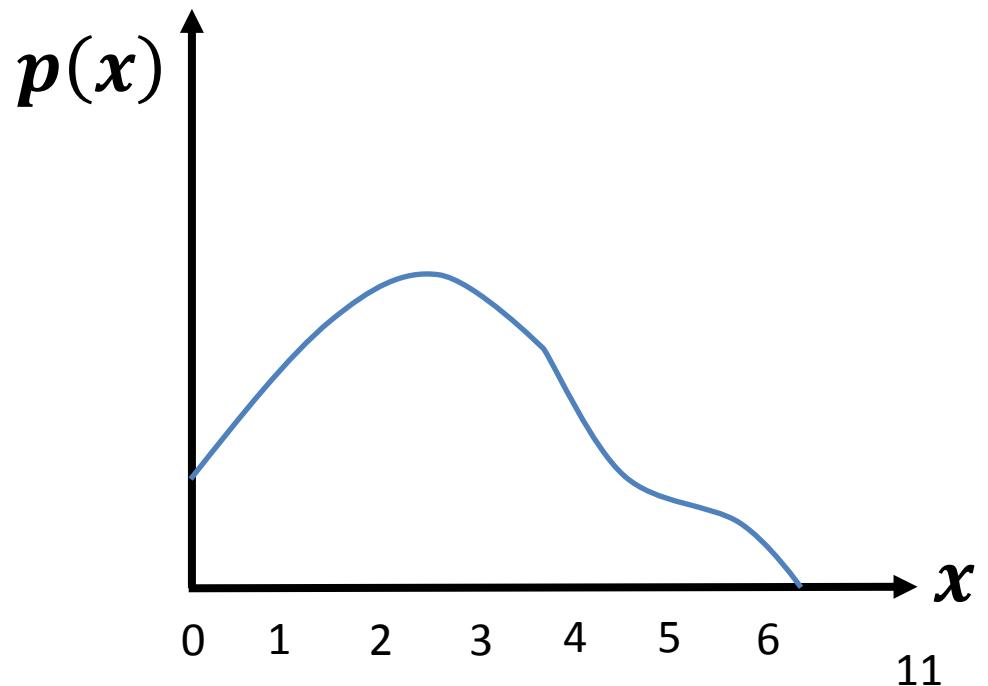
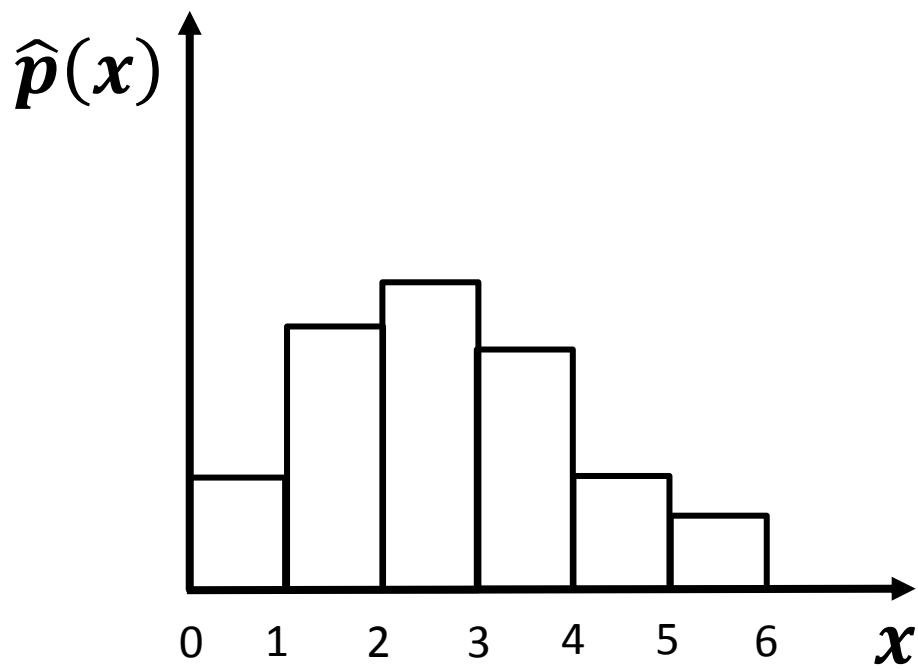
$$p(x) = \frac{k}{M\Delta}$$

Recall:

$$\hat{p}(x) = \frac{m}{M * (\text{size of bin})}$$

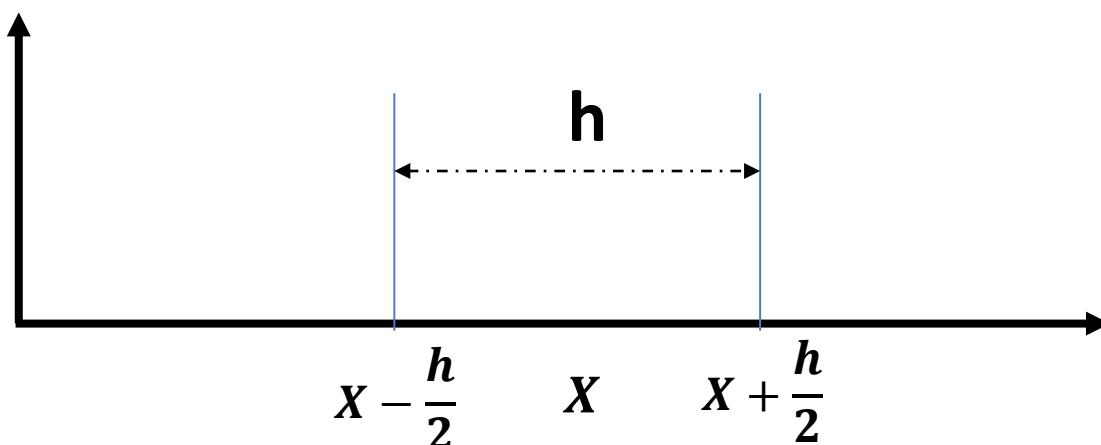
Histogram Analysis

- **Weak** method of estimation
- **Discontinuity** of these density estimates, even though the true densities are assumed to be smooth



Naïve Estimator

- Instead of partitioning X , i.e., feature space, into a number of prespecified ranges, we perform a similar range analysis for every X



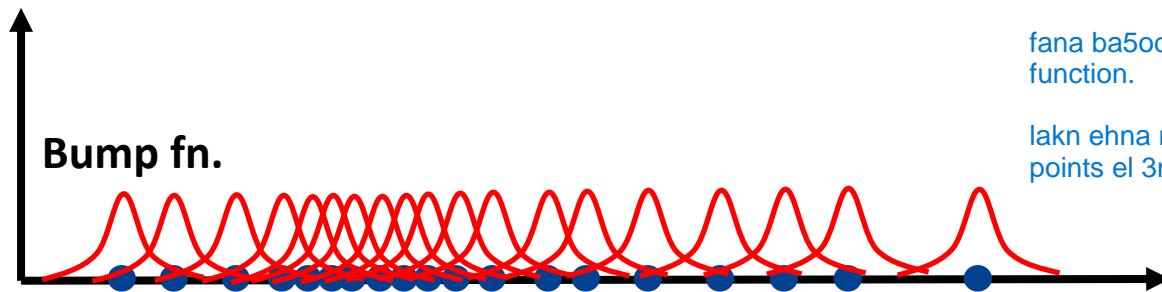
$$\hat{P}(X) = \frac{\#points \text{ falling in } \left[X - \frac{h}{2}, X + \frac{h}{2} \right]}{Mh}$$

Naïve Estimator

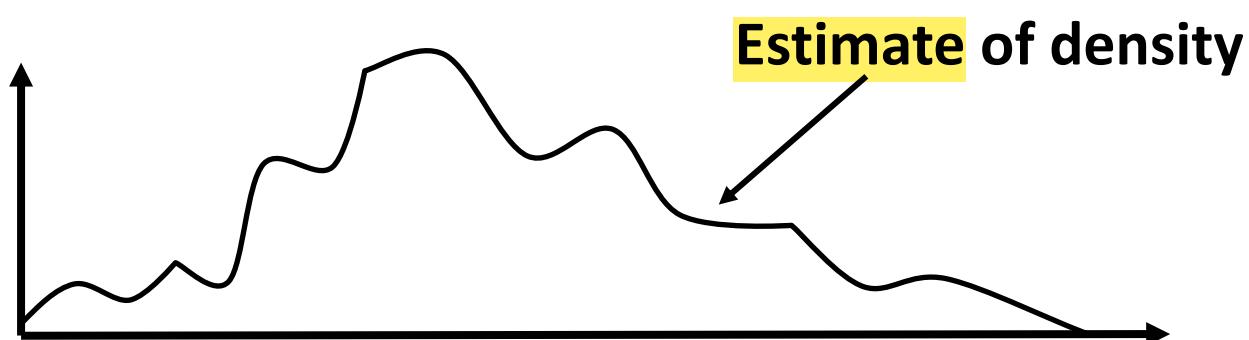
- Drawbacks:
 - Discontinuity of the density estimates
 - All data points are weighted equally regardless of their distance to the estimation point, i.e, X

Kernel Density Estimator

- a.k.a. Parzen Window Density Estimator
- Choose a bump function



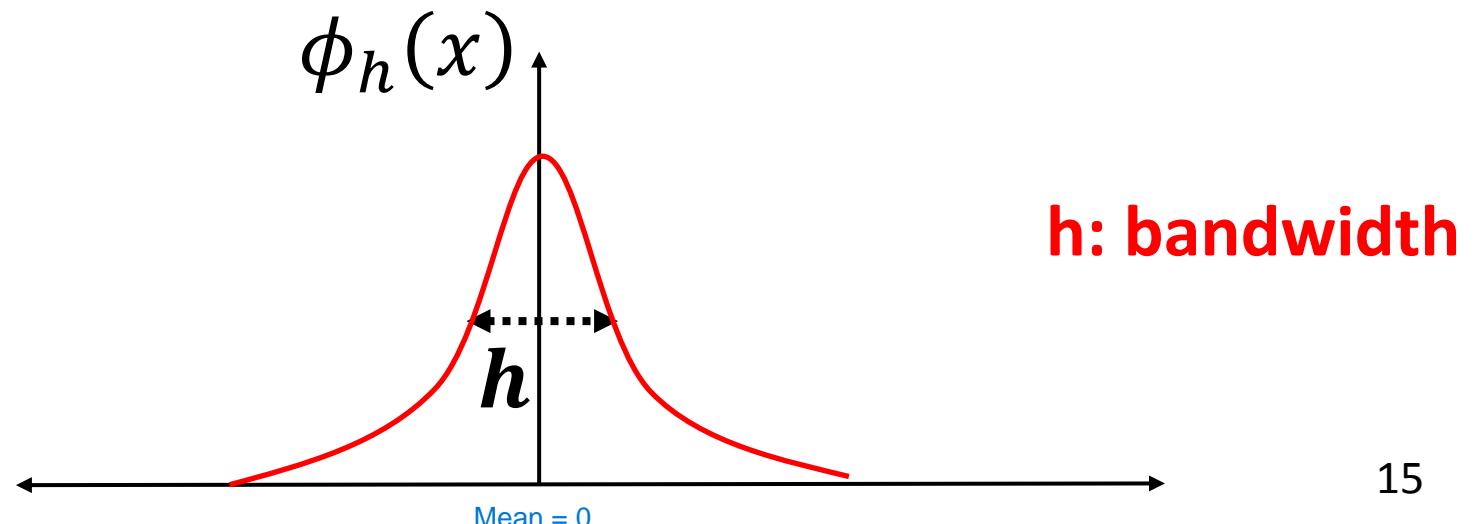
- Summation of bump functions:



Kernel Density Estimator

- Choose bump function as Gaussian with standard deviation (bandwidth) h :

$$\phi_h(x) = \frac{e^{\frac{-x^2}{2h^2}}}{\sqrt{2\pi}h}$$

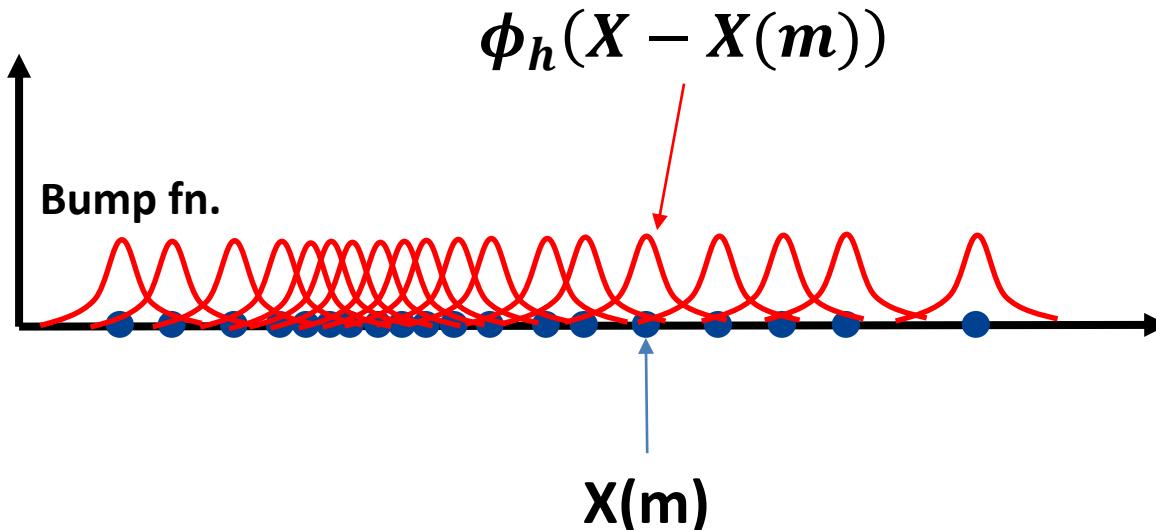


Kernel Density Estimator

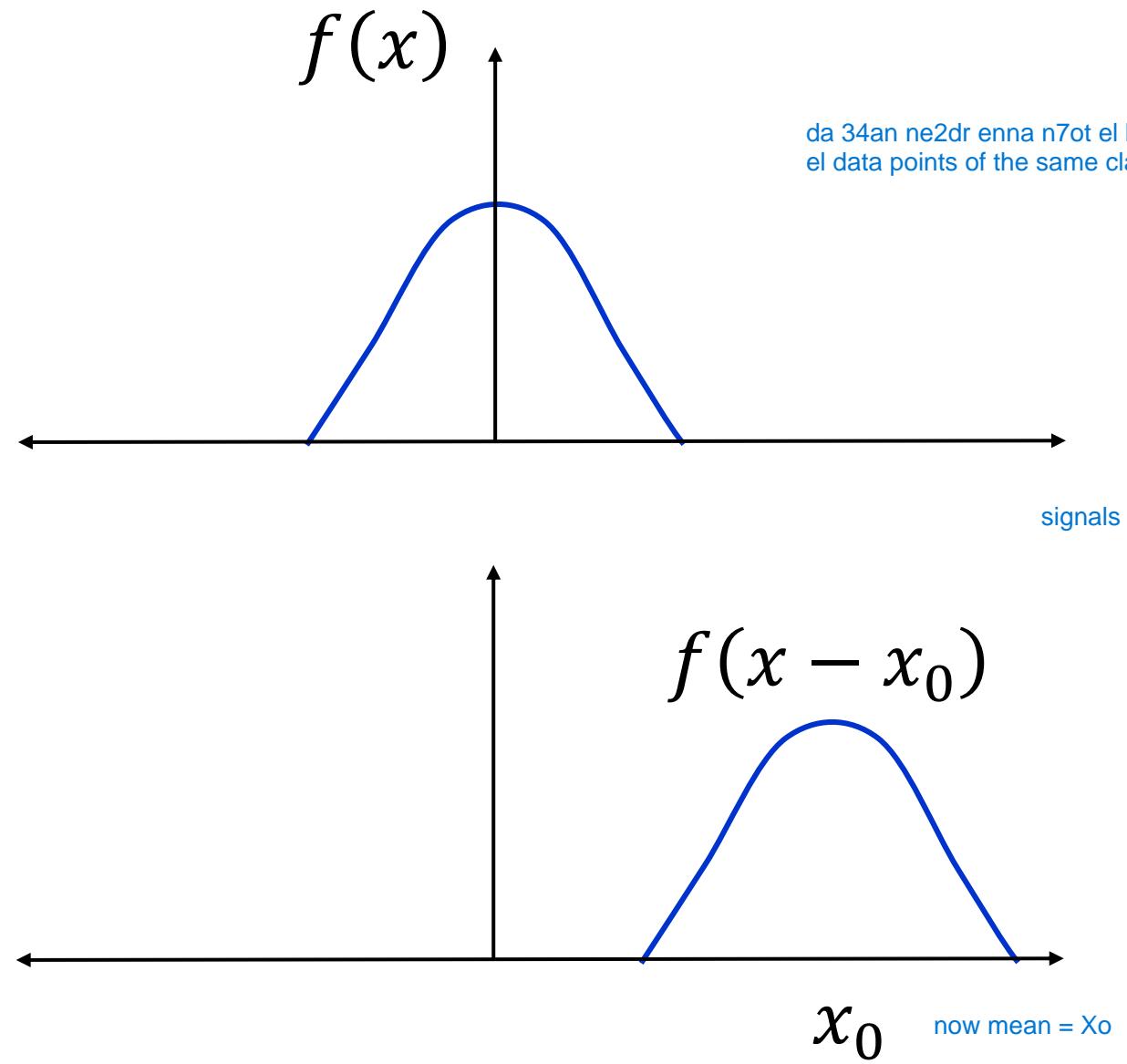
- Choose bump function as Gaussian with standard deviation (bandwidth) h :

$$\phi_h(x) = \frac{e^{\frac{-x^2}{2h^2}}}{\sqrt{2\pi}h}$$

- $X(m)$ are the points generated from the density $P(X)$ that we want to estimate



Kernel Density Estimator



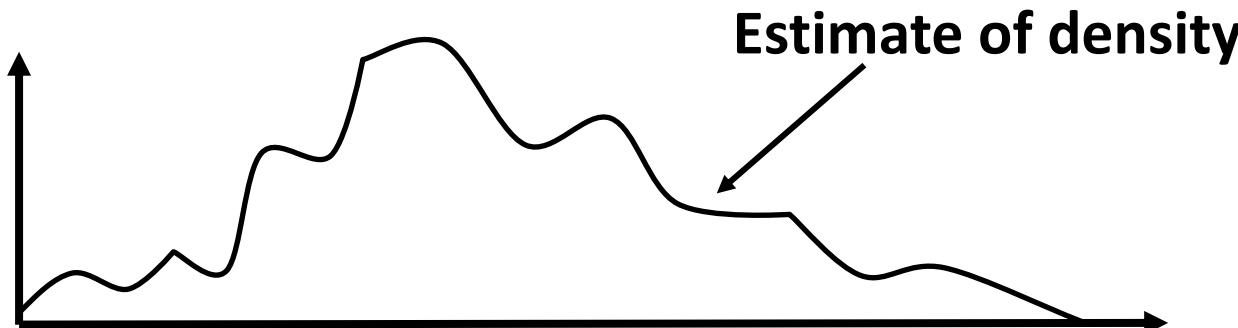
Kernel Density Estimator

- Summation of bump functions:

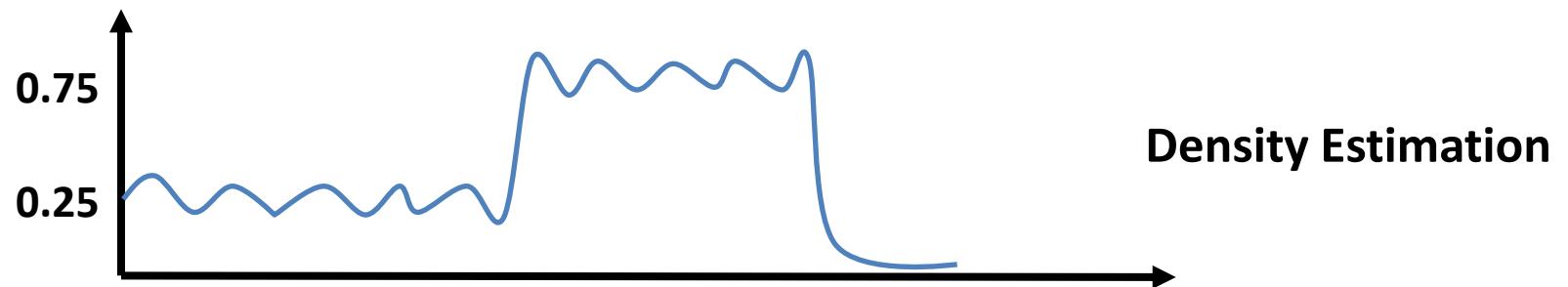
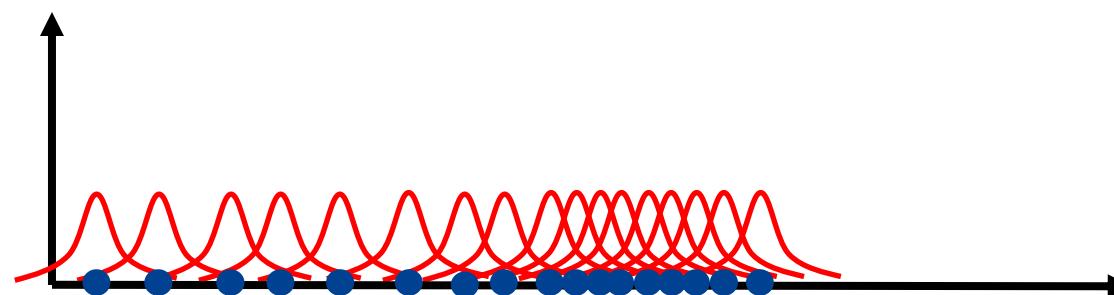
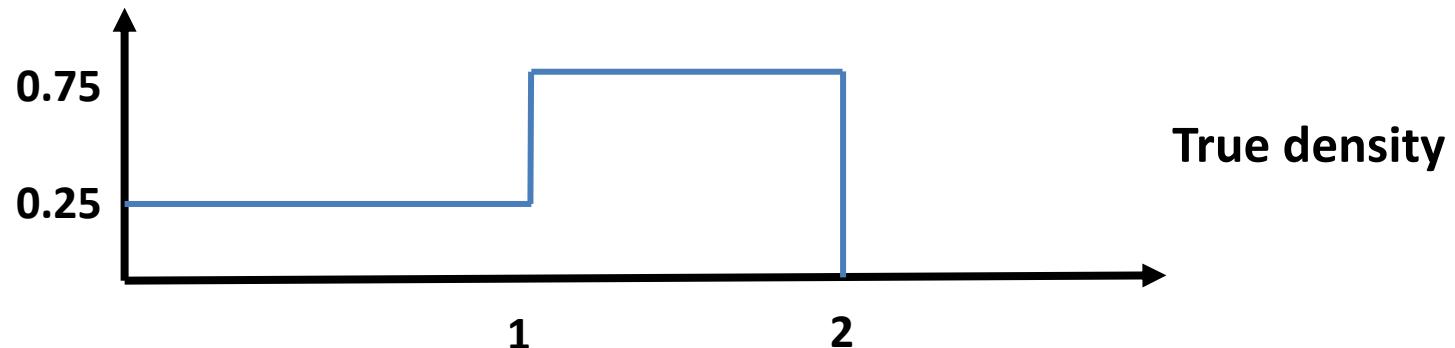
$$\hat{P}(X) = \frac{1}{M} \sum_{m=1}^M \phi_h(X - X(m))$$

P hat mean P estimated
w estimated 34an ehna bngebha
mn el data points w de btb2a 3bara
3n samples msh el 7a2e2a kolaha

Summation over # of generated points



Kernel Density Estimator



Kernel Density Estimator

- ϕ_h does not have to be Gaussian but it is preferred

this is the generic form

$$\phi_h = \frac{1}{h} g\left(\frac{x}{h}\right)$$

where $g(\cdot)$ is any suitable bump function that integrates to 1:

$$\int_{-\infty}^{\infty} g(x) dx = 1$$

density function

symmetric around the mean
(y axis)

w da 34an mt3mlsh favoring bump 3n el tany.
fa tb2a fair.

e.g.

$$g(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}} \rightarrow \phi_h(x) = \frac{e^{-x^2/2h^2}}{\sqrt{2\pi h}}$$

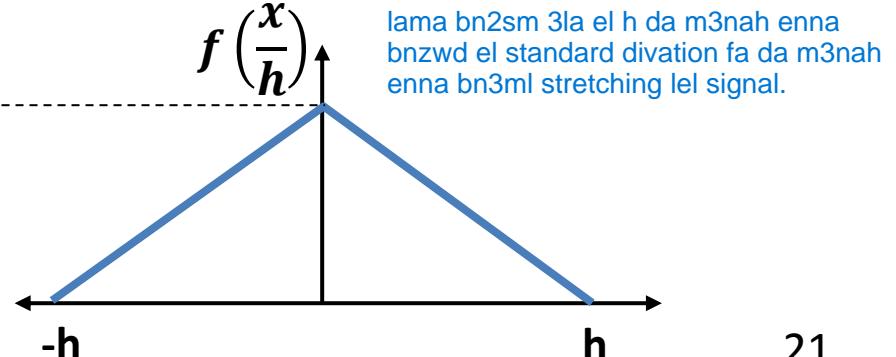
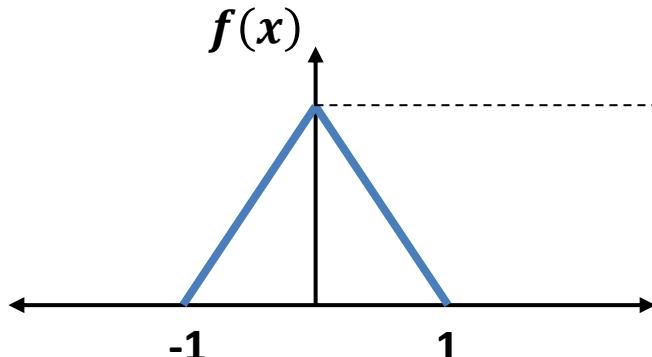
Kernel Density Estimator

- ϕ_h does not have to be Gaussian

$$\phi_h = \frac{1}{h} g\left(\frac{x}{h}\right)$$

where $g(\cdot)$ is any suitable bump function that integrates to 1:

$$\int_{-\infty}^{\infty} g(x)dx = 1$$

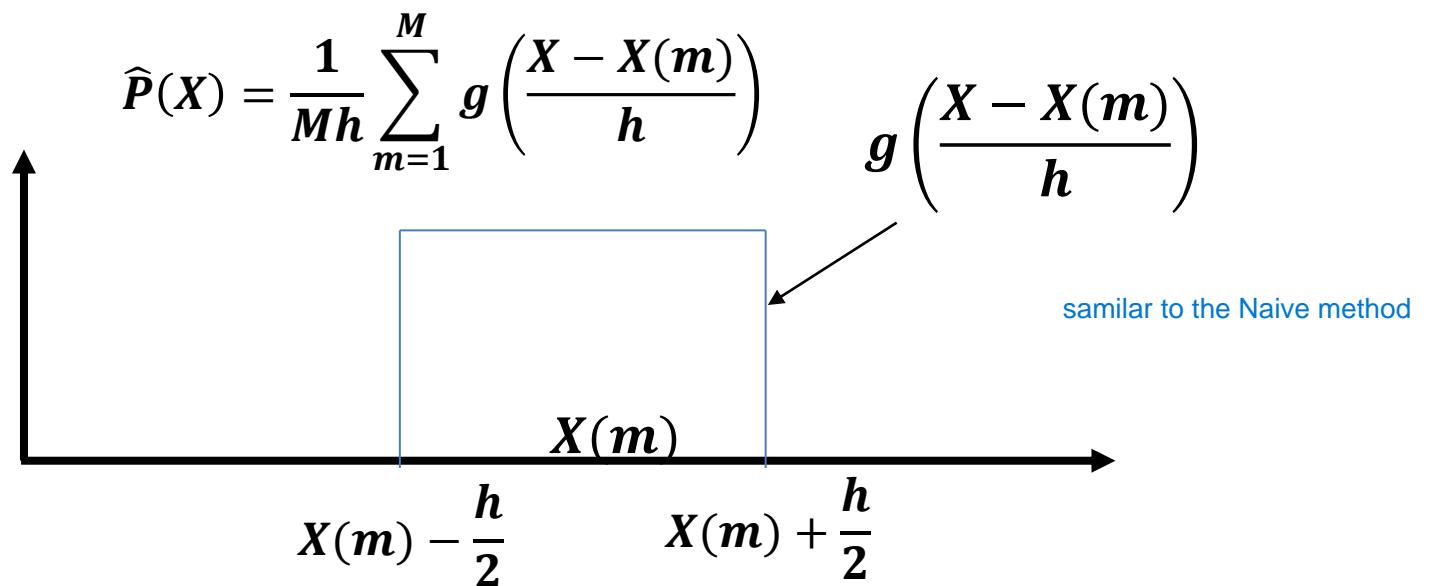


Kernel Density Estimator

- Naïve estimator is equivalent to a Parzen window estimator with:

$$g(x) = \begin{cases} 1, & -\frac{1}{2} \leq x < \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

- In this case:



Kernel Density Estimator

1-D form:

$$\widehat{P}(X) = \frac{1}{M} \sum_{m=1}^M \phi_h(X - X(m))$$

$$= \frac{1}{Mh} \sum_{m=1}^M g\left(\frac{X - X(m)}{h}\right)$$

look at slide 21

$$\int_{-\infty}^{\infty} g(x) dx = 1$$

$$\int_{-\infty}^{\infty} \widehat{P}(x) dx = 1$$

X(m) da 3bara 3n mkan el class el bn3ml 3ndo el shift.
aknk fe el signals w 3auz t3ml shift lel signal ymen fa 34an
n3ml keda bn2ol el shifted version = f(x-xo)
xo hena baa hya el X(m)

Kernel Density Estimator

Multi-dimension Form:

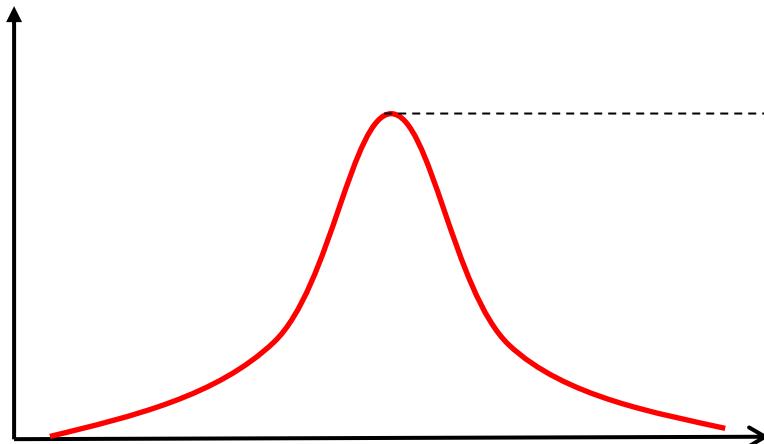
$$\hat{P}(\underline{X}) = \frac{1}{Mh^N} \sum_{m=1}^M g\left(\frac{\underline{X} - \underline{X}(m)}{h}\right)$$

$$\int_{-\infty}^{\infty} g(\underline{X}) d\underline{X} = 1$$

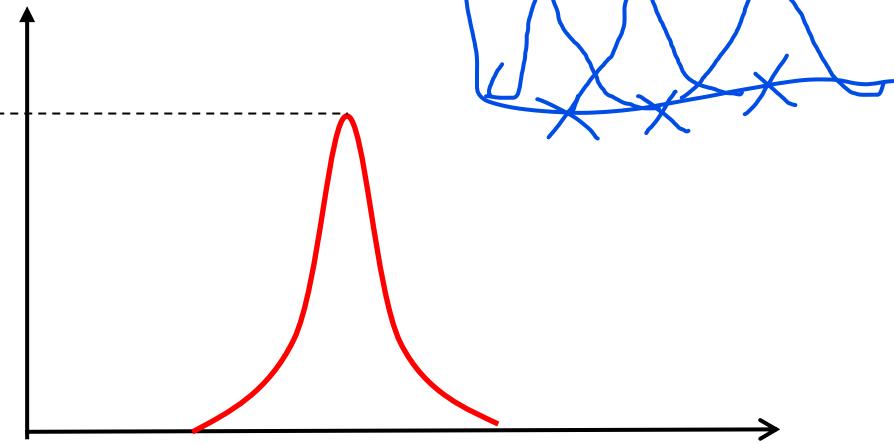
For example: multi-dimension independent Gaussian density:

$$g(\underline{X}) = \frac{e^{-\sum_{i=1}^N \frac{x_i^2}{2}}}{(2\pi)^{N/2}}$$

How to choose h?



Large h



Small h

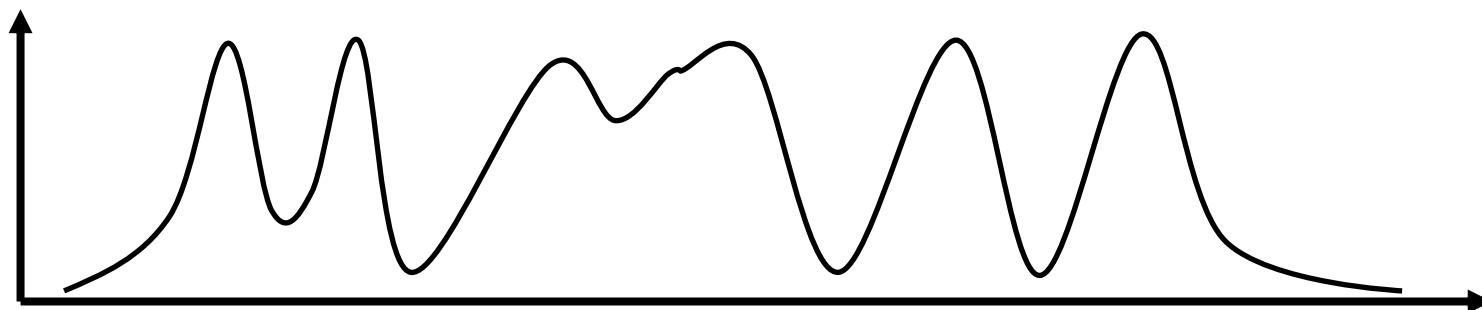
eh faydet el h?

kol ma t2ll el h kol ma btgeb details aktur w vice versa baa.

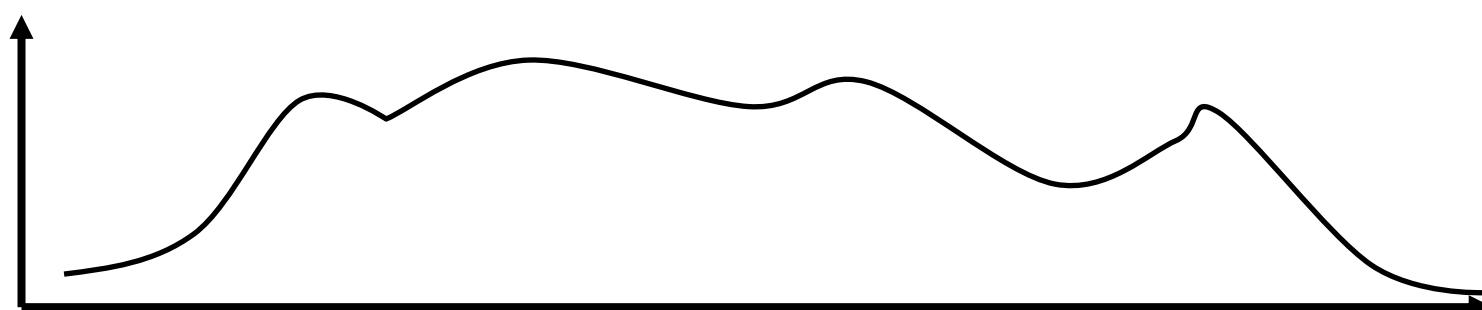
bumby 34an enta btgm3 el signal 3la b3d, ershom gmb b3d
htfhm, ana fhmtha fl mo7dra.

How to choose h ?

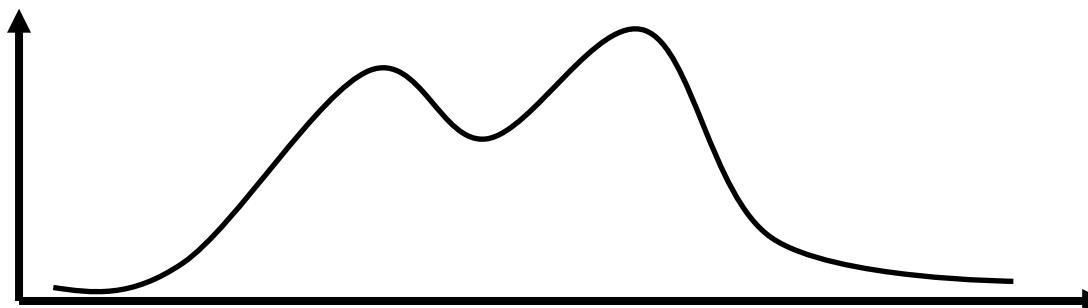
- ✓ Too small $h \rightarrow$ bumpy estimate or non-smooth



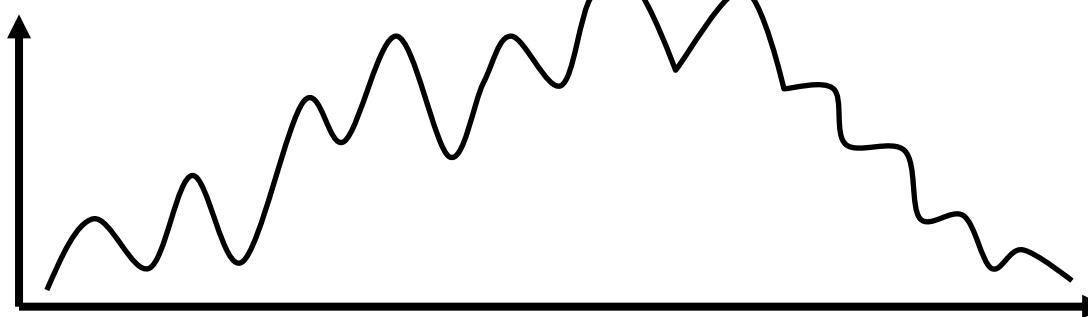
- ✓ Too large $h \rightarrow$ the estimate could be too smooth that **essential details** of the density will be **lost** or smoothed **out**



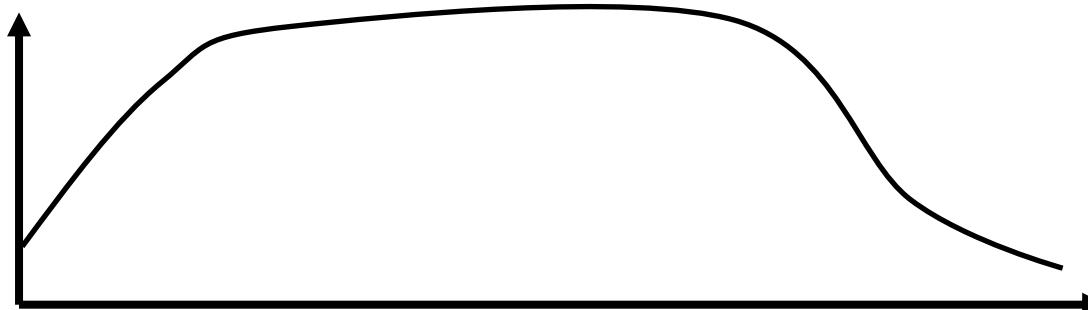
How to choose h ?



True Density



Too small h



Too large h

Optimal h

34an nsahel el denya,
 1. bnfrd enha independet.
 2. bngeb el average bta3 kol dimension
 3. w bngeb el optimal h w nw7do lehom kolohom.

- The optimal H (diagonal bandwidth matrix) can be approximated as :

$$H_i = \sigma_i \left[\frac{4}{(N+2)M} \right]^{\frac{1}{N+4}}$$

of points of the same class

kol ma el Hi yekbr da m3nah enny akbur el width bta3y, 34an da m3nah en el data scattered aslun
normal reference rule

where

$$\sigma_i = \sqrt{[\Sigma_X]_{i,i}}$$

gaya mn el covariance matrix

- Σ_X is the estimated covariance matrix, i.e.,

$$\Sigma_X = \frac{1}{M} \sum_{m=1}^M (\underline{X}(m) - \hat{\mu})(\underline{X}(m) - \hat{\mu})^T$$

$[\Sigma_X]_{i,i} \equiv i^{th}$ diagonal element of Σ_X

$N \equiv$ dimensions

number of dimensions

$$h_{opt} = \frac{1}{N} \sum_{i=1}^N H_i$$

For multi-variate normal kernel & diagonal bandwidth matrix

Acknowledgment

- These slides have been created relying on lecture notes of Prof. Dr. Amir Atiya