Rensselaer

Lecture 6: Non-Parametric Methods – Parzen Estimation

Dr. Chengjiang Long
Computer Vision Researcher at Kitware Inc.
Adjunct Professor at RPI.
Email: **longc3@rpi.edu**

Recap Previous Lecture

$$p(\mathrm{x} \mid D) = \int p(\mathrm{x}, \theta \mid D) d\theta$$

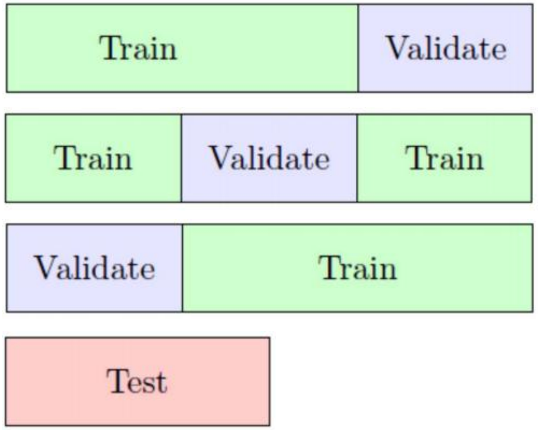
$$= \int p(\mathrm{x} \mid \theta, D) p(\theta \mid D) d\theta = \int p(\mathrm{x} \mid \theta) p(\theta \mid D) d\theta$$

p(x) is completely known given θ, independent of samples in D

$$\Rightarrow$$

$$P(X_1 \ldots X_n | Y) = \prod_i P(X_i | Y)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

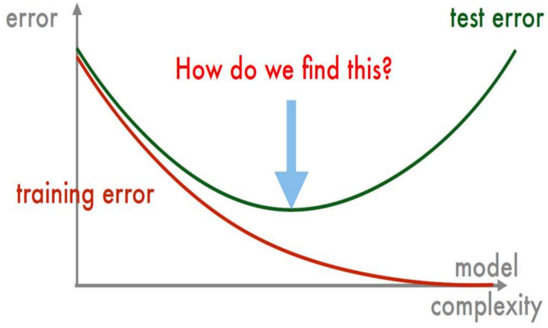


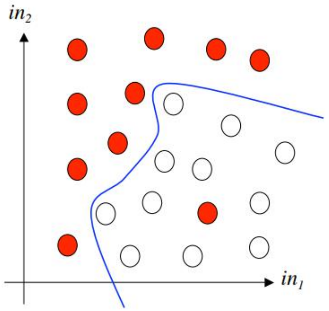
Train | Validate

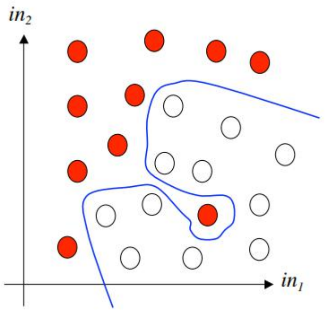
Train | Validate | Train

Validate | Train

Test







Outline

- Parametric and Non-Parametric
- Density Estimation
- Parzen Window Estimation

Outline

- **Parametric and Non-Parametric**
- Density Estimation
- Parzen Window Estimation

3 Styles of Probability Density Estimation

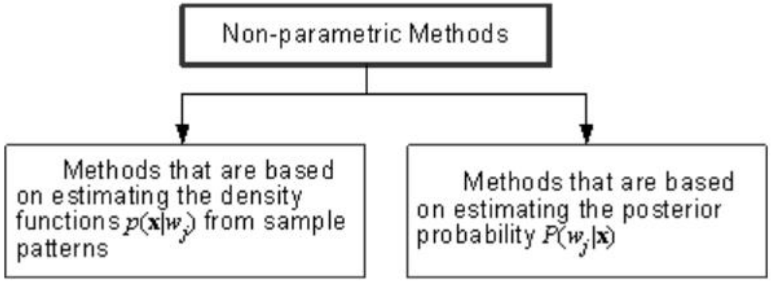
- **Parametric**:
  - Assume a specific functional form for the densities.
  - Parameters of the distribution optimised to fit the data.
  - Good if model fits the data bad if not.
- **Non-parametric**:
  - No assumptions about form of the density function, determined entirely from the data.
  - Number of parameters grows with size of the data-set and so models can quickly become unwieldy and can take long to incorporate new data
- **Semi-parametric**:
  - Mixture models.
  - Tries to achieve best of both worlds by allowing general functional form for densities where number of parameters (and thus flexibility of function) can be extended independently of the size of data-set. Could combine worst of both approaches?

Parametric vs. Non-Parametric

- Parametric
  - ✓ Based on Functions (e.g Normal Distribution)
  - ✓ Unimodal
  - ✓ Only one peak
  - ✓ Unlikely real data confines to function

VS

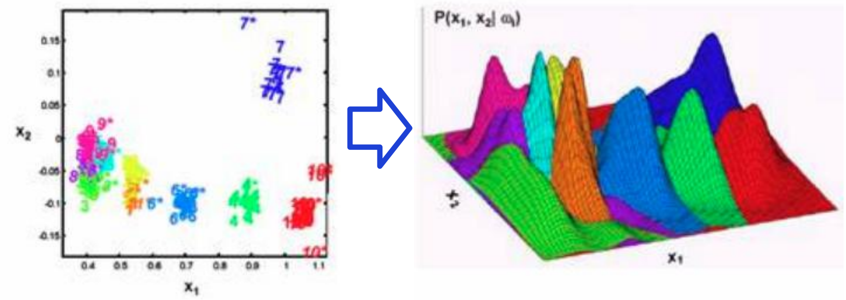
- Non–Parametric
  - ✓ Based on Data
  - ✓ As many peaks as Data has
  - ✓ Methods for both *P(x|w$_j$)* and *P(w$_j$|x)*



Non-parametric Methods

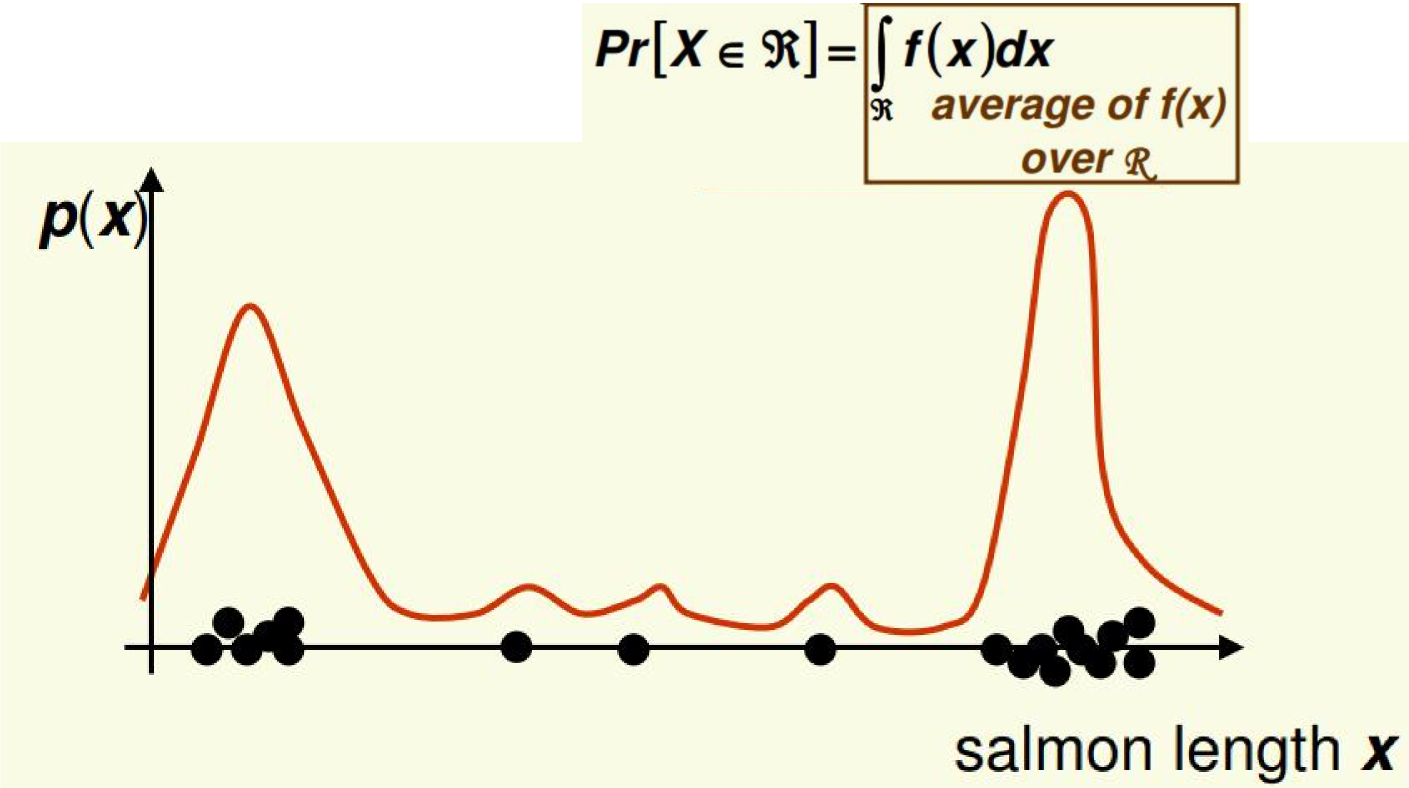
Methods that are based on estimating the density functions $p(\mathbf{x}|w_j)$ from sample patterns

Methods that are based on estimating the posterior probability $P(w_j|\mathbf{x})$



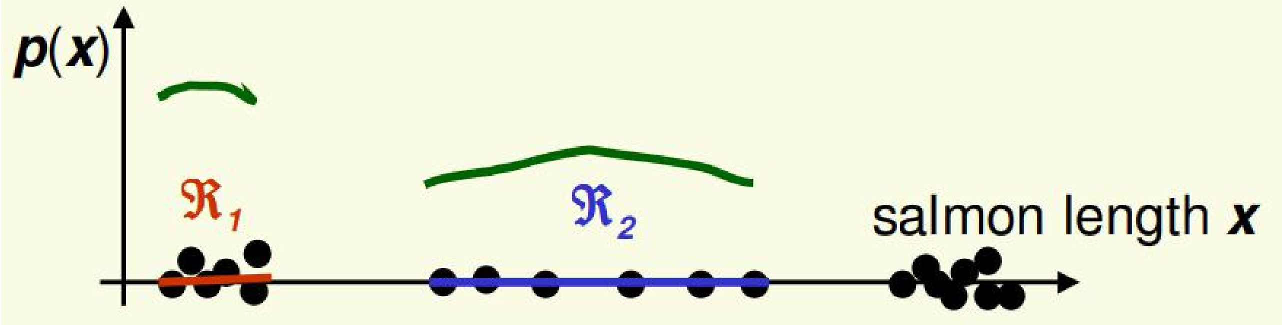
Non-Parametric Techniques: Introduction

- Nonparametric techniques attempt to estimate the underlying density functions from the training data
  - Idea: the more data in a region, the larger is the density function



$$Pr[X \in \Re] = \int_{\Re} f(x)dx$$

Non-Parametric Techniques: Introduction



- How can we approximate Pr[X ∈ ℜ1] and Pr[X ∈ℜ2]?
  - Pr [X ∈ ℜ1] ≈ 6/20, Pr [X ∈ ℜ2] ≈ 6/20
- Should the density curves above ℜ1 and ℜ2 be equally high?
  - No, since is ℜ1 smaller than ℜ2:

$$Pr[X \in \Re_1] = \int_{\Re_1} f(x)dx \approx \int_{\Re_2} f(x)dx = Pr[X \in \Re_2]$$

  - To get density, normalize by region size

Non-Parametric Techniques: Introduction

- Assuming f(x) is basically flat inside $\Re$

$$\frac{\#\,of\ samples\ in\,\Re}{total\,\#\,of\ samples} \approx Pr[X \in \Re] = \int_{\Re} f(y)dy \approx f(x) * Volume(\Re)$$

- Thus, density at a point x inside $\Re$ can be approximated

$$f(x) \approx \frac{\#\,of\ samples\ in\,\Re}{total\,\#\,of\ samples} \frac{1}{Volume(\Re)}$$

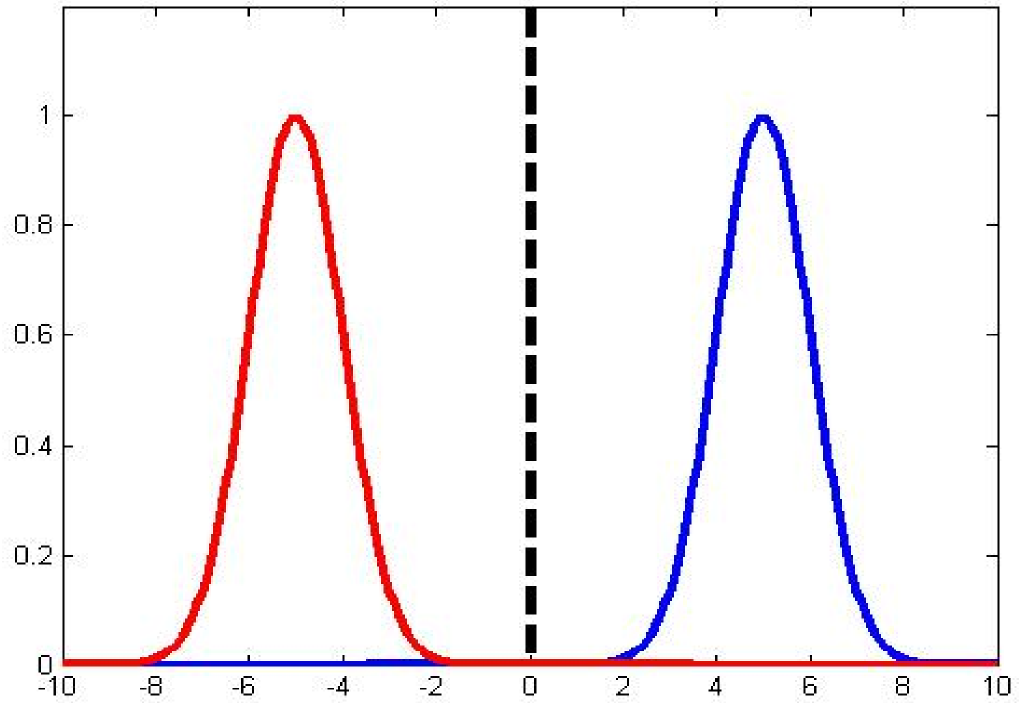
- Now let’s derive this formula more formally.

Outline

- Parametric and Non-Parametric
- **Density Estimation**
- Parzen Window Estimation

Motivation

- Why we need to estimate the probability density?
- If we can estimate p(x), we can estimate the class conditional probabilities P(x, | w_i) and therefore work out optimal (Bayesian) decision boundary.



Binomial Random Variable

- Let us flip a coin n times (each one is called “trial”)
  - Probability of head ρ , probability of tail is $1-\rho$
- Binomial random variable K counts the number of heads in n trials

$$P(K=k)=\binom{n}{k}\rho^{k}(1-\rho)^{n-k}$$

$$\text{where} \quad \binom{n}{k}=\frac{n!}{k!(n-k)!}$$

- Mean is $E(K)=n\rho$
- Variance is $\operatorname{var}(K)=n\rho(1-\rho)$

Density Estimation: Basic Issues

- From the definition of a density function, probability ρ that a vector x will fall in region ℜ is:

$$\rho = Pr[x \in \Re] = \int_{\Re} p(x')dx'$$

- Suppose we have samples x1, x2,…, xn drawn from the distribution p(x). The probability that k points fall in ℜ is then given by binomial distribution:

$$\mathrm{Pr}[K = k] = \binom{n}{k} \rho^k (1 - \rho)^{n-k}$$

- Suppose that k points fall in ℜ, we can use MLE to estimate the value of ρ . The likelihood function is:

$$p(x_1,\ldots,x_n \mid \rho) = \binom{n}{k} \rho^k (1 - \rho)^{n-k}$$

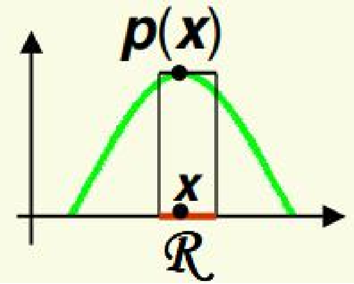
Density Estimation: Basic Issues

$$p(x_1,\ldots,x_n \mid \rho) = \binom{n}{k} \rho^k (1-\rho)^{n-k}$$

- This likelihood function is maximized at $\rho = \frac{k}{n}$
- Thus the MLE is $\hat{\rho} = \frac{k}{n}$
- Assume that p(x) is continuous and that the region ℜ is so small that p(x) is approximately constant in ℜ

$$\int_{\Re} p(x')dx' \cong p(x)V$$

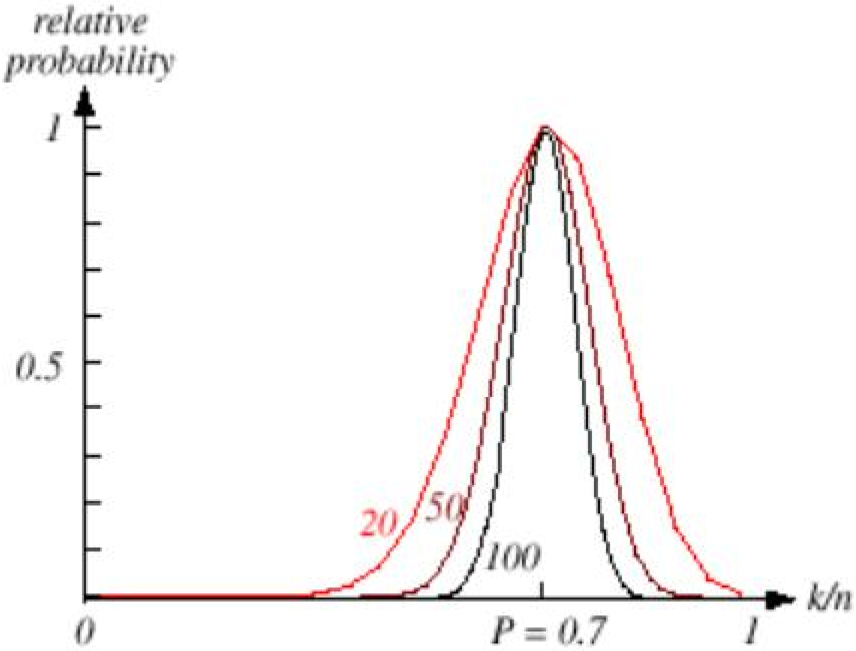
x is in ℜ and V is the volume of ℜ



- Recall from the previous slide: $\rho = \int_{\Re} p(x')dx'$
- Thus p(x) can be approximated: $\boxed{p(x) \approx \frac{k/n}{V}}$

Discussion

- If volume V is fixed, and n is increased towards ∞, P(x) converges to the average p of that volume.

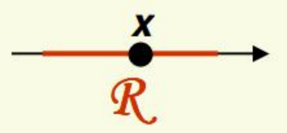


It peaks at the true probability, which is 0.7, and with infinite n, will converge to 0.7.

Density Estimation: Basic Issues

- This is exactly what we had before:

$$p(x) \approx \frac{k/n}{V}$$



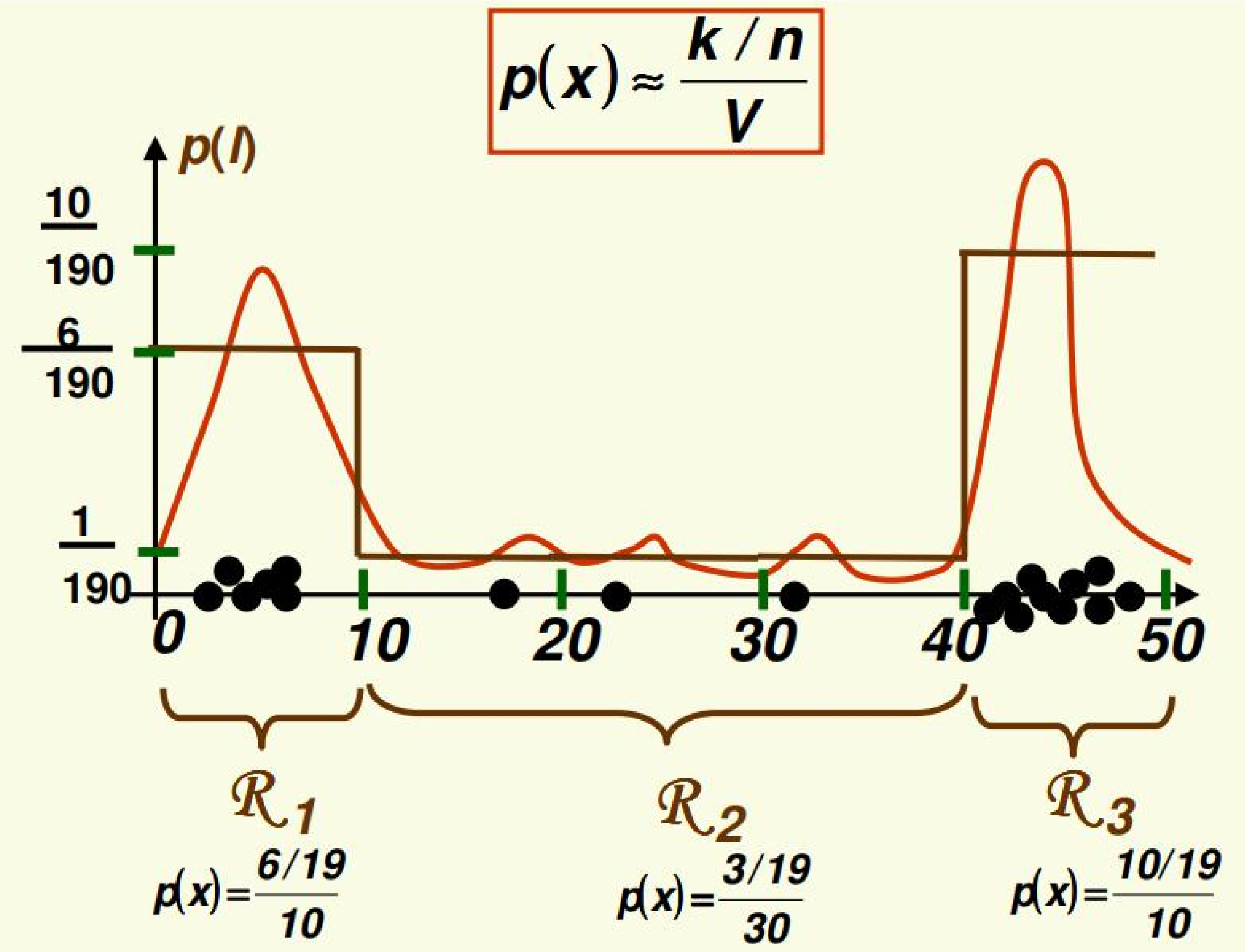
x is inside some region $\Re$
k = number of samples inside $\Re$
n=total number of samples
V=volume of $\Re$

- Our estimate will always be the average of true density over $\Re$

$$p(x) \approx \frac{k/n}{V} = \frac{\hat{\rho}}{V} \approx \frac{\int_{\Re} p(x')dx'}{V}$$

- Ideally, p(x) should be constant inside $\Re$

Density Estimation: Histogram

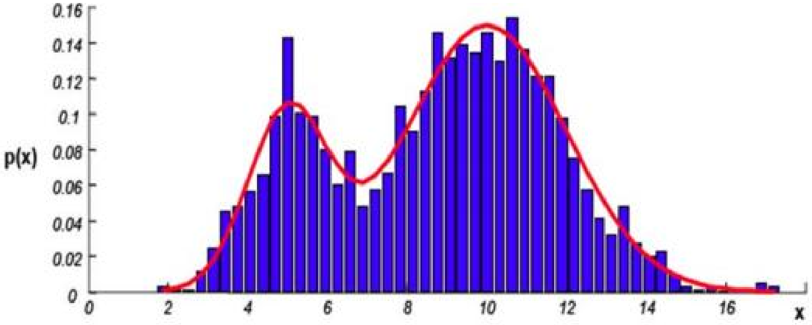


- If regions $\mathfrak{R}_i$'s do not overlap, we have a histogram

Density Estimation: Histogram

- The simplest form of non−parametric density estimation is the histogram
  - – Divide sample space in number of bins
  - – Approximate the density at the center of each bin by the fraction of points that fall into the bin
  - – Two parameters: bin width and starting position of first bin (or other equivalent pairs)

- Drawbacks:
  - – Depends on position of bin centers
    - Often compute two histograms, offset by ½ bin width
  - – Discontinuities as an artifact of bin boundaries
  - – Curse of dimensionality



Density Estimation: Accuracy

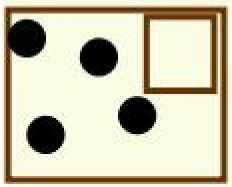
- How accurate is density approximation $p(x) \approx \frac{k/n}{V}$?
- We have made two approximations

1. $\hat{\rho} = \frac{k}{n}$

As n increases, this estimate becomes accurate

2. $\int_{\Re} p(x')dx' \cong p(x)V$

As $\Re$ grows smaller, the estimate becomes more accurate



As we shrink $\Re$ we have to make sure it contains samples, otherwise our estimated p(x) = 0 for x in $\Re$.

- Thus in theory, if we have an unlimited number of samples, to get convergence as we simultaneously increase the number of samples n, and shrink regions $\Re$, but not too much so that $\Re$ still contains a lot of samples.

Density Estimation: Accuracy

$$p(x) \approx \frac{k/n}{V}$$

- In practice, the number of samples is always fixed
- Thus the only available option to increase the accuracy is by decreasing the size of $\mathfrak{R}$ (V gets smaller)
  - If V is too small, p(x)=0 for most x, because most regions will have no samples
  - Thus have to find a compromise for V

    - not too small so that it has enough samples

    - but also not too large so that p(x) is approximately constant inside V

Density Est. with Infinite data

- To get the density at x. Assume a sequence of regions (R1 , R2 , … Rn) that all contain x. In Ri the estimate uses i samples
- Vn is volume of Rn , k_n is the number of samples in Rn .
  - $p_n(x)$ is the n–th estimate for n.

$$p_n(\mathbf{x}) = \frac{k_n/n}{V_n}$$

  - Goal is to get $p_n(x)$ to converge to p(x)

Convergence of $p_n(x)$ to $p(x)$

$$p(x) \cong \frac{k/n}{V}$$

- $p_n(x)$ converges to $p(x)$ if the following is true

$$\lim_{n\to\infty} V_n = 0$$

- Region R covers negligible space

$$\lim_{n\to\infty} k_n = \infty$$

- p(x) is average of infinite samples (unless p(x) = 0)

$$\lim_{n\to\infty} \frac{k_n}{n} = 0$$

- The samples of k, are a negligible amount of the whole set n. n gets bigger faster then k does.

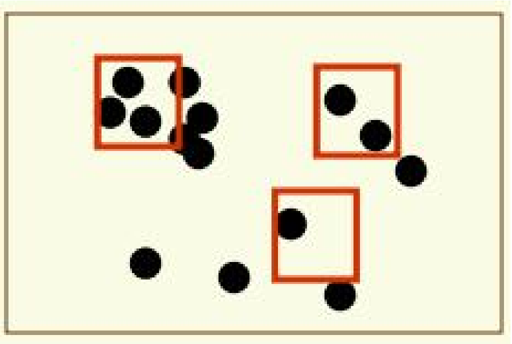
Density Estimation

- If n is fixed, and V approaches zero, V will become so small it has zero samples, or reside directly on a point, making $p(x) \approx 0$ or ∞
- In Practice, can not allow volume to become too small, since data is limited.
  - If you use a non–zero V, estimation will have some variance in k/n from actual.
- In theory, with unlimited data, can get around limitations

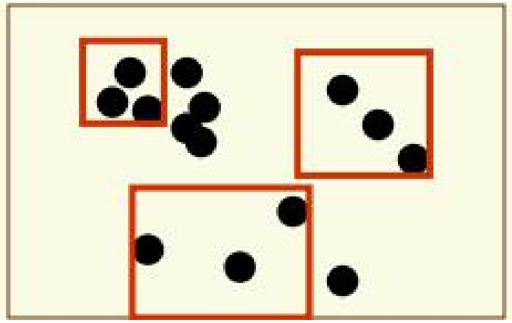
Density Estimation: Two Approaches

$$p(x) \approx \frac{k/n}{V}$$

- Parzen Windows:
  - ✓ Choose a fixed value for volume V and determine the corresponding k from the data.



- k-Nearest Neighbors
  - ✓ Choose a fixed value for k and determine the corresponding volume V from the data

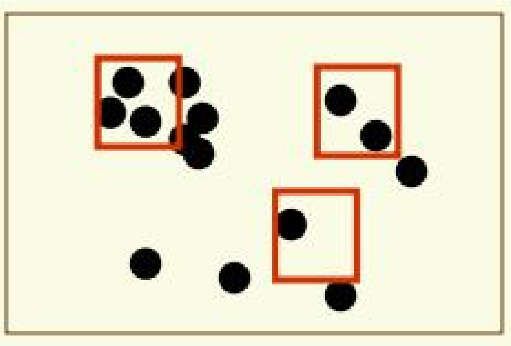


Under appropriate conditions and as number of samples goes to infinity, both methods can be shown to converge to the true p(x)

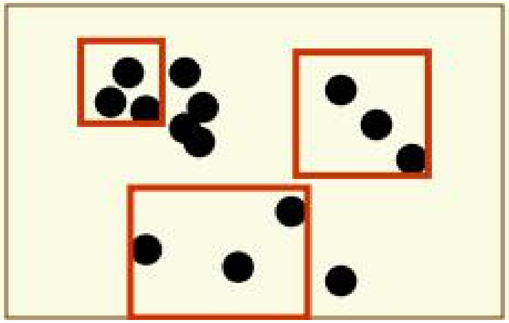
Density Estimation: Two Approaches

$$p(x) \approx \frac{k/n}{V}$$

- Parzen Windows:
  - ✓ Shrink an initial region where $V_n = 1/\sqrt{n}$ and show that
    $$p_n(x) \underset{n\to\infty}{\to} p(x)$$
  - ✓ This is called "the Parzen window estimation method"



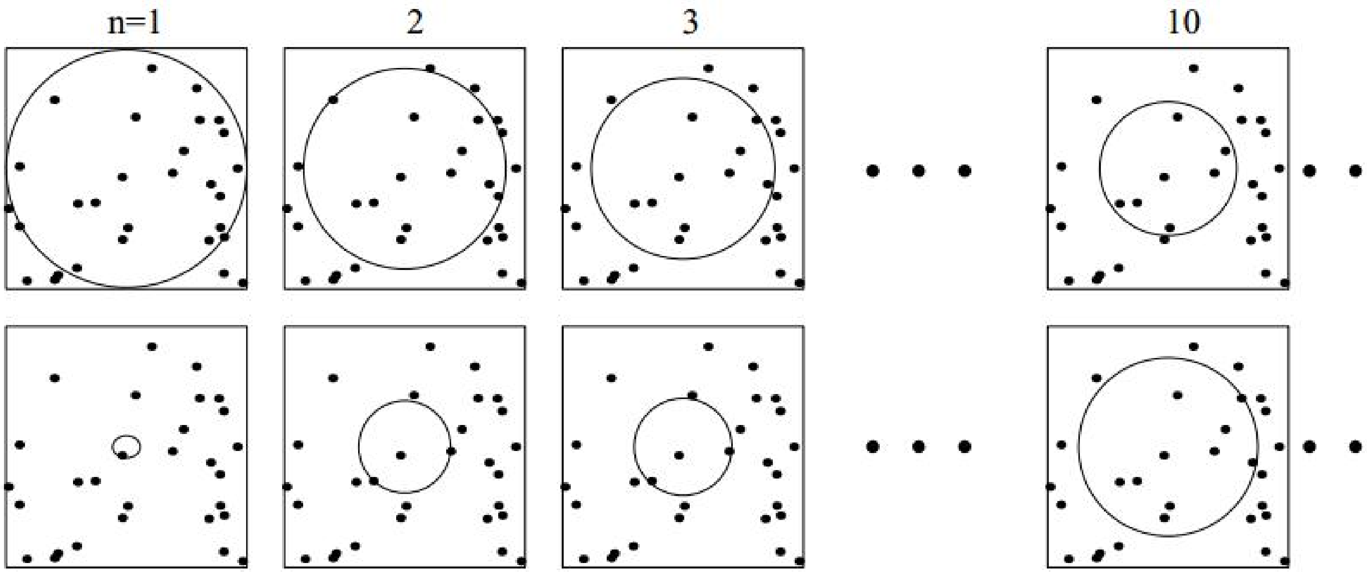
- k-Nearest Neighbors
  - ✓ Specify *kn* as some function of n, such as $k_n = \sqrt{n}$ the volume *Vn* is grown until it encloses *kn* neighbors of x. This is called "the *kn*–nearest neighbor estimation method"



Density Estimation: Two Approaches

$$V_n = 1/\sqrt{n}$$

$$k_n = \sqrt{n}$$

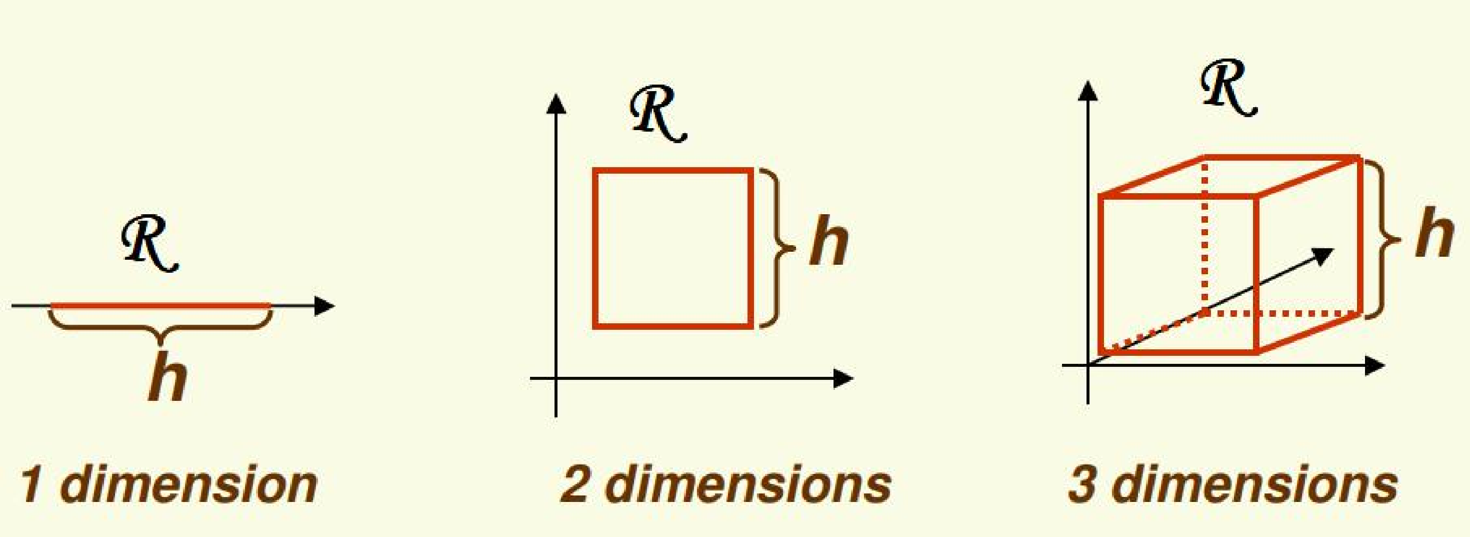


Outline

- Parametric and Non-Parametric
- Density Estimation
- **Parzen Window Estimation**

Parzen Windows

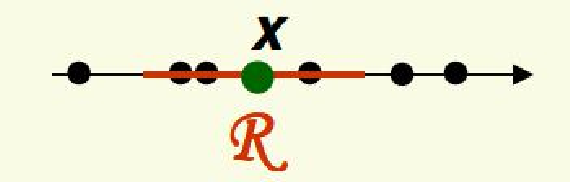
- In Parzen window approach to estimate densities we fix the size and shape of region $\mathfrak{R}$
- Let us assume that the region $\mathfrak{R}$ is a d–dimensional hypercube with side length h thus it's volume is h^d



Parzen Windows

- To estimate the density at point x, simply center the region $\mathfrak{R}$ at x, count the number of samples in $\mathfrak{R}$, and substitute everything in our formula

$$p(x) \approx \frac{k/n}{V}$$

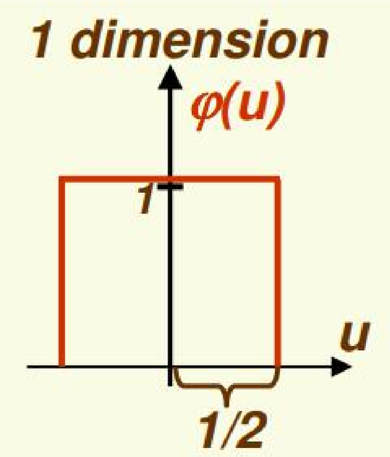


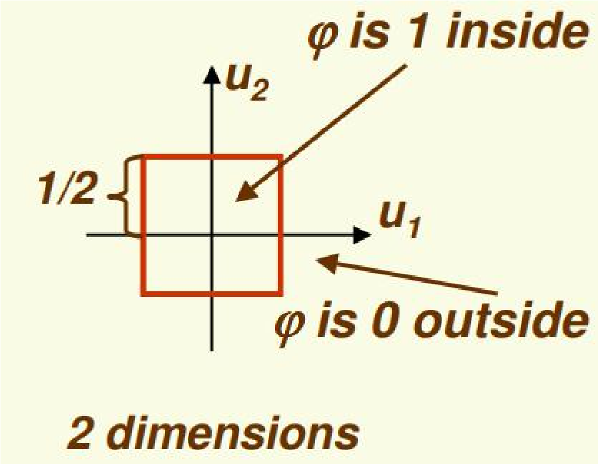
$$p(x) \approx \frac{3/6}{10}$$

Parzen Windows

- We wish to have an analytic expression for our approximate density $\Re$
- Let us define a ***window function***

$$\varphi(u) = \begin{cases} 1 & |u_j| \leq \frac{1}{2} \quad j = 1, \ldots, d \\ 0 & \text{otherwise} \end{cases}$$

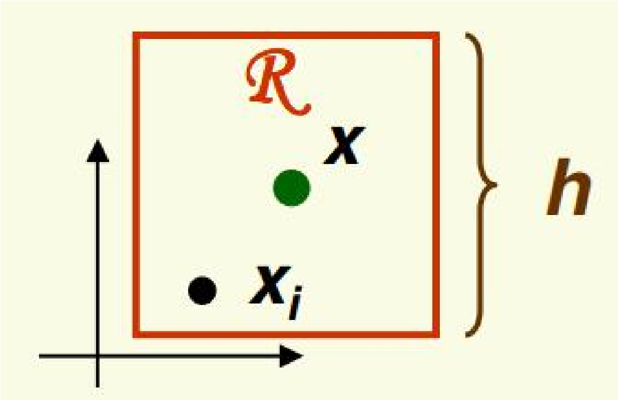




Parzen Windows

- Recall we have samples x1 , x2 ,…, xn . Then

$$\varphi\left(\frac{x - x_i}{h}\right) = \begin{cases} 1 & |x - x_i| \le \dfrac{h}{2} \quad j = 1,\dots,d \\ 0 & \text{otherwise} \end{cases}$$



$$\varphi\left(\frac{x - x_i}{h}\right) = \begin{cases} 1 & \text{if } x_i \text{ is inside the hypercube with width } h \text{ and centered at } x \\ 0 & \text{otherwise} \end{cases}$$

Parzen Windows

- How do we count the total number of sample points x1 , x2 ,…, xn which are inside the hypercube with side h and centered at x ?

$$k = \sum_{i=1}^{i=n} \varphi\left(\frac{x - x_i}{h}\right)$$

- Recall

$$p(x) \approx \frac{k/n}{V}$$

- Thus we get the desired analytical expression for the estimate of density

$$p_\varphi(x) = \frac{1}{n}\sum_{i=1}^{i=n} \frac{1}{h^d} \varphi\left(\frac{x - x_i}{h}\right)$$

Parzen Windows

$$p_\varphi(x) = \frac{1}{n}\sum_{i=1}^{i=n}\frac{1}{h^d}\varphi\left(\frac{x-x_i}{h}\right)$$

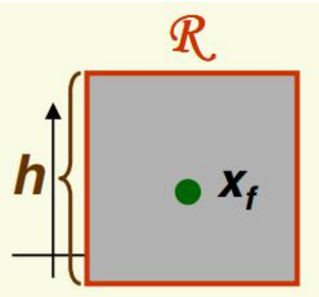
- Let’s make sure $p_\varphi(x)$ is in fact a density

$$p_\varphi(x) \geq 0 \qquad \forall x$$

$$\int p_\varphi(x)dx = \int \frac{1}{n}\sum_{i=1}^{i=n}\frac{1}{h^d}\varphi\left(\frac{x-x_i}{h}\right)dx$$

$$= \frac{1}{h^d n}\sum_{i=1}^{i=n}\overbrace{\int \varphi\left(\frac{x-x_i}{h}\right)dx}^{\text{volume of hypercube}}$$

$$= \frac{1}{n}\frac{1}{h^d}\sum_{i=1}^{i=n} h^d = 1$$

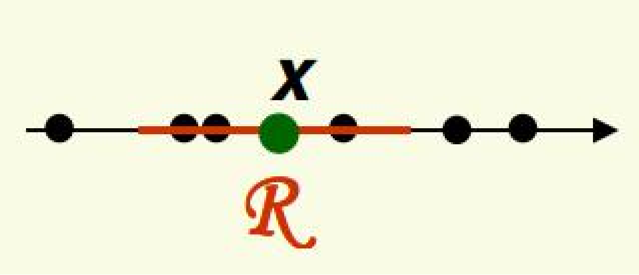


Parzen Windows

$$p(x) \approx \frac{k/n}{V}$$

x is inside some region ℜ
k = number of samples inside ℜ
n=total number of samples
V=volume of ℜ

- To estimate the density at point x, simply center the region ℜ at x, count the number of samples in ℜ , and substitute everything in our formula



$$p(x) \approx \frac{3/6}{10}$$

Parzen Windows

$$p(x) \approx \frac{k/n}{V}$$

x is inside some region ℜ
k = number of samples inside ℜ
n=total number of samples
V=volume of ℜ

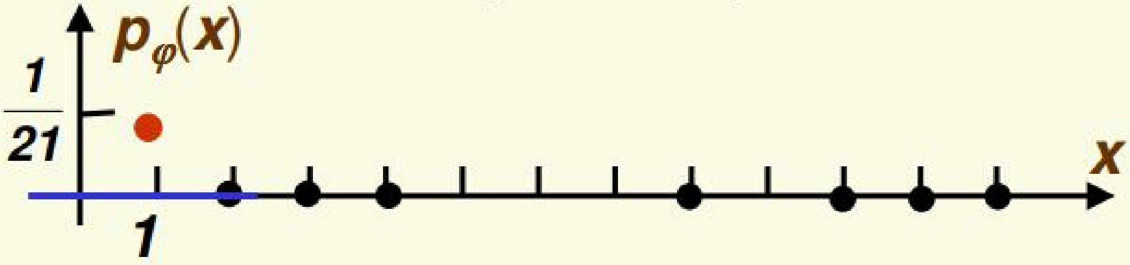
- Formula for Parzen window estimation

$$p_{\varphi}(x) = \frac{\overbrace{\sum_{i=1}^{i=n} \varphi\left(\frac{x - x_i}{h}\right)}^{=k} \Big/ n}{\underbrace{h^d}_{=V}} = \frac{1}{n}\sum_{i=1}^{i=n} \frac{1}{h^d}\,\varphi\left(\frac{x - x_i}{h}\right)$$

Parzen Windows: Example in 1D

$$p_\varphi(x) = \frac{1}{n}\sum_{i=1}^{i=n}\frac{1}{h^d}\varphi\left(\frac{x-x_i}{h}\right)$$

- Suppose we have 7 samples ***D***={2,3,4,8,10,11,12}



- Let window width ***h***=3, estimate density at x=1

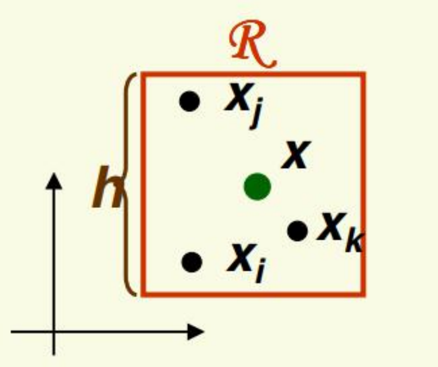
$$p_\varphi(1) = \frac{1}{7}\sum_{i=1}^{i=7}\frac{1}{3}\varphi\left(\frac{1-x_i}{3}\right) = \frac{1}{21}\left[\varphi\left(\frac{1-2}{3}\right)+\varphi\left(\frac{1-3}{3}\right)+\varphi\left(\frac{1-4}{3}\right)+\ldots+\varphi\left(\frac{1-12}{3}\right)\right]$$

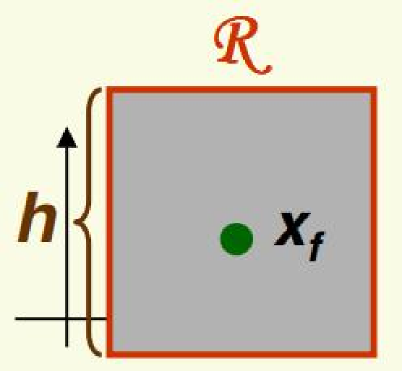
$\left|-\frac{1}{3}\right| \le 1/2$ $\left|-\frac{2}{3}\right| > 1/2$ $\left|-1\right| > 1/2$ $\left|-\frac{11}{3}\right| > 1/2$

$$p_\varphi(1) = \frac{1}{7}\sum_{i=1}^{i=7}\frac{1}{3}\varphi\left(\frac{1-x_i}{3}\right) = \frac{1}{21}[1+0+0+\ldots+0] = \frac{1}{21}$$

Parzen Windows: Sum of Functions

- Fix x, let i vary and ask:
  - For which samples xi is $\varphi\left(\frac{x - x_i}{h}\right) = 1$?





- Now fix f and let x vary and ask:
  - For which samples x is $\varphi\left(\frac{x - x_f}{h}\right) = 1$? For all x in gray box

- Thus $\varphi\left(\frac{x - x_f}{h}\right) = 1$ is simply a function which is 1 inside square of width h centered at xf and 0 otherwise!

Parzen Windows: Sum of Functions

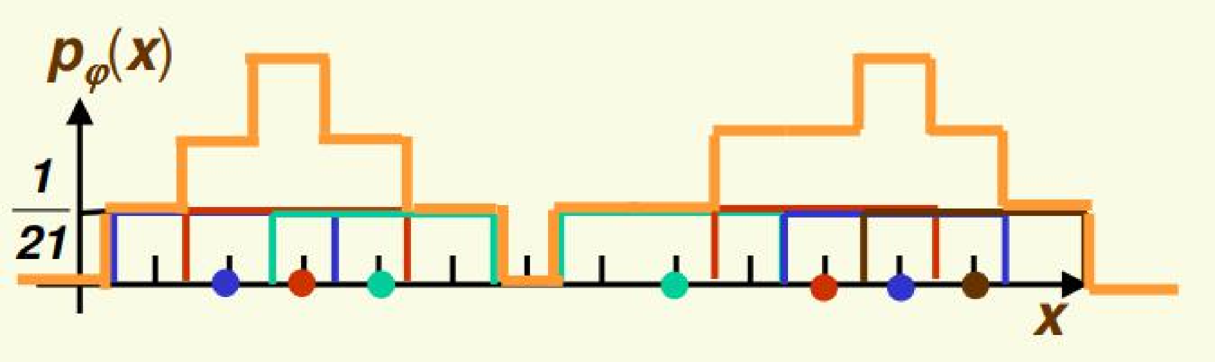
- Now let’s look at our density estimate again

$$p_\varphi(x) = \frac{1}{n}\sum_{i=1}^{i=n}\frac{1}{h^d}\varphi\left(\frac{x - x_i}{h}\right) = \sum_{i=1}^{i=n}\frac{1}{nh^d}\underbrace{\varphi\left(\frac{x - x_i}{h}\right)}_{\boxed{\begin{array}{ll}1 & \text{inside square centered at } x_i \\ 0 & \text{otherwise}\end{array}}}$$

- Thus $p_\varphi(x)$ is just a sum of *n* “box like” functions each of height $\frac{1}{nh^d}$

Parzen Windows: Example in 1D

- Let's come back to our example
  - 7 samples D={2,3,4,8,10,11,12}, h=3

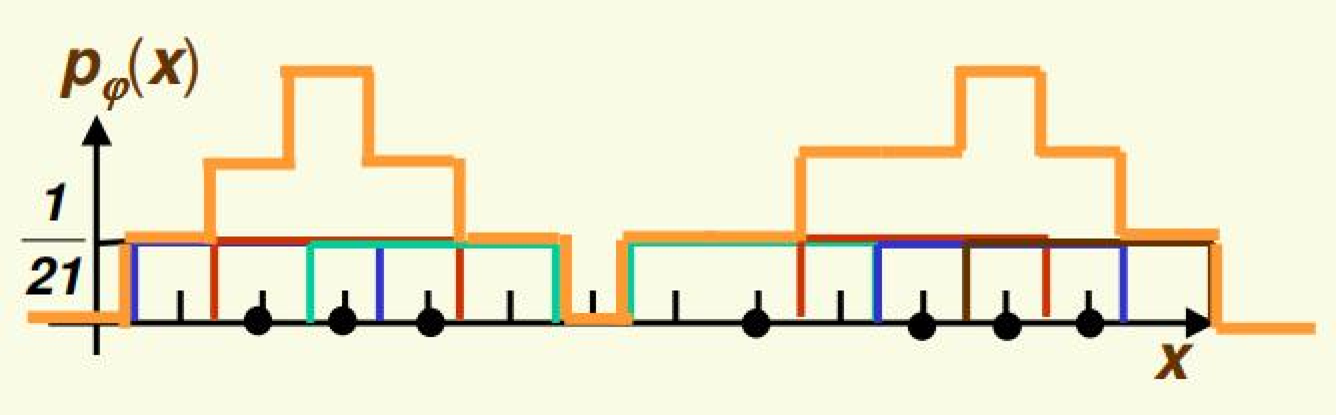


- To see what the function looks like, we need to generate 7 boxes and add them up.
- The width is h=3 and the height, according to previous slide is

$$\frac{1}{nh^d} = \frac{1}{21}$$

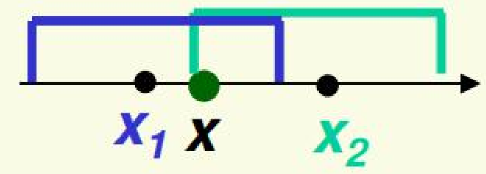
Parzen Windows: Interpolation

- In essence, window function φ is used for interpolation: each sample xi contributes to the resulting density at x if x is close enough to x_i



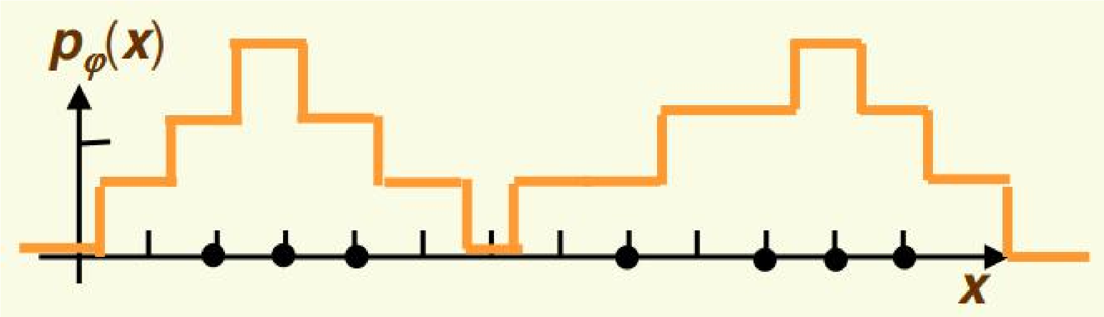
Parzen Windows: Drawbacks of Hypercube

- As long as sample point xi and x are in the same hypercube, the contribution of xi to the density at x isconstant, regardless of how close xi is to x



$$\varphi\left(\frac{x-x_1}{h}\right)=\varphi\left(\frac{x-x_2}{h}\right)=1$$

- The resulting density $p_\varphi(x)$ is not smooth, it has discontinuities

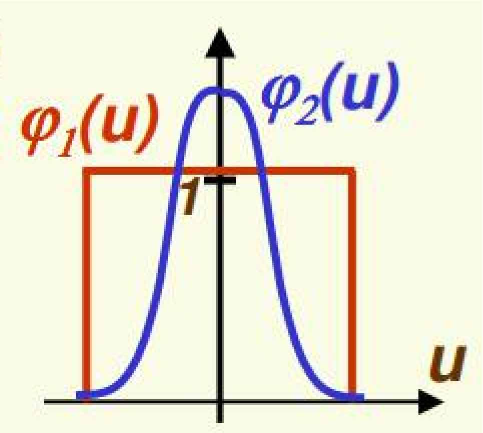


Parzen Windows: general window function φ

$$p_\varphi(x) = \frac{1}{n}\sum_{i=1}^{i=n} \frac{1}{h^d}\varphi\left(\frac{x - x_i}{h}\right)$$

- We can use a general window φ as long as the resulting $p_\varphi(x)$ is a legitimate density, i.e.,

1. $p_\varphi(u) \geq 0$
   - satisfied if $\varphi(u) \geq 0$
2. $\int p_\varphi(x)dx = 1$
   - satisfied if $\int \varphi(u)du = 1$



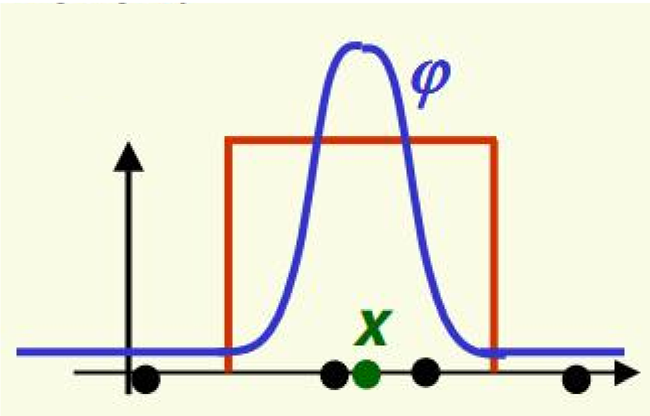
$$\int p_\varphi(x)dx = \frac{1}{nh^d}\sum_{i=1}^{i=n}\int \varphi\left(\frac{x - x_i}{h}\right)dx = \frac{1}{nh^d}\sum_{i=1}^{n}\int h^d \varphi(u)du = 1$$

$$\boxed{\text{change coordinates to } u = \frac{x - x_i}{h}, \quad \text{thus } du = \frac{dx}{h}}$$

Parzen Windows: general window function φ

$$p_\varphi(x) = \frac{1}{n}\sum_{i=1}^{i=n}\frac{1}{h^d}\varphi\left(\frac{x-x_i}{h}\right)$$

- Notice that with the general window φ we are no longer counting the number of samples inside $\mathfrak{R}$
- We are counting the weighted average of potentially every single sample point (although only those within distance h have any significant weight)



- With infinite number of samples, and appropriate conditions, it can still be shown that

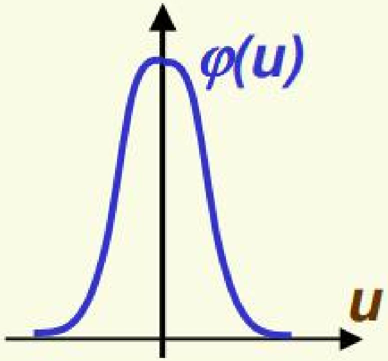
$$p_\varphi^n(x) \to p(x)$$

Parzen Windows: Gaussian φ

$$p_\varphi(x) = \frac{1}{n}\sum_{i=1}^{i=n}\frac{1}{h^d}\varphi\left(\frac{x - x_i}{h}\right)$$

- A popular choice for φ is N(0,1) density

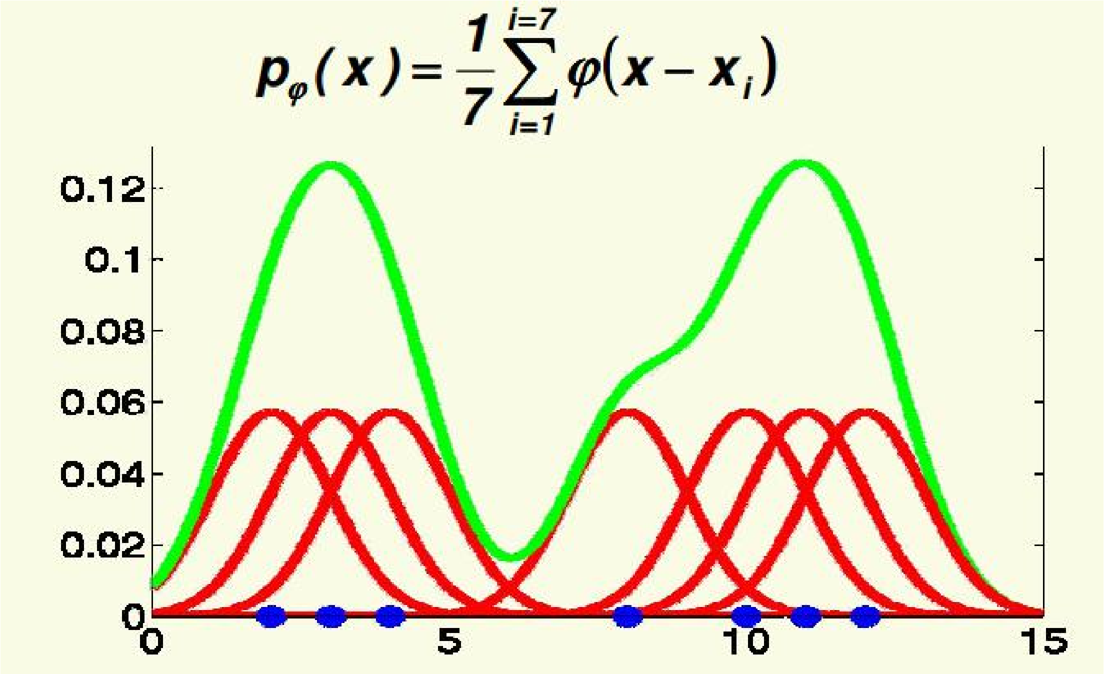
$$\varphi(u) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2}$$



- Solves both drawbacks of the “box” window
  - Points x which are close to the sample point xi receive higher weight
  - Resulting density $p_\varphi(x)$ is smooth

Example with General φ

- Let's come back to our example
  - 7 samples D={2,3,4,8,10,11,12}, h=1



$$p_\varphi(x) = \frac{1}{7}\sum_{i=1}^{i=7}\varphi(x - x_i)$$

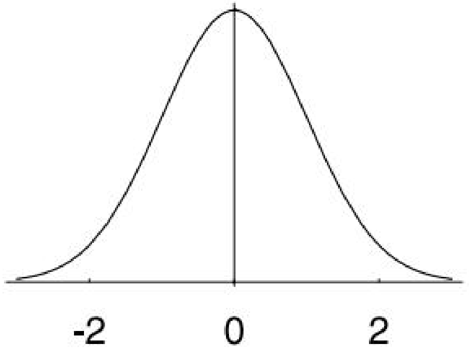
- $p_\varphi(x)$ is the sum of of 7 Gaussians, each centered at one of the sample points, and each scaled by 1/7

Did We Solve the Problem?

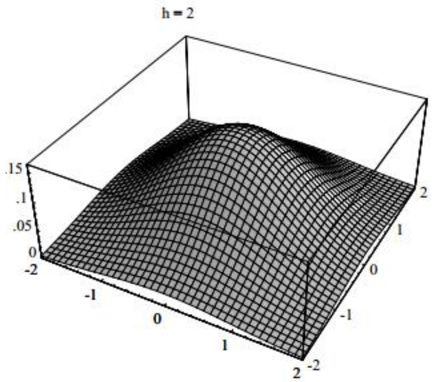
- Let’s test if we solved the problem
  - ✓ 1. Draw samples from a known distribution
  - ✓ 2. Use our density approximation method and compare with the true density

- We will vary the number of samples n and the window size h

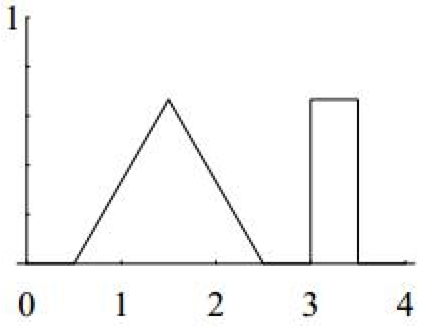
- We will play with 3 distributions



Univariate normal density

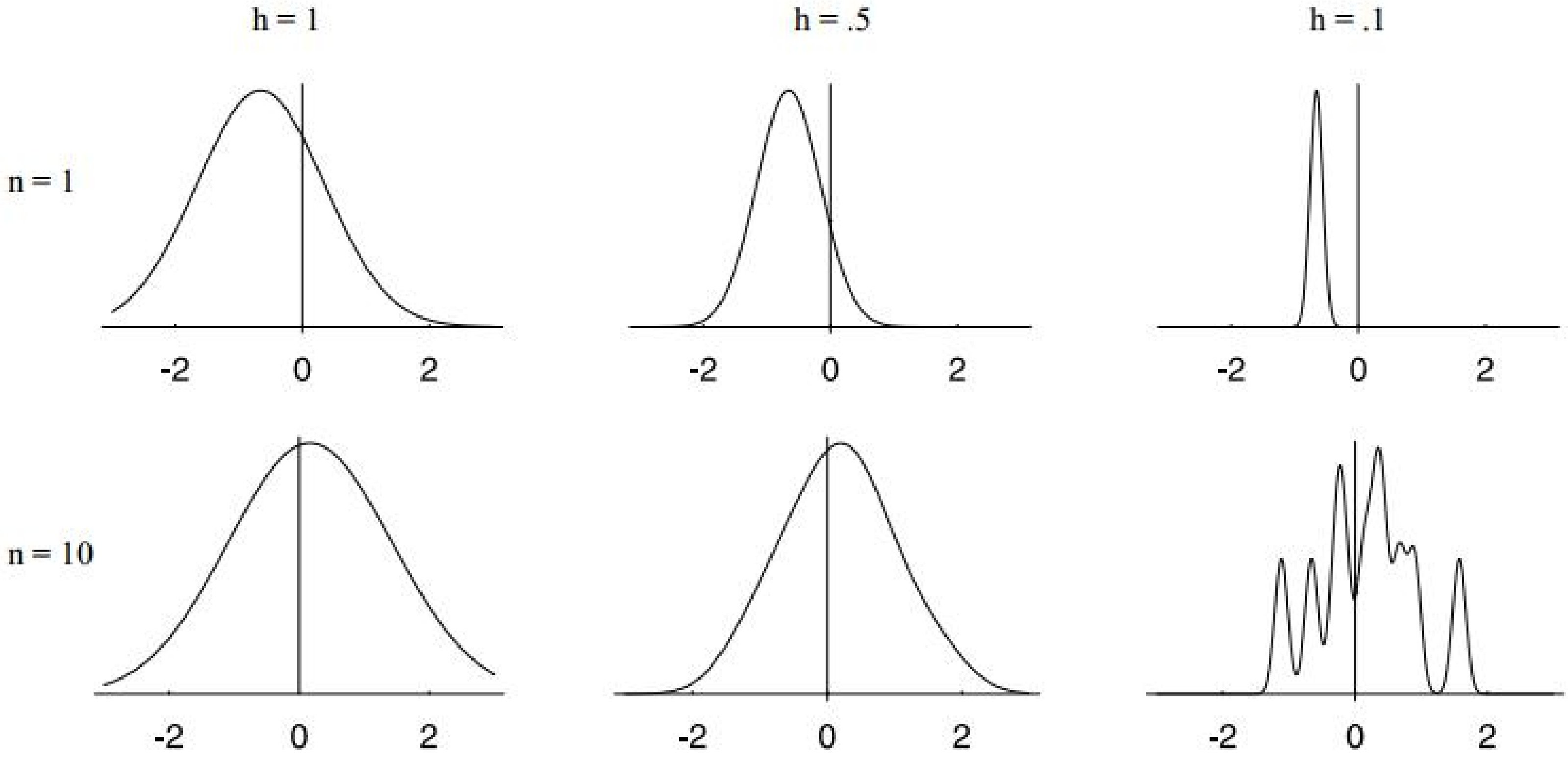


Bivariate normal density



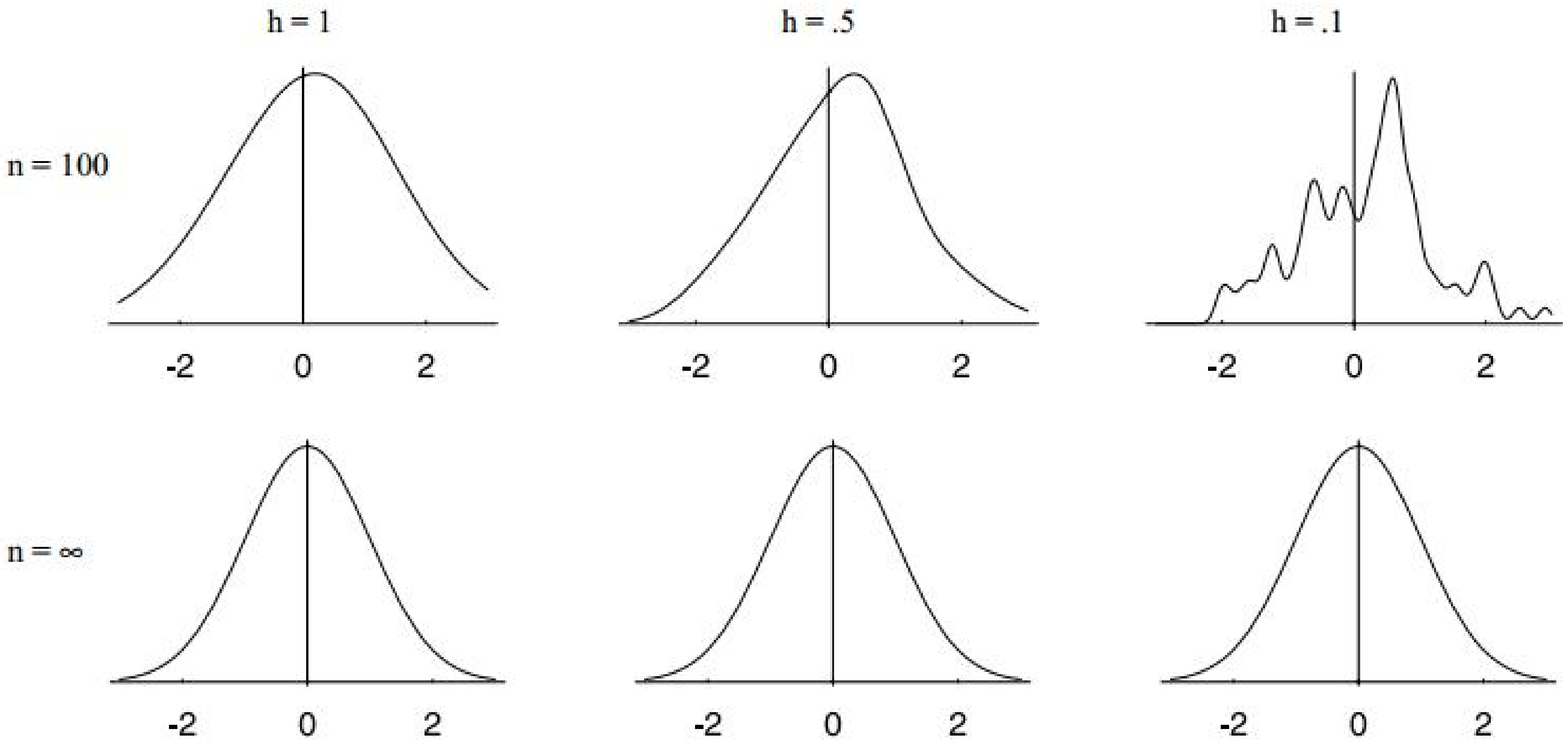
Mixture of Uniform and Triangle

Parzen-window estimates of a univariate normal density (1)



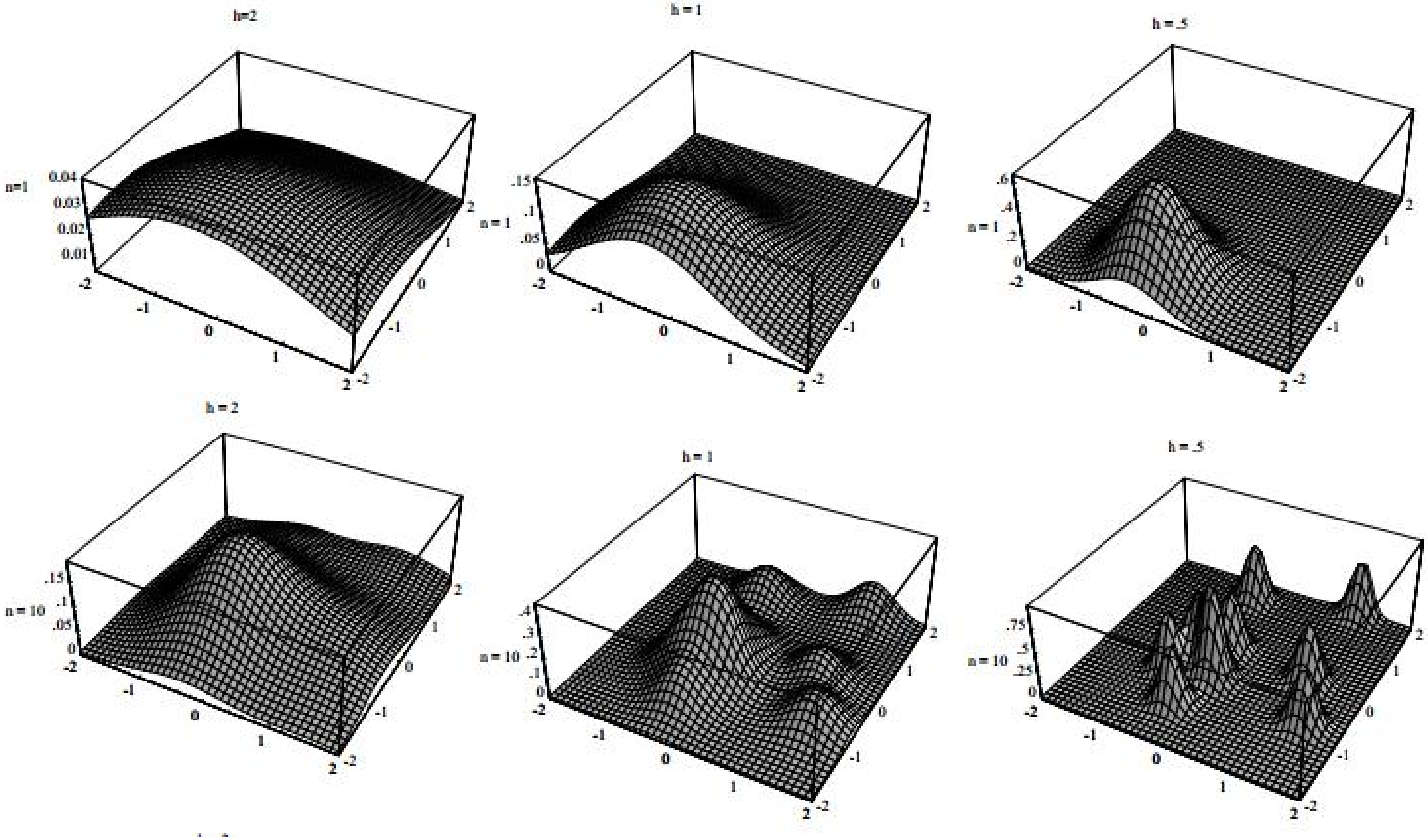
$$p_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\varphi\Big(\frac{x - x_i}{h_n}\Big)$$

Parzen-window estimates of a univariate normal density (2)



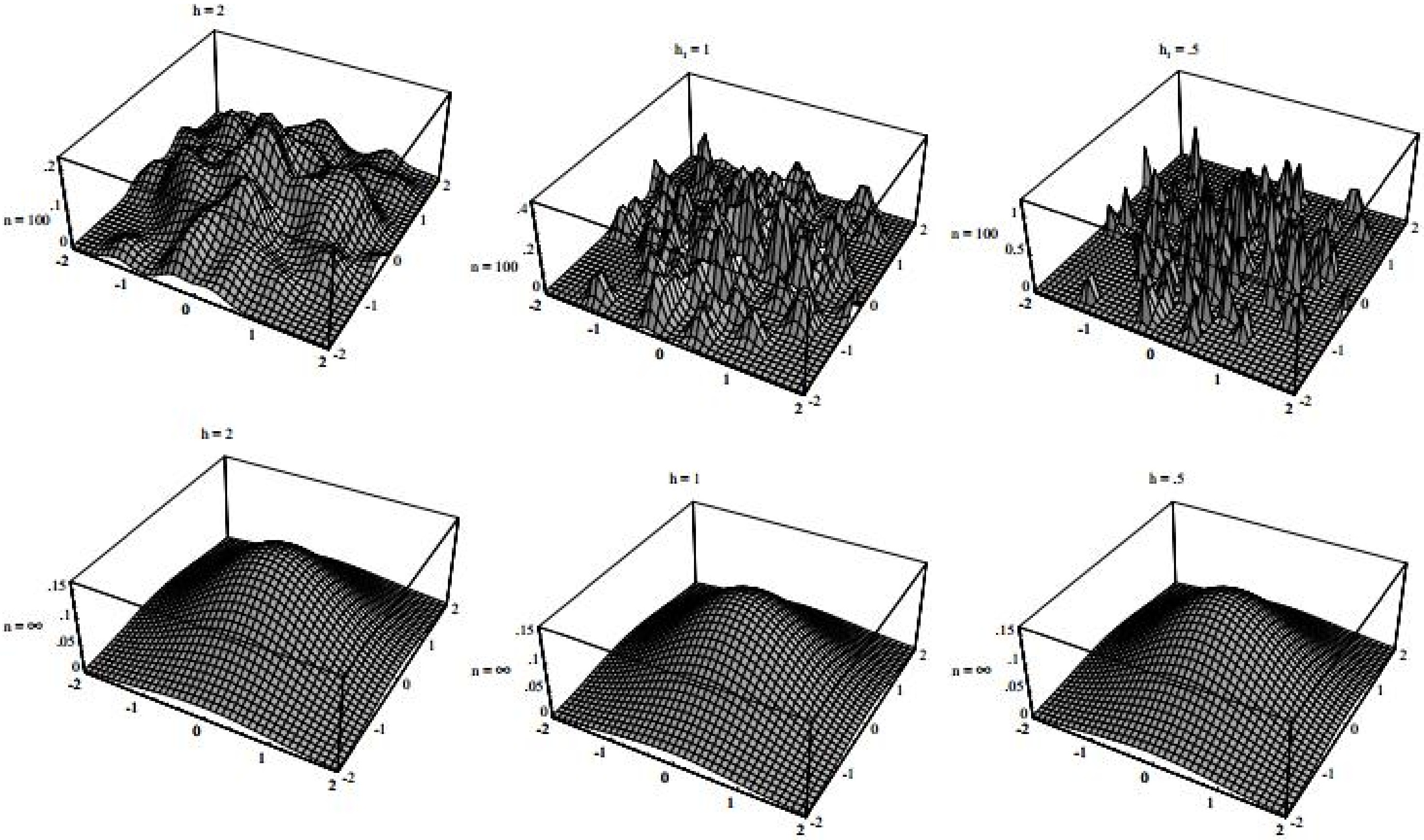
$$p_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\varphi\left(\frac{x-x_i}{h_n}\right)$$

Parzen-window estimates of a bivariate normal density (1)



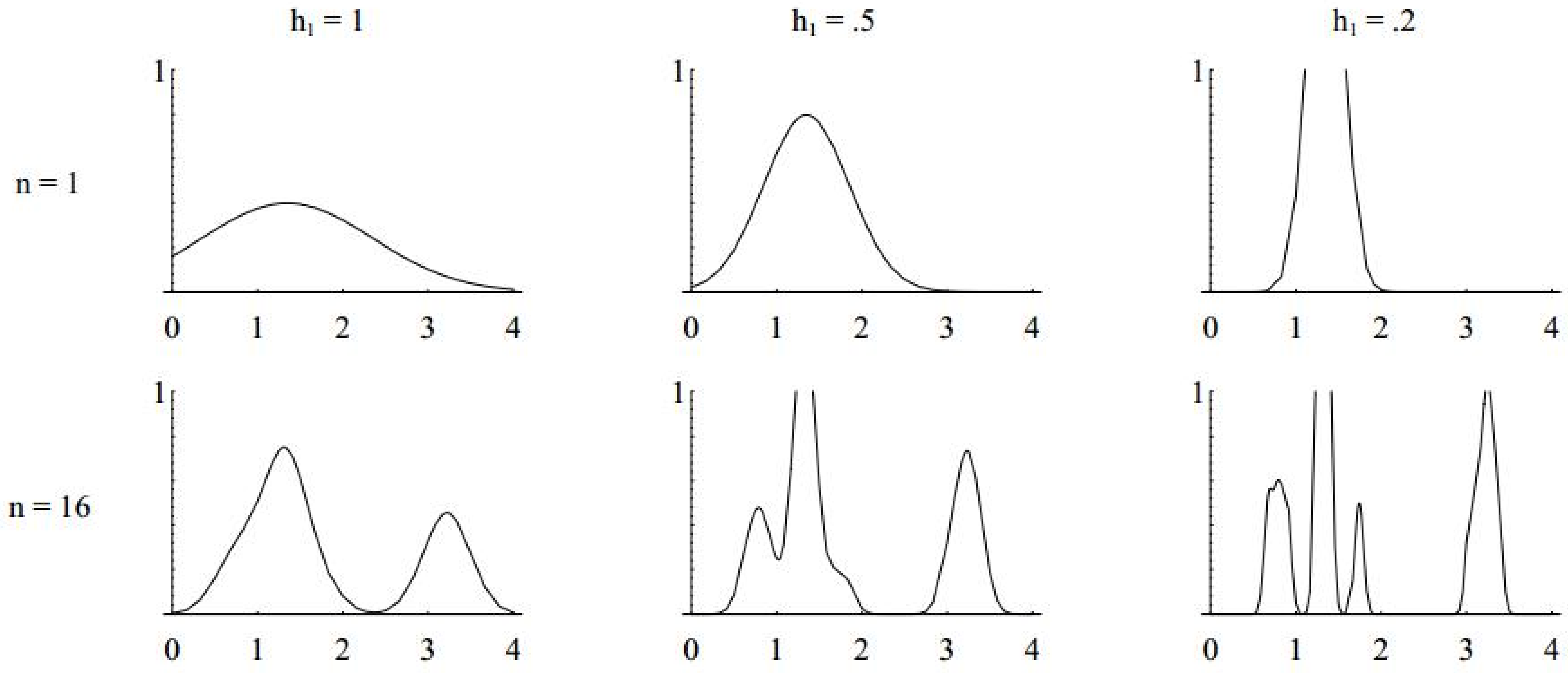
$$p_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\varphi\left(\frac{x - x_i}{h_n}\right)$$

Parzen-window estimates of a bivariate normal density (2)



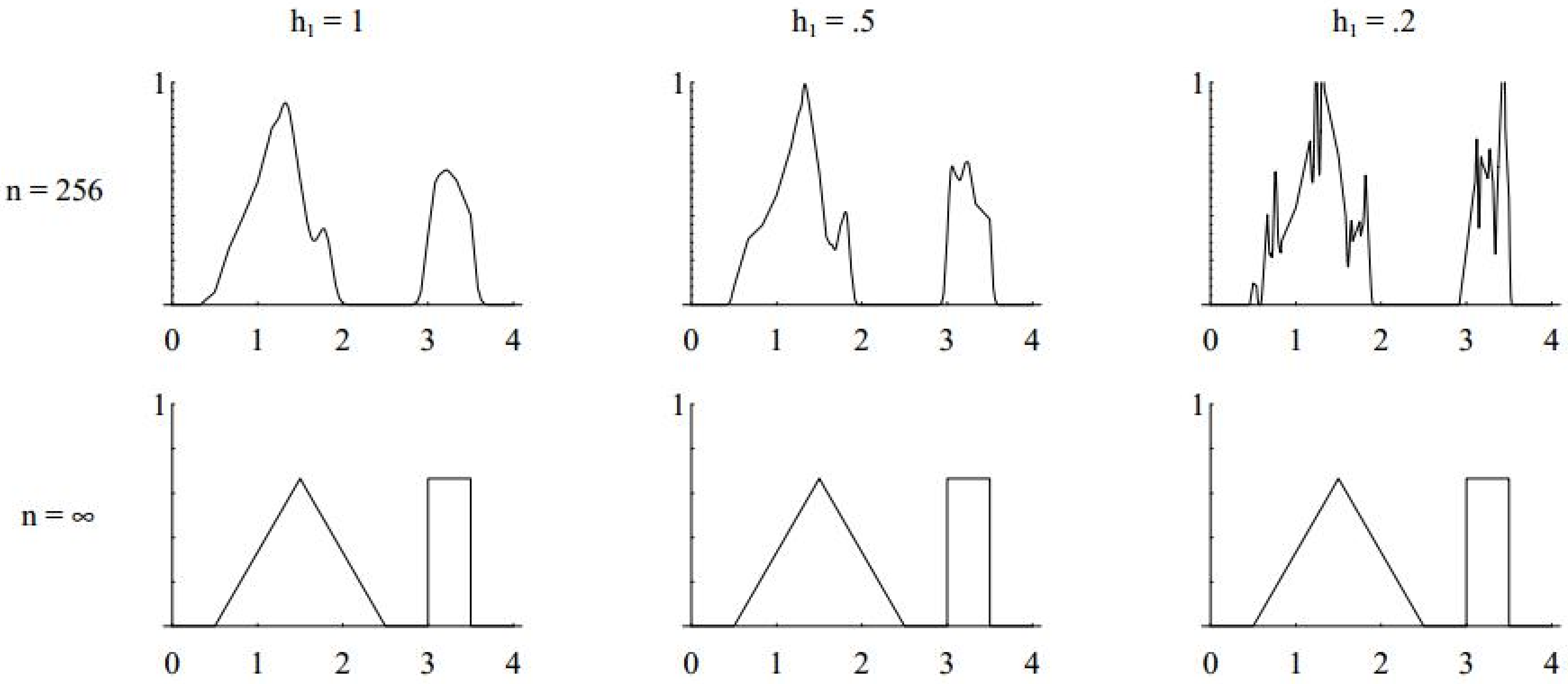
$$p_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\varphi\Big(\frac{x - x_i}{h_n}\Big)$$

Parzen-window estimates of a bimodal distribution (1)



Unknown density, mixture of a uniform and a triangle density

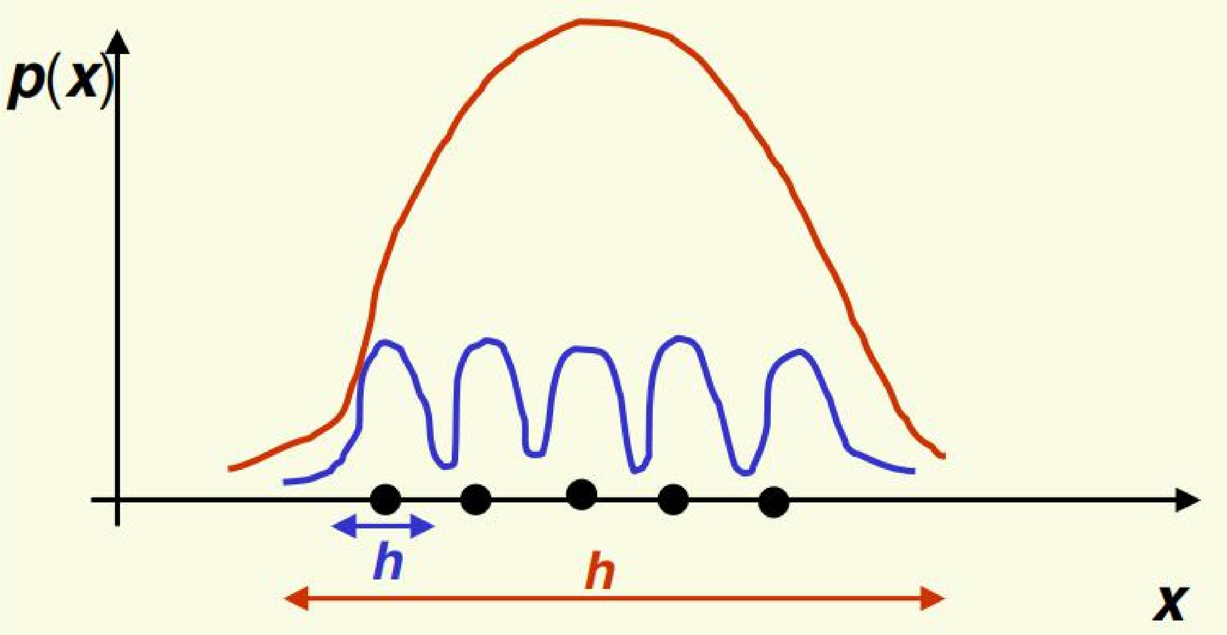
Parzen-window estimates of a bimodal distribution (2)



Unknown density, mixture of a uniform and a triangle density

Parzen Windows: Effect of Window Width h

- By choosing h we are guessing the region where density is approximately constant
- Without knowing anything about the distribution, it is really hard to guess were the density is approximately constant



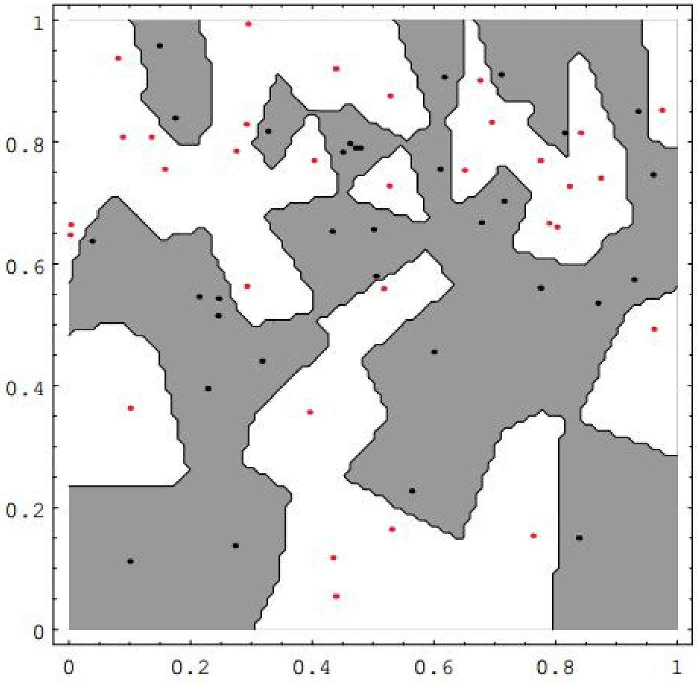
Parzen Windows: Effect of Window Width h

- If h is small, we superimpose n sharp pulses centered at the data
  - Each sample point xi influences too small range of x
  - Smoothed too little: the result will look noisy and not smooth enough
- If h is large, we superimpose broad slowly changing functions,
  - Each sample point xi influences too large range of x
  - Smoothed too much: the result looks oversmoothed or "out of focus"
- Finding the best h is challenging, and indeed no single h may work well
  - May need to adapt h for different sample points
- However we can try to learn the best h to use from the test data

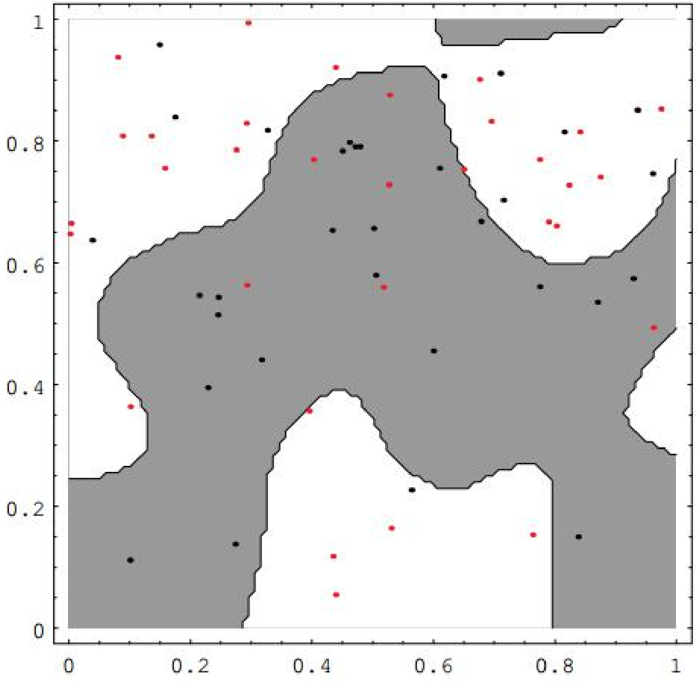
Parzen Windows: Classification Example

- In classifiers based on Parzen window estimation:
  - – We estimate the densities for each category and classify a test point by the label corresponding to the maximum posterior
  - – The decision region for a Parzen window classifier depends upon the choice of window function as illustrated in the following figure

Parzen Windows: Classification Example



- For **small** enough window size h is classification on training data is be perfect
- However decision boundaries are complex and this solution is not likely to generalize well to novel data



- For **larger** window size h, classification on training data is not perfect
- However decision boundaries are simpler and this solution is more likely to generalize well to novel data

Parzen Windows: Summary

- Advantages
  - Can be applied to the data from any distribution
  - In theory can be shown to converge as the number of samples goes to infinity
- Disadvantages
  - Number of training data is limited in practice, and so choosing the appropriate window size h is difficult
  - May need large number of samples for accurate estimates
  - Computationally heavy, to classify one point we have to compute a function which potentially depends on all samples
  - Window size h is not trivial to choose

Q & A