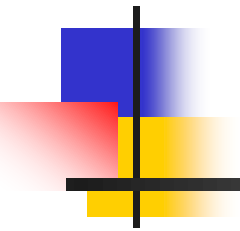


Nonparametric Techniques

Introduction

- Neither probability distribution nor discriminant function is known
- All we have is labeled data

=> Estimate the probability distribution from the labeled data

Introduction

- Fundamental techniques
- Parzen Windows
- Nearest Neighbors

Introduction

- Parametric methods– known "forms of the density functions"
- The sad part → in most PR applications this assumption is NOT sure
- What is needed → to examine *nonparametric* procedures/techniques
  - that can be used with arbitrary distributions, and
  - without the assumption that the forms of density functions are known

Introduction

- Nonparametric techniques:
  - Parzen windows:
    => Estimate likelihood $p(\mathbf{x}|C_j)$
  - Nearest neighbors
    => Bypass likelihood and go directly to posterior estimation $p(\mathbf{x}|C_j)$

5.1 INTRODUCTION

In Chapter 4, we considered discrete random variables—that is, random variables whose set of possible values is either finite or countably infinite. However, there also exist random variables whose set of possible values is uncountable. Two examples are the time that a train arrives at a specified stop and the lifetime of a transistor. Let X be such a random variable. We say that X is a *continuous*[†] random variable if there exists a nonnegative function f, defined for all real $x \in (-\infty, \infty)$, having the property that, for any set B of real numbers,[‡]

$$P\{X \in B\} = \int_B f(x)\,dx \tag{1.1}$$

The function f is called the *probability density function* of the random variable X. (See Figure 5.1.)

In words, Equation (1.1) states that the probability that X will be in B may be obtained by integrating the probability density function over the set B. Since X must assume some value, f must satisfy

$$1 = P\{X \in (-\infty, \infty)\} = \int_{-\infty}^{\infty} f(x)\,dx$$

All probability statements about X can be answered in terms of f. For instance, from Equation (1.1), letting $B = [a, b]$, we obtain

$$P\{a \le X \le b\} = \int_a^b f(x)\,dx \tag{1.2}$$

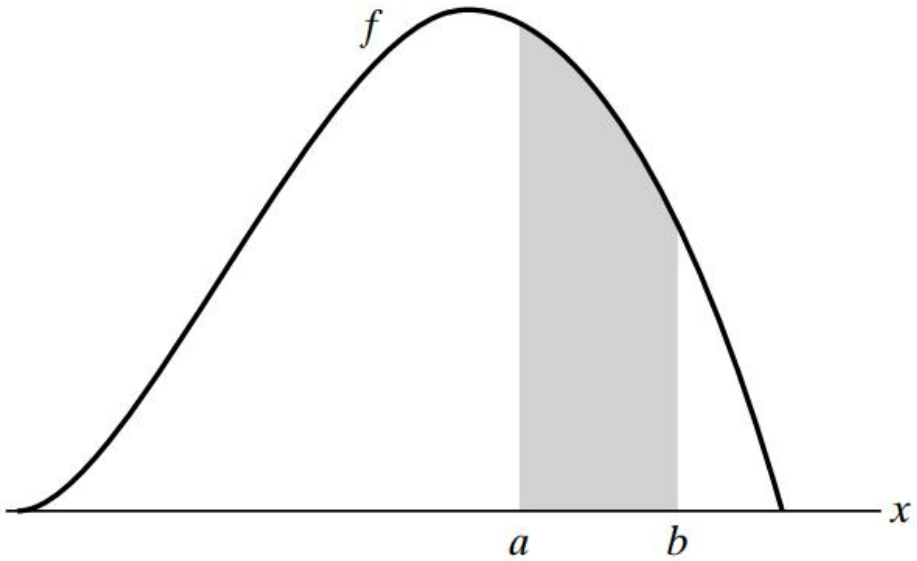


FIGURE 5.1: Probability density function f.

If we let $a = b$ in Equation (1.2), we get

$$P\{X = a\} = \int_a^a f(x)\,dx = 0$$

In words, this equation states that the probability that a continuous random variable will assume any fixed value is zero. Hence, for a continuous random variable,

$$P\{X < a\} = P\{X \leq a\} = F(a) = \int_{-\infty}^{a} f(x)\,dx$$

EXAMPLE 1a

Suppose that X is a continuous random variable whose probability density function is given by

$$f(x) = \begin{cases} C(4x - 2x^2) & 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

(a) What is the value of C?

(b) Find $P\{X > 1\}$.

Solution. **(a)** Since f is a probability density function, we must have $\int_{-\infty}^{\infty} f(x)\,dx = 1$, implying that

$$C\int_0^2 (4x - 2x^2)\,dx = 1$$

or

$$C\left[2x^2 - \frac{2x^3}{3}\right]\Bigg|_{x=0}^{x=2} = 1$$

or

$$C = \frac{3}{8}$$

Hence,

(b) $P\{X > 1\} = \int_1^{\infty} f(x)\,dx = \frac{3}{8}\int_1^2 (4x - 2x^2)\,dx = \frac{1}{2}$

■

EXAMPLE 1b

The amount of time in hours that a computer functions before breaking down is a continuous random variable with probability density function given by

$$f(x) = \begin{cases} \lambda e^{-x/100} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

What is the probability that

(a) a computer will function between 50 and 150 hours before breaking down?
(b) it will function for fewer than 100 hours?

Solution. (a) Since

$$1 = \int_{-\infty}^{\infty} f(x)\,dx = \lambda \int_{0}^{\infty} e^{-x/100}\,dx$$

we obtain

$$1 = -\lambda(100)e^{-x/100}\Big|_0^{\infty} = 100\lambda \quad \text{or} \quad \lambda = \frac{1}{100}$$

Hence, the probability that a computer will function between 50 and 150 hours before breaking down is given by

$$\begin{aligned} P\{50 < X < 150\} &= \int_{50}^{150} \frac{1}{100} e^{-x/100}\,dx = -e^{-x/100}\Big|_{50}^{150} \\ &= e^{-1/2} - e^{-3/2} \approx .384 \end{aligned}$$

(b) Similarly,

$$P\{X < 100\} = \int_{0}^{100} \frac{1}{100} e^{-x/100}\,dx = -e^{-x/100}\Big|_0^{100} = 1 - e^{-1} \approx .633$$

In other words, approximately 63.3 percent of the time, a computer will fail before registering 100 hours of use. ■

EXPECTATION AND VARIANCE OF CONTINUOUS RANDOM VARIABLES

In Chapter 4, we defined the expected value of a discrete random variable X by

$$E[X] = \sum_x xP\{X = x\}$$

If X is a continuous random variable having probability density function $f(x)$, then, because

$$f(x)\,dx \approx P\{x \leq X \leq x + dx\} \quad \text{for } dx \text{ small}$$

it is easy to see that the analogous definition is to define the expected value of X by

$$E[X] = \int_{-\infty}^{\infty} xf(x)\,dx$$

EXAMPLE 2a

Find $E[X]$ when the density function of X is

$$f(x) = \begin{cases} 2x & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Solution.

$$\begin{aligned} E[X] &= \int xf(x)\,dx \\ &= \int_0^1 2x^2\,dx \\ &= \frac{2}{3} \end{aligned}$$

■

EXAMPLE 2a

Find $E[X]$ when the density function of X is

$$f(x) = \begin{cases} 2x & \text{if} \quad 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Solution.

$$\begin{aligned} E[X] &= \int x f(x)\, dx \\ &= \int_0^1 2x^2\, dx \\ &= \frac{2}{3} \end{aligned}$$

EXAMPLE 2b

The density function of X is given by

$$f(x) = \begin{cases} 1 & \text{if} \quad 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Find $E[e^X]$.

EXPONENTIAL RANDOM VARIABLES

A continuous random variable whose probability density function is given, for some $\lambda > 0$, by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

is said to be an *exponential* random variable (or, more simply, is said to be exponentially distributed) with parameter λ. The cumulative distribution function $F(a)$ of an exponential random variable is given by

$$\begin{aligned} F(a) &= P\{X \leq a\} \\ &= \int_0^a \lambda e^{-\lambda x}\, dx \\ &= -e^{-\lambda x}\Big|_0^a \\ &= 1 - e^{-\lambda a} \quad a \geq 0 \end{aligned}$$

Note that $F(\infty) = \int_0^\infty \lambda e^{-\lambda x}\, dx = 1$, as, of course, it must. The parameter λ will now be shown to equal the reciprocal of the expected value.

EXAMPLE 5a

Let X be an exponential random variable with parameter λ. Calculate (a) $E[X]$ and (b) $\text{Var}(X)$.

Solution. (a) Since the density function is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

we obtain, for $n > 0$,

$$E[X^n] = \int_0^\infty x^n \lambda e^{-\lambda x}\, dx$$

Integrating by parts (with $\lambda e^{-\lambda x} = dv$ and $u = x^n$) yields

$$\begin{aligned} E[X^n] &= -x^n e^{-\lambda x}|_0^\infty + \int_0^\infty e^{-\lambda x} n x^{n-1}\, dx \\ &= 0 + \frac{n}{\lambda} \int_0^\infty \lambda e^{-\lambda x} x^{n-1}\, dx \\ &= \frac{n}{\lambda} E[X^{n-1}] \end{aligned}$$

Letting $n = 1$ and then $n = 2$ gives

$$E[X] = \frac{1}{\lambda}$$

$$E[X^2] = \frac{2}{\lambda} E[X] = \frac{2}{\lambda^2}$$

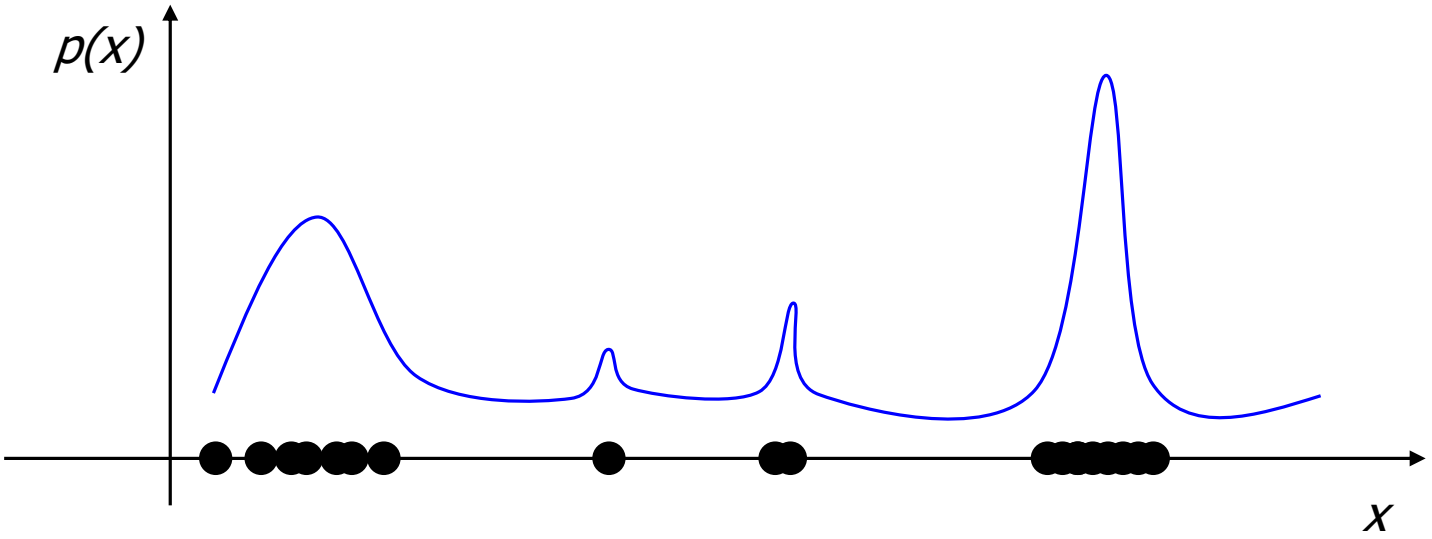
(b) Hence,

$$\text{Var}(X) = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

Thus, the mean of the exponential is the reciprocal of its parameter λ, and the variance is the mean squared. ■

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



$P_r(x \in R) = \int_{\mathcal{R}} p(\mathbf{x}')\, d\mathbf{x}'$

Avarage of p(x) on R

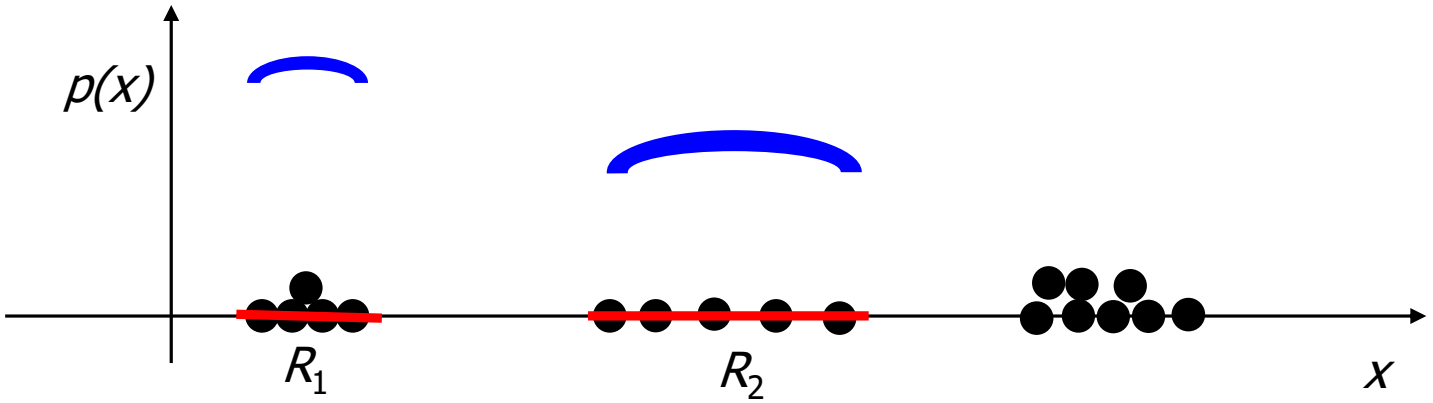
Introduction

$$P_r(x \in R) = \int_{\mathcal{R}} p(\mathbf{x}')\, d\mathbf{x}'$$

- How can we estimate $P_r(x \in R_1)$, $P_r(x \in R_2)$?

  $=>????$

- The density curves above R1 and R2 are equally high?
  - No, since R1 smaller than R2
  - To get density, normalize by region size



Fundamental techniques

- **The fundamental techniques**
  - **The fact that: the probability *P* that a vector x will fall in a region *R***

$$P = \int_R p(x')dx' \approx p(x) * V$$

P is smoothed/averaged version of density function *p(x)* ➔ *p* can be estimated by estimating P

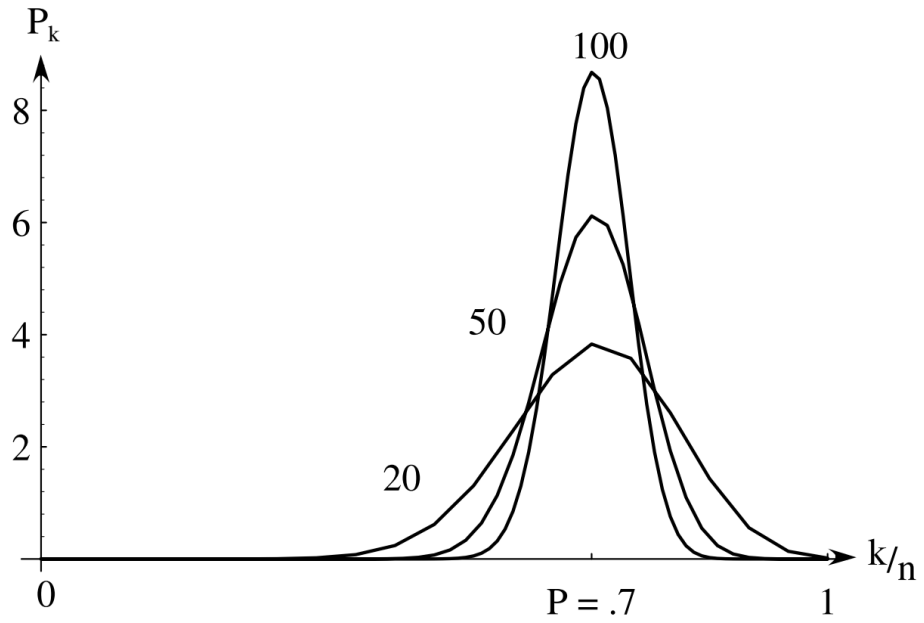
- **If *k* of *n* i.i.d samples are drawn to fall in R → probability follows the binomial law**

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

- **The expected value for *k***

$$\mathcal{E}[k] = nP$$

Density Estimation

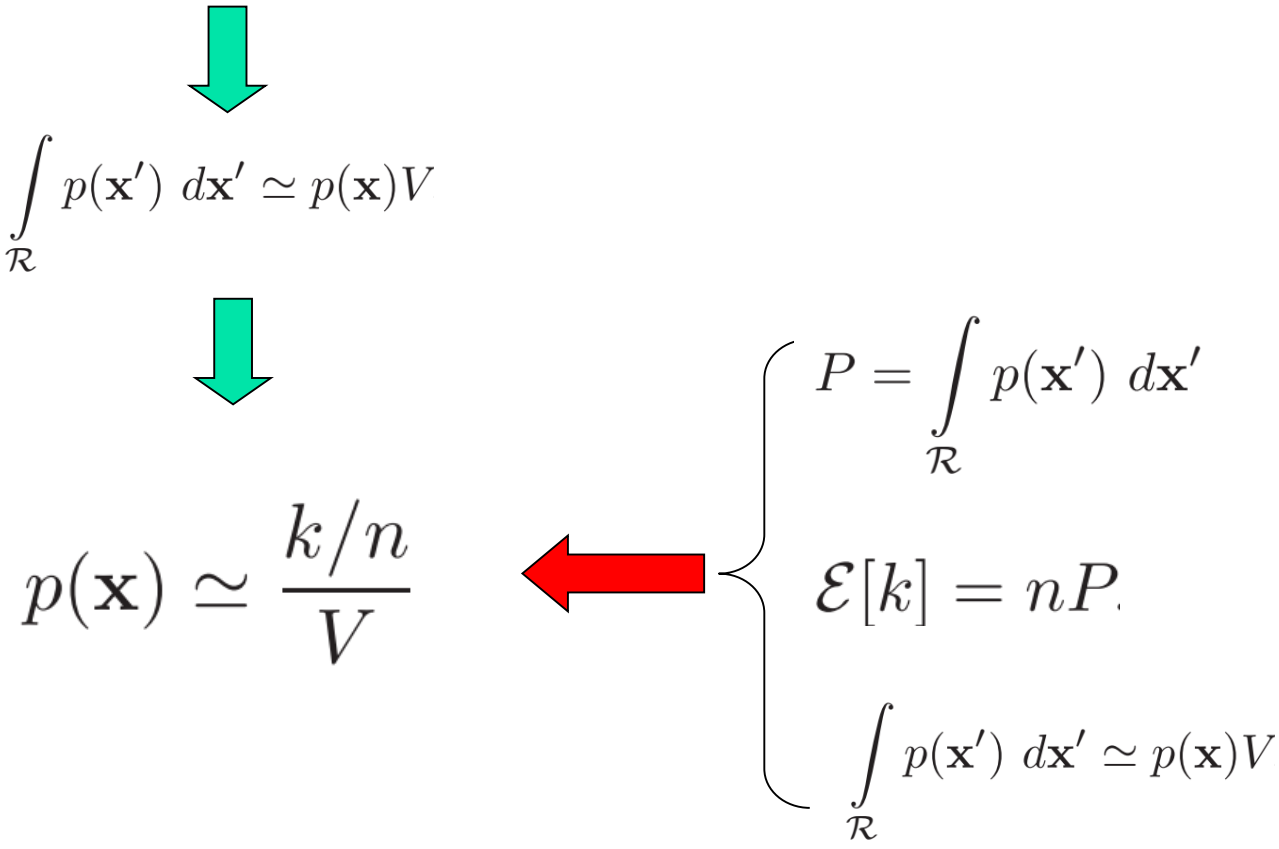


The probability P_k of finding k patterns in a volume where the space averaged probability is P as a function of k/n. Each curve is labelled by the total number of patterns n. For large n, such binomial distributions peak strongly at $k/n = P$ (here chosen to be 0.7)

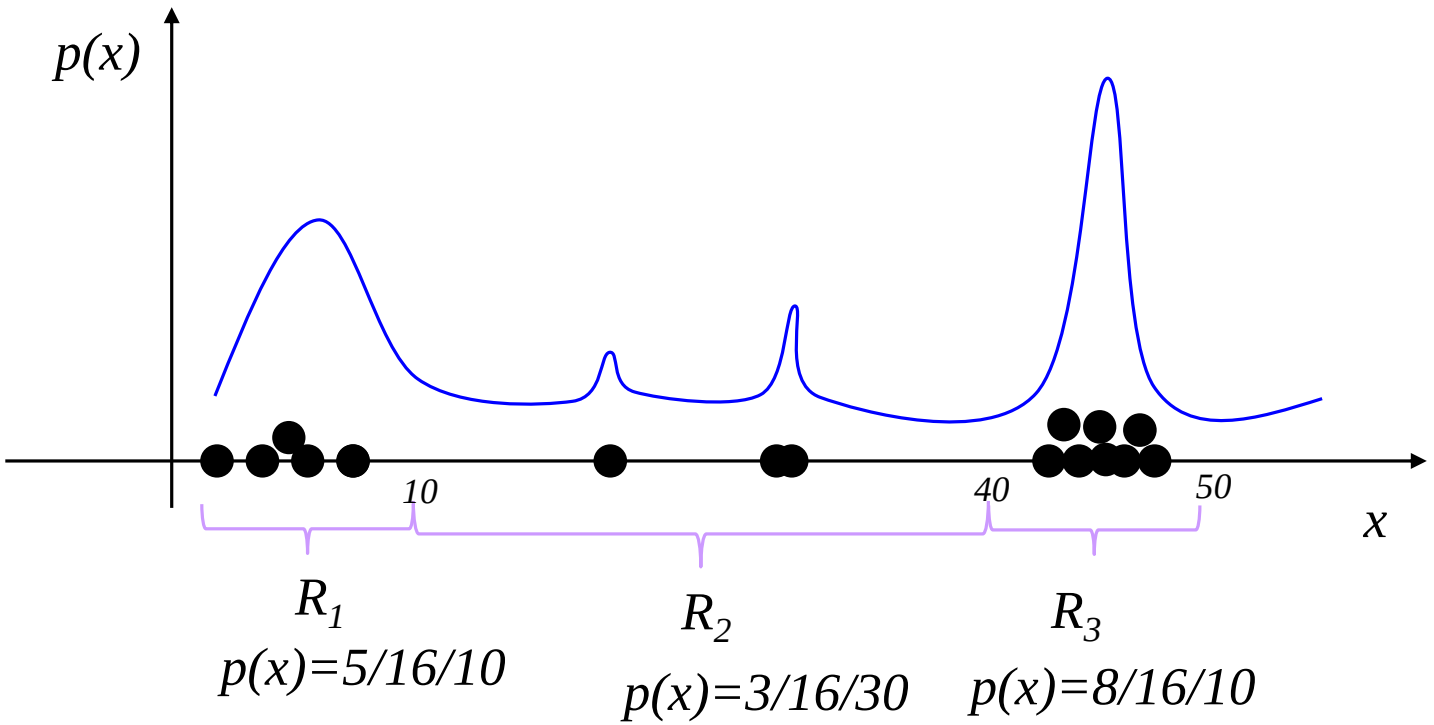
the ratio k/n will be a very good estimate for the probability P

Density Estimation

For continuous *p*(x); very small *R*, *p* does not vary appreciably within *R*



Density Estimation



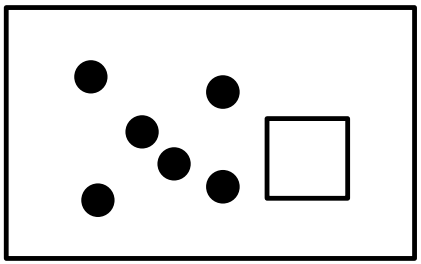
If regions R_i's do not overlap, we have a histogram

Density Estimation

Problem with: $p(\mathbf{x}) \simeq \frac{k/n}{V}$

- **How accurate is density approximation**?
- Two approximations:
  - *k/n* → converge (in probability): as n increase, this estimate becomes more accurate.

- $\int_R p(x')dx' \approx p(x) * V$, as *R* shrinks smaller, the estimate becomes more accurate. As we shrink *R* we have to make sure it contains samples, otherwise our estimated $\boldsymbol{p}(\boldsymbol{x}) = 0$ for all $\boldsymbol{x}$ in R



Density Estimation

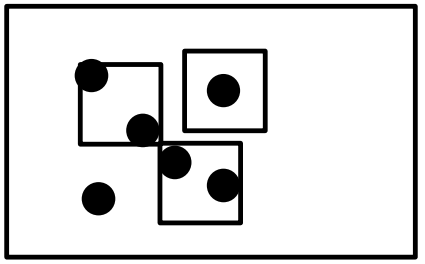
Thus in theory, if we have an unlimited number of samples, to we get convergence as we simultaneously increase the number of samples ***n***, and shrink region *R*, but not too much so that *R* still contains a lot of samples

Density Estimation

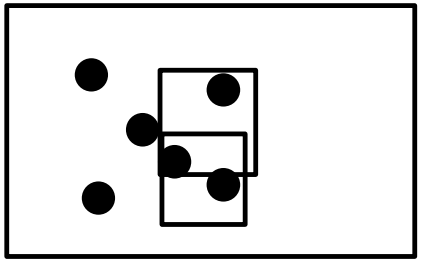
- In practice, the number of samples is fixed.
- Thus, an available option is to decrease the size of R (V).
  - If V is too small, $p(x)=0$ for most x, because most regions will have no samples
  - => Have to find a compromise V
    - not too small so that it has enough samples
    - but also not too large so that $\boldsymbol{p}(\boldsymbol{x})$ is approximately constant inside $\boldsymbol{V}$

Density Estimation

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



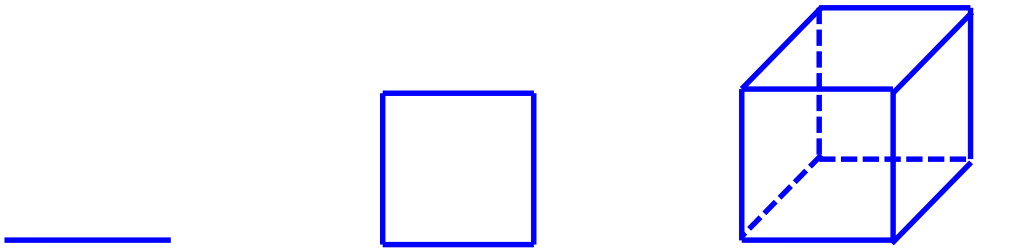
- K-nearest neighbors
  - Choose a fixed value for ***k*** and determine the corresponding volume ***V*** from the data



- **Both the methods converge, BUT**
- **Their finite-sample behavior can't be easily predicted**

Parzen Windows

- The size and shape of region ***R*** are fixed
- Let us assume that the region ***R*** is a d-dimensional hypercube with side length h thus it's volume is h^d

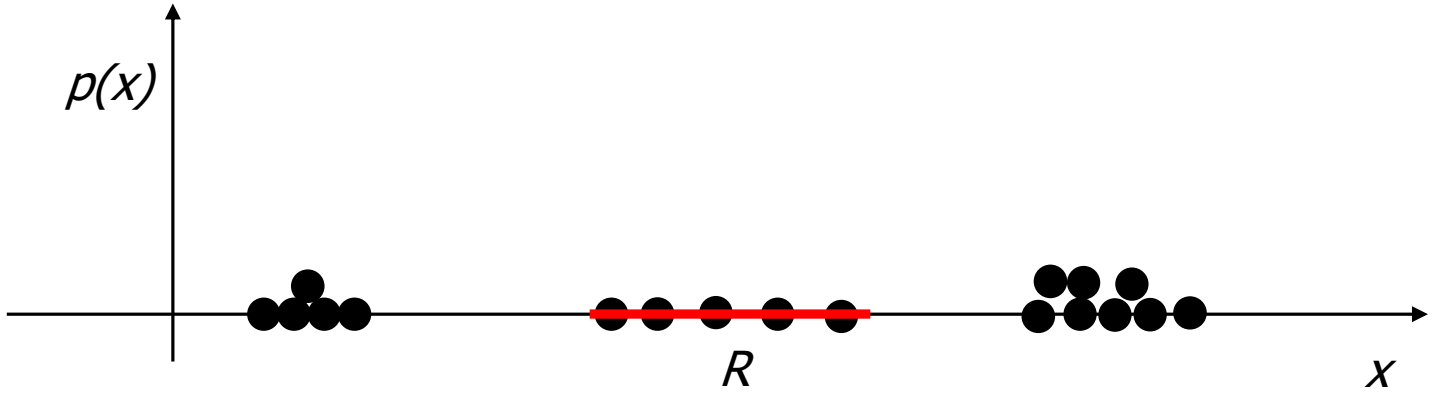


1 dimension 2 dimensions 3 dimensions

Parzen Windows

- 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

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

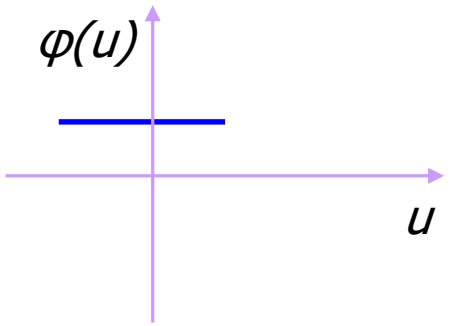


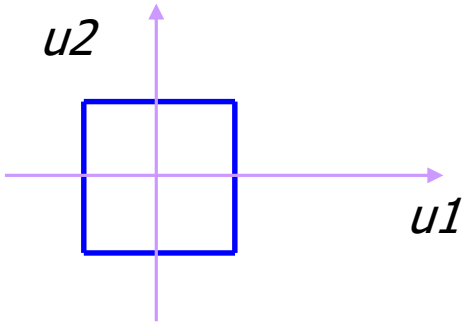
P(x)=5/18/10

Parzen Windows

- Let us define a window function

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



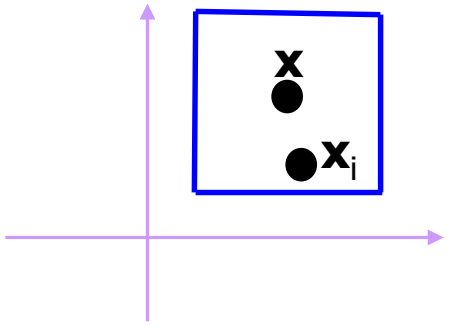


Parzen Windows

- If we have samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$

then

$$\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) = \begin{cases} 1 & \text{if } \left|\mathbf{x}-\mathbf{x}_i\right| \le \frac{h}{2} \\ 0 & \text{otherwise} \end{cases}$$



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

Parzen Windows

- How do we count the total number of sample points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ which are inside the hypercube with size h and centered at $\mathbf{x}$

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

- we also have $p(\mathbf{x}) \simeq \frac{k/n}{V}$

- Thus, the estimation of density:

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

Parzen Windows

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

- Is it a density?
  - $p_\varphi(\mathbf{x}) >= 0 \ \forall \mathbf{x}$
  - $$\int p_\varphi(x)dx = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^d}\int\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)\mathrm{dx} = \frac{1}{nh^d}\sum_{i=1}^{n}\int\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)$$
    $$= \frac{1}{nh^d}\sum_{i=1}^{n}h^d = 1$$

Parzen Windows

$$p_{\varphi}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^{d}}\varphi(\frac{\mathbf{x}-\mathbf{x}_{i}}{h})$$

- Example: Assume that we have 8 samples:

  D={2, 2, 3, 4, 9, 13, 14, 15}

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

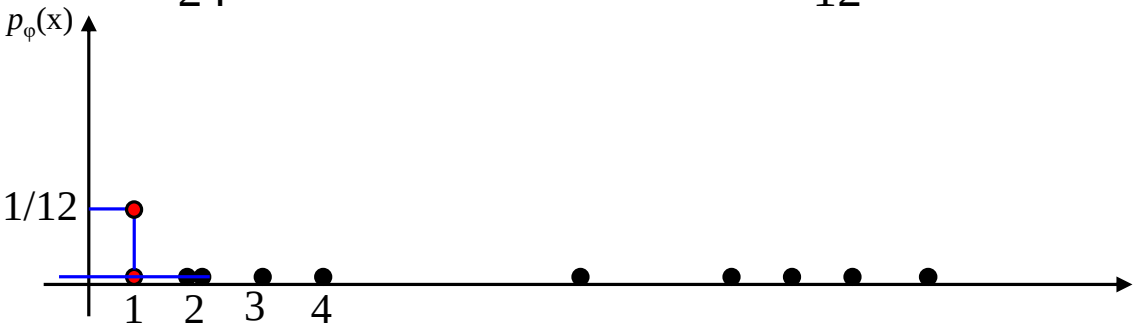
Parzen Windows

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

D={2, 2, 3, 4, 9, 13, 14, 15}

$$p_\varphi(1) = \frac{1}{8}\sum_{i=1}^{8}\frac{1}{3^1}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{3}\right) = \frac{1}{24}\left(\varphi\left(\frac{1-\mathbf{x}_1}{3}\right)+\varphi\left(\frac{1-\mathbf{x}_2}{3}\right)+\ldots+\varphi\left(\frac{1-\mathbf{x}_8}{3}\right)\right)$$

$$= \frac{1}{24}(1+1+0+0+0+0+0+0) = \frac{1}{12}$$



Parzen Windows

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

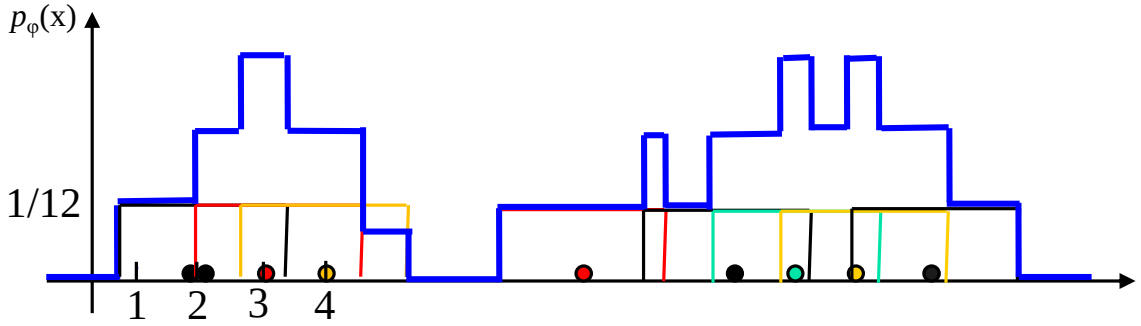
1 inside square centered at $\mathbf{x}_i$
0 otherwise

Thus $p_\varphi(x)$ is just a sum of n "box like" functions each of height $1/(nh^d)$

Parzen Windows

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

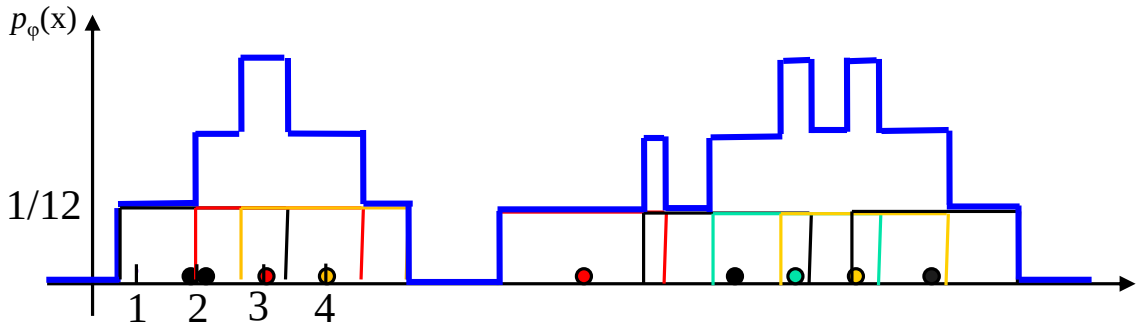
D={2, 2, 3, 4, 9, 13, 14, 15}



- We need to generate 8 boxes and add them up
- The width is h=3 and the height is $1/(nh^{d})$=1/24

Parzen Windows

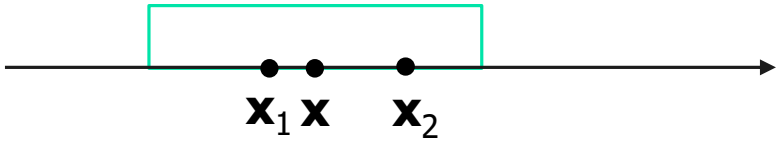
$$p_\varphi(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^d}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)$$



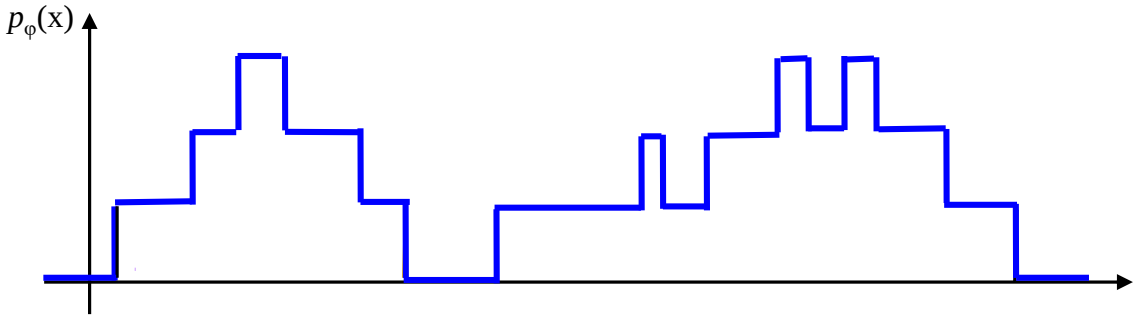
In essence, the window function φ is used for interpolation: each sample $\mathbf{x}_i$ contributes to the resulting density at $\mathbf{x}$ if $\mathbf{x}$ is close enough to $\mathbf{x}_i$

Parzen Windows

- Drawbacks of Hybecube φ
  - As long as sample $\mathbf{x}_i$ and $\mathbf{x}$ are in the same hypercube, the contribution of $\mathbf{x}_i$ to the density at $\mathbf{x}$ is constant, regardless of how close $\mathbf{x}_i$ is to $\mathbf{x}$.



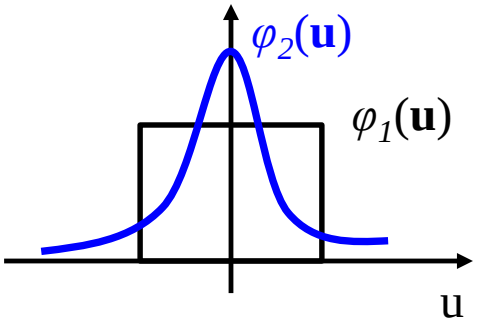
  - The resulting density $p_\varphi(\mathbf{x})$ is not smooth, it has discontinuities.



Parzen Windows: General φ

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

- We can use a general window φ as long as the resulting $p_\varphi(\mathbf{x})$ is a legitimate density, i.e.
  - $p_\varphi(\mathbf{x})>=0$
    - ✓ satisfied if $\varphi(\mathbf{u})>=0$
  - $\int p_\varphi(x) = 1$
    - ✓ satisfied if $\int \varphi(\mathbf{u}) = 1$



$$\int p_\varphi(x)\,\mathrm{d}\mathbf{x} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^d}\int\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)d\mathbf{x}$$

$$= \frac{1}{nh^d}\sum_{i=1}^{n}\int h^d\varphi(\mathbf{u})d\mathbf{u}$$

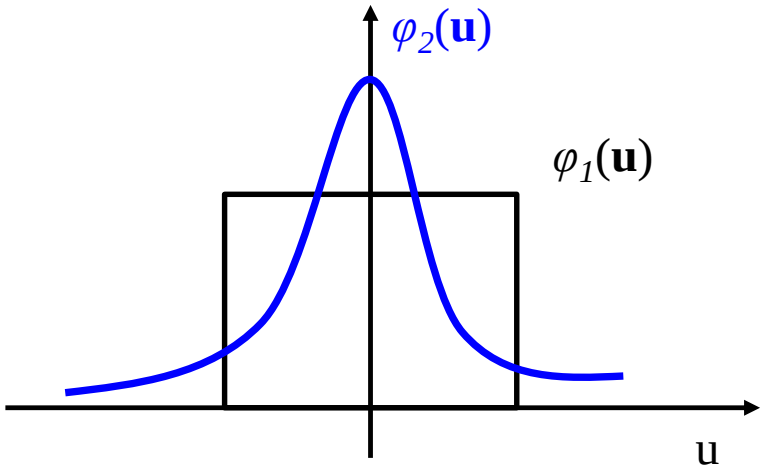
$$= 1$$

$$\mathbf{u} = \frac{\mathbf{x}-\mathbf{x}_i}{h}$$

Parzen Windows: General φ

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

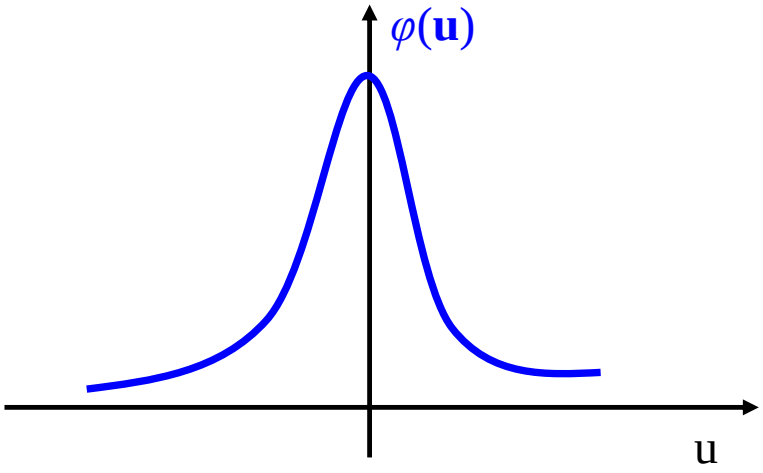
- We are no longer counting the number of samples inside *R*.
- We are counting the weighted average of potentially every single sample point



Parzen Windows: General φ

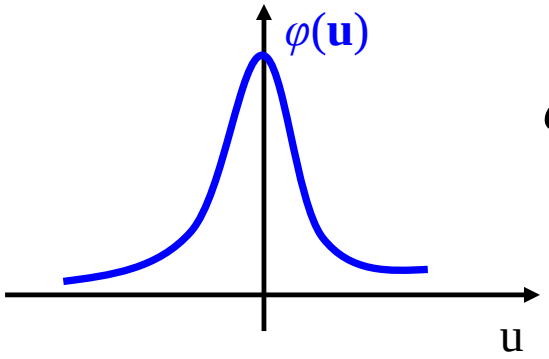
$$p_\varphi(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^d}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)$$

- Popular choice for φ if $N(0,1)$ density



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

Parzen Windows: General φ

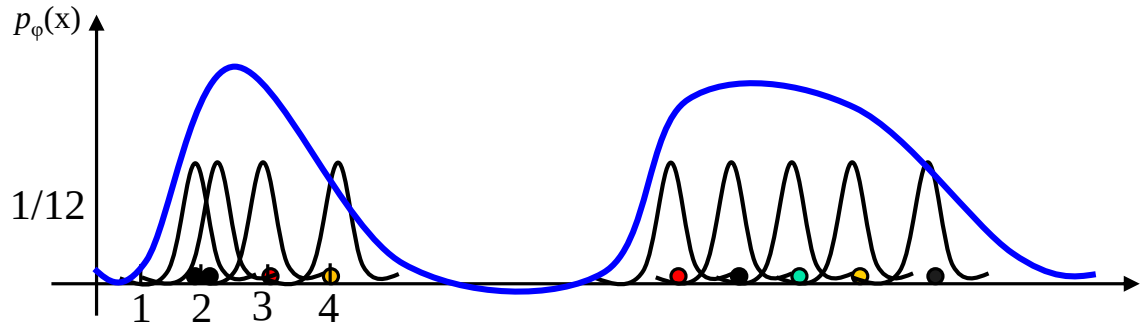


$$\varphi(\mathbf{u}) = \frac{1}{\sqrt{2\pi}} e^{-\mathbf{u}^2/2}$$

- Solves both drawbacks of the "box" window:
  - Points $\boldsymbol{x}$ which are close to the sample point $\mathbf{x}_i$ receive higher weight
  - Resulting density $p_\varphi(\mathbf{x})$ is smooth

Parzen Windows

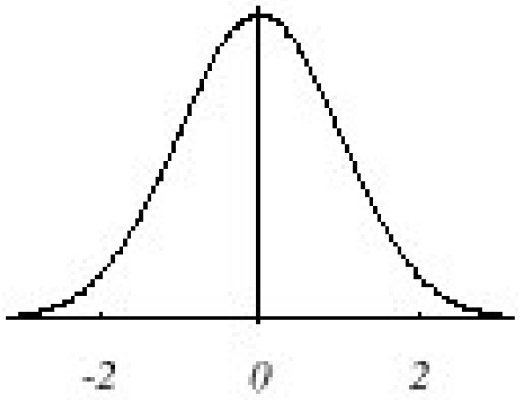
$$p_\varphi(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^d}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)$$



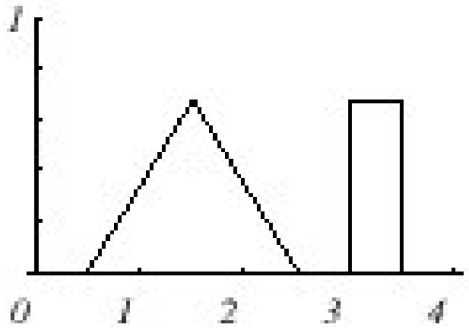
$p_\varphi(\mathbf{x})$ is the sum of 9 Gaussians, each centered at one of the sample points, and each scaled by 1/9

Parzen Windows

- Let's test if we solved the problem
  - Draw samples from a known distribution
  - 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 2 distributions:

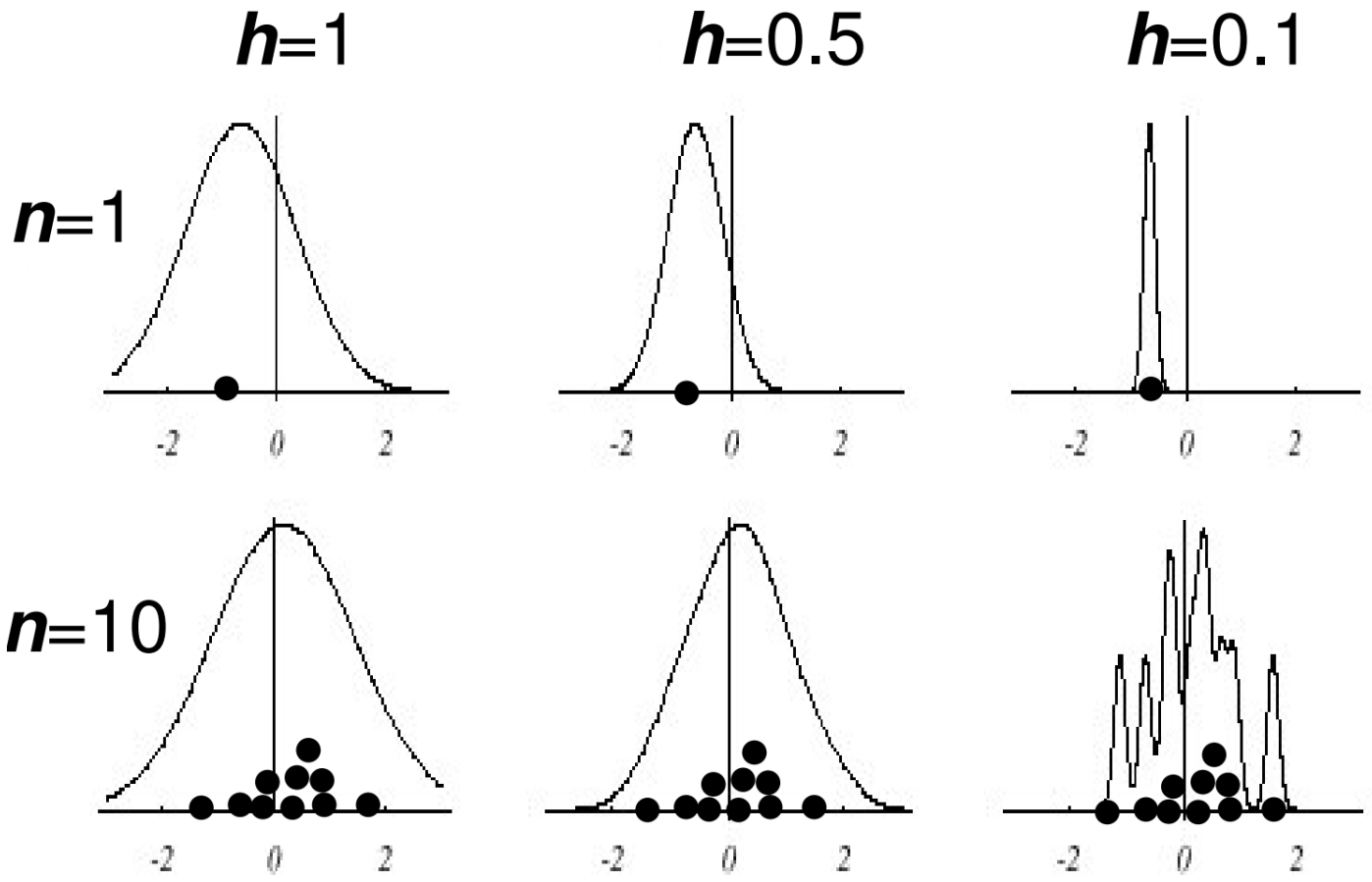


$N(0,1)$



triangle and uniform mixture

Parzen Windows: True density N(0,1)



Parzen Windows: True density N(0,1)

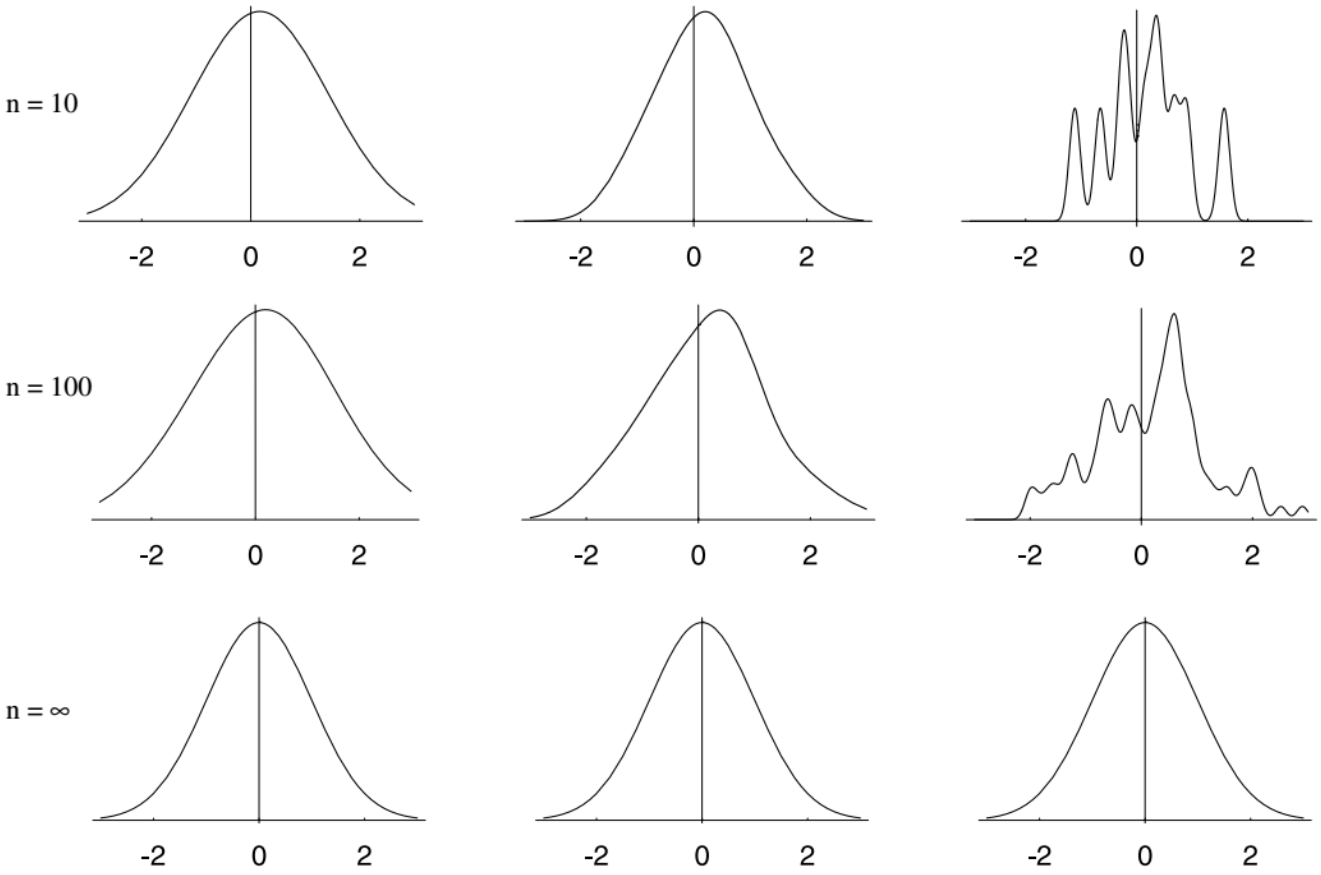
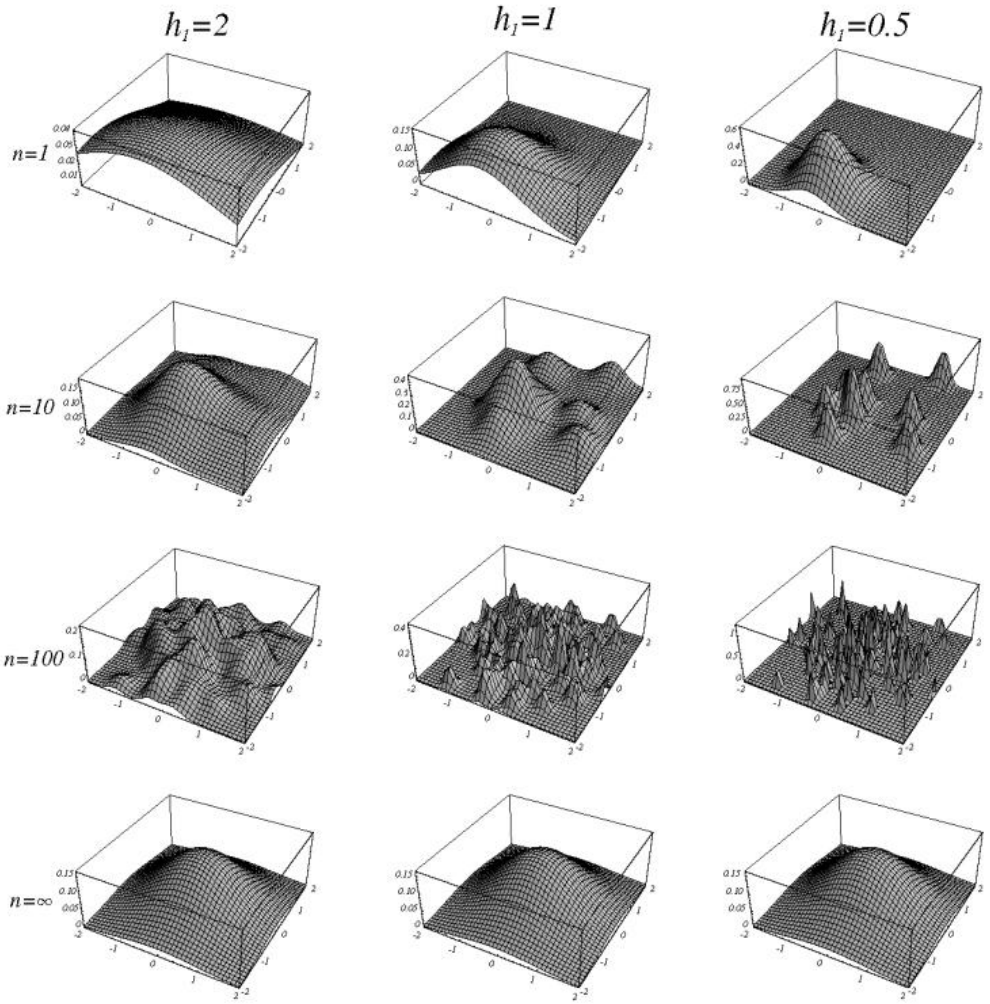
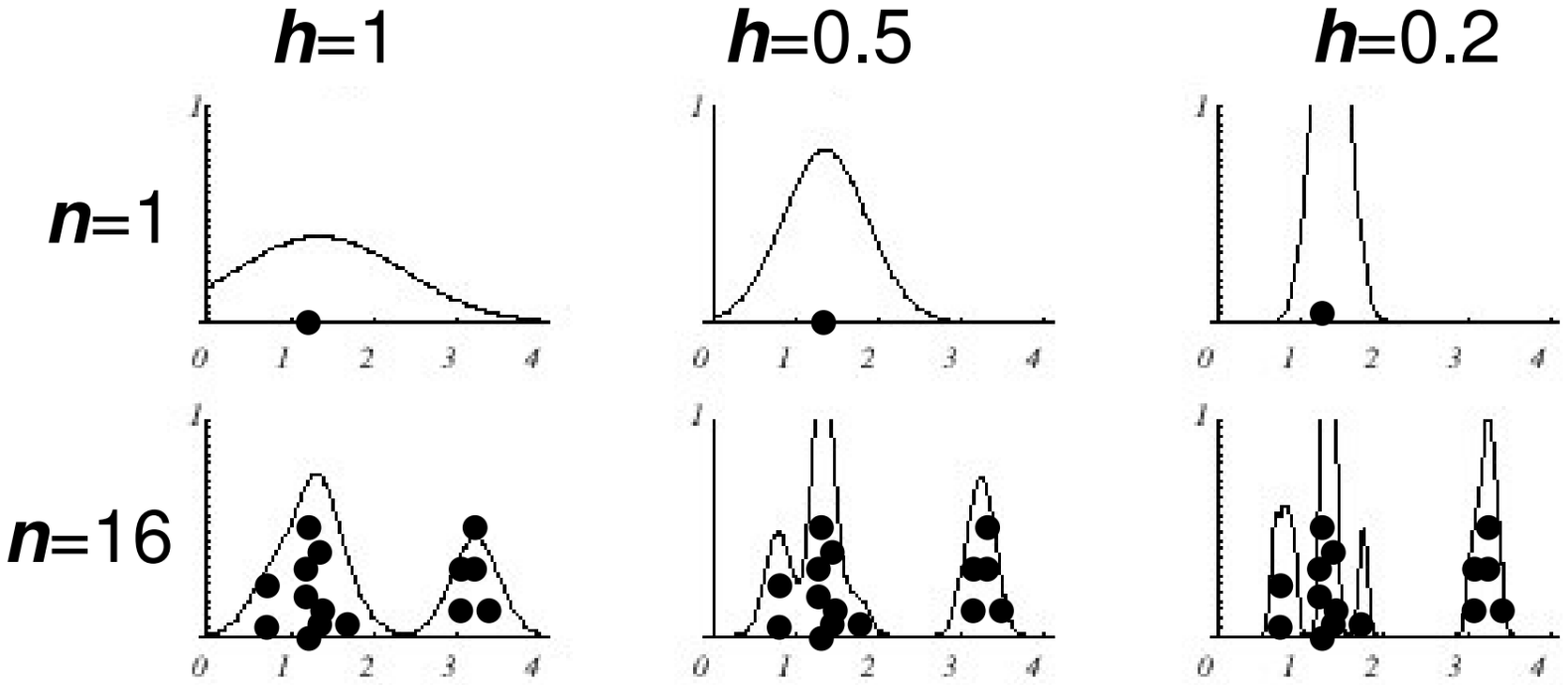


Figure 4.5: Parzen-window estimates of a univariate normal density using different window widths and numbers of samples. The vertical axes have been scaled to best show the structure in each graph. Note particularly that the $n = \infty$ estimates are the same (and match the true generating function), regardless of window width h.

Parzen Windows



Parzen Windows: triangle and uniform mixture



Parzen Windows: triangle and uniform mixture

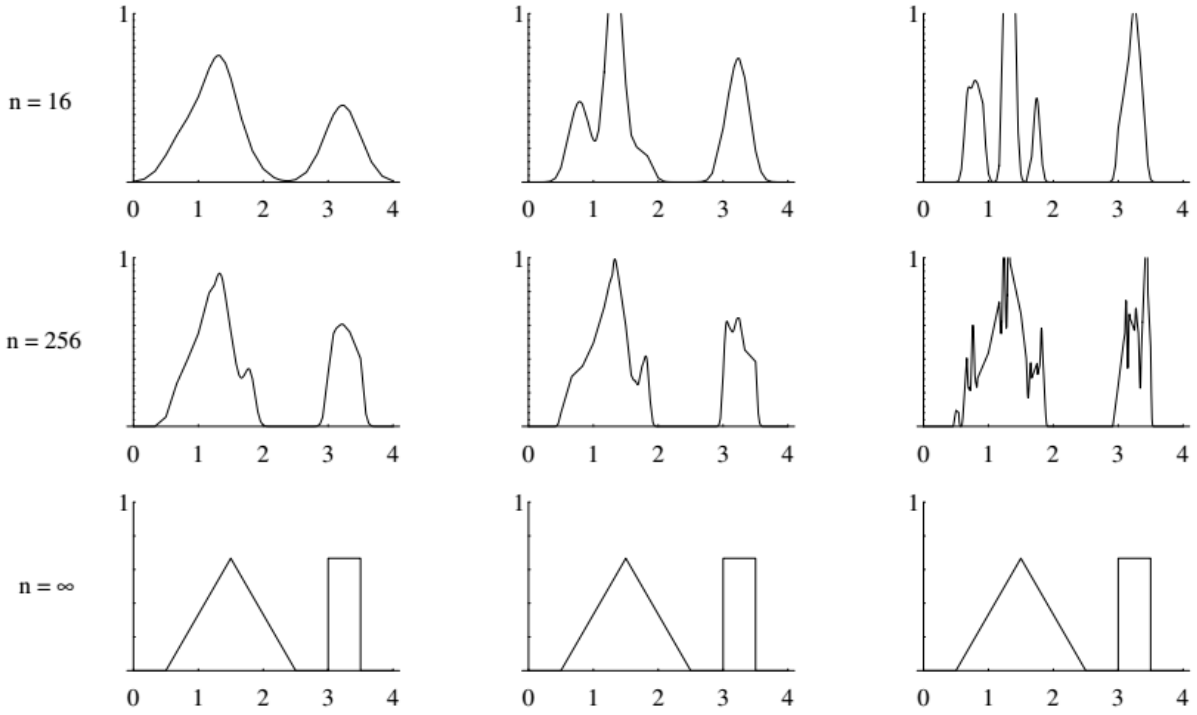


Figure 4.7: Parzen-window estimates of a bimodal distribution using different window widths and numbers of samples. Note particularly that the $n = \infty$ estimates are the same (and match the true generating distribution), regardless of window width h.

Parzen Windows: Effects of window width *h*

- If ***h*** is small, we superimpose ***n*** sharp pulses centered at the data
  - Each sample point $\mathbf{x}_i$ 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 $\mathbf{x}_i$ influences too large range of ***x***
  - Smoothed too much: the result looks oversmoothed
- Finding the best ***h*** is challenging, and indeed no single ***h*** may work well

- 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: Classifier

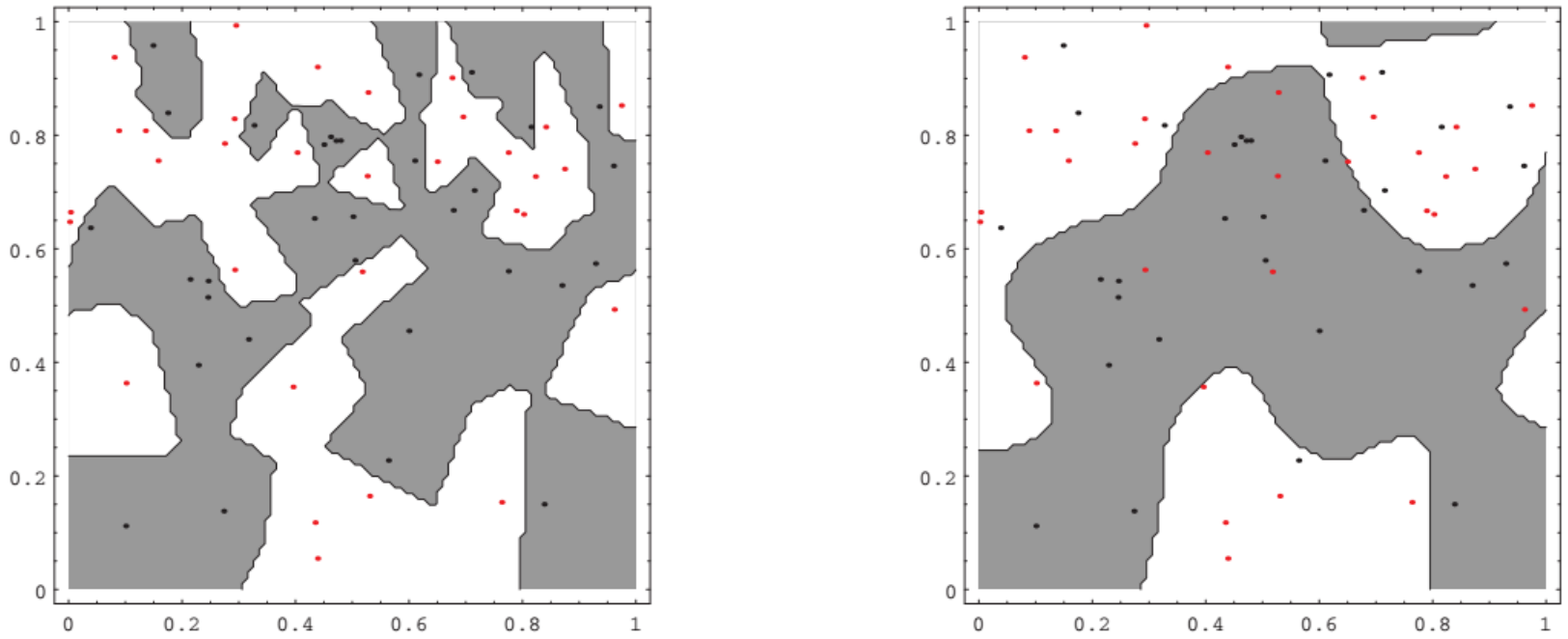


Figure 4.8: The decision boundaries in a two-dimensional Parzen-window dichotomizer depend on the window width h. At the left a small h leads to boundaries that are more complicated than for large h on same data set, shown at the right. Apparently, for this data a small h would be appropriate for the upper region, while a large h for the lower region; no single window width is ideal overall.

Parzen Windows: Summary

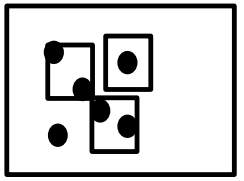
- Advantages
  - Can be applied to the data from any distribution
  - In theory can be shown to converge as the number of samples goes to infinity.
- Disadvantages
  - Number of training data is limited in practice, and so choosing the appropriate window size ***h*** is difficult
  - May need large number of samples for accurate estimates
  - Computationally heavy, to classify one point we have to compute a function which potentially depends on all samples

Density Estimation

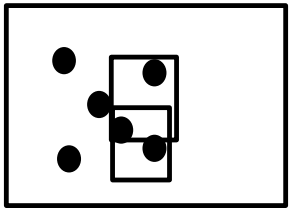
K-nearest neighbor approach

Density Estimation: *k*NN

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



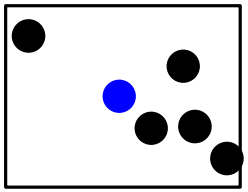
- Parzen Windows
  - Choose a fixed value for volume ***V*** and determine the corresponding ***k*** from the data.

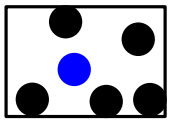


- K-nearest neighbors
  - Choose a fixed value for ***k*** and determine the corresponding volume ***V*** from the data

Density Estimation: *k*NN

- K-nearest neighbors:
  - Let the cell volume be a function of the training data
  - Center a cell about x and let it grows until it captures *k* samples
  - *k* is called the *k* nearest-neighbors of **x**
- 2 possibilities may occur:
  - Density is high near *x*; therefore the cell will be small which provides a good resolution.
  - Density is low; therefore the cell will grow large and stop until higher density regions are reached

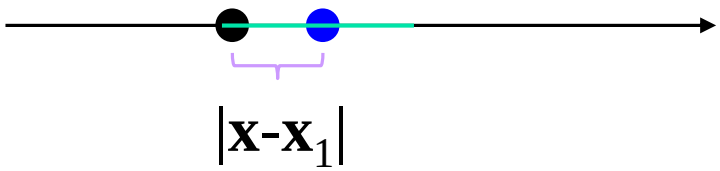




Density Estimation: *k*NN

- Let's look at 1-D example:
  - We have one sample, i.e. n=1

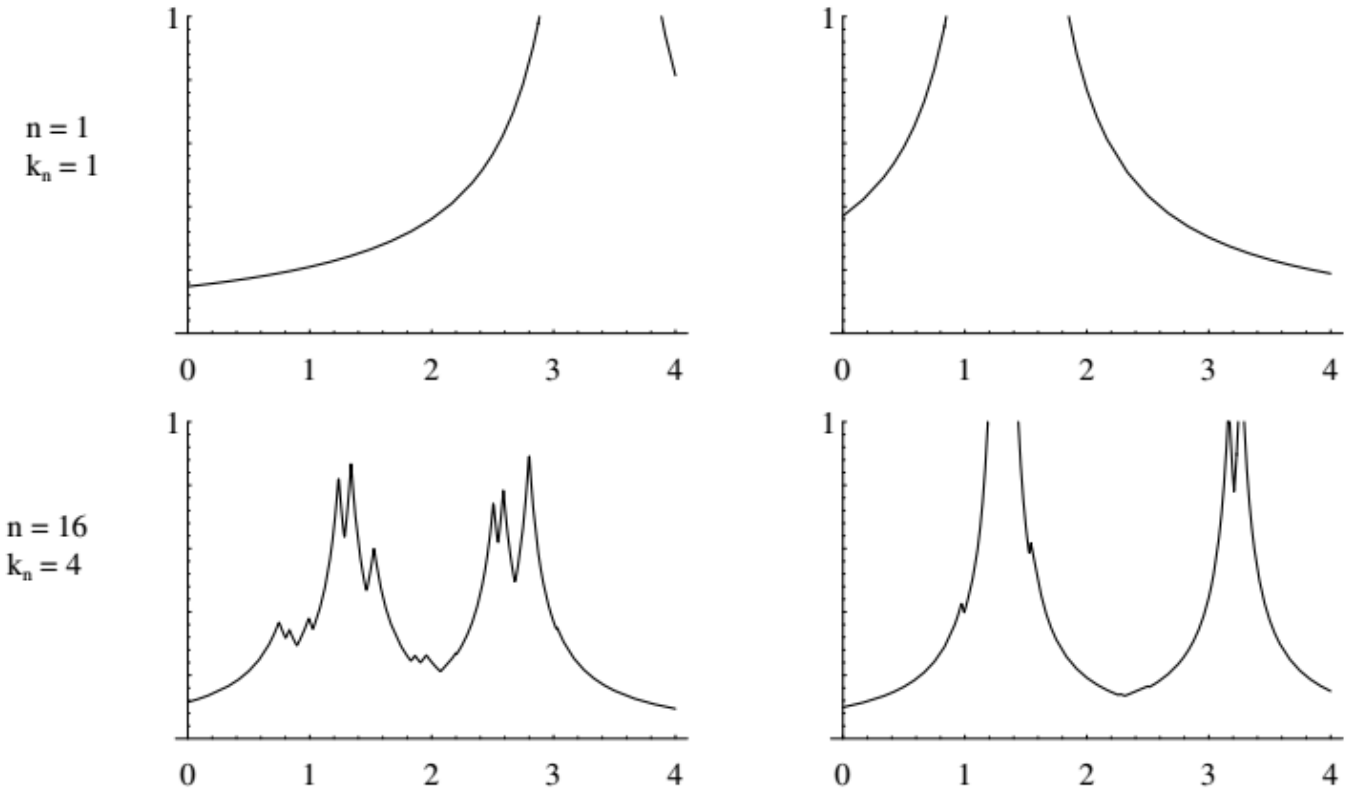
$$p(x) \approx \frac{k/n}{V} = \frac{1}{2|\mathbf{x}-\mathbf{x}_1|}$$



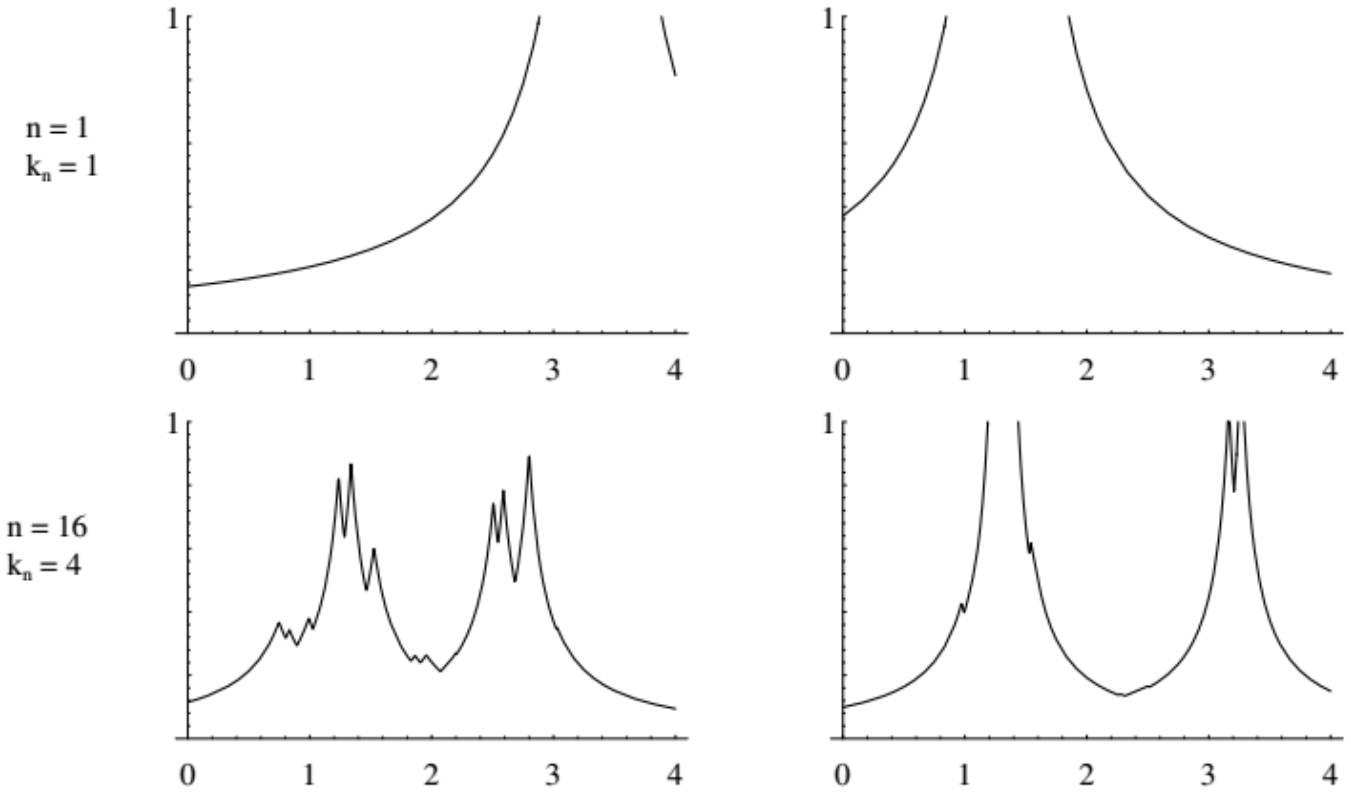
  - But the estimated $\boldsymbol{p}(\boldsymbol{x})$ is not even close to a density function

$$\int_{-\infty}^{\infty} \frac{1}{2|\mathbf{x}-\mathbf{x}_1|} d\mathbf{x} = \infty \neq 1$$

Density Estimation: *k*NN



Density Estimation: *k*NN



Density Estimation: *k*NN

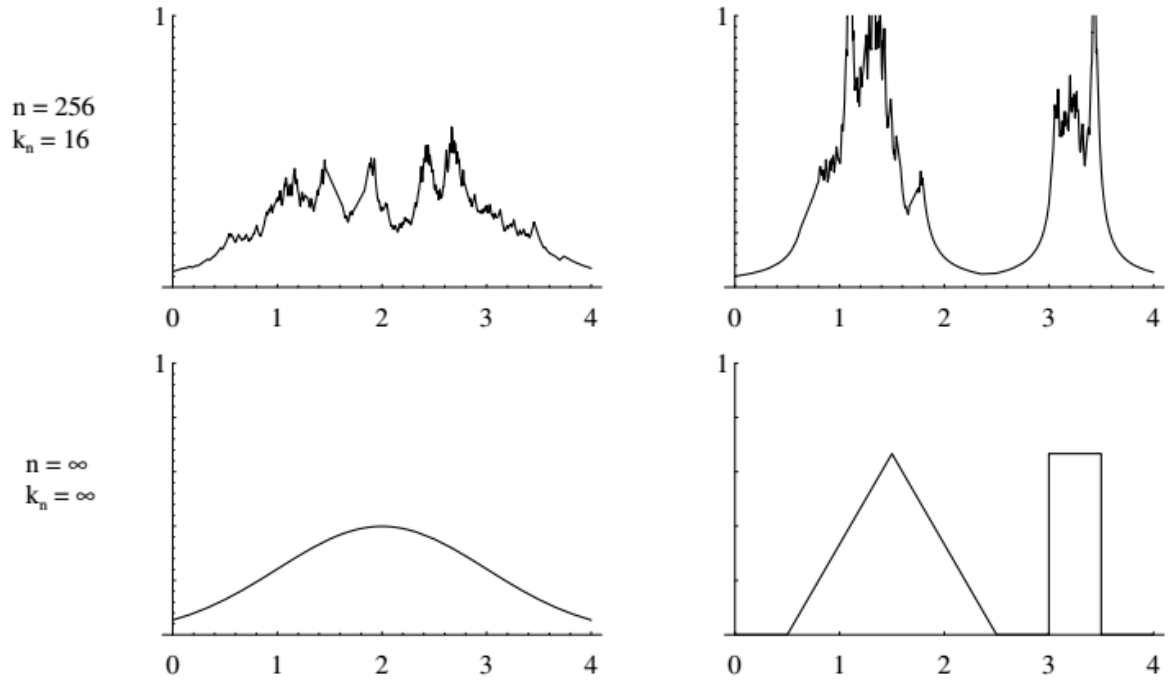


Figure 4.12: Several k-nearest-neighbor estimates of two unidimensional densities: a Gaussian and a bimodal distribution. Notice how the finite n estimates can be quite "spiky."

Density Estimation: *k*NN

- Instead of approximating the density $\boldsymbol{p}(\boldsymbol{x})$, we can use kNN method to approximate the posterior distribution $\boldsymbol{P}(\boldsymbol{C}_i|\mathbf{x})$
- We don't even need $\boldsymbol{p}(\mathbf{x})$ if we can get a good estimate on $\boldsymbol{P}(\boldsymbol{C}_\mathrm{i}|\boldsymbol{x})$

Density Estimation: *k*NN

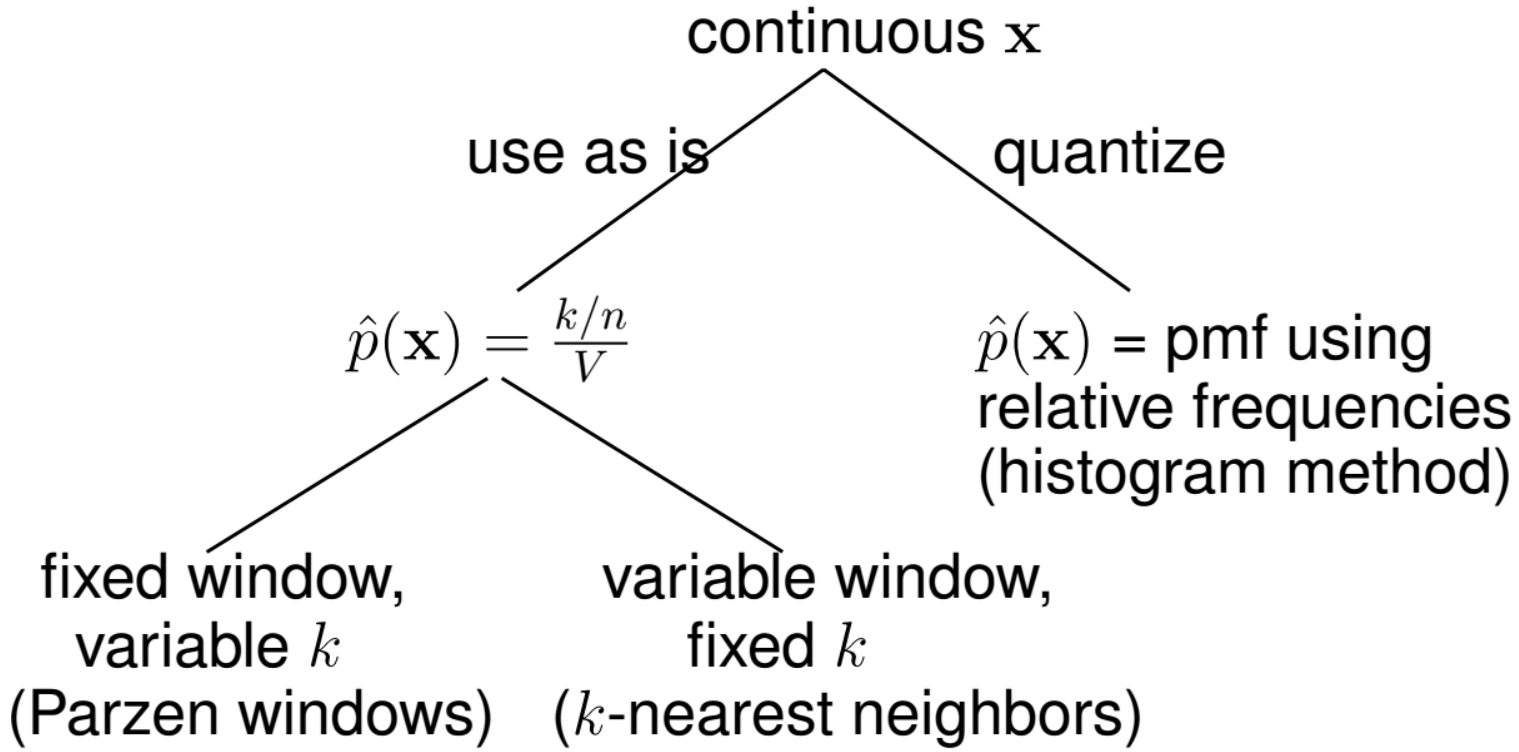
- Posterior probabilities can be estimated from a set of n labeled samples and can be used with the Bayesian decision rule for classification.
- Suppose that a volume V around $\mathbf{x}$ includes k samples, k_i of which are labeled as belonging to class w_i.
- As estimate for the joint probability $p(\mathbf{x}, w_i)$ becomes

$$p_n(\mathbf{x}, w_i) = \frac{k_i/n}{V}$$

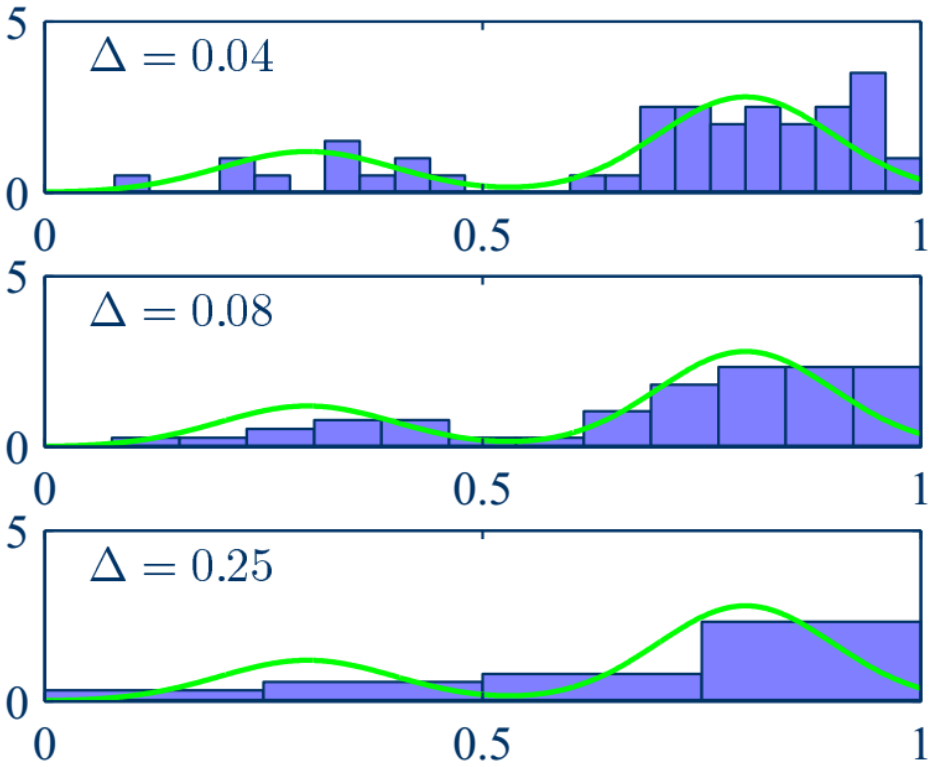
and gives an estimate for the posterior probability

$$P_n(w_i|\mathbf{x}) = \frac{p_n(\mathbf{x}, w_i)}{\sum_{j=1}^{c} p_n(\mathbf{x}, w_j)} = \frac{k_i}{k}.$$

Density Estimation



Density Estimation



An illustration of the histogram approach to density estimation, in which a data set of 50 points is generated from the distribution shown by the green curve. Histogram density estimates are shown for various values of the cell volume (Δ).

Density Estimation

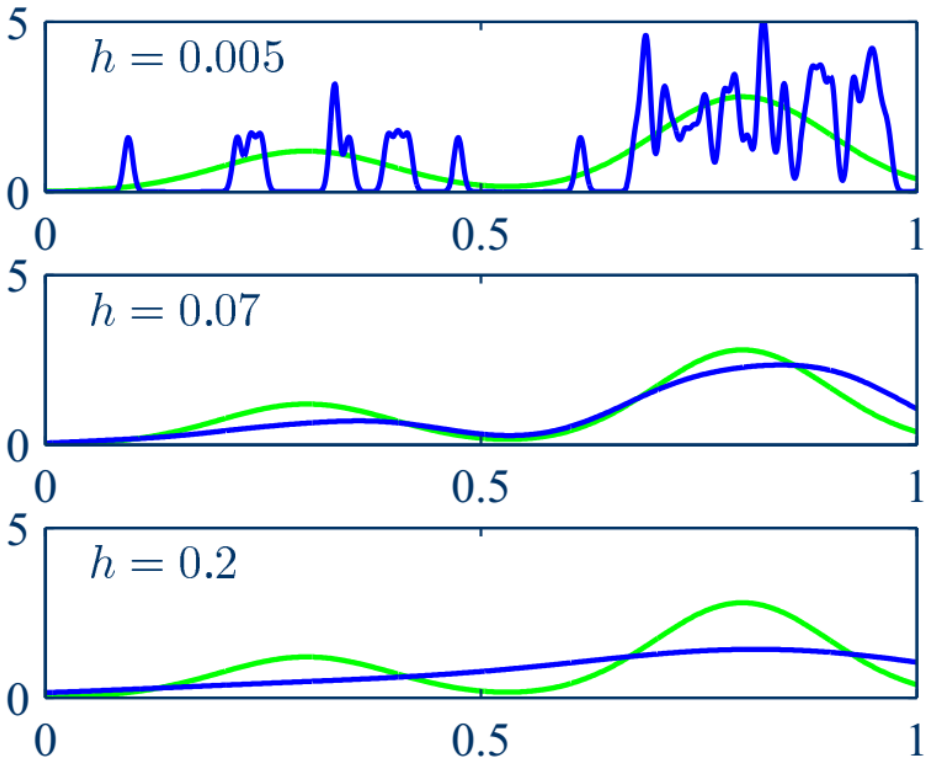


Illustration of the Parzen density model. The window width (h) acts as a smoothing parameter. If it is set too small (top), the result is a very noisy density model. If it is set too large (bottom), the bimodal nature of the underlying distribution is washed out. An intermediate value (middle) gives a good estimate.

Density Estimation

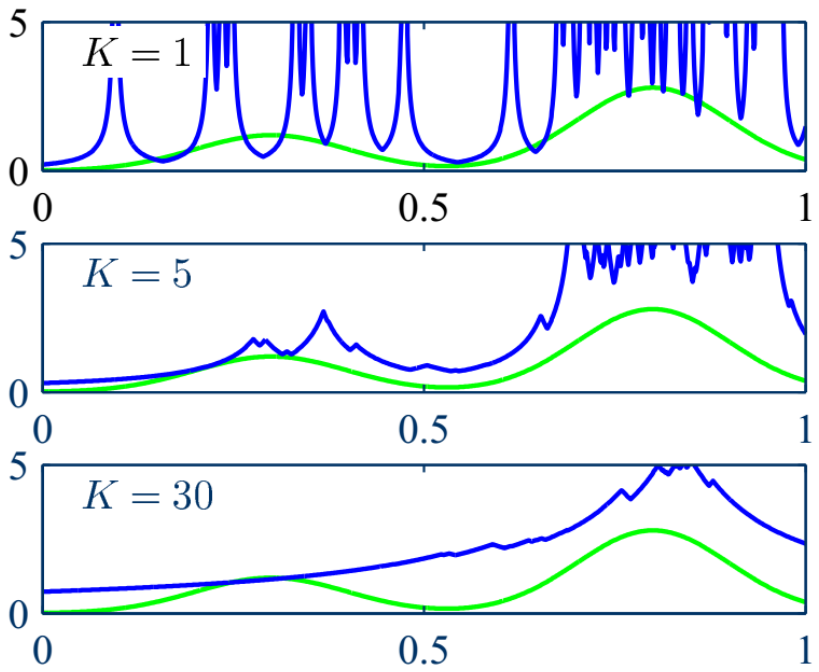
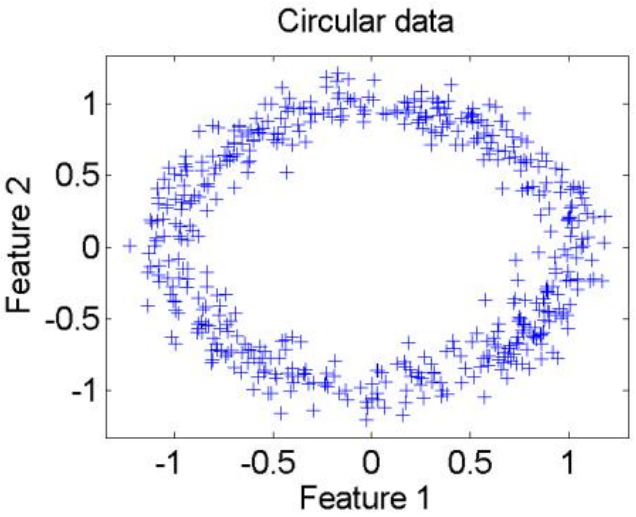
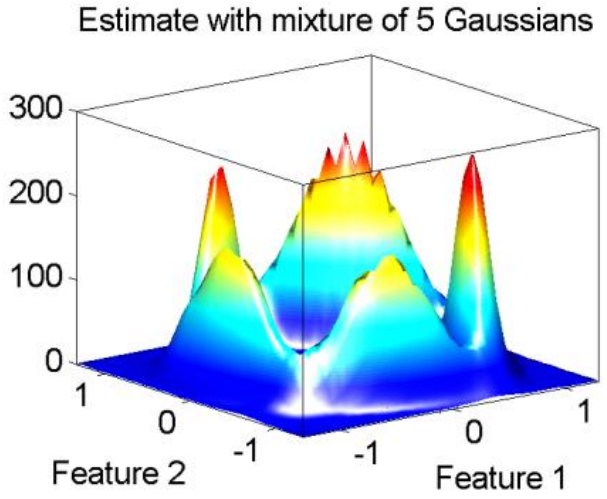
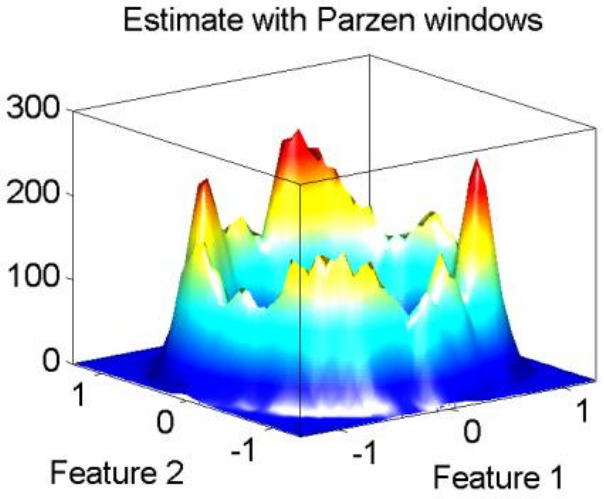


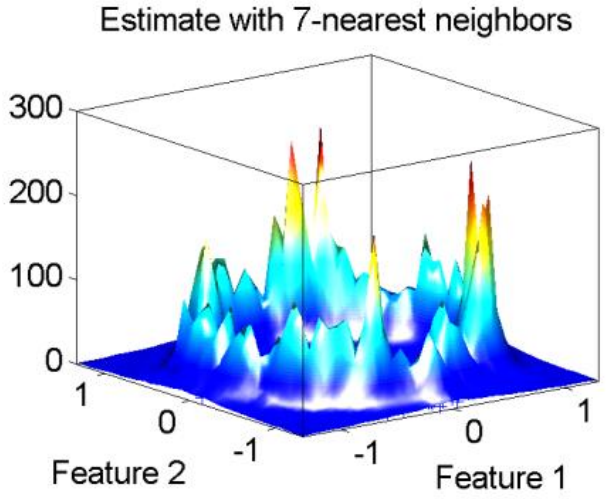
Illustration of the *k*-nearest neighbor density model. The parameter *k* governs the degree of smoothing. A small value of *k* (top) leads to a very noisy density model. A large value (bottom) smoothes out the bimodal nature of the true distribution.

Density Estimation









Density Estimation