Chapter 4 (Part 1): Non-Parametric Classification (Sections 4.1-4.3)

- Introduction
- Density Estimation
- Parzen Windows

Introduction

- All Parametric densities are unimodal (have a single local maximum), whereas many practical problems involve multi-modal densities

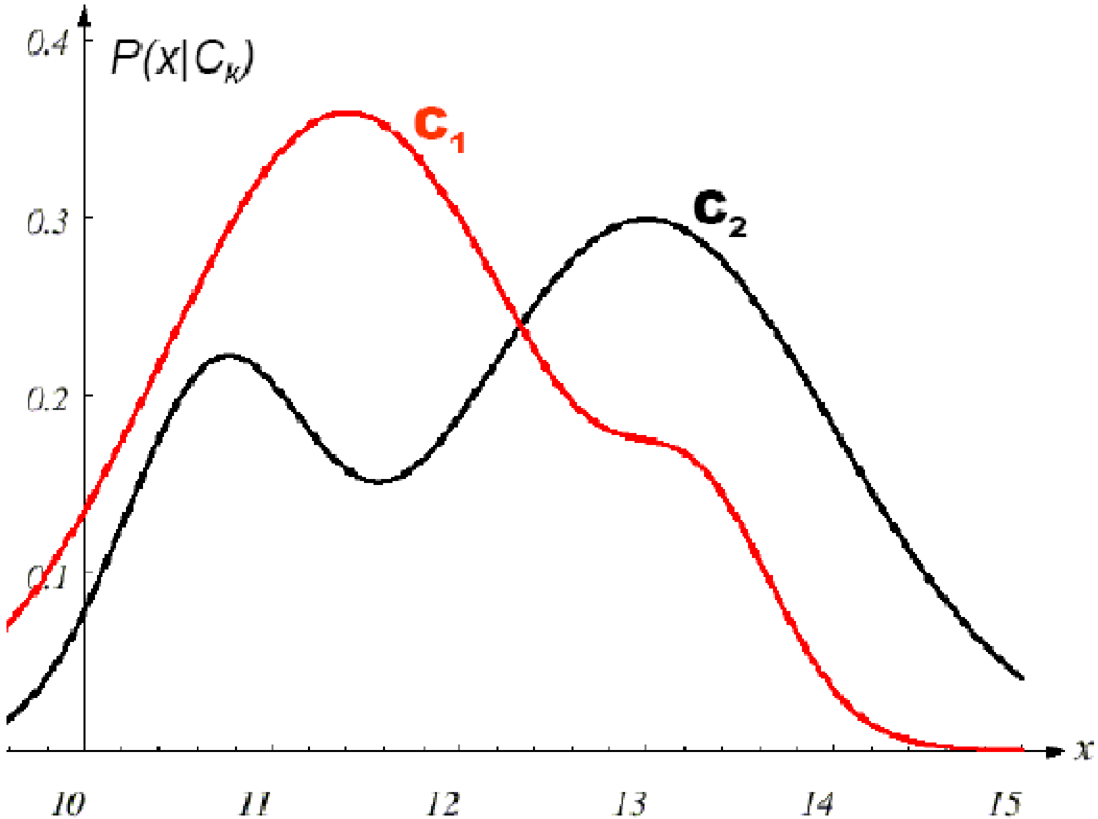
- Nonparametric procedures can be used with arbitrary distributions and without the assumption that the forms of the underlying densities are known

- There are two types of nonparametric methods:

  - Estimating P$(x \mid w_j)$
  - Bypass probability and go directly to a-posteriori probability estimation $P\ (w_j / x)$

ESTIMATION PROBLEM

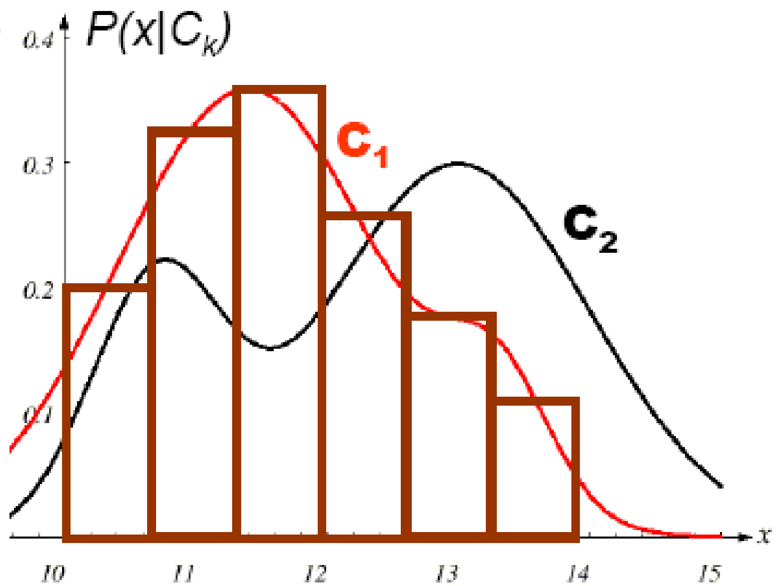
- **Problem:** estimate the model of probability function *P(x)* given a finite number of data points $X_1, X_2, \ldots, X_n$ (while assuming that the estimation is driven entirely by the data).



HISTOGRAMS

- **Idea:** Histograms based on the training set are the simplest methods for approximating (directly) the probability density functions. Histograms are "smoothed" by averaging over a local region of the feature space.
- **Method:** Divide the sample space into a number of bins and approximate the density at the centre of each bin by the fraction of points in the training data that fall into the corresponding bin.
- If *Count(x)* is the number of samples (out of total *n*) in the same bin as *x* and *Width(x)* is the width of the bin containing *x*, then:

$$P_H(x) = \frac{1}{n}\frac{Count(x)}{Width(x)}$$



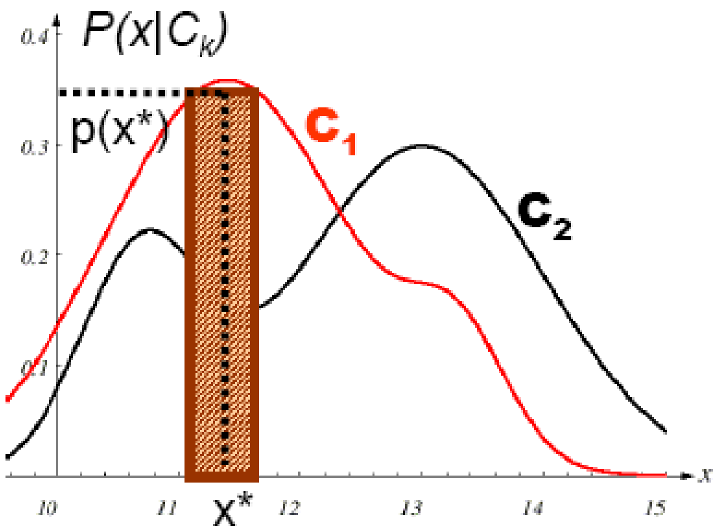
- **Issues:**
  - Artificial discontinuities (at bin boundaries) due to bin width and locations.
  - Problems with selecting the bin size (to limit the number of bins).

MATHEMATICAL MODEL

- Probability that feature vector *x* falls within the region *R* is: $P(x) = \int_{x \in R} p(x)dx$
- **Approximations** for the integral:
  - If p(x) is continuous and does not vary significantly within region *R*, then *P(x)* can be approximated with the product between the average value of density function *p(x)* within the region and the area/volume of the region.
  - If samples are independent and identically distributed (i.i.d), then *P(x)* can be approximated with the rate of samples falling into the region (i.e. *k/n*, when *k* is the number of samples in region *R*).

$$P(x) = \int_R p(x)dx \approx p(x^*)V \approx \frac{k}{n}$$

- **Density estimate:** $\boxed{p(x^*) \approx \frac{k}{n \times V}}$
- **Conclusion:** estimate becomes more accurate as the number of sample points *n* increases while the region's volume *V* shrinks.



ESTIMATION PROCEDURE

- Relative probability (i.e. density) of x can be estimated by forming a sequence of regions R_1, R_2, … containing x, where R_1 is used with one sample, R_2 is used with two, and so on. Given n samples, let k_n be the number of samples falling within region R_n having volume V_n. Then:

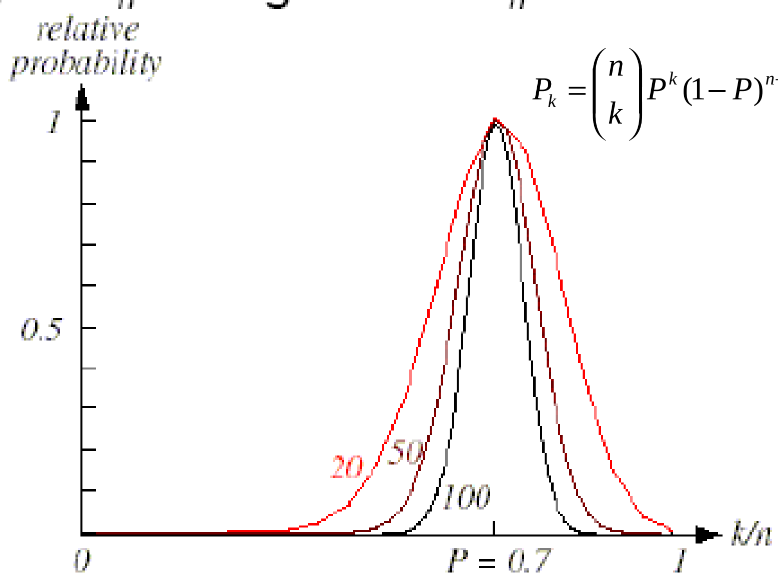
$$p_n(x) = \frac{k_n}{n \times V_n}$$

- **Convergence conditions** required for $p_n(x) \rightarrow p(x)$:

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

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

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



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

Suppose that the true probability was chosen to be 0.7. Each curve is labelled by the total number of patterns n sampled, and is scaled to give the same maximum (at the true probability). For large n, the curve peaks strongly at the true probability. At the limit, when $n \rightarrow \infty$, the curve approaches a delta function, and the estimate is guaranteed to yield the true probability.

$E\{k/n\} = p \quad var(k/n) = p(1-p)/n$

DENSITY ESTIMATION

- **General expression** for non-parametric density estimation is:

$$p(x) = \frac{k}{n \times V}$$

  - n = total number of samples (i.e. data points)
  - V = volume of the region R surrounding x
  - k = number of samples inside R (of volume V)

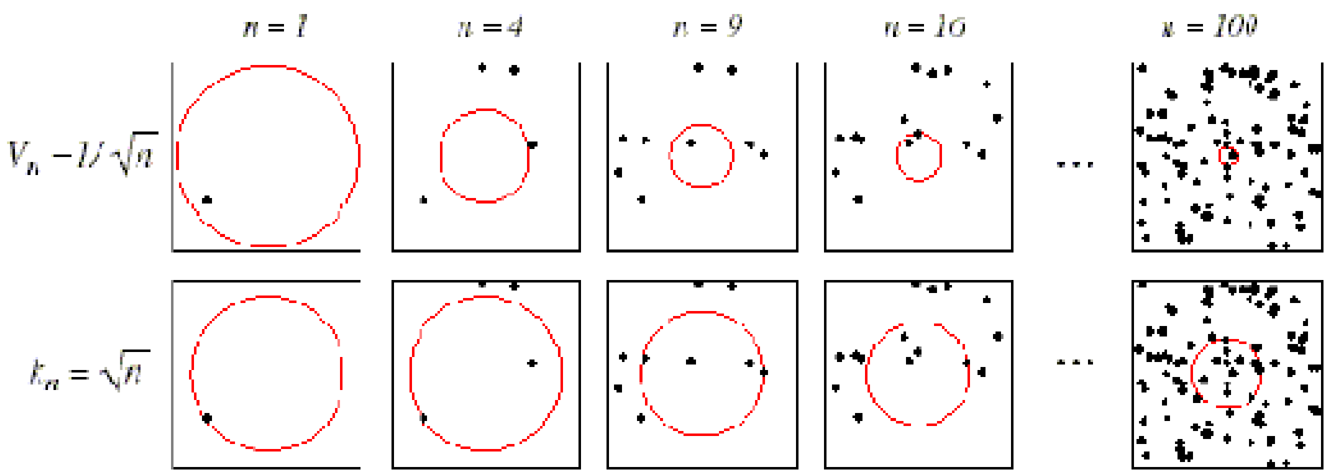
- **Analogy:** $p(x)$ is analogue to the physical density while k/n is analogue to the mass of samples within R.

- **Practical notes:**
  - The total number of training samples n is usually fixed.
  - In order to improve estimation accuracy, V should approach zero, but then the region R would enclose no samples. Therefore, V should be:
    - Large enough to include enough samples within R.
    - Small enough to support the assumption that $p(x)$ is constant throughout R.

- **Two approaches:**
  - **Kernel Density Estimation** (**KDE**) or **Parzen-window** method: choose a fixed value for V and determine k from training data.
  - **k-Nearest-Neighbour** (**kNN**): choose a fixed value of k and determine the corresponding volume V from training data.

TWO APPROACHES

- There are two leading methods for estimating the density at a point at the centre of an estimation region (shown below as a square).
  - **Parzen-window** estimation method: shrink the region by specifying the volume V_n as some function of n, such as $V_n = n^{-½}$.
  - ***k_n*-nearest-neighbour** estimation method: decrease the volume in a data-dependent way, for instance letting the volume enclose some number $k_n = n^{½}$ of sample points.
- The sequences in both cases represent random variables that generally converge and allow the true density at the test point to be calculated.



Parzen Windows

- Parzen-window approach to estimate densities assume that the region R_n is a d-dimensional hypercube

$$V_n = h_n^d \quad (h_n : \text{length of the edge of } \Re_n)$$

$$\text{Let } \varphi(u) \text{ be the following window function:}$$

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

- $\varphi((x-x_i)/h_n)$ is equal to unity if x_i falls within the hypercube of volume V_n centered at x and equal to zero otherwise.

The number of samples in this hypercube is:

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

By substituting k_n in equation (7), we obtain the following estimate:

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

$P_n(x)$ estimates $p(x)$ as an average of functions of x and the samples (x_i) $(i = 1,\dots,n)$. These functions φ can be general!

- Illustration

  - The behavior of the Parzen-window method

    - Case where $p(x) \to N(0,1)$

      Let $\varphi(u) = (1/\sqrt{2\pi}) \exp(-u^2/2)$ and $h_n = h_1/\sqrt{n}$ $(n>1)$

      (h_1: known parameter)

      Thus:

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

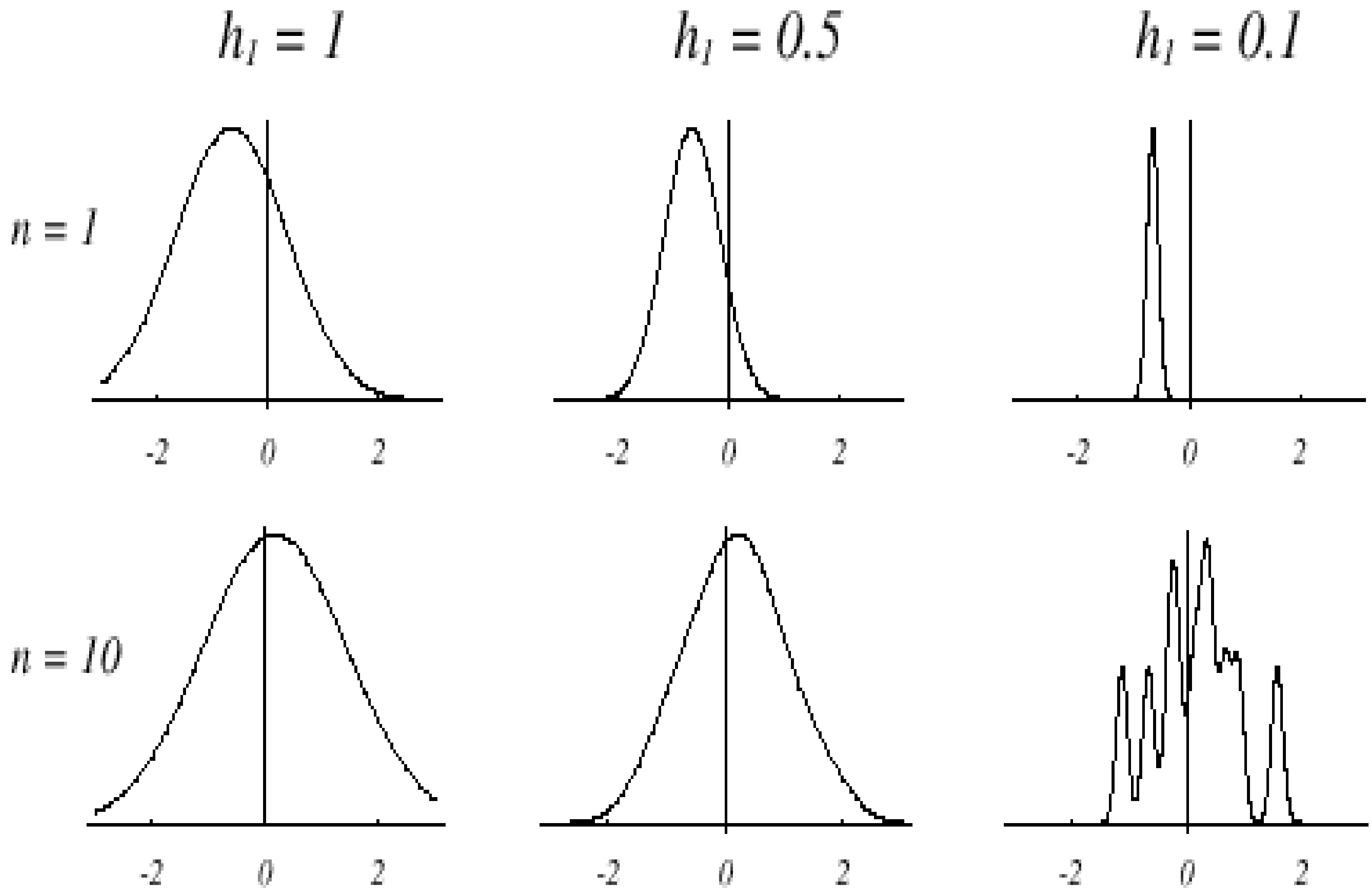
      is an average of normal densities centered at the samples x_i.

Numerical results:

For $n = 1$ and $h_1 = 1$

$$p_1(x) = \varphi(x - x_1) = \frac{1}{\sqrt{2\pi}} e^{-1/2}(x - x_1)^2 \rightarrow N(x_1, 1)$$

For $n = 10$ and $h = 0.1$, the contributions of the individual samples are clearly observable !



h1 = 1
h1 = 0.5
h1 = 0.1
n = 1
n = 10
-2
0
2

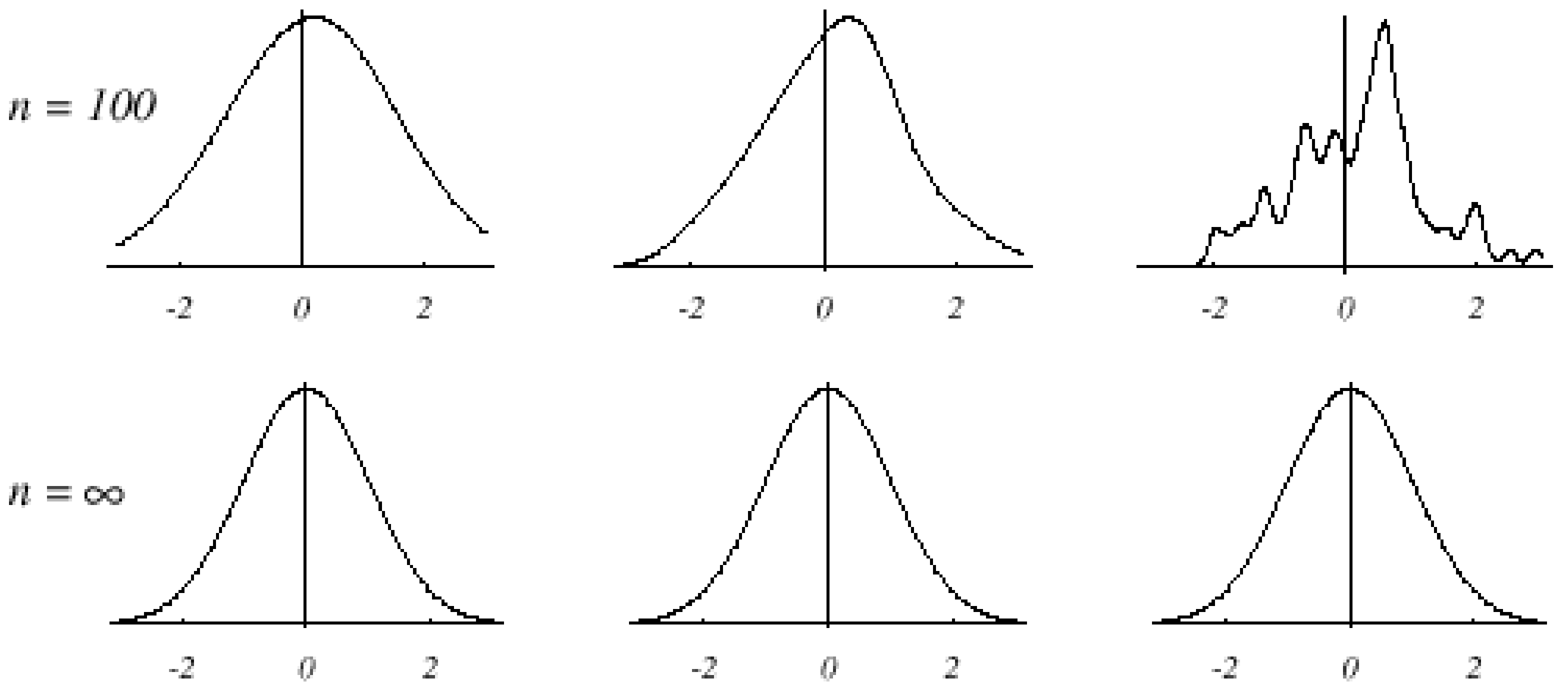
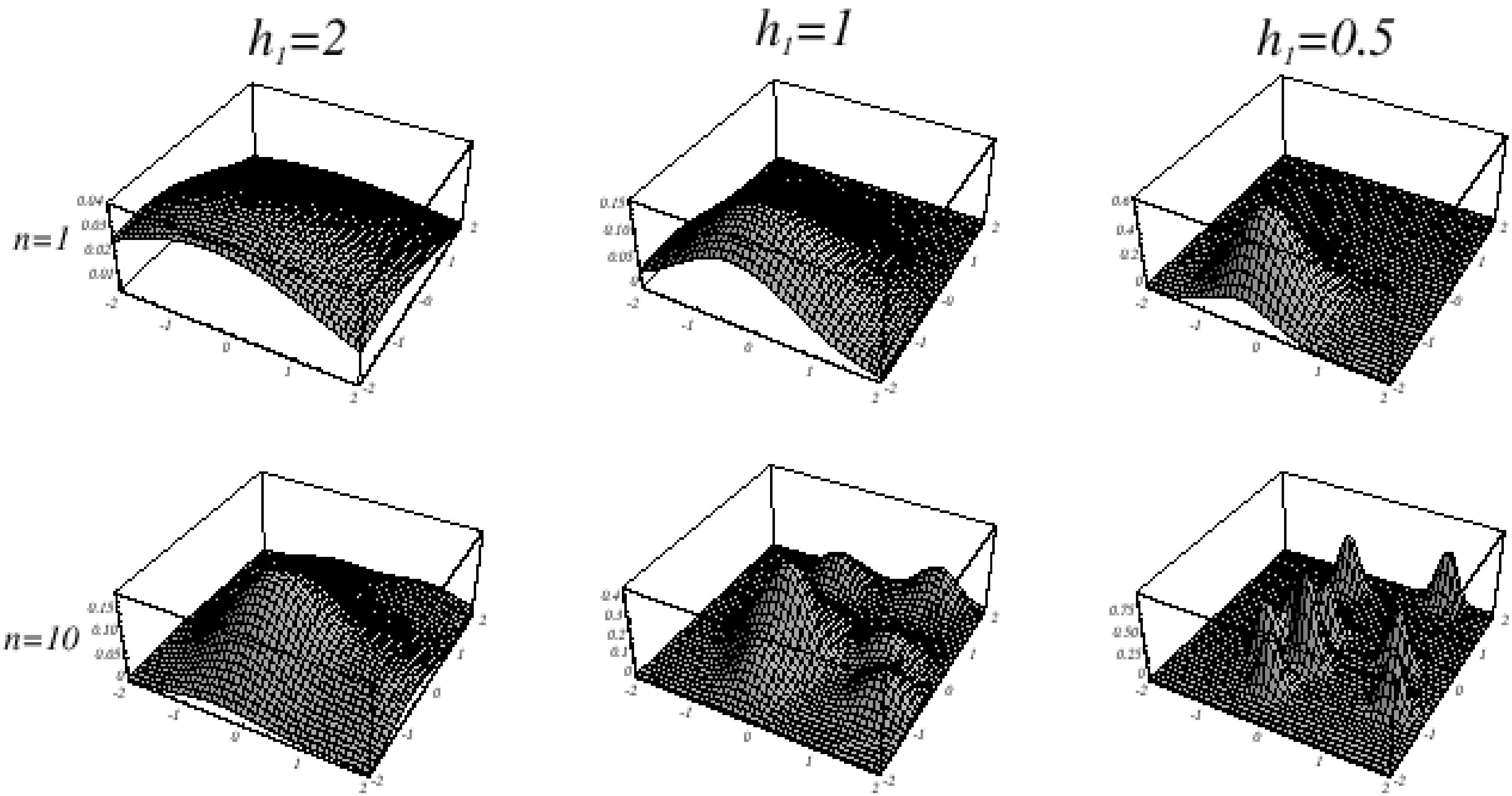


FIGURE 4.5. Parzen-window estimates of a univariate normal density using different window widths and numbers of samples. The vertical axes have been scaled to best show the structure in each graph. Note particularly that the $n = \infty$ estimates are the same (and match the true density function), regardless of window width. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Analogous results are also obtained in two dimensions as illustrated:



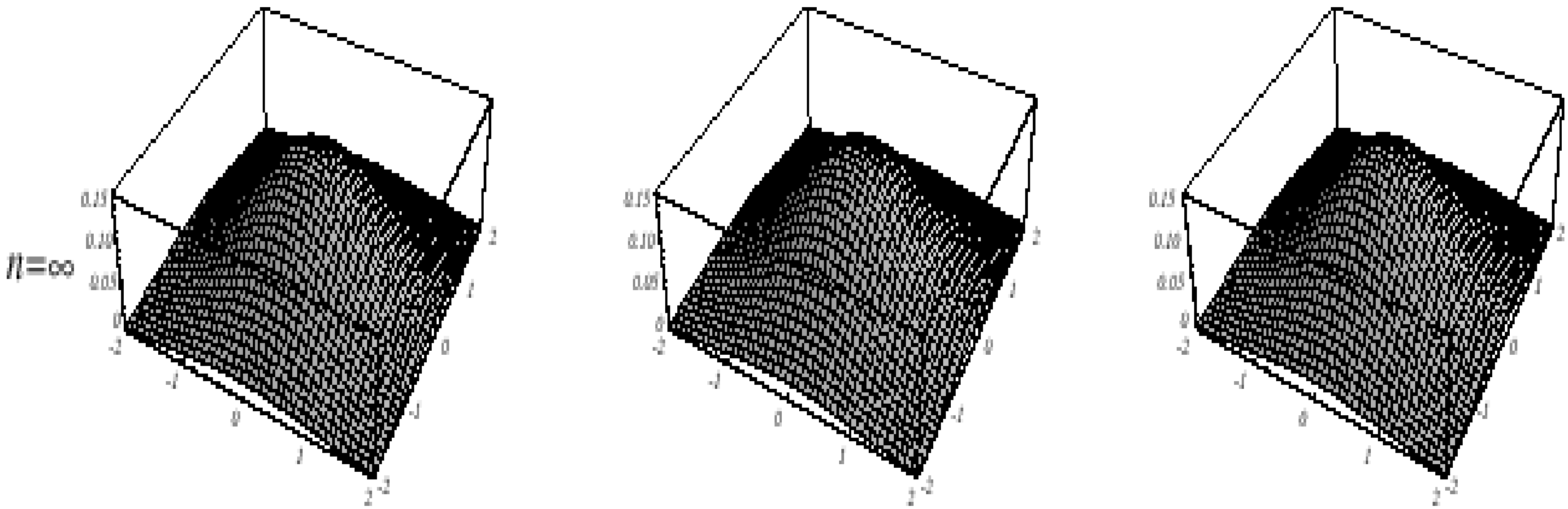
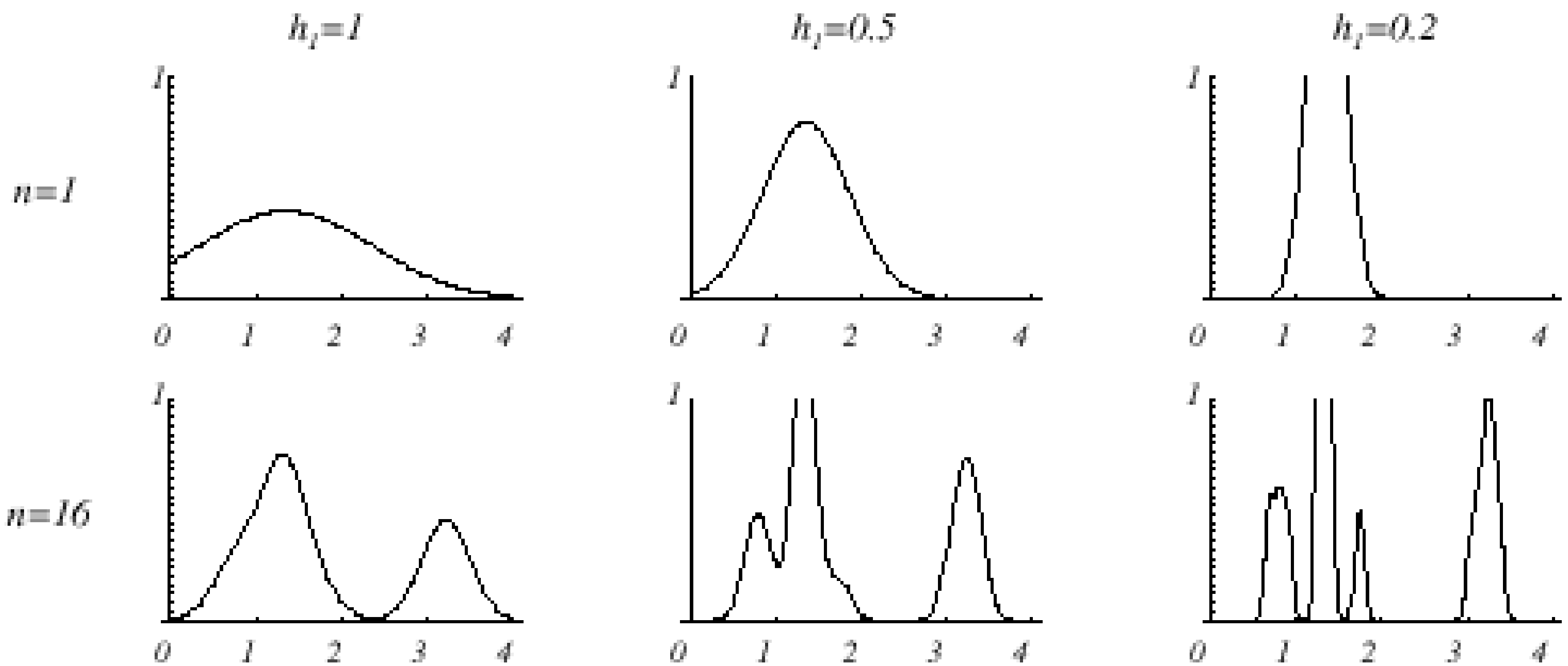


FIGURE 4.6. Parzen-window estimates of a bivariate normal density using different window widths and numbers of samples. The vertical axes have been scaled to best show the structure in each graph. Note particularly that the $n = \infty$ estimates are the same (and match the true distribution), regardless of window width. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Case where $p(x) = l_1.U(a,b) + l_2.T(c,d)$ (unknown density) (mixture of a uniform and a triangle density)



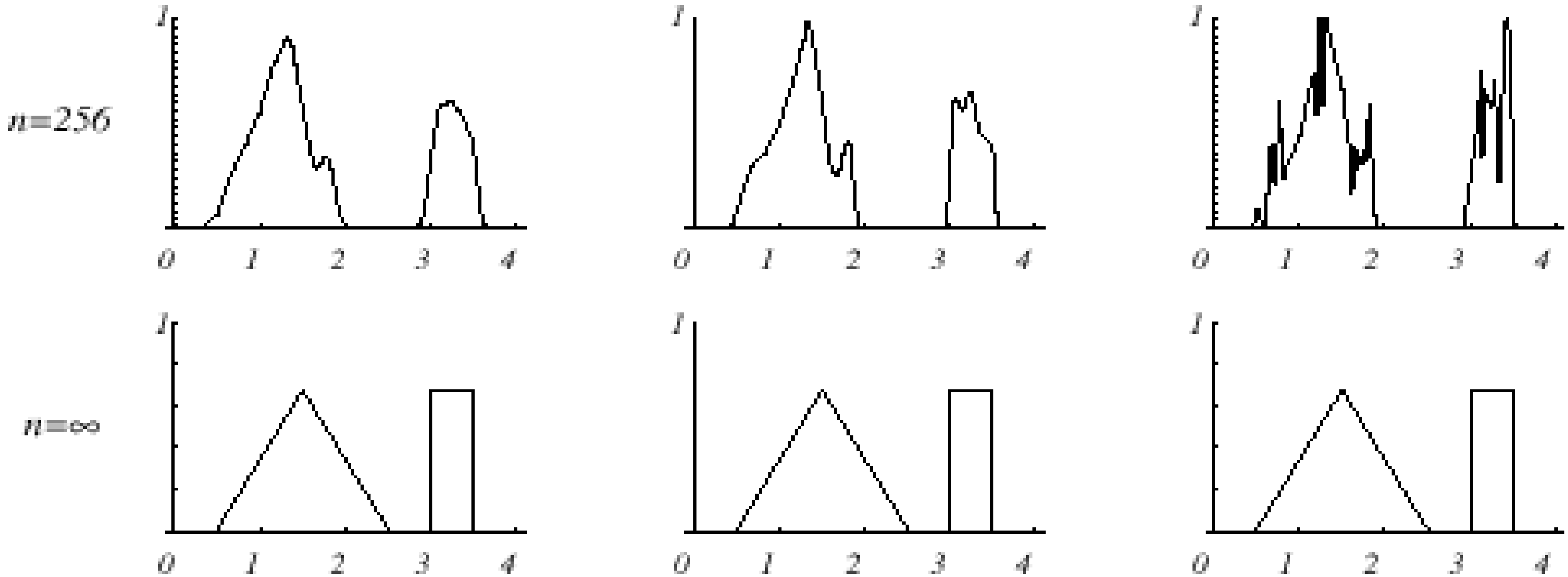


FIGURE 4.7. Parzen-window estimates of a bimodal distribution using different window widths and numbers of samples. Note particularly that the $n = \infty$ estimates are the same (and match the true distribution), regardless of window width. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- **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.

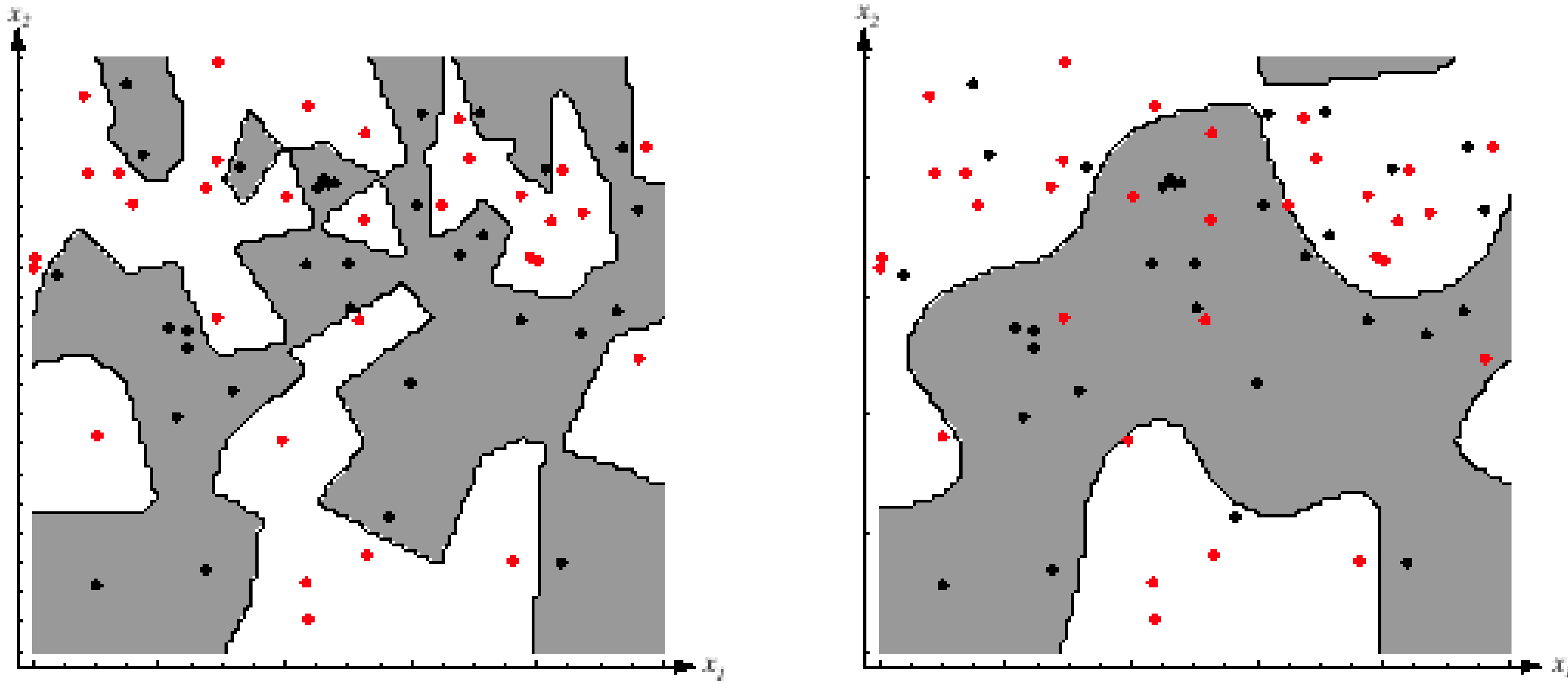


FIGURE 4.8. The decision boundaries in a two-dimensional Parzen-window dichotomizer depend on the window width h. At the left a small h leads to boundaries that are more complicated than for large h on same data set, shown at the right. Apparently, for these data a small h would be appropriate for the upper region, while a large h would be appropriate for the lower region; no single window width is ideal overall. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.