Nonparametric Techniques for Density Estimation (DHS Ch. 4)

- **Introduction**
- **Density Estimation Procedure**
- **Concept of General Techniques**
- **Parzen Window Estimation & Example**
- **K_n-Nearest Neighbor Estimation & Example**

Introduction

Suppose you don't know the form of the densities

Try to estimate $p(x|S_k)$ (=likelihood) or $P(S_k|x)$ (= a posteriori)

➔ Estimation of probability density functions.

Consider sample vector $x_1, x_2, \ldots, x_J$, drawn from a class independently with probability density $p(x)$.

The probability that a vector x lies in a region *R* is

$$P = \int_R p(x')dx'$$

P is a smoothed or averaged version of the density function p(x')

[Consider Binomial Case]

Probability that k of the vectors lie in *R* is binomial (if samples drawn i.i.d.)

$$P_k = \binom{J}{k} P^k (1-P)^{J-k}$$ probability of k samples out of J fall in *R*.

P = probability that it lies in *R*

1-P = probability that it doesn't lie in *R*

Mean $E\{k\} = JP$

k/J = reasonable estimate for P. P=k/J

Assume region *R* is small and has a volume V (if we can find).

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

$$p(x) \approx P/V$$

Then, leads to an estimate

$$p(x) \approx (k/J) / V = k/(JV)$$

Would like to take a limit to V-> to reduce smoothing of p(x') but number of samples is finite.

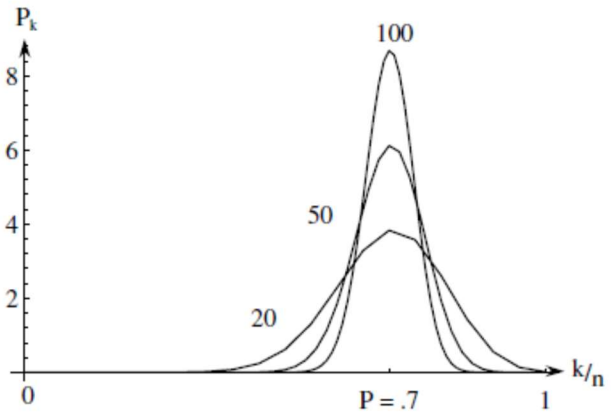


Figure 4.1: The probability P_k of finding k patterns in a volume where the space averaged probability is P as a function of k/n. Each curve is labelled by the total number of patterns n. For large n, such binomial distributions peak strongly at $k/n = P$ (here chosen to be 0.7).

Some Problems

If we fix the volume and take more training samples => the ratio k/J will converge to P, but p(x) will be space-averaged value of p(x). (check out Fig. 4.1 above)

If V is getting big, p(x) gets smaller.

If we shrink V to zero and fix the number of n samples, *R* becomes too small to enclose any samples in V, making p(x) close to zero.

Estimation Procedure to Estimate the Density of x

1. Form a sequence of regions R_1, R_2, …
2. Region R_j is employed for j samples
3. Let V_j be the volume of R_j
4. Let k_j be the number of samples falling in R_j
5. The j-th estimate of p(x) is $p_j(x)=[k_j/j]/ V_j$.
   $p_j(x)$ will converge to p(x) if:

   (1) $\lim_{j->\infty} V_j = 0$

   (2) $\lim_{j->\infty} k_j = \infty$

   (3) $\lim_{j->\infty} k_j / j = 0$

Many ways of satisfying these conditions:

1. Shrink the regions, say $V_j = 1/\sqrt{j}$ (Parzen window)

2. Let $k_j = \sqrt{j}$, and let the volume grow to enclose k_j neighbors of x. (nearest neighbor)

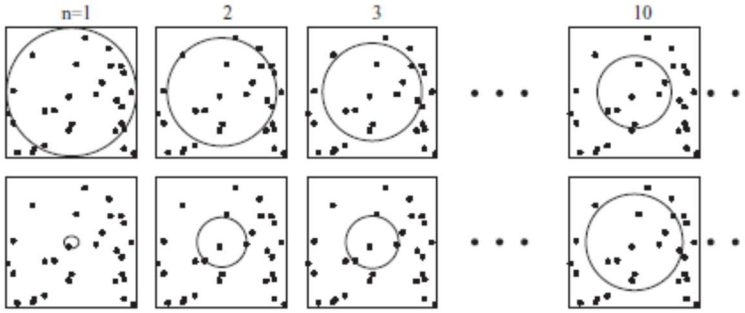
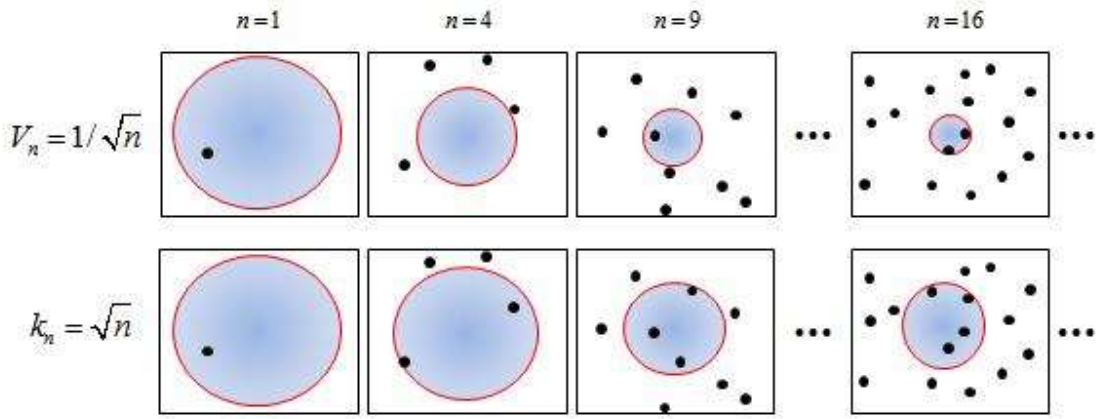


Figure 4.2: Two methods for estimating the density at a point x (at the center of each square) are to xxx.



Concepts of General Techniques

Estimate p(x) from samples x_j

Technique (1): Histogram, fixed bin size and location

Count number samples that fall into each bin. (crude estimate)

Technique (2): Fixed bin size, variable bin location (i.e., sliding bins)

Count number samples that fall into region centered at x, for each x.

Technique (3): Bin locations set by samples, bin shape is a parameter.

Each sample x_i gives rise to a window function centered about x_i. Estimate p(x) by summing over window functions.

Window function $\Delta(x-x_i)$

$$p_j(x) = 1/j \sum_{i=1}^{j} \Delta(x - x_j)$$

(2) and (3) are equivalent for certain choices of window functions Δ.

Two Popular Techniques in Nonparametric Techniques

1) **Parzen Window Estimation**
2) **Nearest Neighbor Estimation**

---

1) Parzen Window Estimation (DHS 4.3)

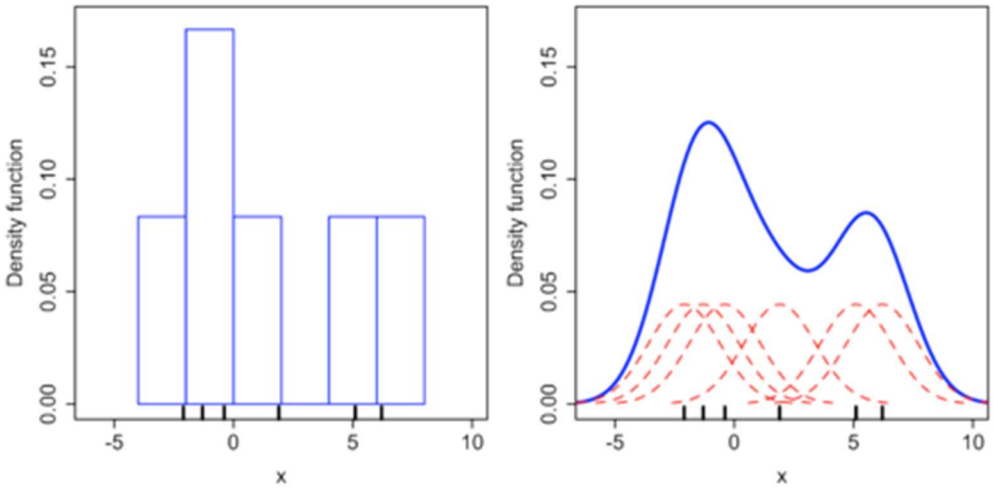
Define a window function $\Delta(\underline{u})=\Delta(\underline{x}-\underline{x}_i)$

Estimate $p(\underline{x})$. Given a sample $\underline{x}=\underline{x}_i$, $p(\underline{x}_i)$ is nonzero, and if $p(\underline{x})$ is continuous, $p(\underline{x})$ is nonzero for $\underline{x}$ close to $\underline{x}_i$

Use window function $\Delta(\underline{x}-\underline{x}_i)$ centered at $\underline{x}_i$. Δ should be non-increasing.

Estimate of $p(\underline{x})$ is

$$p_j(\underline{x}) = (1/j) \sum_{i=1}^{j} \Delta(\underline{x}-\underline{x}_i) \qquad \text{(Parzen window estimate)}$$



To ensure that $p_j(x)$ represents a density, require:

(*) $\Delta(\underline{u}) \geq 0$

$\int \Delta(\underline{u}) d\underline{u} = 1$

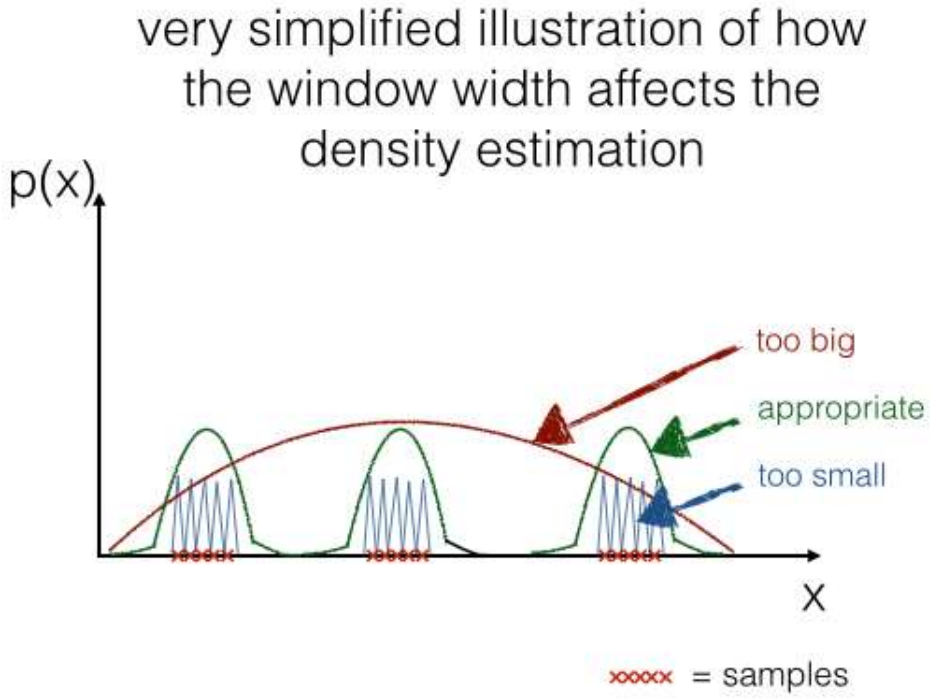
Let $\Delta_j(\underline{x}) = (1/V_j)\, \Phi(x)$

If $V_j = h_j^d$ (d is dimension)

Choice of scale or width of $\Delta_j(\underline{x})$ is important. h_j affects both the amplitude and the width.

Small width => high resolution in $p_j(\underline{x})$, but noisy

Large width => $p_j(\underline{x})$ will be over-smoothed.


very simplified illustration of how the window width affects the density estimation
p(x)
too big
appropriate
too small
X
= samples

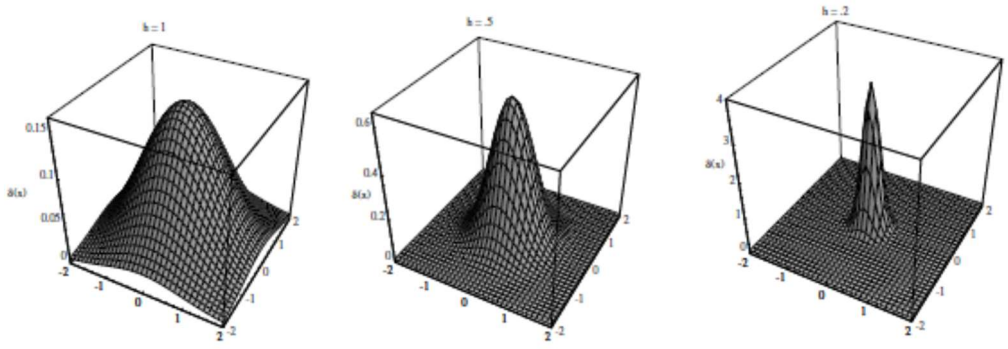


Figure 4.3: Examples of two-dimensional circularly symmetric normal Parzen windows $\varphi(\mathbf{x}/h)$ for three different values of h. Note that because the $\delta_k(\cdot)$ are normalized, different vertical scales must be used to show their structure.

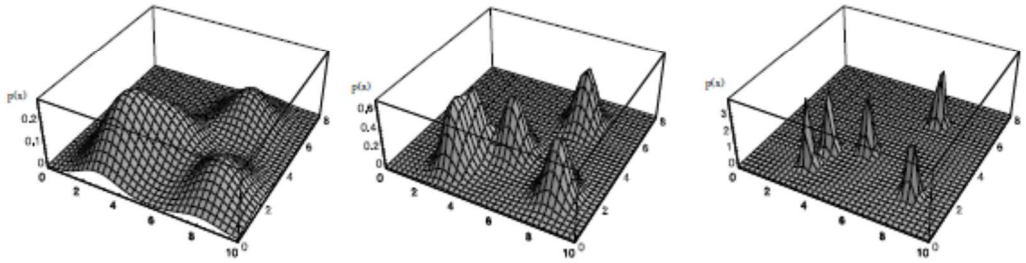


Figure 4.4: Three Parzen-window density estimates based on the same set of five samples, using the window functions in Fig. 4.3. As before, the vertical axes have been scaled to show the structure of each function.

Choice of h_j or V_j affects on $p_j(x)$. If V_j is too large, the estimate will suffer from too little resolution. If V_j is too small, the estimate will suffer from too much statistical variability.

With a limited number of samples, the best is to accept compromise.

With an unlimited number of samples, let V_j slowly approach zero as j increases and have $p_j(x)$ converges to the unknown density p(x).

Parzen Window Example (DHS 4.3.3)

Unknown density p(x) is normal.

$p(x) = N(N, m, \sigma^2)$

zero-mean, unit-variance, univariate normal density

Choose a window function:

$\Phi(u) = 1/(\text{sqrt}(2\pi)) \exp \{ (-1/2)u^2 \}$

$\Delta_j(\underline{x}-\underline{x}_i)=(1/h_j)\, \Phi\, [(\underline{x}-\underline{x}_j)/h_j]$

$$= 1/(\text{sqrt}(2\pi)h_j) \exp \{ (-1/2)\left(\frac{x-x_i}{h_j} \right)^2 \}$$

Window width = $h_j = h_1/\text{sqrt}(j)$

h_1 is a parameter at our disposal.

$$p_j(x) = (1/j) \sum_{i=1}^{j} \Delta_j(x-x_i)$$

$$p_j(x) = \frac{1}{j} \sum_{i=1}^{j} \frac{1}{\sqrt{2\pi} h_j} \exp\left\{ -\frac{1}{2} \left(\frac{x-x_i}{h_j} \right)^2 \right\}$$

(x_i is an observed sample)

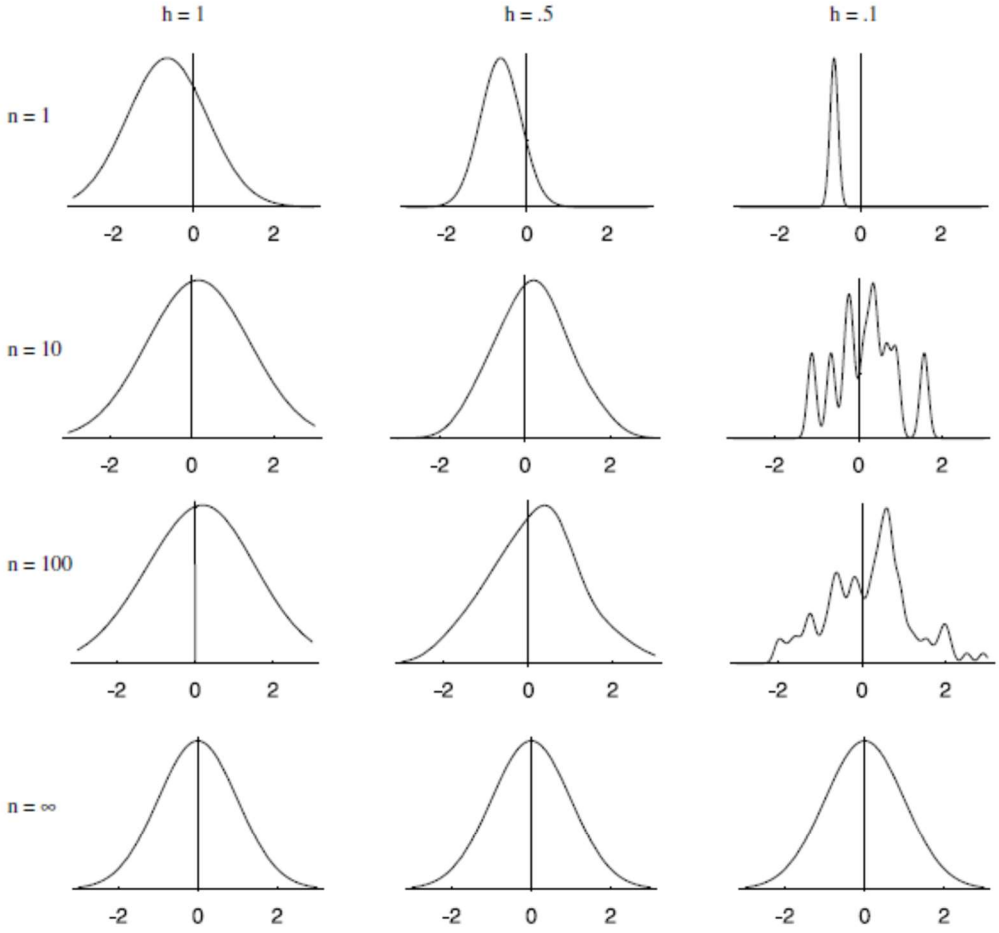


Figure 4.5: Parzen-window estimates of a univariate normal density using different window widths and numbers of samples. The vertical axes have been scaled to best show the structure in each graph. Note particularly that the $n = \infty$ estimates are the same (and match the true generating function),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

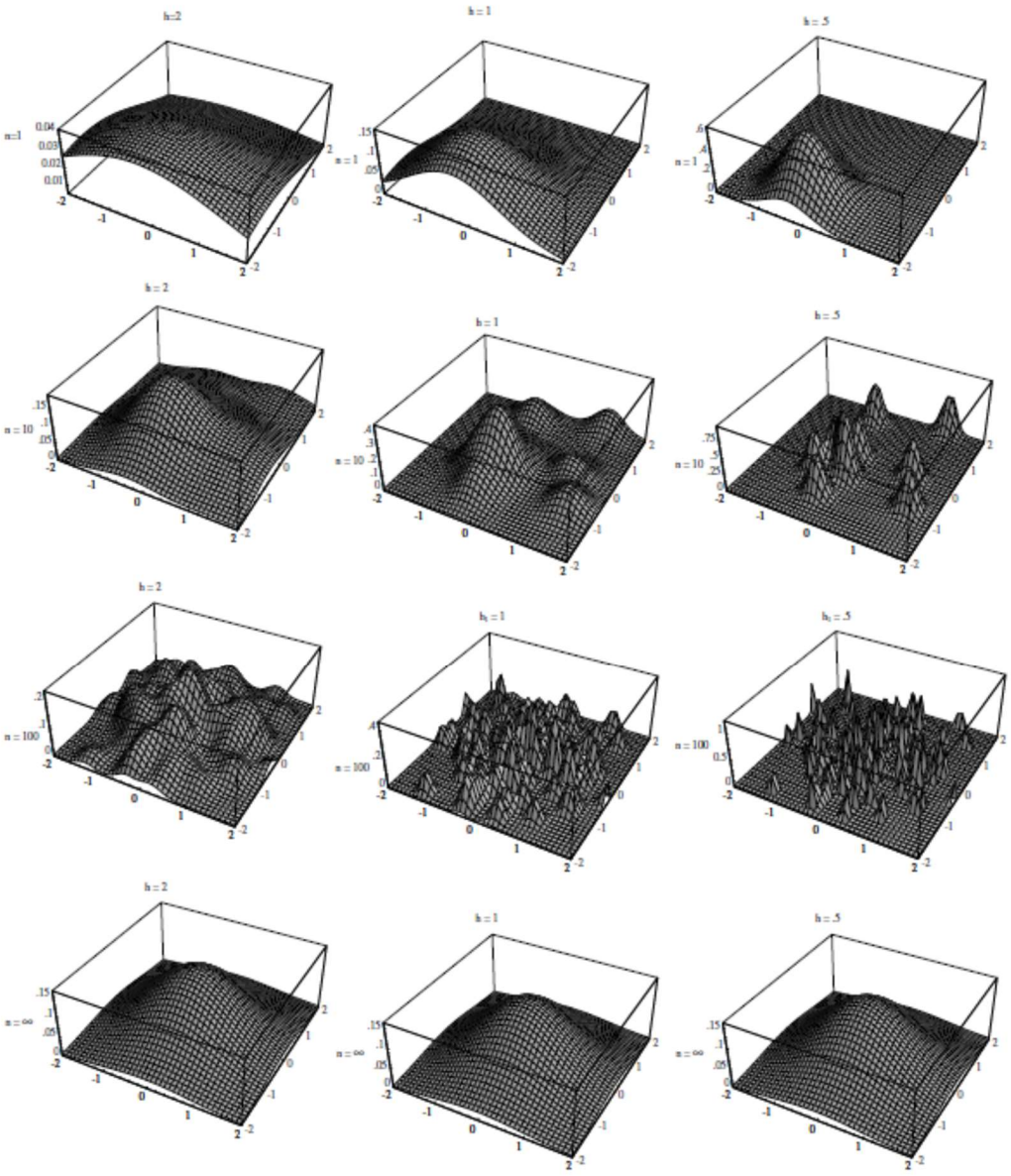


Figure 4.6: Parzen-window estimates of a bivariate normal density using different window widths and numbers of samples. The vertical axes have been scaled to best show the structure in each graph. Note particularly that the $n = \infty$ estimates are the same (and match the true generating distribution), regardless of window width h.

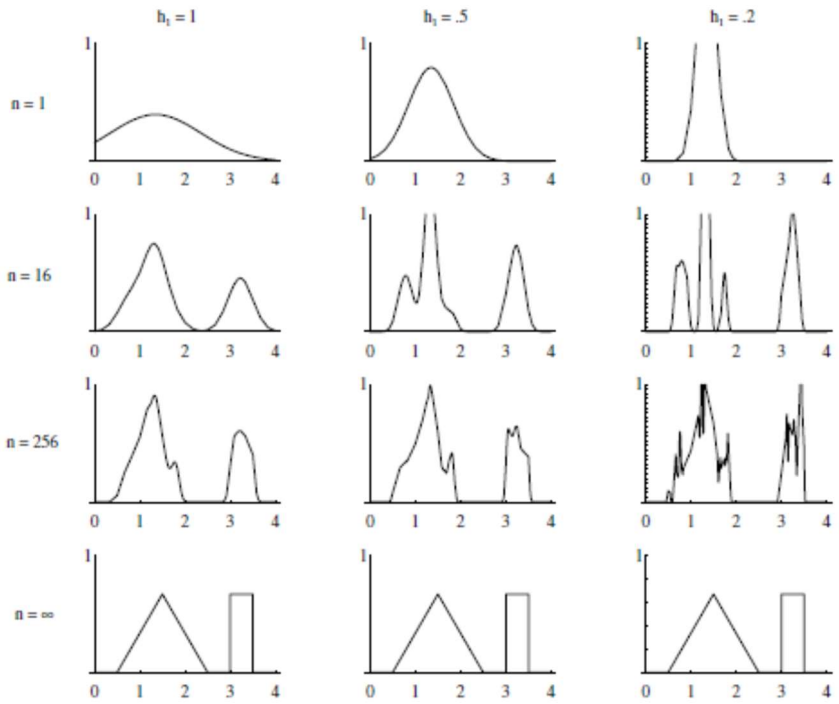


Figure 4.7: Parzen-window estimates of a bimodal distribution using different window widths and numbers of samples. Note particularly that the $n = \infty$ estimates are the same (and match the true generating distribution), regardless of window width h.

Classification Example

Classifier based on Parzen-window estimation.

Estimate the densities for each category and classify a test point by the label corresponding to the maximum posterior.

Figure 4.8 Decision regions for a Parzen-window classifier depend upon the choices of window function.

In general, the training error can be made arbitrarily low by making the window width sufficiently small. But a low training error does not guarantee a small test error.

Curse of dimensionality: demand for a larger number of samples grows exponentially with the dimensionality of the feature space.

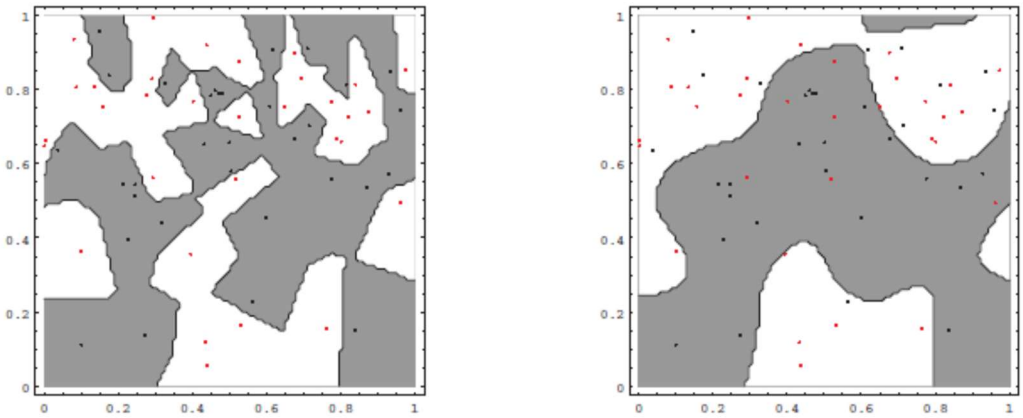


Figure 4.8: The decision boundaries in a two-dimensional Parzen-window dichotomizer depend on the window width h. At the left a small h leads to boundaries that are more complicated than for large h on same data set, shown at the right. Apparently, for this data a small h would be appropriate for the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

Parzen window techniques: advantages and disadvantages

Advantage

- Generality: No a prior assumptions (except continuity of p(x)). Given enough samples, it is guaranteed to converge to correct density p(x).

Disadvantages

- Number of samples required is generally quite large
- Number of samples required grows exponentially with the number of dimensions in feature space.
- Choice of sizes of regions V_j is important.

Choosing the window function (DHS 4.3.6)

One of problems in Parzen-window approach is the choice of the sequence of cell-volume sizes V_1, V_2, … or overall window size.

If $V_j = V_1/\text{sqrt}(j)$, the results of any finite j will be sensitive to the choice of the initial volume V_1

If V_1 is too small, most of the volume will be empty

If V_1 is too large, important spatial variations in p(x) could be lost due to averaging.