

# Non-parametric Methods

Selim Aksoy

Department of Computer Engineering

Bilkent University

saksoy@cs.bilkent.edu.tr

CS 551, Spring 2009



# Introduction

- ▶ Density estimation with parametric models assumes that the forms of the underlying density functions are known.
- ▶ However, common parametric forms do not always fit the densities actually encountered in practice.
- ▶ In addition, most of the classical parametric densities are unimodal, whereas many practical problems involve multimodal densities.
- ▶ Non-parametric methods can be used with arbitrary distributions and without the assumption that the forms of the underlying densities are known.



# Non-parametric Density Estimation

- ▶ Suppose that  $n$  samples  $\mathbf{x}_1, \dots, \mathbf{x}_n$  are drawn i.i.d. according to the distribution  $p(\mathbf{x})$ .
- ▶ The probability  $P$  that a vector  $\mathbf{x}$  will fall in a region  $\mathcal{R}$  is given by

$$P = \int_{\mathcal{R}} p(\mathbf{x}') d\mathbf{x}'.$$

- ▶ The probability that  $k$  of the  $n$  will fall in  $\mathcal{R}$  is given by the binomial law

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

- ▶ The expected value of  $k$  is  $E[k] = nP$  and the MLE for  $P$  is  $\hat{P} = \frac{k}{n}$ .



# Non-parametric Density Estimation

- ▶ If we assume that  $p(\mathbf{x})$  is continuous and  $\mathcal{R}$  is small enough so that  $p(\mathbf{x})$  does not vary significantly in it, we can get the approximation

$$\int_{\mathcal{R}} p(\mathbf{x}') d\mathbf{x}' \simeq p(\mathbf{x})V$$

where  $\mathbf{x}$  is a point in  $\mathcal{R}$  and  $V$  is the volume of  $\mathcal{R}$ .

- ▶ Then, the density estimate becomes

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



# Non-parametric Density Estimation

- ▶ Let  $n$  be the number of samples used,  $\mathcal{R}_n$  be the region used with  $n$  samples,  $V_n$  be the volume of  $\mathcal{R}_n$ ,  $k_n$  be the number of samples falling in  $\mathcal{R}_n$ , and  $p_n(\mathbf{x}) = \frac{k_n/n}{V_n}$  be the estimate for  $p(\mathbf{x})$ .
- ▶ If  $p_n(\mathbf{x})$  is to converge to  $p(\mathbf{x})$ , three conditions are required:

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

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

$$\lim_{n \rightarrow \infty} \frac{k_n}{n} = 0.$$



# Histogram Method

- ▶ A very simple method is to partition the space into a number of equally-sized cells (*bins*) and compute a *histogram*.

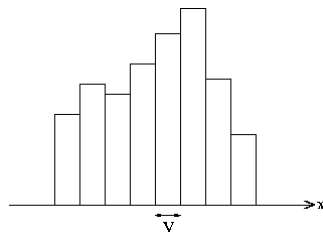


Figure 1: Histogram in one dimension.

- ▶ The estimate of the density at a point  $\mathbf{x}$  becomes

$$p(\mathbf{x}) = \frac{k}{nV}$$

where  $n$  is the total number of samples,  $k$  is the number of samples in the cell that includes  $\mathbf{x}$ , and  $V$  is the volume of that cell.



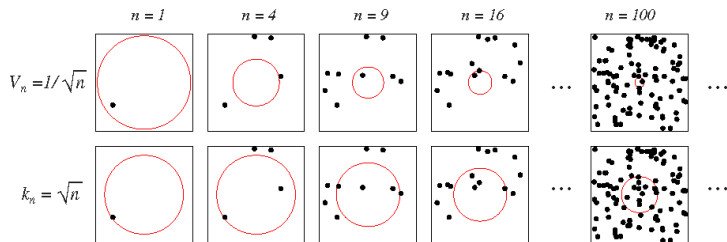
# Histogram Method

- ▶ Although the histogram method is very easy to implement, it is usually not practical in high-dimensional spaces due to the number of cells.
- ▶ Many observations are required to prevent the estimate being zero over a large region.
- ▶ Modifications for overcoming these difficulties:
  - ▶ Data-adaptive histograms,
  - ▶ Independence assumption (naive Bayes),
  - ▶ Lancaster models,
  - ▶ Dependence trees.



# Non-parametric Density Estimation

- ▶ Other methods for obtaining the regions for estimation:
  - ▶ Shrink regions as some function of  $n$ , such as  $V_n = 1/\sqrt{n}$ . This is the *Parzen window* estimation.
  - ▶ Specify  $k_n$  as some function of  $n$ , such as  $k_n = \sqrt{n}$ . This is the *k-nearest neighbor* estimation.



**Figure 2:** Methods for estimating the density at a point, here at the center of each square.



# Parzen Windows

- Suppose that  $\varphi$  is a  $d$ -dimensional window function that satisfies the properties of a density function, i.e.,

$$\varphi(\mathbf{u}) \geq 0 \quad \text{and} \quad \int \varphi(\mathbf{u}) d\mathbf{u} = 1.$$

- A density estimate can be obtained as

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

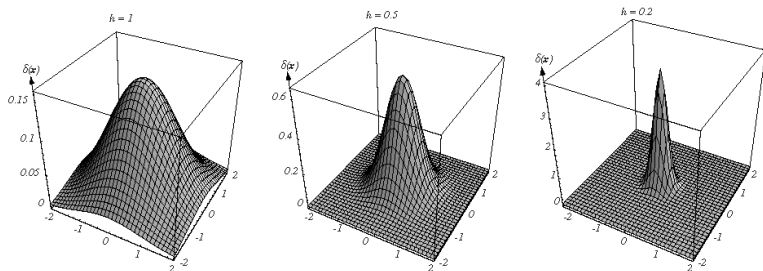
where  $h_n$  is the window width and  $V_n = h_n^d$ .



# Parzen Windows

- The density estimate can also be written as

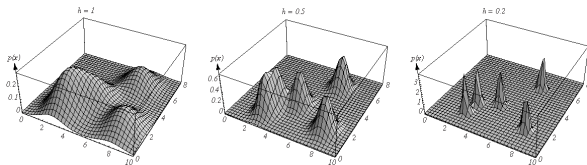
$$p_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \delta_n(\mathbf{x} - \mathbf{x}_i) \quad \text{where} \quad \delta_n(\mathbf{x}) = \frac{1}{V_n} \varphi\left(\frac{\mathbf{x}}{h_n}\right).$$



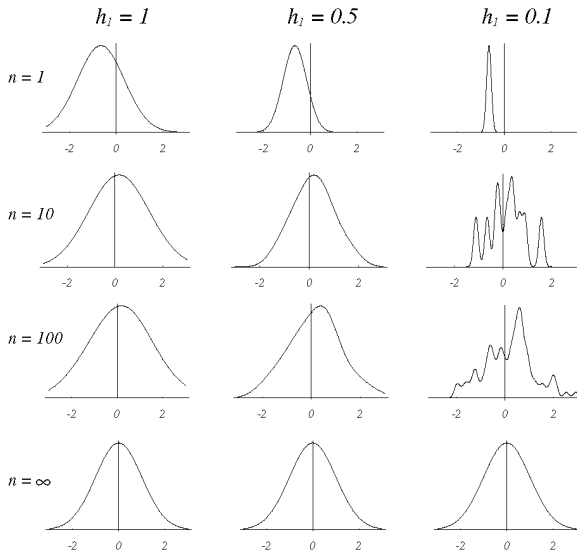
**Figure 3:** Examples of two-dimensional circularly symmetric Parzen window functions for three different values of  $h_n$ . The value of  $h_n$  affects both the amplitude and the width of  $\delta_n(\mathbf{x})$ .

# Parzen Windows

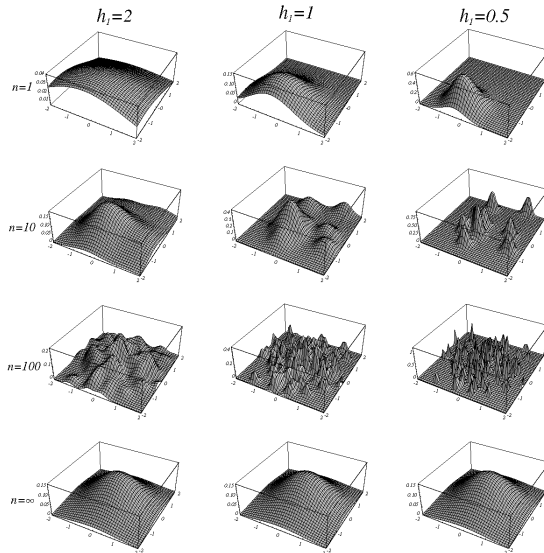
- ▶ If  $h_n$  is very large,  $p_n(\mathbf{x})$  is the superposition of  $n$  broad functions, and is a smooth “out-of-focus” estimate of  $p(\mathbf{x})$ .
- ▶ If  $h_n$  is very small,  $p_n(\mathbf{x})$  is the superposition of  $n$  sharp pulses centered at the samples, and is a “noisy” estimate of  $p(\mathbf{x})$ .
- ▶ As  $h_n$  approaches zero,  $\delta_n(\mathbf{x} - \mathbf{x}_i)$  approaches a Dirac delta function centered at  $\mathbf{x}_i$ , and  $p_n(\mathbf{x})$  is a superposition of delta functions.



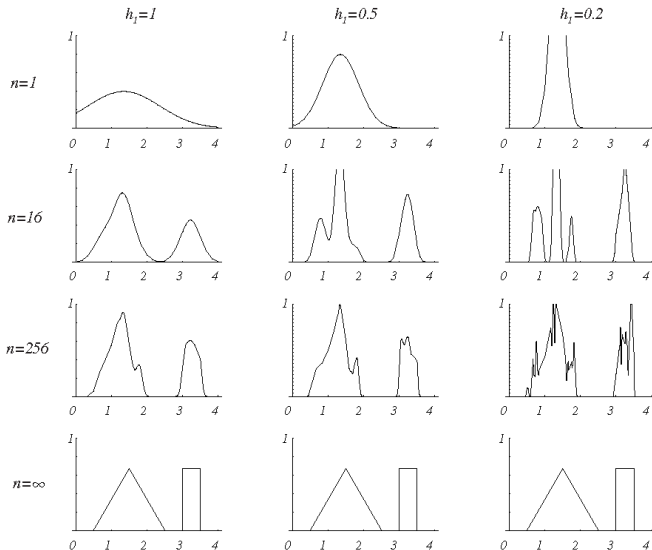
**Figure 4:** Parzen window density estimates based on the same set of five samples using the window functions in the previous figure.



**Figure 5:** Parzen window estimates of a univariate Gaussian density using different window widths and numbers of samples where  $\varphi(u) = N(0, 1)$  and  $h_n = h_1/\sqrt{n}$ .



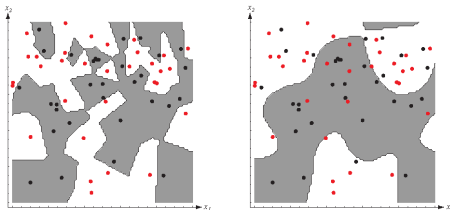
**Figure 6:** Parzen window estimates of a bivariate Gaussian density using different window widths and numbers of samples where  $\varphi(\mathbf{u}) = N(\mathbf{0}, \mathbf{I})$  and  $h_n = h_1/\sqrt{n}$ .



**Figure 7:** Estimates of a mixture of a uniform and a triangle density using different window widths and numbers of samples where  $\varphi(u) = N(0, 1)$  and  $h_n = h_1/\sqrt{n}$ .

# Parzen Windows

- ▶ Densities estimated using Parzen windows can be used with the Bayesian decision rule for classification.
- ▶ The training error can be made arbitrarily low by making the window width sufficiently small.
- ▶ However, the goal is to classify novel patterns so the window width cannot be made too small.



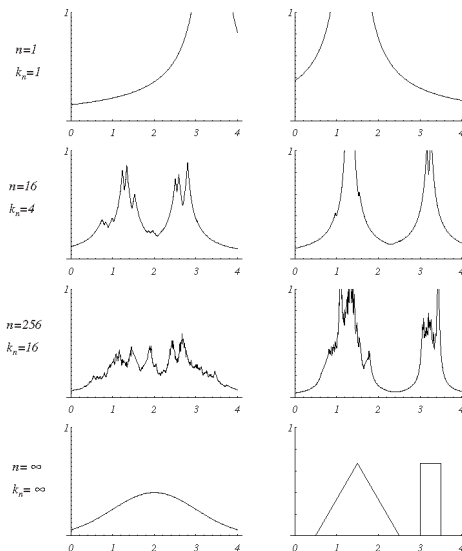
**Figure 8:** Decision boundaries in 2-D. The left figure uses a small window width and the right figure uses a larger window width.

# $k$ -Nearest Neighbors

- ▶ A potential remedy for the problem of the unknown “best” window function is to let the estimation volume be a function of the training data, rather than some arbitrary function of the overall number of samples.
- ▶ To estimate  $p(\mathbf{x})$  from  $n$  samples, we can center a volume about  $\mathbf{x}$  and let it grow until it captures  $k_n$  samples, where  $k_n$  is some function of  $n$ .
- ▶ These samples are called the  $k$ -nearest neighbors of  $\mathbf{x}$ .
- ▶ If the density is high near  $\mathbf{x}$ , the volume will be relatively small. If the density is low, the volume will grow large.







**Figure 9:**  $k$ -nearest neighbor estimates of two 1-D densities: a Gaussian and a bimodal distribution.

# $k$ -Nearest Neighbors

- ▶ 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

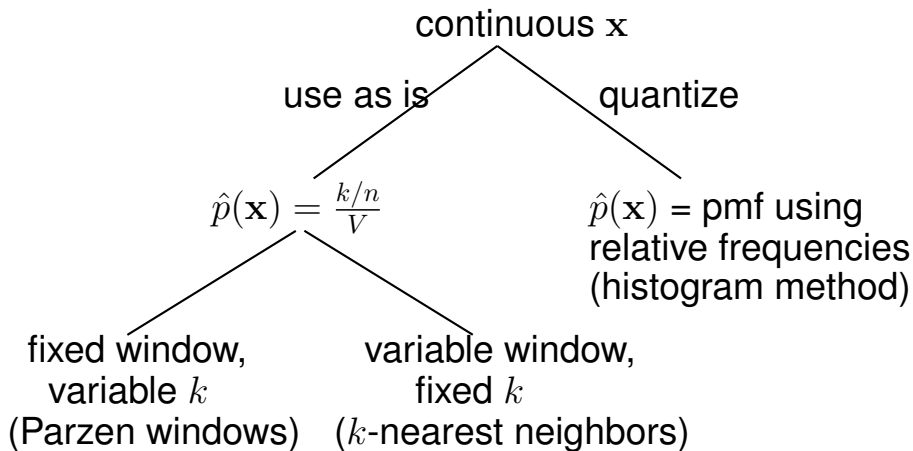
$$p_n(\mathbf{x}, w_i) = \frac{k_i/n}{V}$$

and gives an estimate for the posterior probability

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



# Non-parametric Methods



# Non-parametric Methods

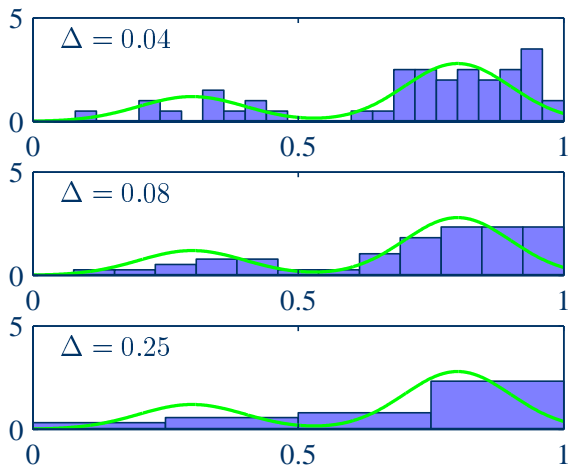
## ► Advantages:

- No assumptions are needed about the distributions ahead of time (generality).
- With enough samples, convergence to an arbitrarily complicated target density can be obtained.

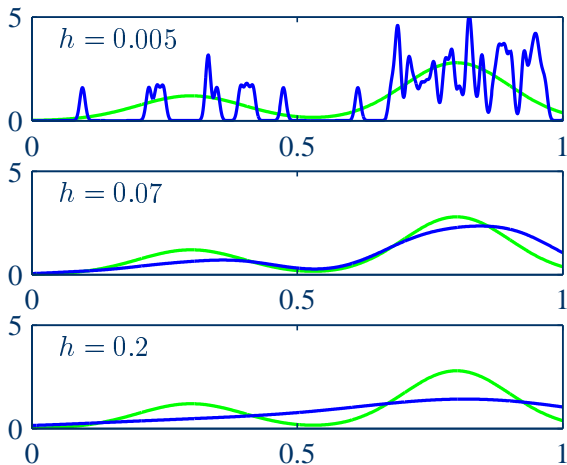
## ► Disadvantages:

- The number of samples needed may be very large (number grows exponentially with the dimensionality of the feature space).
- There may be severe requirements for computation time and storage.

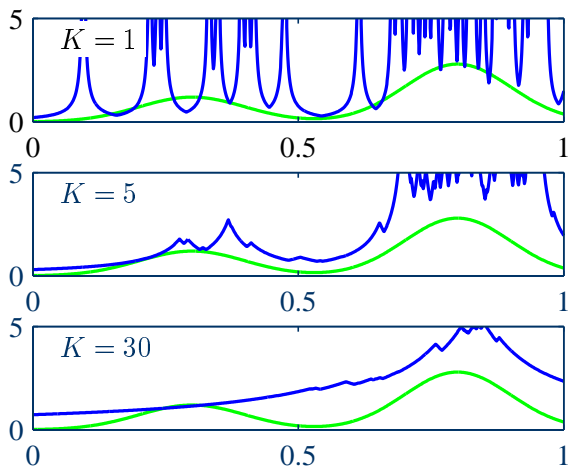




**Figure 10:** An illustration of the histogram approach to density estimation, in which a data set of 50 points is generated from the distribution shown by the green curve. Histogram density estimates are shown for various values of the cell volume ( $\Delta$ ).



**Figure 11:** 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.



**Figure 12:** 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) smooths out the bimodal nature of the true distribution.

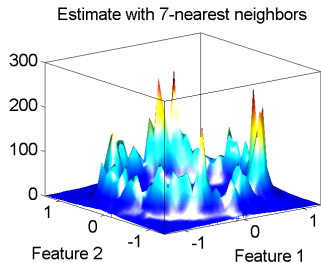
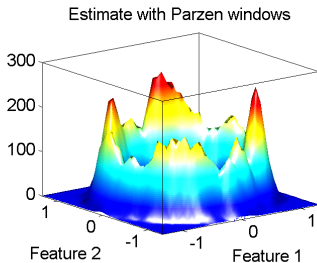
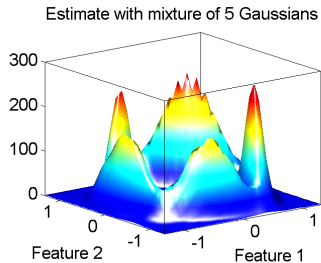
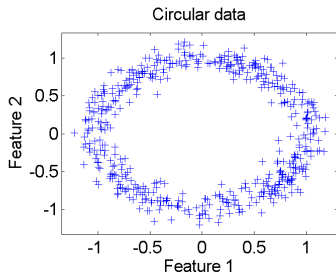


Figure 13: Density estimation examples for 2-D circular data.



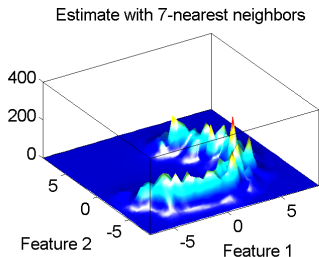
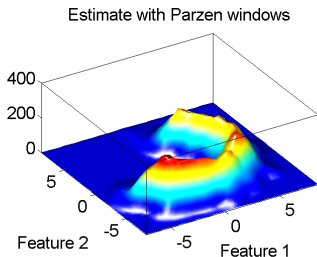
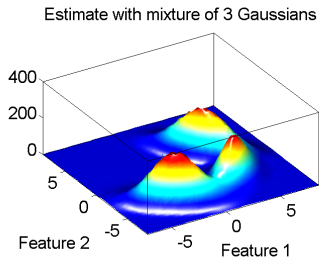
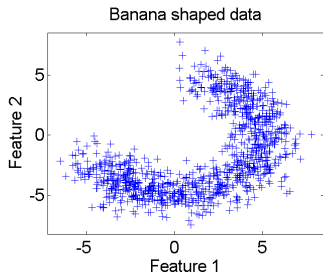


Figure 14: Density estimation examples for 2-D banana shaped data.