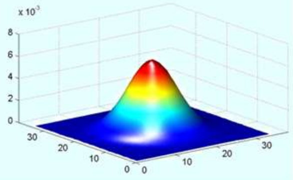
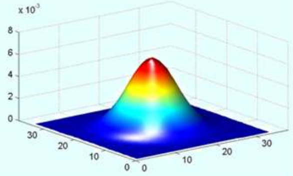


Non-parametric methods for PDF estimation

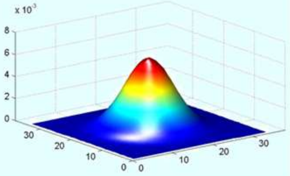


Bibliography

Density estimation for statistics & data analysis
Silverman, B.W. (1986) CRC press

See Google books

http://books.google.es/books?vid=ISBN0412246201



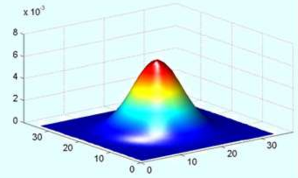
PDF estimation: goal

Obtain an estimation of the PDF from the data in the sample

Types of methods:

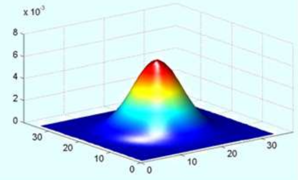
- **Parametric:** they start from a functional form of the PDF (e.g. a physical model) and fit its parameters using the sample (e.g. maximum likelihood)

- **Non parametric:** try to model the PDF without making any hypothesis about its functional form



For simplicity we will study the univariate case:

- Random variable x
- PDF $f(x)$
- Probability $$P(a < x < b) = \int_a^b f(x)\,dx$$
- Sample: $x_1, \ldots, x_n$

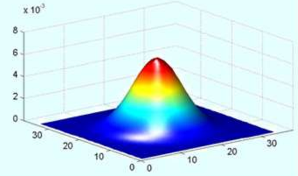


Histograms

- It's the oldest method and the most used
- Given an origin x_0 and a width *h* we count the objects inside each interval $[x_0+ih, x_0+(i+1)h]$ $i=0,1,2, \ldots$
- We estimate the PDF as

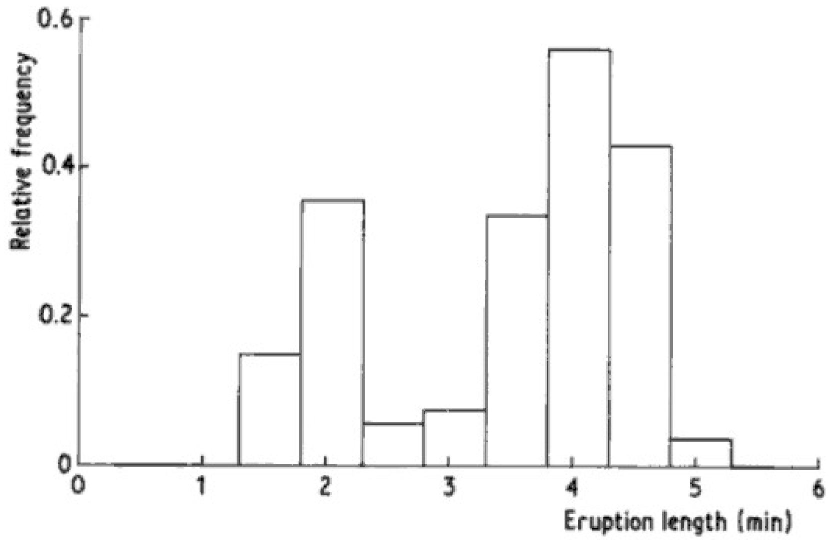
$$\hat{f}(x) = \frac{N_{obj}\left(x_0 + ih, x_0 + (i+1)h\right)}{nh}$$

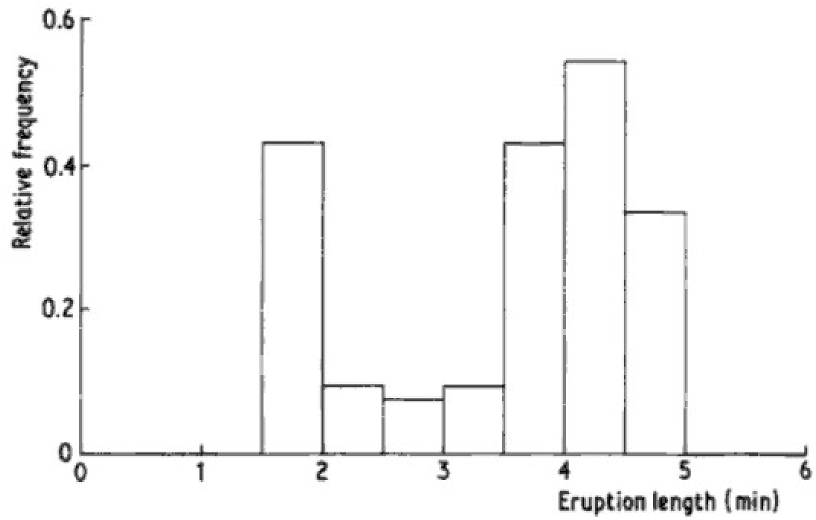
where *n* is the total number of objects

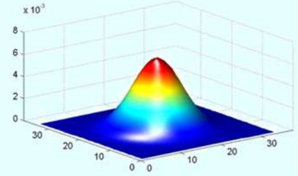


Disadvantages:

- Very convenient for 1D data, but unpractical for multidimensional data
- The estimation of the PDF is discontinuous
- Is very dependent of the origin and width chosen







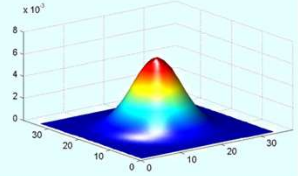
Simple estimator

- Starting from the definition of the PDF

$$f(x) = \lim_{h \to 0} \left(\frac{1}{2h} P(x - h < X < x + h) \right)$$

- We fix a small width *h* and define a “natural” estimator of the PDF

$$\hat{f}(x) = \frac{N_{obj}(x-h, x+h)}{2nh}$$



- It is usually expressed in the following way:

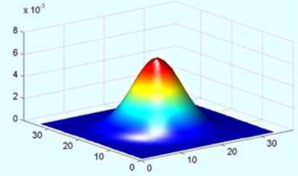
Weight function

$$w(\alpha) = \begin{cases} 1/2 & |\alpha| < 1 \\ 0 & \end{cases}$$

Estimator

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} w\left(\frac{x - x_i}{h} \right)$$

It is the sum over the sample of “boxes” centered in each one of the observations; is a kind of histogram where each point is the center of a sampling interval.

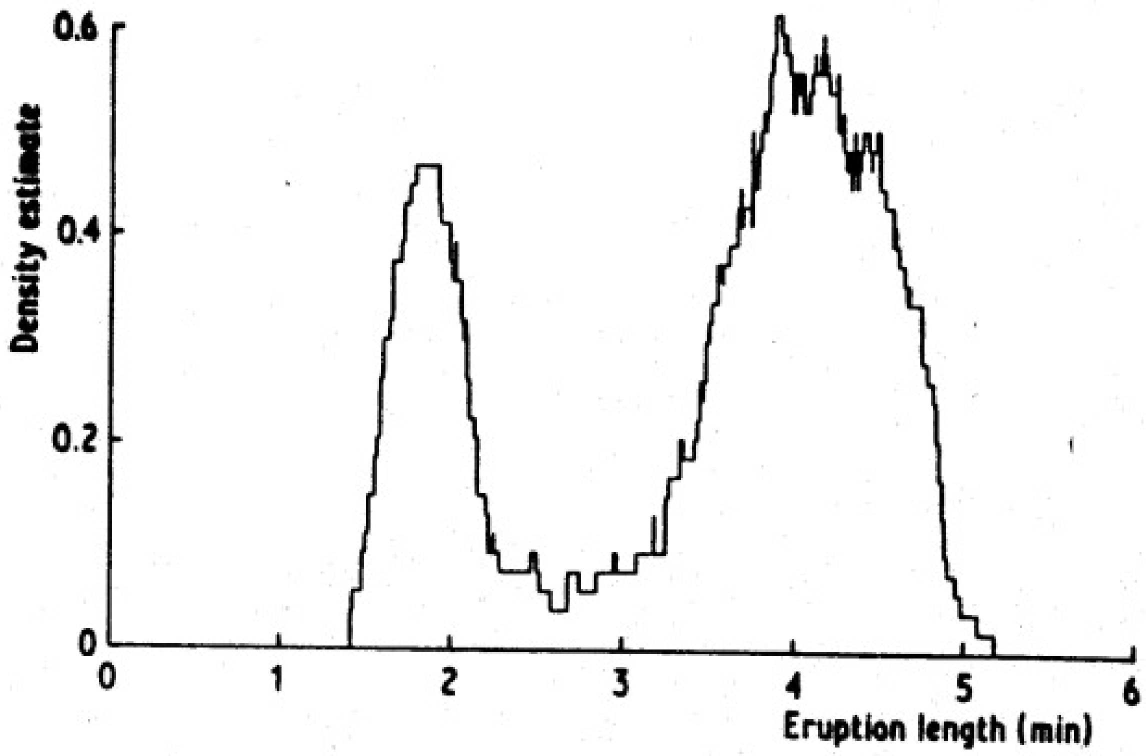


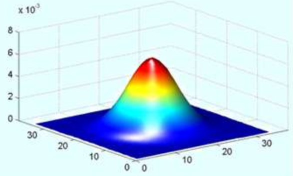
Advantages:

- It does not depend on any origin

Disadvantages:

- Depends of the chosen width *h*
- The estimation of the PDF is discontinuous





Kernel estimator

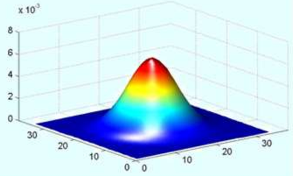
- Starting from the simple estimator, the weight function is substituted by a *kernel function* $\kappa(x)$ such that

$$\int_{-\infty}^{\infty} \kappa(x)\, dx = 1 \qquad \text{(normalized)}$$

and the estimator of the PDF is

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \kappa\left(\frac{x - x_i}{h} \right)$$

where *h* is the window width or smoothing parameter

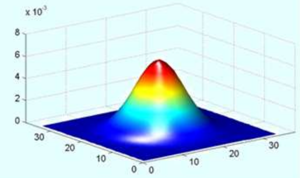


Advantages:

- This PDF estimator has all the continuity and differentiability properties of the kernel function

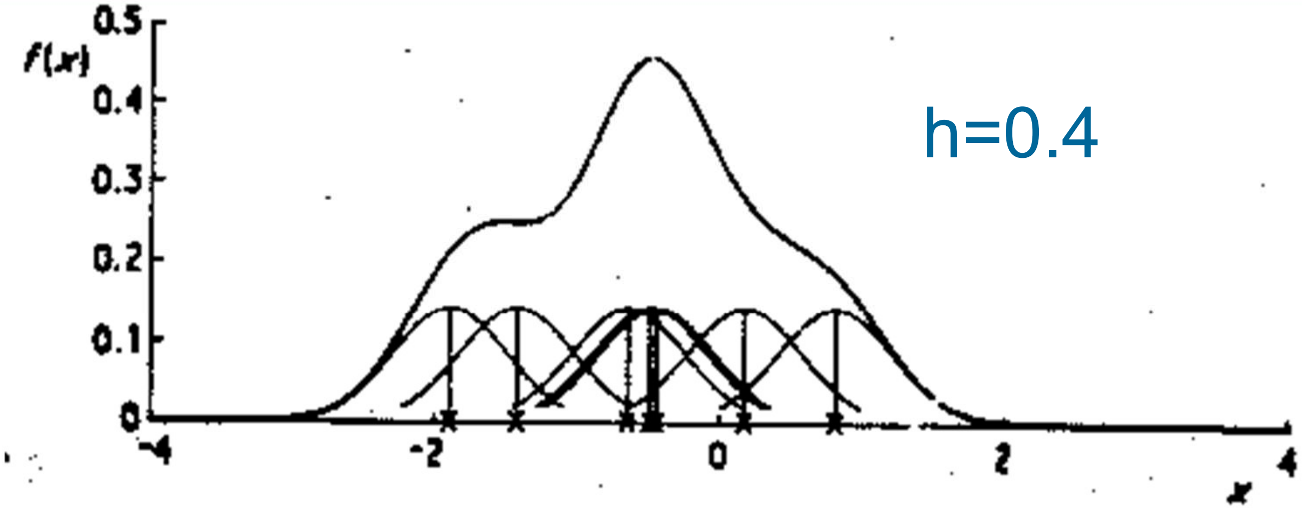
Disadvantages:

- Depends of the chosen width *h*
  - For sparse samples small values of *h* tend to produce spurious spikes
  - For large values of h the details of the PDF are lost (smoothed out)

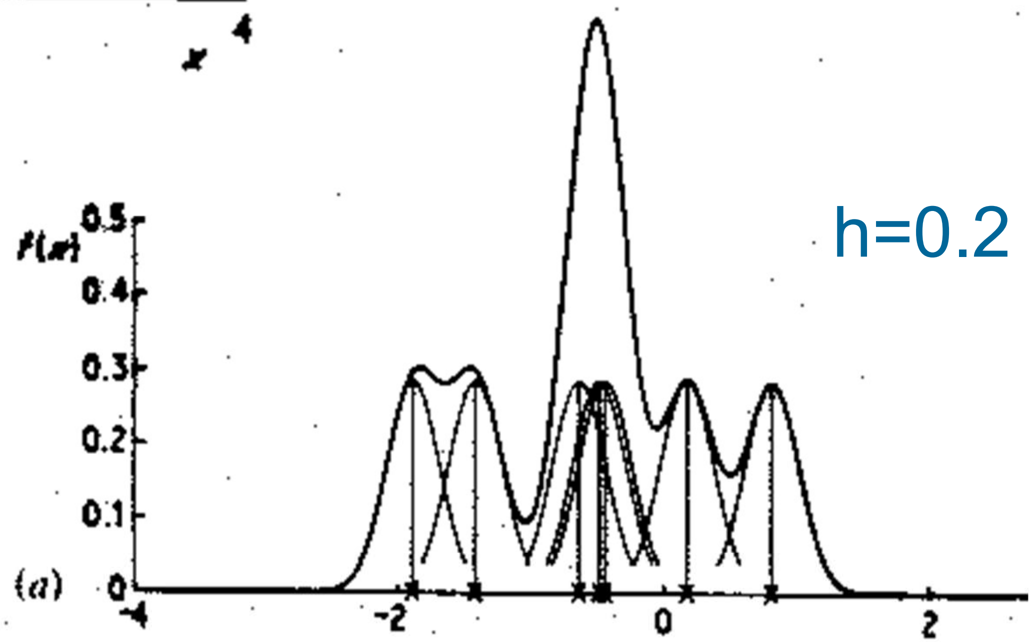


Example:

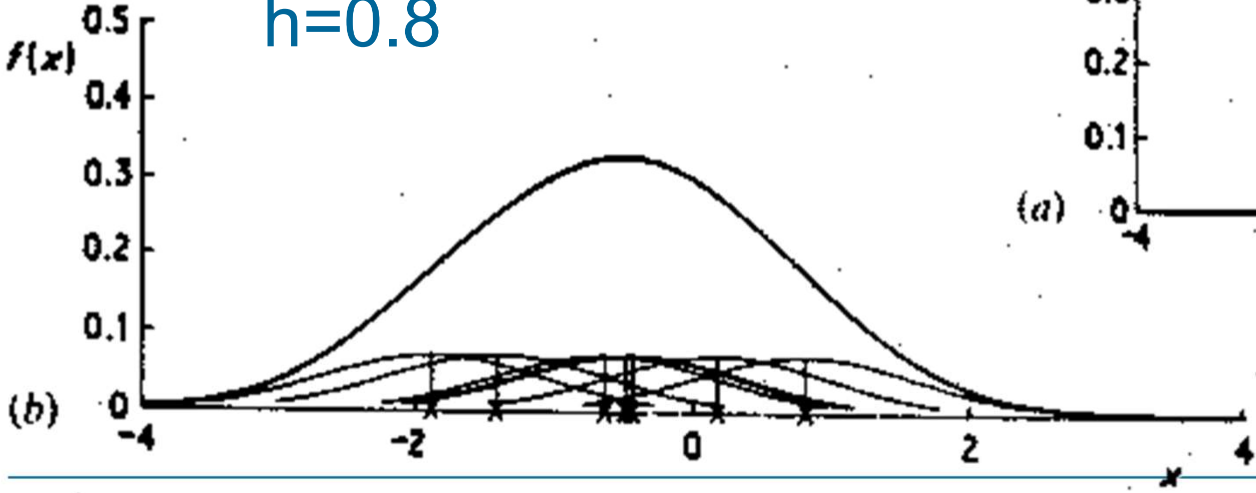
A gaussian kernel function: we are representing the PDF as a sum of gaussians centered in each one of the observations and a width determined by the *h* parameter h=0.4

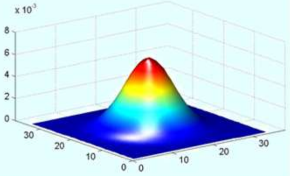

f(x)
0.5
0.4
0.3
0.2
0.1
0
-4
-2
0
2
4
x

h=0.2


f(x)
0.5
0.4
0.3
0.2
0.1
(a)
0
-4
-2
0
2

h=0.8

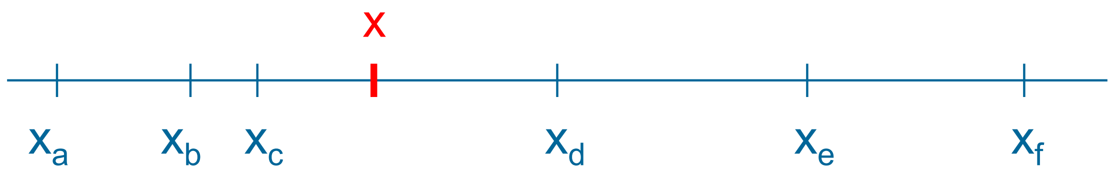

f(x)
0.5
0.4
0.3
0.2
0.1
(b)
0
-4
-2
0
2
4
x



Nearest Neighbor estimator

- Aims to adapt the smoothing to the local density of the data

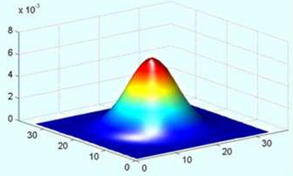
- For each point x of the possible range of values calculate and sort the distances to the points in the sample



$$d_1 \leq d_2 \leq \ldots \leq d_k \leq d_{k+1} \ldots$$

- Choose the smoothing factor *k*, a small integer, typically

$$k \cong n^{1/2}$$



- Define the PDF estimator as

$$\hat{f}(x) = \frac{k}{2n\,d_k(x)}$$

The idea is that with this definition the density of observations in an interval $2d_k(x)$ around x is just *k*:

$$k = n\,2d_k(x)\,f(x)$$

It’s like defining a simple estimator where the “box” is variable and chosen so that there are *k* observations in it.

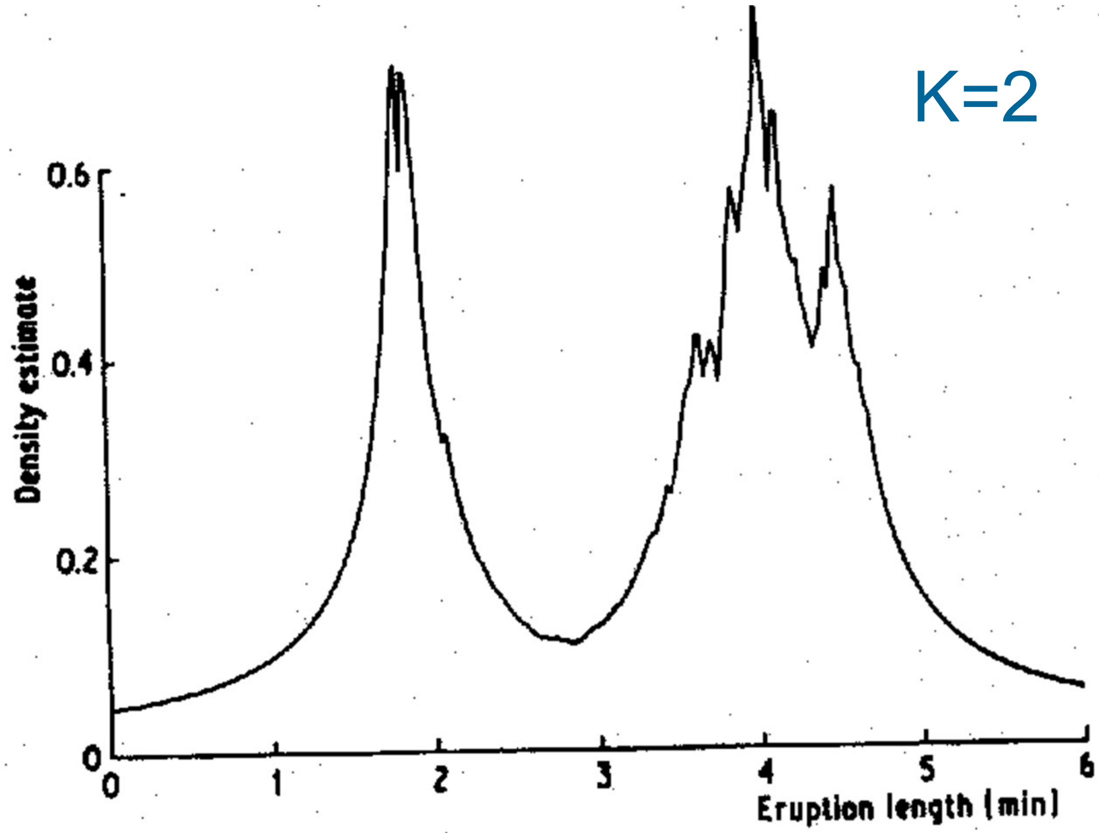
Advantages:

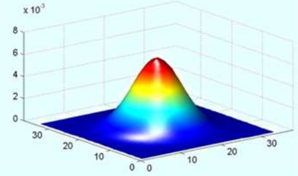
- It does not depend on the origin and *k* can be determined in a “natural” way as $n^{1/2}$

Disadvantages:

- It’s continuous, but it
- may not be derivable

K=2



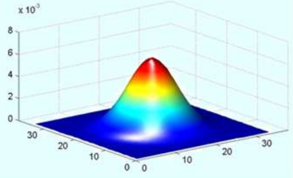


NN estimator with kernel

- We use the kernel functions in the NN method

$$\hat{f}(x) = \frac{1}{n\, d_k(x)} \sum_{i=1}^{n} \kappa\left(\frac{x - x_i}{d_k(x)} \right)$$

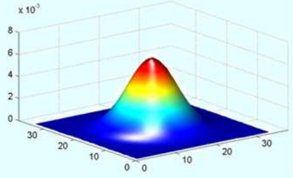
It’s like the kernel estimator but with a variable *h* parameter that is self-adapted to each point.



Advantages:

- Does not depend of any origin and k can be determined in a “natural” way as n1/2

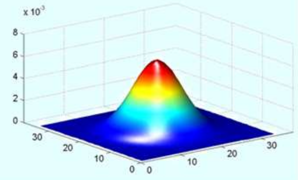
- Continuous and derivable



Question: which kernel function should we use?

Table 3.1 *Some kernels and their efficiencies*

Kernel	$K(t)$	Efficiency (exact and to 4 d.p.)
Epanechnikov	$\frac{3}{4}(1-\frac{1}{5}t^2)/\sqrt{5}$ for $\lvert t\rvert<\sqrt{5}$, 0 otherwise	1
Biweight	$\frac{15}{16}(1-t^2)^2$ for $\lvert t\rvert<1$, 0 otherwise	$\left(\frac{3087}{3125}\right)^{1/2} \approx 0.9939$
Triangular	$1-\lvert t\rvert$ for $\lvert t\rvert<1$, 0 otherwise	$\left(\frac{243}{250}\right)^{1/2} \approx 0.9859$
Gaussian	$\frac{1}{\sqrt{2\pi}} e^{-(1/2)t^2}$	$\left(\frac{36\pi}{125}\right)^{1/2} \approx 0.9512$
Rectangular	$\frac{1}{2}$ for $\lvert t\rvert<1$, 0 otherwise	$\left(\frac{108}{125}\right)^{1/2} \approx 0.9295$

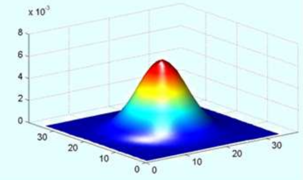


The properties of the kernel methods are well known:

- Biases
- Error of the estimator
- Optimisation techinques to choose the best smoothing factor

See Silverman (1986) for a discussion. Specifically see the **discrepancy measures between an estimator and the PDF**

$$MSE_x(\hat{f}) = E\left\{\left(\hat{f}(x) - f(x)\right)^2\right\}$$



Extension of the kernel methods to multivariate data

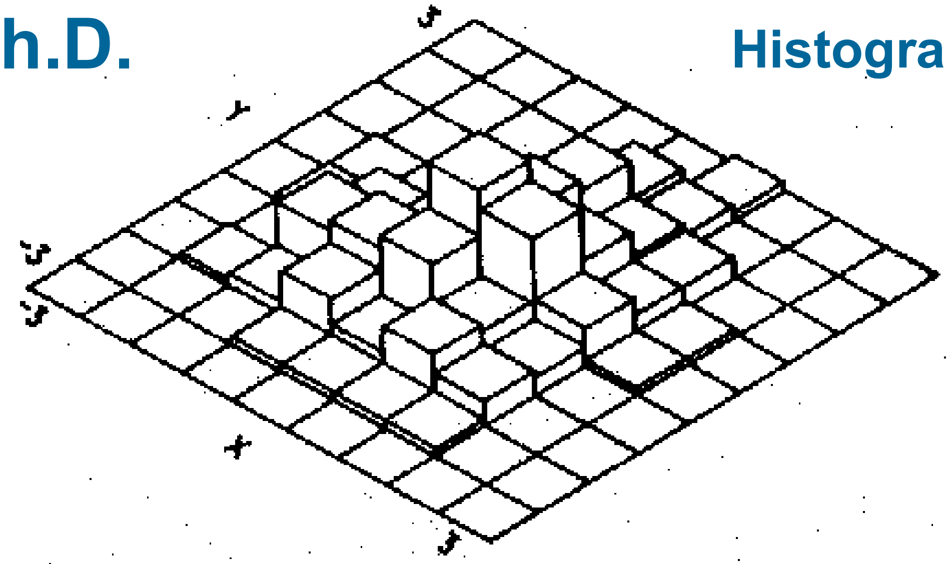
$$\int_{\Re^n} \kappa(\vec{x})\, d\vec{x} = 1$$

$$\hat{f}(\vec{x}) = \frac{1}{n\,h^d} \sum_{i=1}^{n} \kappa\left(\frac{\vec{x} - \vec{x}_i}{h} \right)$$

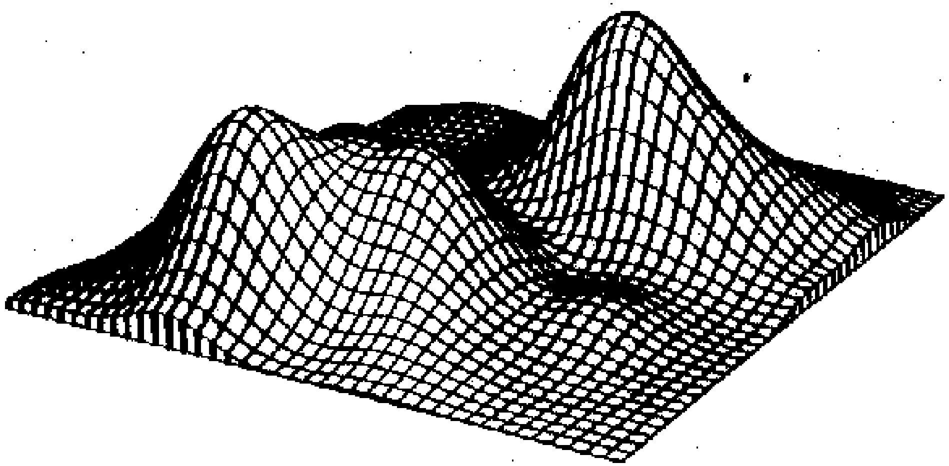
The use of a single smoothing coefficient is usually enough if the data have been standardized

Example from R. Asiain Ph.D.

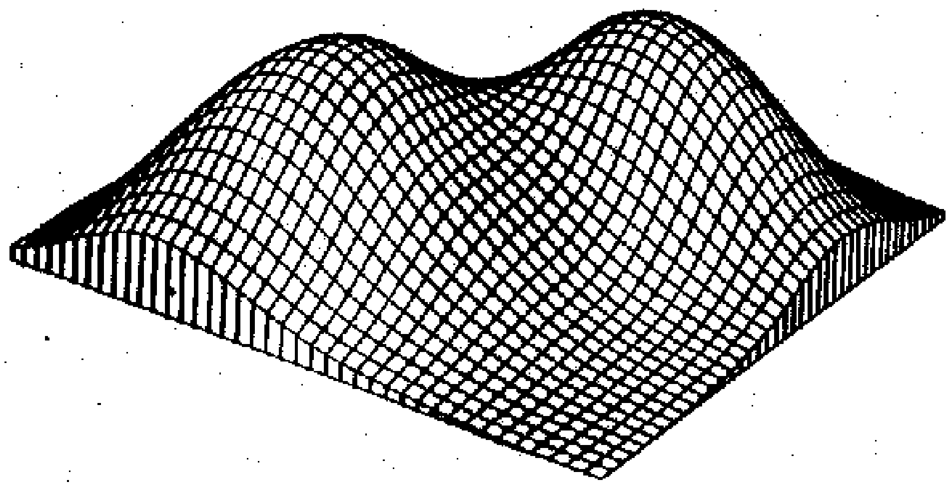
Bivariate normal distributions



Histogram



Kernel estimation h=1.2



Kernel estimation h=2.2