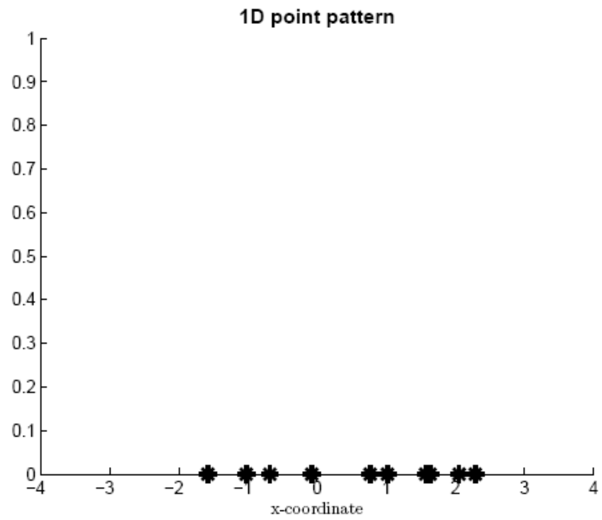
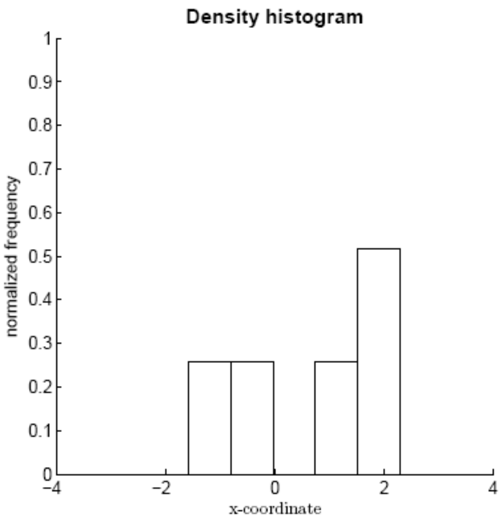
Intensity or Density Estimation in 1D

- Consider a hypothetical 1D point pattern comprised of N = 10 events (left) and estimate their local intensity, i.e., a 1D profile of average # of events per unit area:

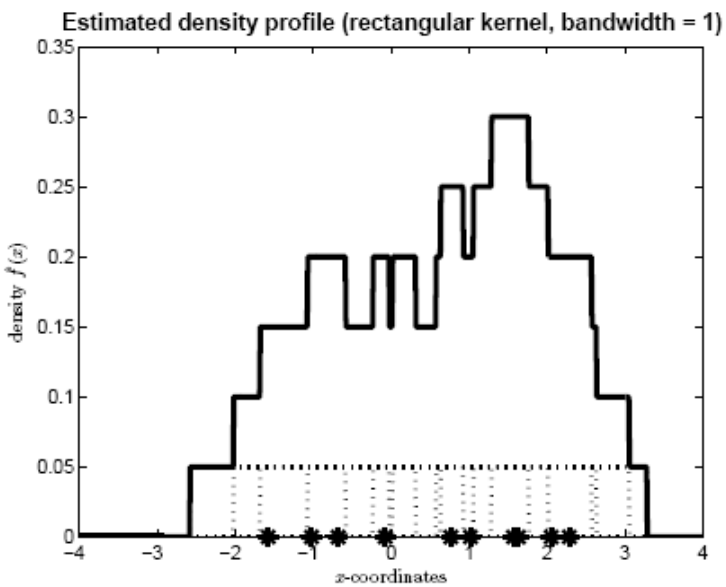


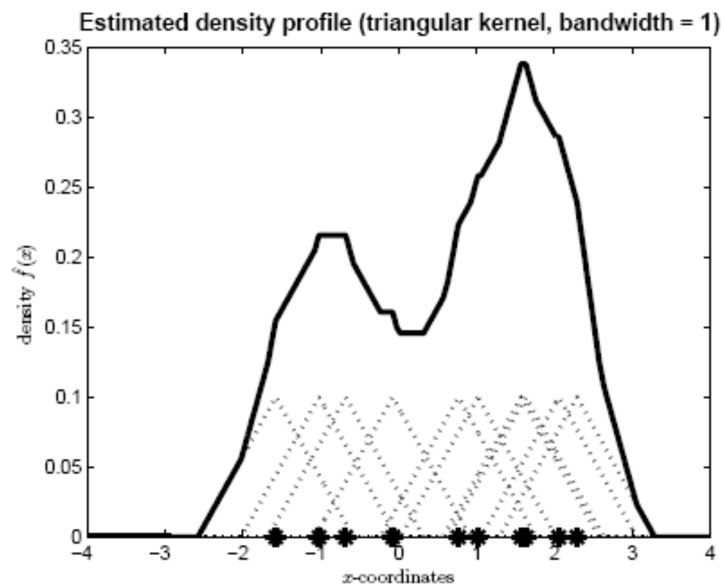


- Statistical analogy
  - The objective is to describe the density of x-coordinates, and this problem has been treated extensively in the non-parametric density estimation literature; a first-cut at such a density profile is provided by the density histogram plot (right).
  - **In other words**: the set of N x-coordinates of events in a 1D point pattern can be viewed as N values of an attribute, here the x-coordinate . . .

Density Estimation Preview

- Key concepts
  - density estimation via a histogram calls for deciding on: (i) the number of attribute classes (bins), and (ii) their centers in the abscissa
  - instead of choosing a limited # of bins, choose as many bins as the # of events in the data set
  - bars in a histogram are rectangular, but nothing prevents us from using other shapes to build a density profile

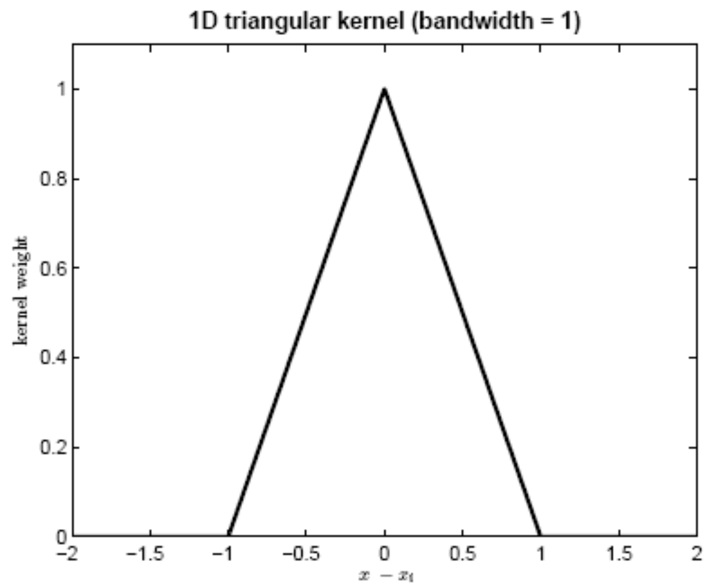




- Another key concept: Each bar (left) or triangle (right) can be regarded as the influence of an observed event to the likelihood of seeing other events around that observed one

1D or Univariate Kernel I

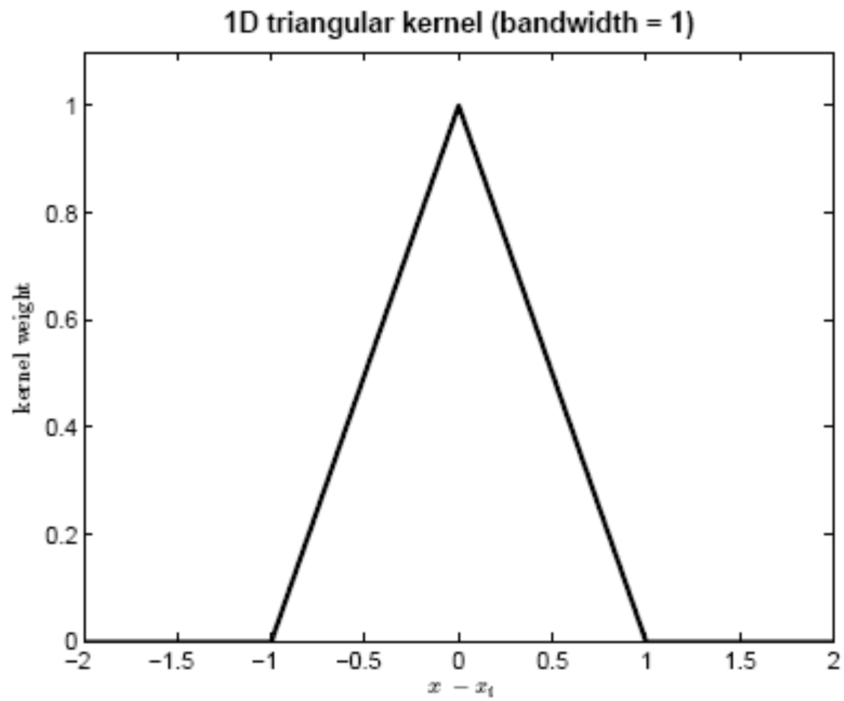
- Kernel function
  - Analytical expression for likelihood of a particular x-coordinate, or in other words for probability of observing an event at the particular x-coordinate, given presence of an event at coordinate x_i



- Kernel characteristics I
  - function of distance $h_i = |x-x_i|$ between arbitrary point at location x and event at location x_i : $k(x,x_i) = k(|x-x_i|) = k_i(h)$, where h is the distance between an arbitrary location x and the kernel center, here the event location x_i (assumed to be at x = 0 on the graph)
  - typically all ***N*** kernels are assumed the same, i.e., $k_i(h) = k(h)$, for all i

1D or Univariate Kernel II

- Kernel characteristics II



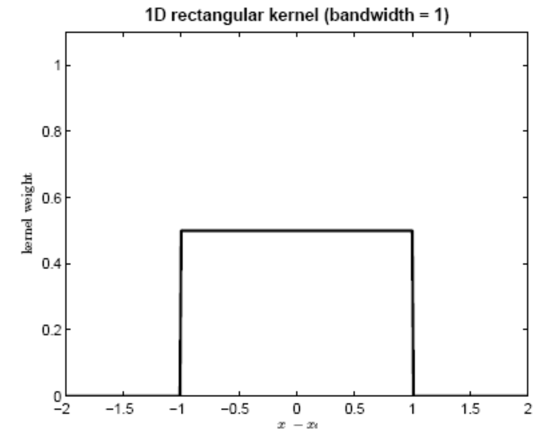
  - kernels are (typically symmetric) probability density functions (PDFs), hence non-negative and integrating to 1: $k(h) \geq 0$, and $\int k(h)dh = 1$
  - as PDFs, kernels have a mean (0, since the abscissa quanties distance from an event) and positive finite variance: $\int hk(h)dh = 0$ and $0 < \int h^2 k(h)dh < \infty$

- Relation to density estimation
  - Instead of fixing the # of bins and their origin (as done with histograms) we can
  - estimate the local density f(x) at an arbitrary x-value as a weighted **sum** of N values $k(x-x_i)$; each such value belongs to a different kernel $k_i(h)$ centered at a x_i location/coordinate

Some 1D Kernel Functions I

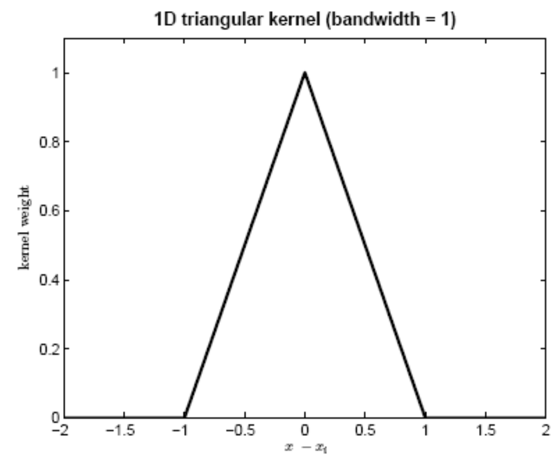
- rectangular or uniform or Parzen:

$$k(h) = \begin{cases} 1/2 & \text{if } h \in [-1\ 1] \\ 0 & \text{if not} \end{cases}$$



- triangular:

$$k(h) = \begin{cases} 1 - |h| & \text{if } h \in [-1\ 1] \\ 0 & \text{if not} \end{cases}$$

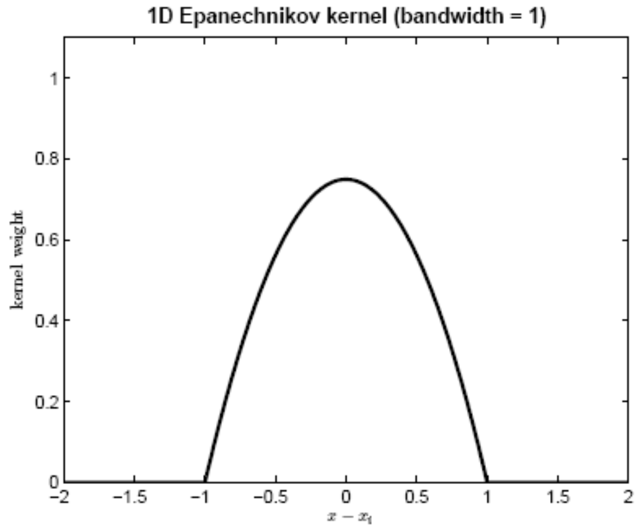


- For a set of P values $\{x_p;\ p = 1 \quad P\}$ discretizing a 1D segment, and for a particular datum coordinate x_i, the function $k(x_p - x_i)$ can be evaluated P times, and the resulting kernel "profile" can be stored in a (Px1) array $k_i = [k(x_p - x_i), p=1 \quad P]^T$

Some 1D Kernel Functions II

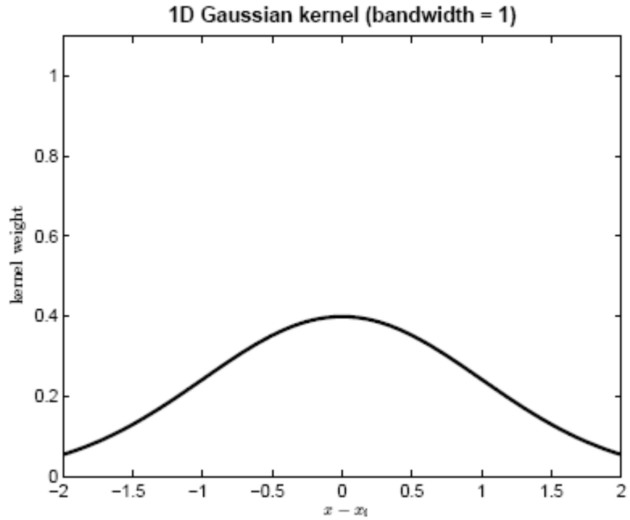
- Quadratic or Epanechnikov:

$$k(h) = \begin{cases} \frac{3}{4}(1-h^2) & \text{if } h \in [-1\ 1] \\ 0 & \text{if not} \end{cases}$$



- Gaussian:

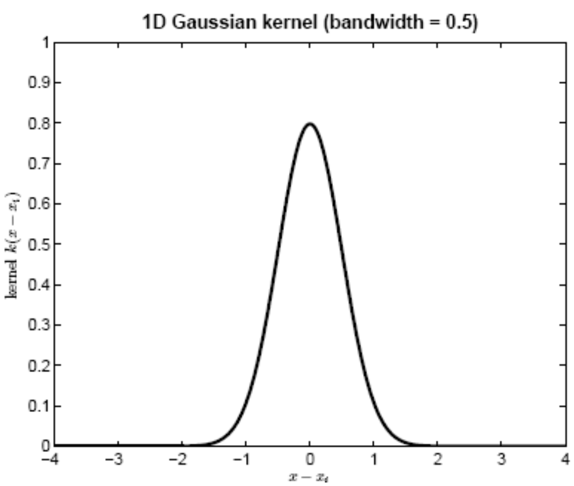
$$k(h) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}h^2)$$

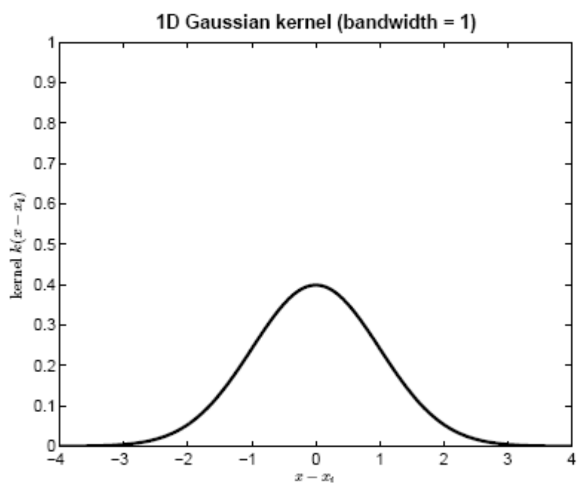


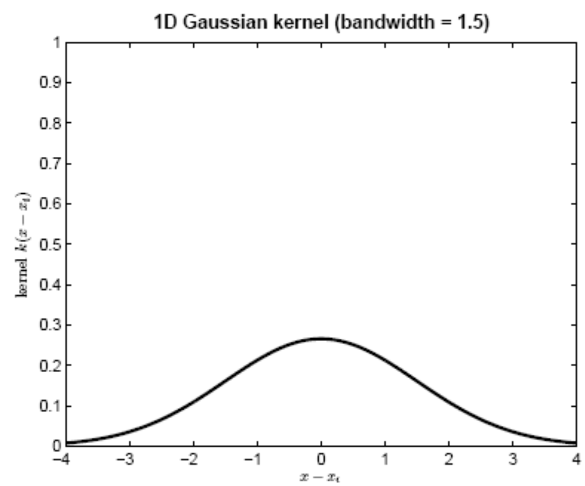
Kernels that reach 0 asymptotically, e.g., Gaussian, are called non-compact kernels

Scaled 1D Kernels I

- Alternative view of a kernel
  - A kernel function $k(x-x_i)$ quantifies the “influence" of a particular event at coordinate x_i to its surroundings, i.e., to all other x-locations
- Scaling the kernel
  - The influence of an event at x_i to all x-coordinates can be altered by scaling the associated kernel function $k(x-x_i)$;
    i.e., by dividing the function argument $x-x_i$ by a constant **b** (called the kernel **bandwidth**); in order to ensure that the new kernel is a PDF, i.e., integrates to 1, divide the output of this new function by b







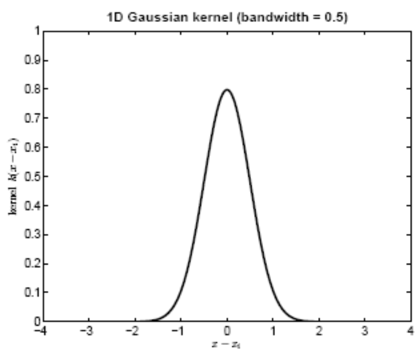
E.g.: scaled or non-standard Gaussian PDF: $k(h; b) = \frac{1}{b\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{h}{b})^2]$, where b is the standard deviation

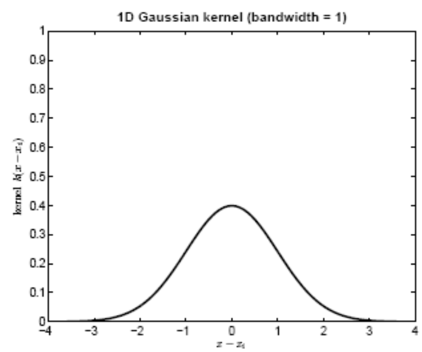
Scaled 1D Kernels II

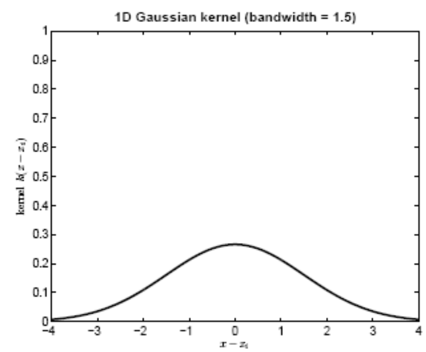
- Scaled kernel function
  - Divide the argument (distance from event) of the kernel function by a scalar b:

$$k(x - x_i; b) = \frac{1}{b} k\left(\frac{x - x_i}{b}\right)$$

- Transformation of PDFs
  - Let X be a RV with PDF $f_X(x)$ and Y be another RV defined as Y = (1/b)X, i.e., y = x/b.
  - The PDF $f_Y(y)$ of RV Y can be computed as:
    - $f_Y(y) = (1/b)f_X(x/b)$; if the original PDF $f_X(x)$ has std deviation 1, the new PDF $f_Y(y)$ has std deviation b







  - For a set of P values $\{x_p\ ; p = 1 \quad P\}$ discretizing a 1D segment, and for a particular datum coordinate x_i the scaled function $k(x_p-x_i; b)$ can be evaluated P times, and the resulting discrete kernel stored in a (Px1) array $k_i(b) = [k(x_p-x_i;b);\ p = 1 \quad P]^T$

Flowchart

1D Kernel Density Estimation Flowchart

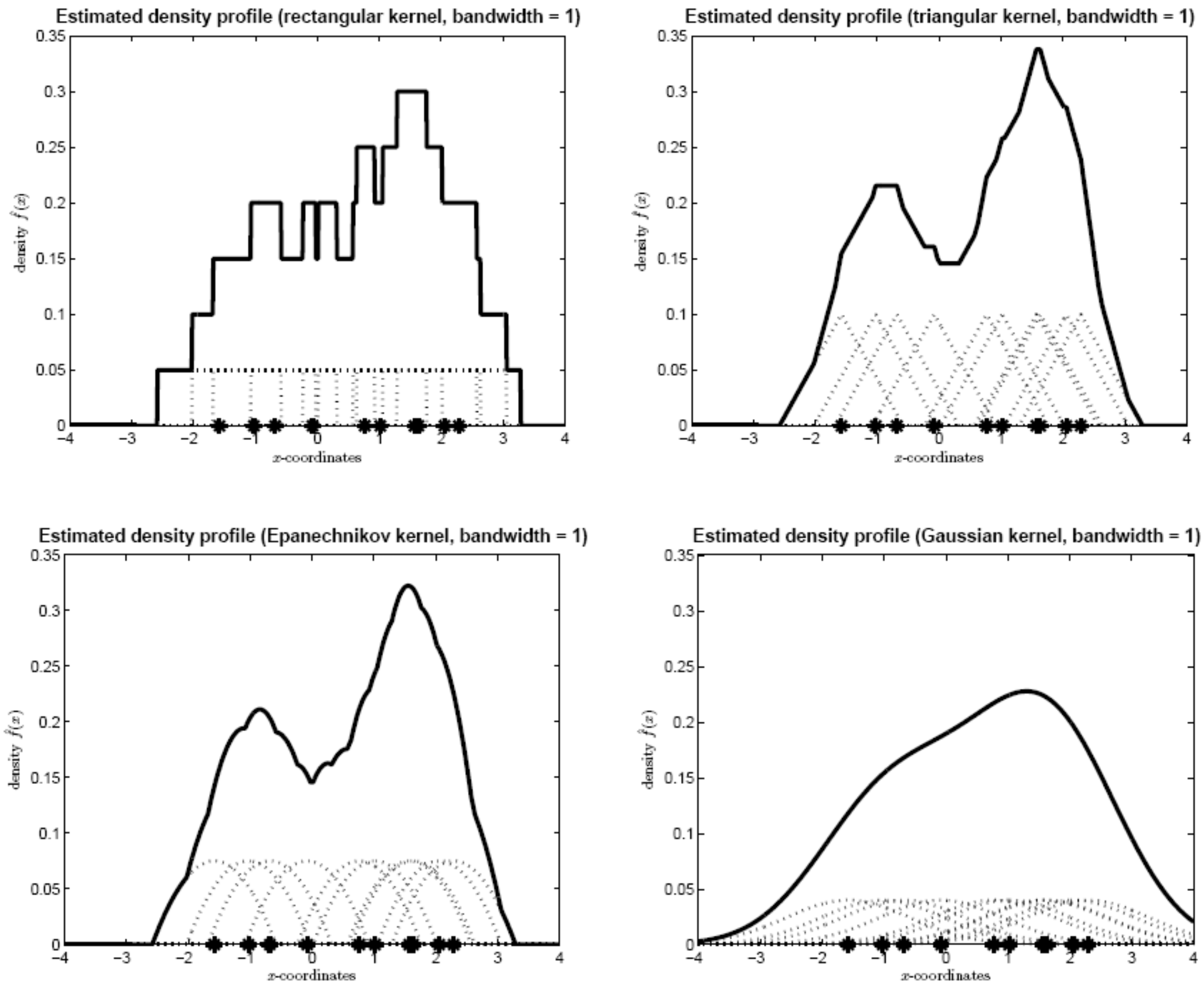
- 1. choose a kernel function $k(x-x_i)$, i.e., a PDF, and a bandwidth parameter b controlling kernel extent and consequently the “smoothness" of the final estimated density profile f(x); this amounts to choosing a scaled kernel function $k(x-x_i;b)$
- 2. discretize 1D segment, i.e., choose a set of P x-coordinates $\{x_p;\ p = 1 \quad P\}$ at which the density function f(x) will be estimated
- 3. for each datum coordinate xi , evaluate the scaled kernel function $k(x_p-x_i;b)$ for all P x-values; this yields N scaled kernel profiles $\{k_i(b);\ i = 1 \quad N\}$ each one stemming from a particular event coordinate x_i
- 4. for each discretization coordinate x_p, compute estimated density $f(x_p)$ as the sum of the N scaled kernel values $k(x_p-x_i;\ b)$, after weighting each such value by 1/N:

$$\hat{f}(x_p) = \sum_{i=1}^{N} \frac{1}{N}\frac{1}{b} k\left(\frac{x_p - x_i}{b}\right) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{b} k\left(\frac{x_p - x_i}{b}\right)$$

Output

- A (Px1) vector k(b) with estimated density values f(x) at the specified x-coordinates; the N scaled & weighted kernels $\{(1/N)k_i(b);\ i = 1 \quad N\}$ can be regarded as N elementary profiles whose super-position builds up the final estimated density profile

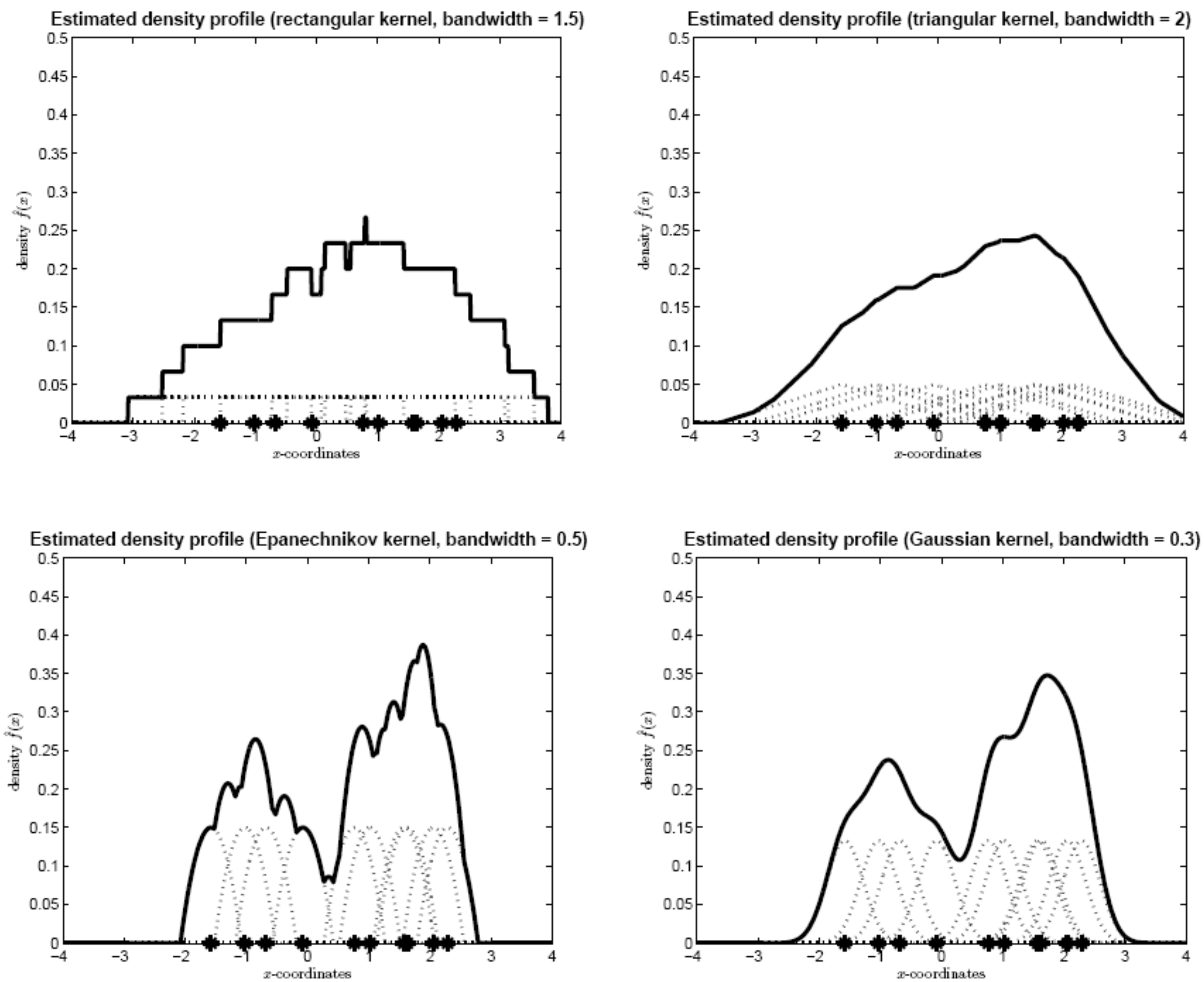
1D Kernel Density Estimation Examples I



Dotted lines depict scaled and weighted (by $1/n$) kernel profiles

- Rules exist for choosing an “optimal" bandwidth parameter, typically based on a pre-supposed distribution type, e.g., Gaussian, for the N data Estimated density profiles are more sensitive to choice of bandwidth parameter b than to choice of kernel type

1D Kernel Density Estimation Examples II



Dotted lines depict scaled and weighted (by $1/N$) kernel profiles

- The smaller the bandwidth, the spikier (noisier) the resulting estimated density profile; too large a bandwidth leads to over-smoothed (with no interesting details) density profiles

Separable 2D Kernels

- Two 1D Gaussian kernels

$$k_x(x - x_i; b_x) = \frac{1}{b_x\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - x_i}{b_x}\right)^2\right]$$

$$k_y(y - y_i; b_y) = \frac{1}{b_y\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y - y_i}{b_y}\right)^2\right]$$

event location $u_i = (x_i;y_i)$, arbitrary location u = (x; y), kernel bandwidths b_x and b_y

- 2D composite kernel

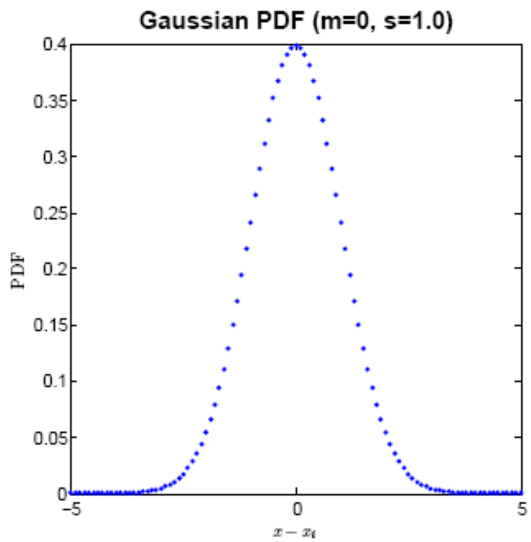
$$k(x - x_i, y - y_i; b_x, b_y) = \frac{1}{2\pi b_x b_y} \exp\left[-\frac{1}{2}\left(\frac{x - x_i}{b_x}\right)^2 - \frac{1}{2}\left(\frac{y - y_i}{b_y}\right)^2\right]$$

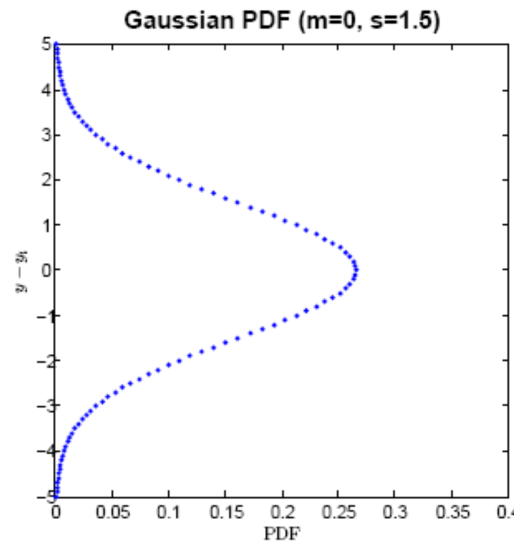
bivariate Gaussian PDF for 2 independent RVs, a product of 2 univariate Gaussian PDFs

- Separability
  - Any (scaled or not) 2D kernel that can be derived as a product of 2 elementary1D kernels is called **separable**

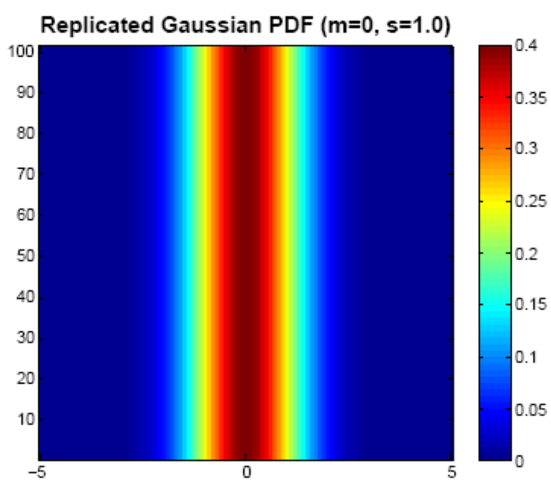
Constructing A Separable 2D Kernel

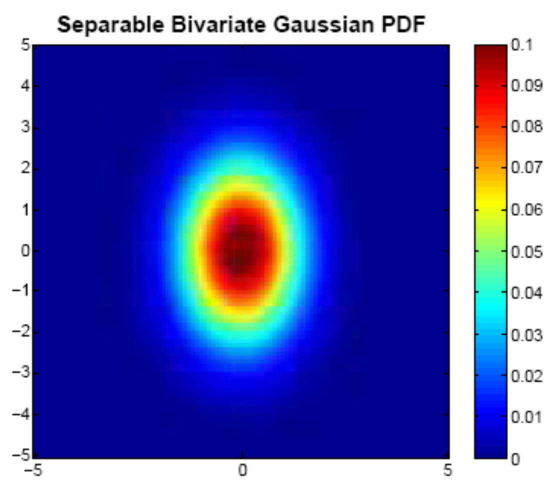
- Two 1D Gaussian kernels for the x- and y-dimensions

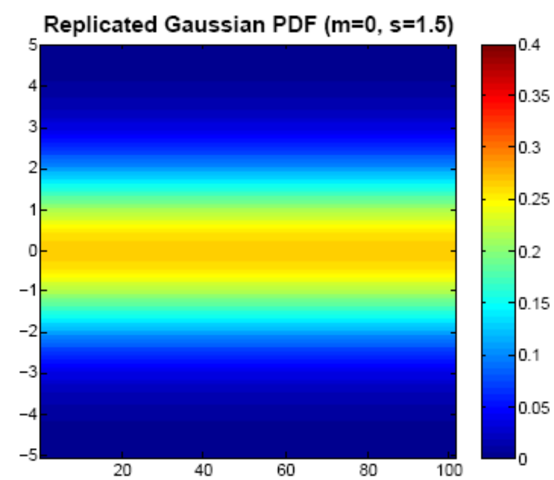




- Replicated 1D Gaussian kernels and 2D separable composite

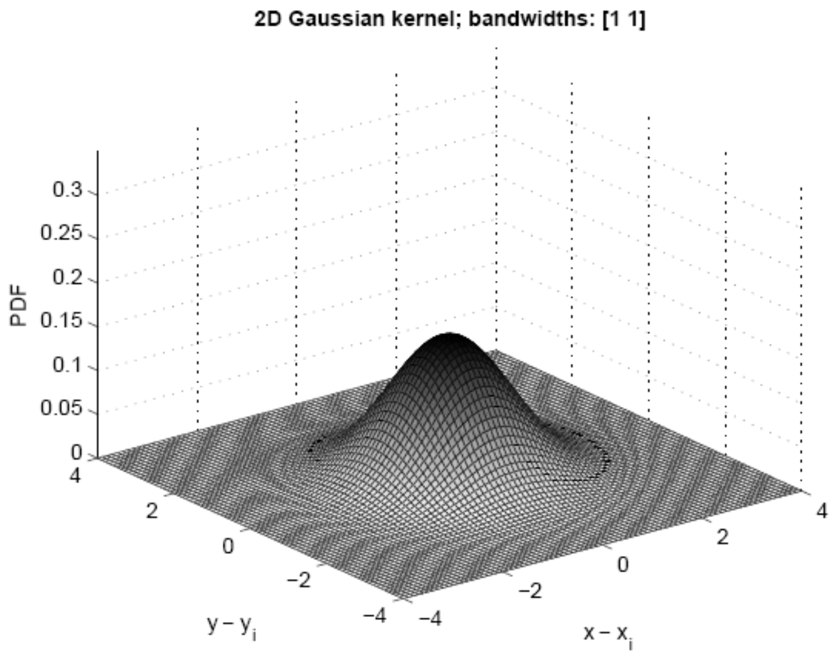


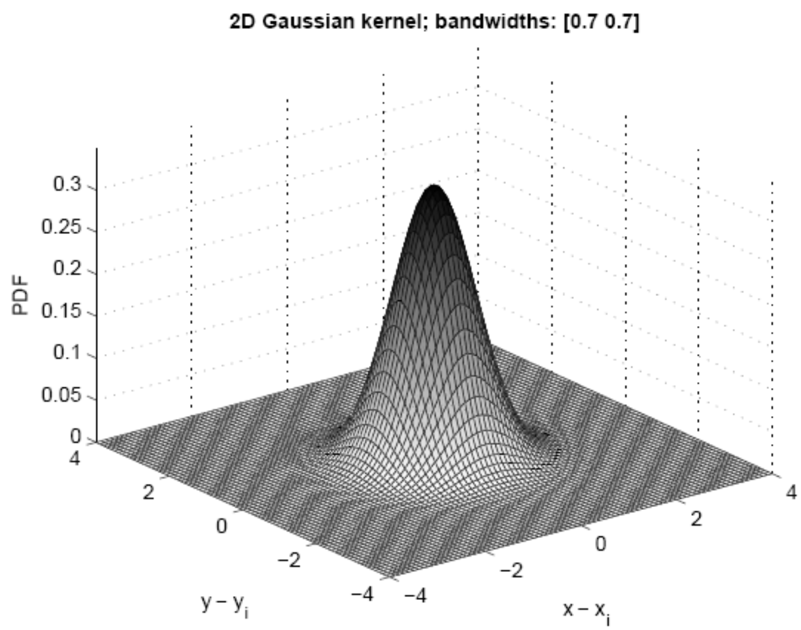


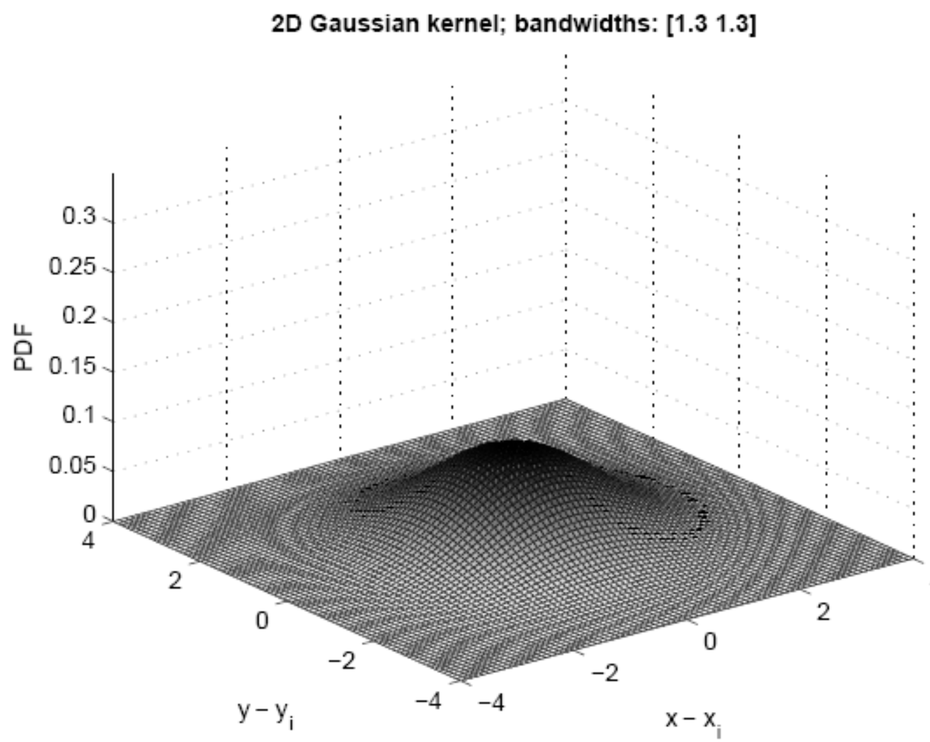


Anisotropic kernel = multidimensional kernel with different bandwidths along different directions

2D Gaussian Kernel Examples







Isotropic kernel = multidimensional kernel with same bandwidth along different directions