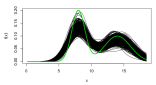
Non parametric statistical estimation

Marcello Chiodi

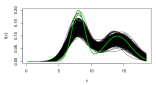
marcello.chiodi@unipa.it http://dssm.unipa.it/chiodi

Dipartimento di Scienze Economiche Aziendali e Statistiche (SEAS)
Università di Palermo

Stuttgard, February 2019

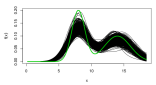


Outline I



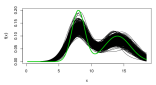
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Outline II



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Outline III



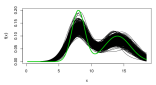
dSEAS dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Outline IV

- Example
- Kernel multivariate variables bands (hints)

13 The problem of dimensionality

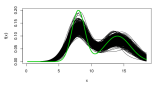
- Dimensionality
- Examples
- Asymptotic results for $d-$variates kernels



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Objectives

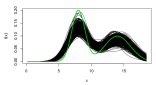
- Main objective of the course is to provide the fundamental tools of **non-parametric statistics** .
- Estimate univariate and multivariate density functions
- Study of dependence relationships among variables.
- At the end of the course the student should be able to describe real data sets using the techniques learned.



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Preliminary Knowledge

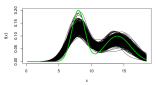
- mathematical analysis and probability Calculus
- Matrix algebra
- R statistical programming environment
- Numerical optimization techniques
- Simulation techniques
- Theory of inference (in particular, properties of estimators and linear regression)



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Teaching Material

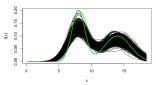
- Slides (where)
  pause or given in the classroom
- An integral part of the course materials will be all the exercises performed, the datasets, the R code and R packages used in the course
- Some specific references will be provided during the course
- Thanks to Dr. Antonio Abbruzzo, Prof. Giada Adelfio and Dr. Mariangela Sciandra for exercises.
-



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Software used in lectures and tutorials

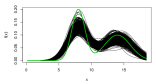
- We will use the open source software R.
- using the R software, using different packages dedicated;
- R routines written for my courses;
- Some theoretical topics of the lectures will be addressed through PC (in particular through simulations) . It will be useful in the classroom that students bring their laptops in the classroom, even for theoretical lessons .
- Exercises on the topics of the course: analysis of real cases .



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Topics

- Introductory real problems .
- Parametric and non-parametric statistics.
- Non parametric estimation of univariate density functions. Kernel estimators.
- Non parametric estimation of multivariate density functions.
- Non-parametric regression : kernel estimators, splines and local polynomial regression.
- A little introduction on GAM (generalized additive models)



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Parametric and non-parametric hypothesis

- The term **nonparametric** was born in the context of hypothesis testing , to indicate inference problems that concern *distributions (or functions)* and not *parameters* .
- (**parametric** Example) :

$$\Omega : \; X \sim \mathcal{N}(\mu, \sigma^2) \tag{1}$$

$$H_0 : \; \mu = \mu_0 \qquad H_1 : \; \mu \neq \mu_0$$

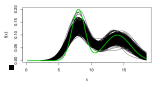
The hypothesis concerns the parameter of a population.

- Another **parametric** example:

$$\Omega : Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$$

$$H_0 : \; \beta = \beta_0 \qquad H_1 : \; \beta \neq \beta_0$$

The hypothesis is still about the parameter of a population.



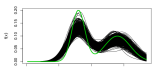
dSEAS dipartimento di scienze economiche aziendali e statistiche department of economics business and statistics

Nonparametric Hypothesis

Instead we see two examples of **nonparametric hypothesis:**

1. $H_0 : X \sim \mathcal{N}(\mu, \sigma^2)$ $H_1 : X \sim \mathcal{D}(\cdot)$, $\mathcal{D} \neq \mathcal{N}$
   The latter hypothesis does not apply to a parameter, but to a distribution:
   Is X normally distributed ?
2. $H_0 : F(X) = F(Y)$ $H_1 : F(X) \neq F(Y)$
   Also this hypothesis does not concern parameters, but two distributions :
   Do X and Y have the same distribution ?

Note that in the last example is not specified what is the joint distribution for X and Y.



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Parametric and non-parametric estimation

- **parametric** Example :

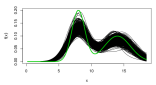
$$\Omega : \; X \sim \mathcal{N}(\mu, \sigma^2) \tag{2}$$

We want to estimate the parameters μ, σ^2 of a population, given a sample of size n $\{x_1, x_2, \ldots, x_i, \ldots, x_n\}$.
(using *Maximum Likelihood* generally is the optimal strategy)

- Another **parametric** example:

$$\Omega : Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$$

We want again to estimates the parameters α, β given a sample y_i, x_i $i = 1, 2, \ldots, n$.



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

non-parametric estimation

- **non-parametric** Example :

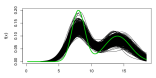
$$\Omega : \; X \sim \mathcal{D}(x) \tag{3}$$

We want to estimate the distribution $\mathcal{D}(x)$ (or better its density) of a population, given a sample of size n.

- Another **non-parametric** example (non-parametric regression):

$$\Omega : \mathrm{E}[Y_i] = g(x_i) \qquad (\mathrm{E}[Y|X = x_i])$$

We want to estimates the regression function $g(\cdot)$ without specifying the kind of function (linear, quadratic, etc.). given a sample of size n of values $y_i, x_i \;\; i = 1, 2, \ldots, n$



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Real data

The dots are the residences of dead people in a period of about 20 years in a town, for a specific cause *(red: cases)* and other causes *(blue: controls)*. Are the spatial distributions of the points similar (and thus the residence does not affect the probability of a particular cause of death)?

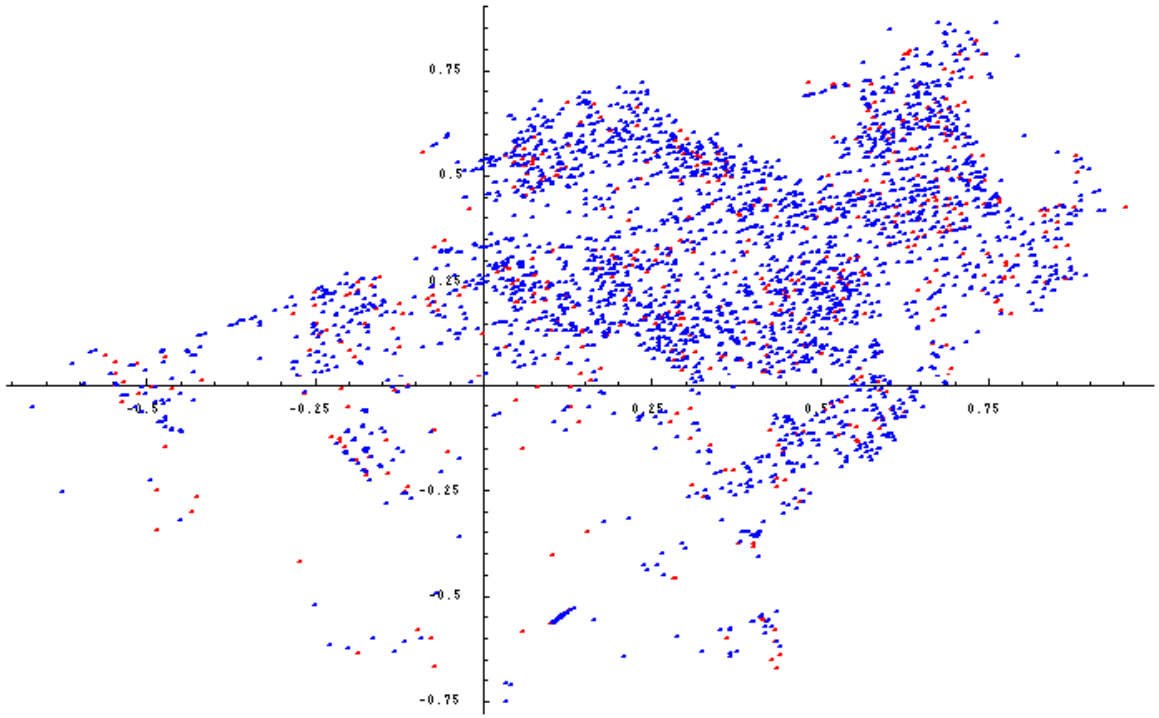


Figure: Spatial distribution of the residences of cases (red) and controls (blue)

Comparison of two spatial distributions

Density estimation for controls

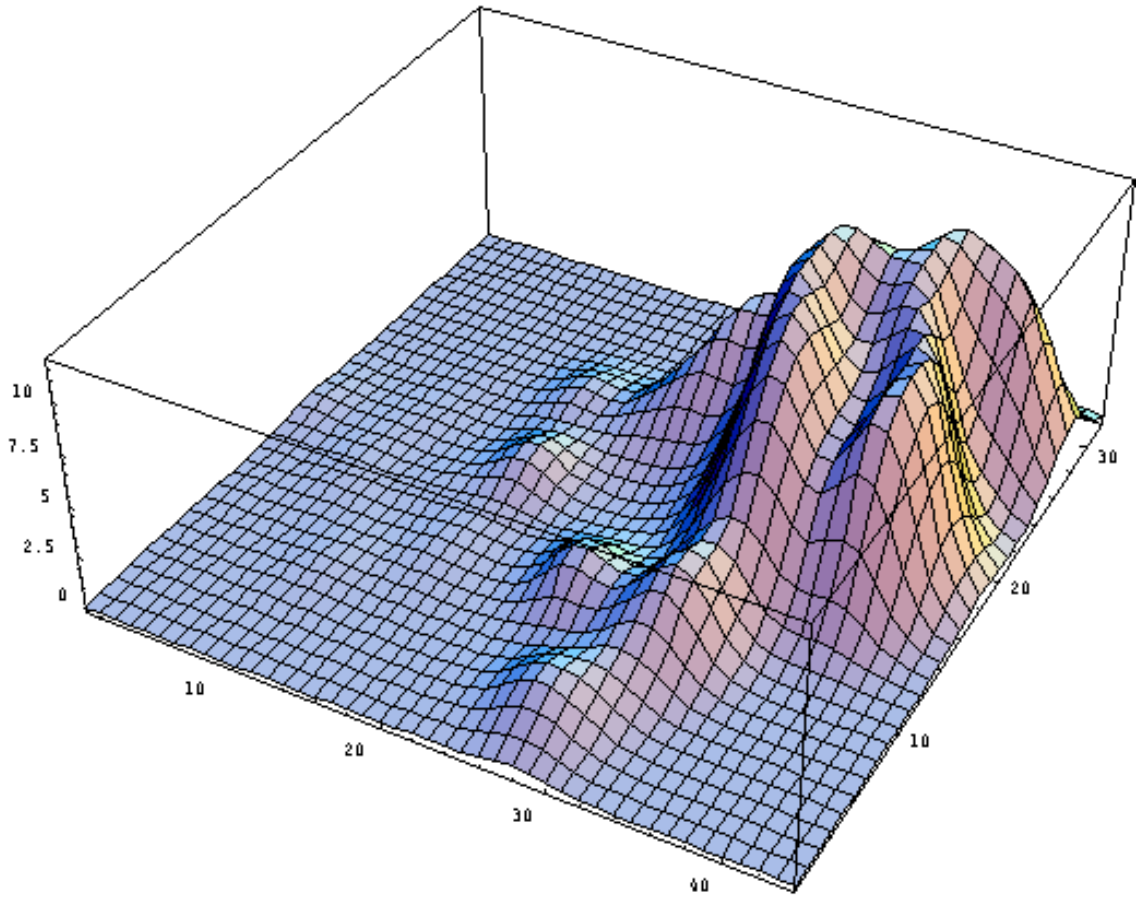
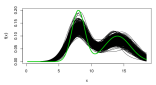


Figure: Density estimation for controls



Preliminary real problems

Density estimation for *cases*: Is this density similar to the previous one (*controls*)?

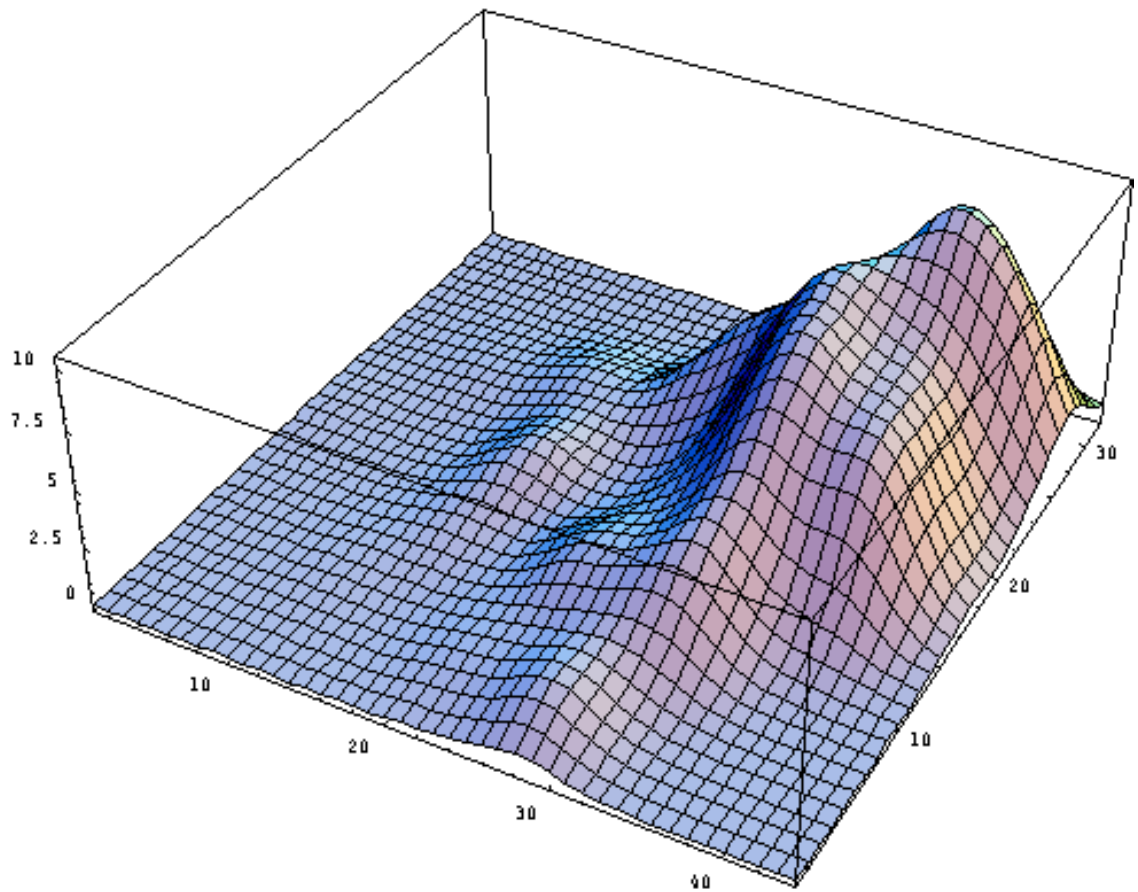


Figure: Density estimation for cases

Screening among firms

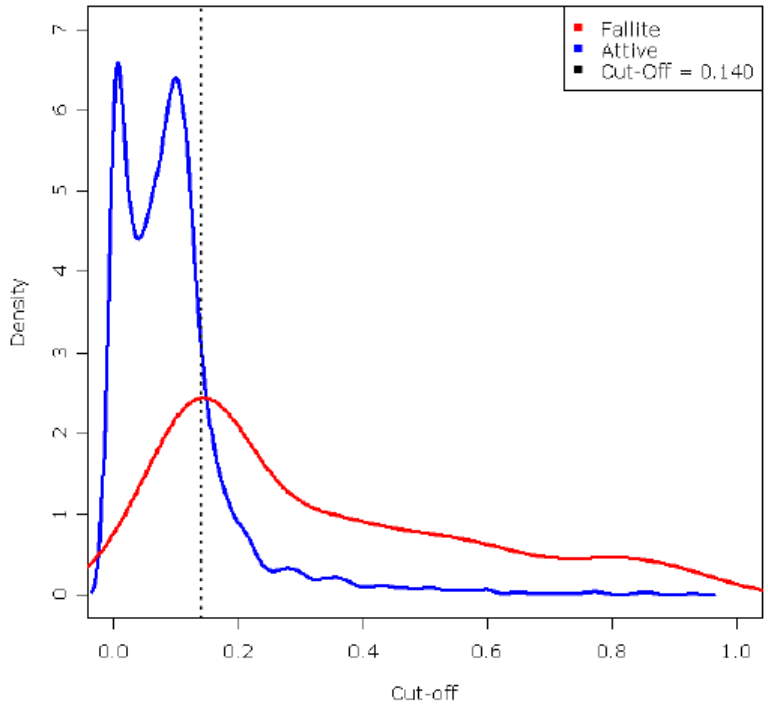


Figure: Screening among firms (good versus default)

Which financial indicator should be preferred?

A seismic sequence

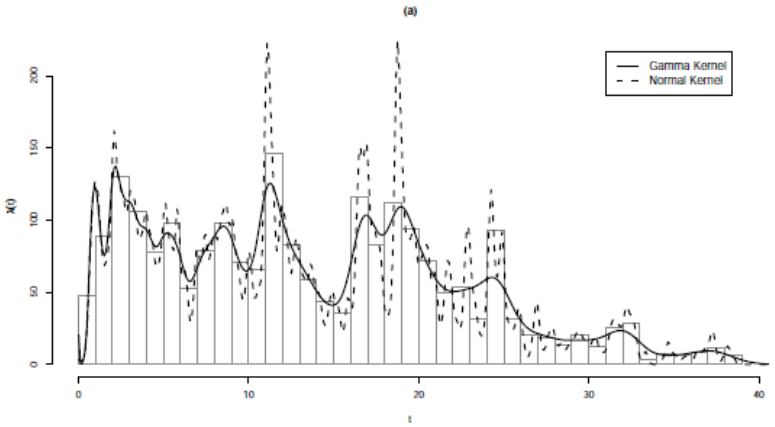
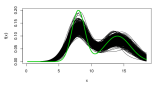


Figure: Time intensity of a seismic sequence (Umbria 1997)



A seismic sequence

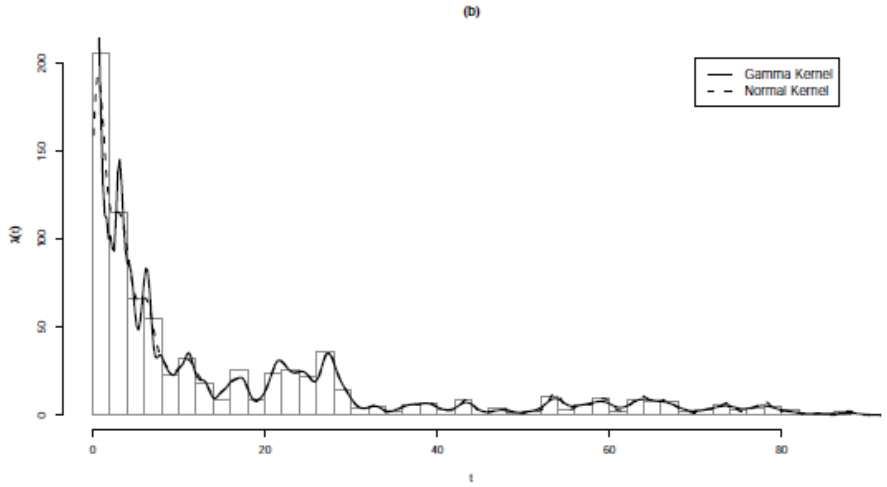
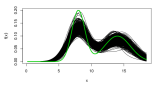


Figure: Time intensity of a seismic sequence (Palermo 2002)



Spatial intensity

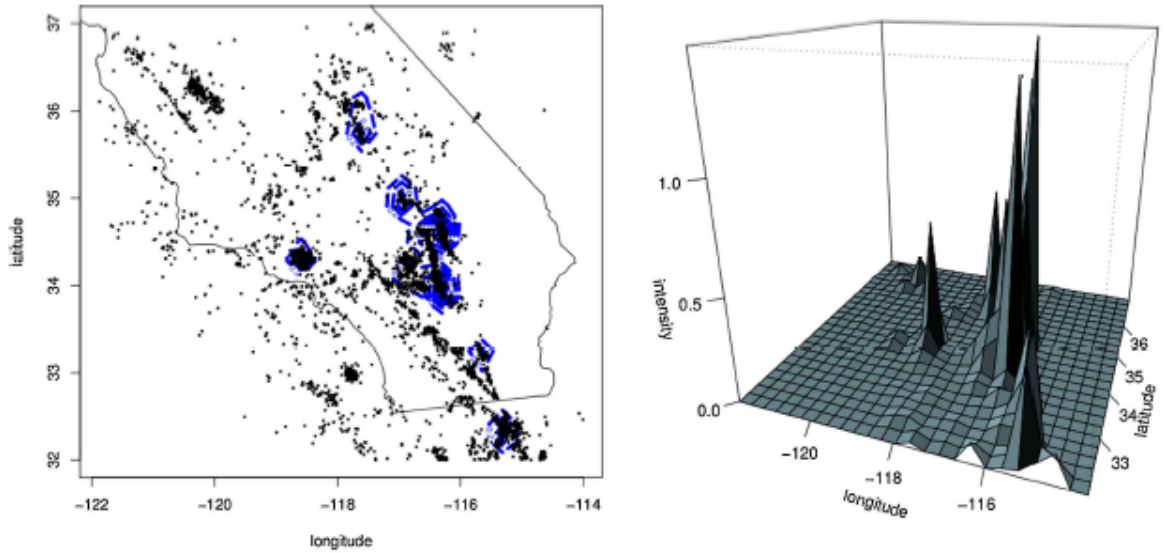
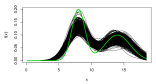


Figure: Spatial seismic intensity function (California seismic catalog)



Space time intensity

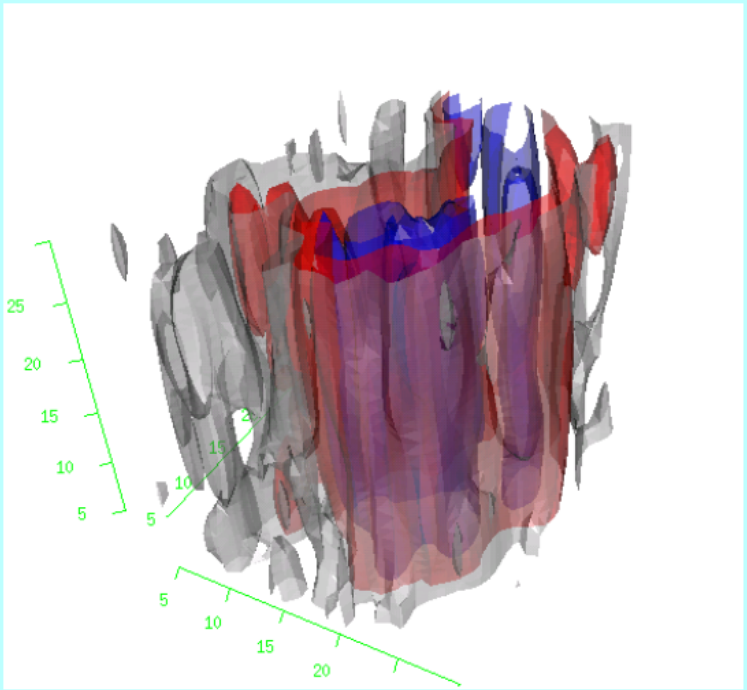
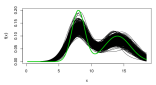


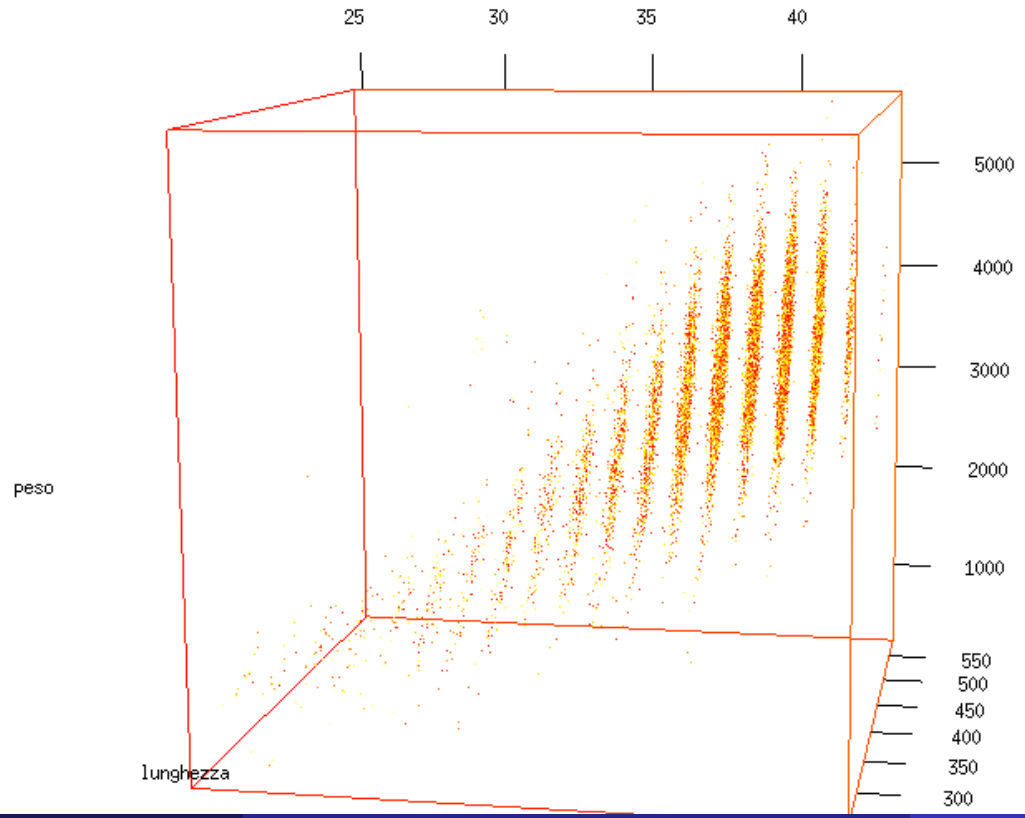
Figure: Space-time seismic intensity function (3-d contour surfaces) (Sicilian seismic catalog)



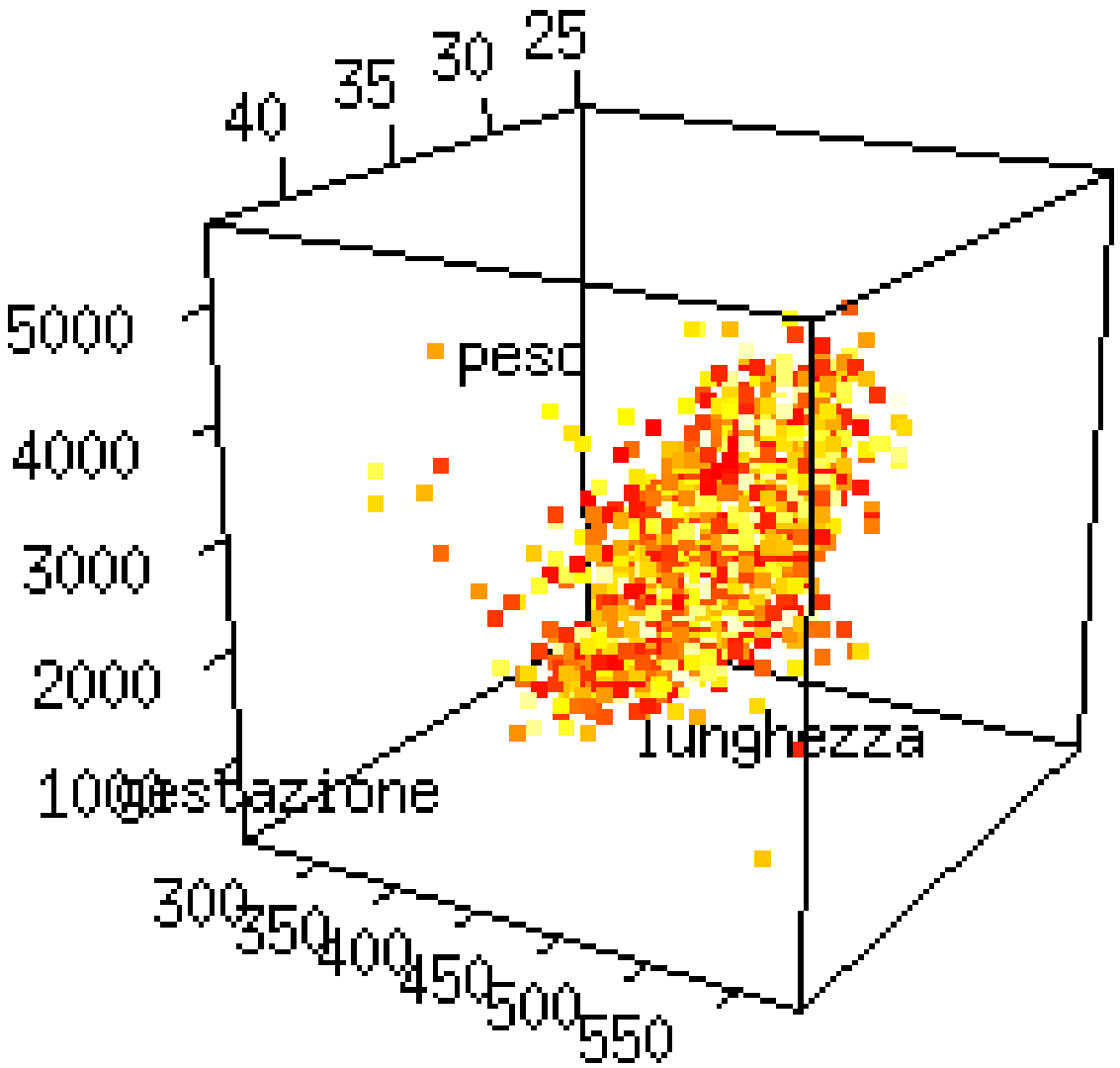
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Non parametric regression

(description of data: Anthropometric measurements of a sample of newborns; weight vs. weeks of gestation and height)
Can we fit a non linear surface (quadratic, exponential, something else)?
Or even a surface with undefined parametric form (and possibly different for different groups)



weight vs. weeks of gestation and height



Residual plot from a linear regression fit: Estimation of of deviations from linearity

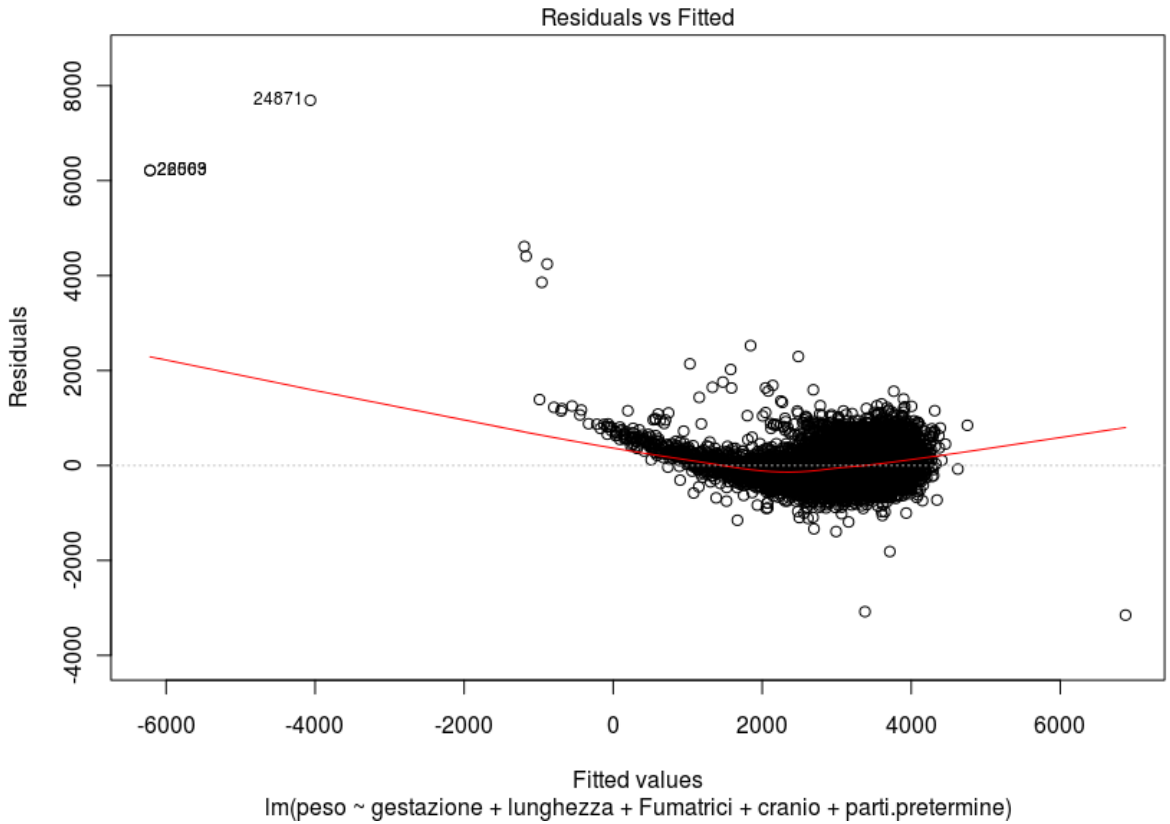
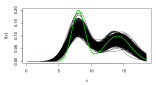


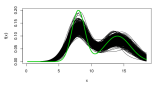
Figure:



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Essential references

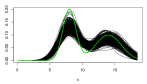
- Bowman, A.; Azzalini, A. (1997) *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations* Oxford Statistical Science Series
- Ruppert, D., Wand, M.P., Carrol, R. J (2003). *Semiparametric Regression*. Cambridge University
- Silverman, B.W. (1998). *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall/CRC.
- Takezawa K. (2005) *Introduction to Nonparametric Regression*. John Wiley & Sons.
- Wand, M.P; Jones, M.C. (1995) *Kernel Smoothing*. London: Chapman & Hall/CRC.



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Density function Estimators: Motivations

- To get ideas on parametric model
- To compare groups
- To have information about asymmetry, bimodality, etc..
- To estimate density (or intensity) for inhomogeneous (spatially or temporally) phenomena
- To integrate classical parametric methods: e.g. to get ideas on the distribution of nadom errors by examining the empirical distribution of residuals.
  Review of the examples of introductory lesson.



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Density function estimation of univariate random variables

distribution function

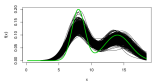
The definition of a density function can be derived from the distribution function:

$$F_X(x) = \text{Prob}\{X \leq x\}$$

density function

if $F(x)$ is differentiable we can write:

$$f_X(x) = \frac{d\left(F_X(x)\right)}{d\,x}$$



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

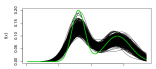
Estimation of the distribution function

The distribution function can be estimated from a sample of n observations $x_i (i = 1, 2, \ldots, n)$ from *(iid)* random variables [1] through the empirical distribution function

empirical distribution function

$$\hat{F}_X(x) = \frac{\#(x_i \leq x)}{n}$$

(it can be shown that this estimator can be improved, for example with $\hat{F}_X(x) = \frac{\#(x_i \leq x)+0.5}{n+1}$, for now this approximation is sufficient)
This estimate can be always calculated, and it leads to a step function.



[1] iid: independent and identically distributed (random variables)

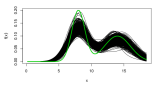
Estimation of the distribution function. Examples with R

empirical distribution function

$$\hat{F}_X(x) = \frac{\#(x_i \leq x)}{n}$$

example code: ese1_NP2013.R

./ese1_NP2013.R



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Distribution function estimation: examples

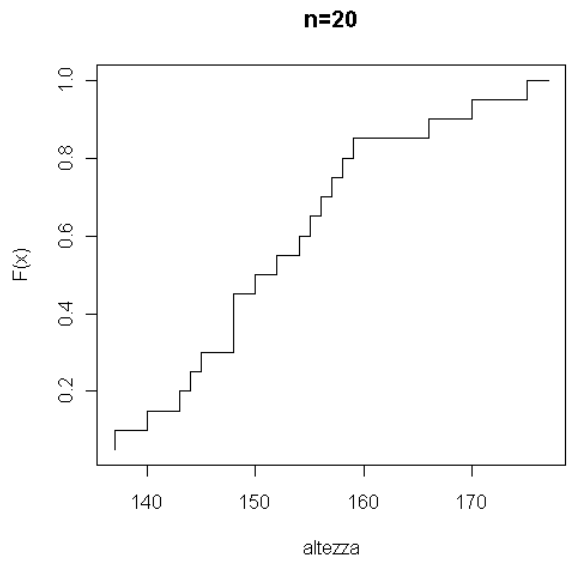


Figure: distribution function $n = 20$

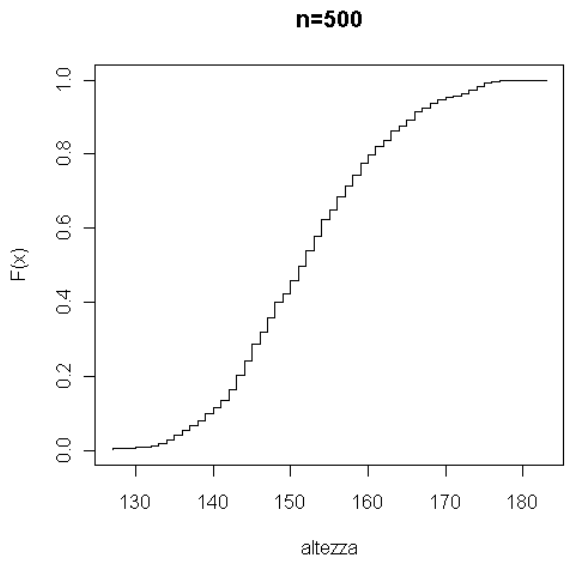
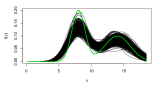


Figure: distribution function $n = 500$

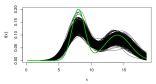


dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Density function estimation

(example) Obviously this method can not be used for density functions.

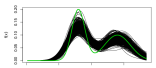
- We can not define a similar estimator for $f(x)$ starting directly from sample data,
  (unless we use an estimator that distributes a mass of $\frac{1}{n}$ to each observation, which in general is not really useful)
- (while *($\hat{F}_X(x)$ is a distribution function! it is the distribution function of the sample)*
- We can not even get $\hat{f}_X(x)$ deriving $\hat{F}_X(x)$, due to discontinuity.



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

(A not so useful) density function estimation

For previous samples density estimates based on histograms of individual observations are shown.
(observations are rounded up to the nearest centimeter, and therefore there are observations with a frequency greater than one)
[1] 143 144 142 137 144 148 145 135 144 143 136 146 144 134 155 150 137 138
[19] 143 145 143 148 138 148 143 140 145 142 140 137 142 133 142 141 154 138
[37] 150 144 145 139 142 147 142 140 133 145 139 151 151 140 143 133 138 144
[55] 130 147 141 151 141 134 136 144 143 143 136 141 145 145 145 144 127 138
...



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

(A not so useful) density function estimation

Two samples of 20 and 500 observations have been drawn from a bigger data set

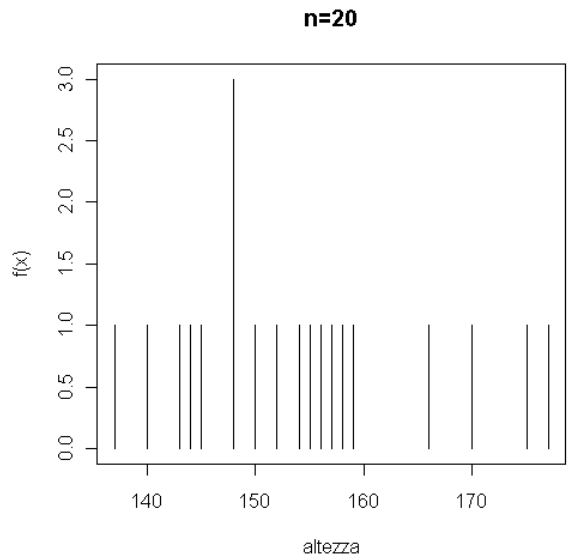


Figure: density function $n = 20$

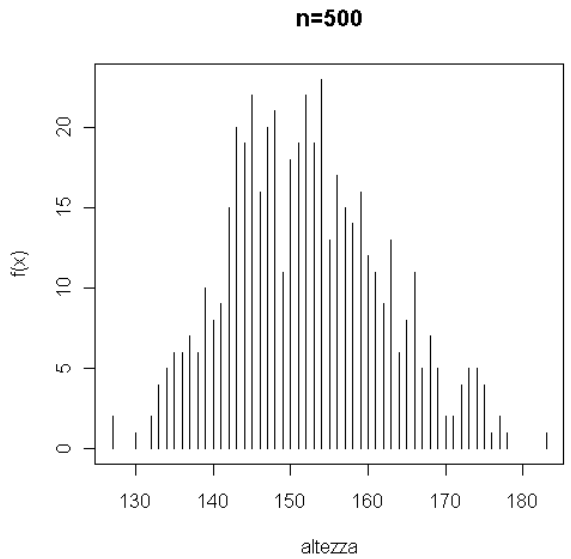
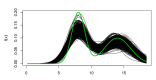


Figure: density function $n = 500$



Density function estimation for large samples

With large samples things goes better (but in the example there are rounding problems...)

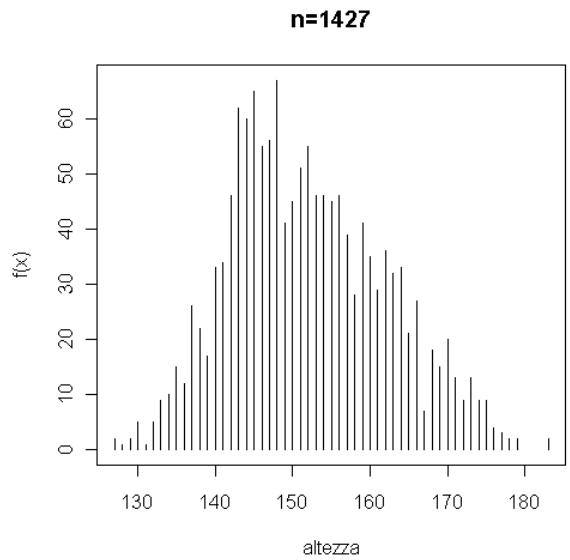


Figure: density function $n = 1,427$

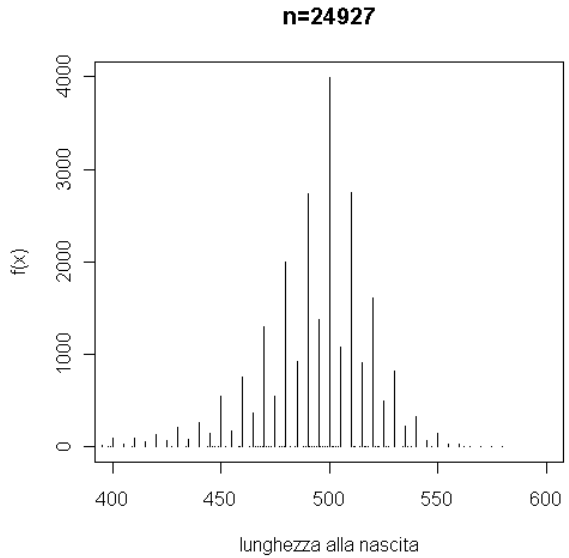
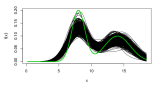
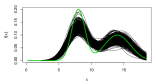


Figure: density function $n = 24,927$ (heights at birth)



Large samples

- With very large samples of observations (we will clarify the term **very large ...**) there are fewer problems, because histograms can have many classes, each with many observations.
- The sampling distribution of the number of observations in each class follows a binomial distribution.
- In the following slides, every time we talk about $f(x)$, we assume that the random variable X has got a density function



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Example with large samples

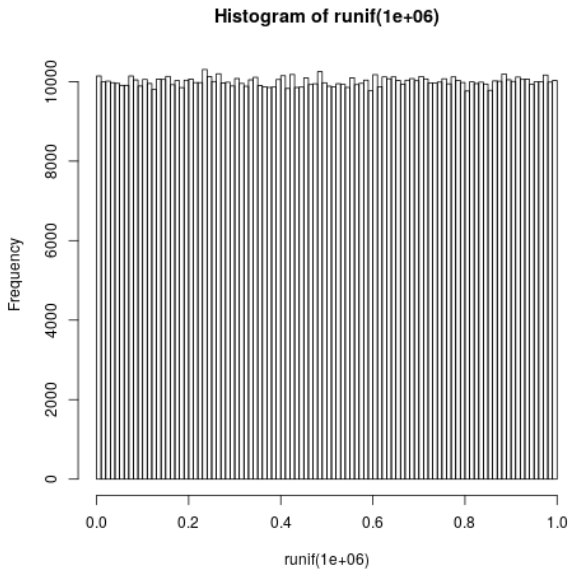


Figure: histogram of 1,000,000 uniform random numbers, with 100 intervals

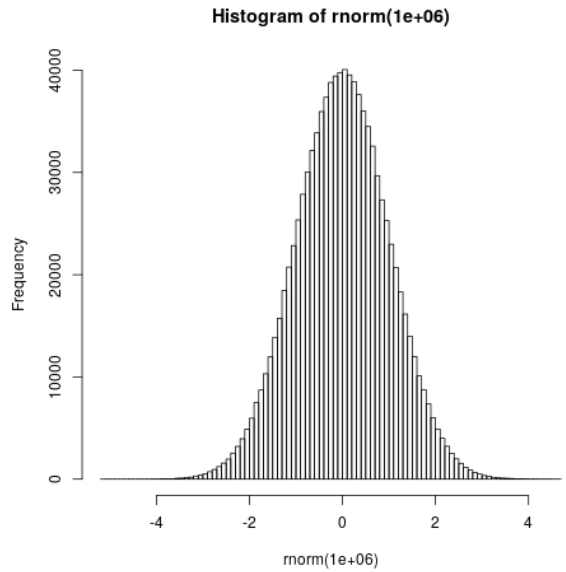
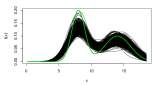


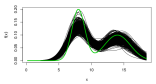
Figure: histogram of 1,000,000 normal random numbers, with 100 intervals



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Difficulties in the definition of $\hat{f}(x)$

- The difficulty in the definition of a density function estimator of a continuous random variable lies on the conceptual difference between the density and distribution functions for a random variable.
- While the latter is always well defined as the probability that X less than or equal to a certain value x (as well as the empirical equivalent), the density is defined as the derivative of that function;
- (**The continuity of $F(x)$ clearly has not an empirical correspondence** , because a large a set of observations will always be finite and non-dense).

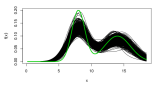


$f(x)$ as a probability approximation

- The probability for a continuous variable is defined for **intervals** and **not for points;**
- We know the conceptual difficulties of defining the probability that X has an exact value x, which is always zero, because $X = x$ is an event of null measure!
- Each of the sample values has a null probability of being drawn

Remember that

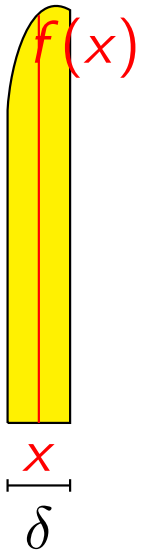
$$\int_{-\infty}^{+\infty} f(x)\ dx = 1$$



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

$f(x)$ as an approximation of probability

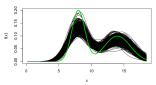
Another difficulty arises even if we try to see $f(x)$ as a function of a probability in the following way:



$$\blacksquare = \text{Prob}\left\{x - \tfrac{\delta}{2} < X \leq x + \tfrac{\delta}{2}\right\}$$

$$= \int_{x-\frac{\delta}{2}}^{x+\frac{\delta}{2}} f(x)\, dx = f(\epsilon)\delta \qquad (\approx f(x)\delta)$$

$\exists\ \epsilon$ within the interval (mean value theorem)



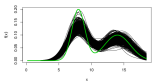
Approximations in the density estimates

it is known that the probability that X belongs to an interval can be approximated by the density at the point:

$$\text{Prob}\left\{x - \frac{\delta}{2} < X \leq x + \frac{\delta}{2}\right\} \approx \delta f(x)$$

- We could approximate the density using (49) for the probability in a small interval around x :

$$f(x) \approx \frac{\text{Prob}\left\{x - \frac{\delta}{2} < X \leq x + \frac{\delta}{2}\right\}}{\delta}$$



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

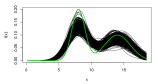
Approximations in the density estimates

- the probability on the right can be estimated from a sample of n observations:

$$\text{Prob}\left\{x - \frac{\delta}{2} < X \leq x + \frac{\delta}{2}\right\} \approx \frac{\#\left\{x - \frac{\delta}{2} < x_i \leq x + \frac{\delta}{2}\right\}}{n}$$

and then:

$$\hat{f}(x) = \frac{\#\left\{x - \frac{\delta}{2} < x_i \leq x + \frac{\delta}{2}\right\}}{\delta\ n}$$

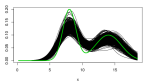


dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Approximations in the estimate of $f(x)$

This leads to two orders of approximation:

- To approximate the theoretical probability in an interval $\text{Prob}\left\{x - \frac{\delta}{2} < X \leq x + \frac{\delta}{2}\right\}$ through the density at the central point $(f(x) \cdot \delta)$
- To estimate this probability through sample observations $\left(\frac{\#\left\{x - \frac{\delta}{2} < x_i \leq x + \frac{\delta}{2}\right\}}{n}\right)$



δ small or large?

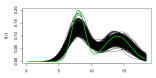
It is easy to see that δ affects differently the two approximations:

Role of δ

- The theoretical probability in an interval is well approximated by the density at the central point ($f(x) \cdot \delta$) if δ **is small**
- The probability is estimated better when we have more observations in the interval, and then, with the same n, when δ **is large**

Role of δ in the approximations

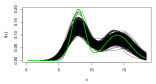
This diversity of behavior of two characteristics of a density estimator with respect to δ will characterize large part of this course.
This trade-off will appear again in more complex and general settings



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

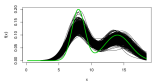
Local and integral properties

It should be noted that the approximations we have seen so far are *local*, that is, are related to a fixed value of x.
We will extend these concepts to *global or integral* measures to evaluate the behavior of $\hat{f}$ over the whole domain of X



What to do in general?

- If we have few observations? What can we do, without a parametric model? If we had a parametric model we could compute $f(x;\hat{\theta})$
- Or, despite we have many observations, the range of the variable is not well covered (very asymmetric distributions, or suspected presence of multimodality ...)
- Despite having many observations how can we get a continuous estimate $\hat{f}_X(x)$ (possibly with some of its derivatives?) (the histogram is not, however, a continuous function)
- Let's try to work *directly* on data to obtain $\hat{f}_X(x)$, without using $\hat{F}_X(x)$ (we exclude the inefficient possibility of to approximating $\hat{F}_X(x)$ by means of a continuous function and then deriving it)



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

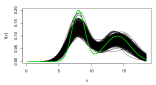
Density function estimation through incremental ratio approximation

A first approximation is given by the histograms used (whose properties will not be studied in these lectures, e.g. with reference to how to select the number of classes) **which however provides a discontinuous estimate of the density.**
We can proceed in general considering that:

$$f_X(x) = \lim_{h \to 0} \frac{\text{Prob}\{x - h < X \leq x + h\}}{2h}$$

and then we can, at least formally, use as an approximation, for fixed h: ($2h = \delta$ of previous slides)

$$\hat{f}_X(x) = \frac{\#(x - h < x_i \leq x + h)}{2\ n\ h}$$



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

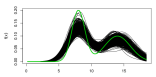
approximation of the Incremental ratio as average of contributions

Each point x_i contributes to the estimate of f if it is far from x less than h. So we can write $\hat{f}(x)$ as an average of the contribution of each point:

$\hat{f}(x)$ as an average

$$\hat{f}_X(x) = \frac{1}{2\ n\ h} \sum_{i=1}^{n} I(x - h < x_i \leq x + h) \tag{4}$$

($I(\cdot)$ is the indicator function `I(TRUE)=1`)
The key aspect of the reasoning, that will lead us also to very elaborate estimators, lies entirely in this **approximation of the incremental ratio as the average of the contributions of the individual observations**

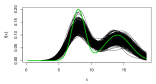


dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Similarities with the histogram

Note:
In a histogram with k classes the reasoning is similar, but while in (4) each interval is constructed around each value of x, in the histogram $\hat{f}$ is estimated with the same value for all x belonging to the j-th class $j = 1, 2, \ldots, k$, with $[a_{j-1}, a_j]$ fixed extremes:

$$\hat{f}_X(x,\ x \in [a_{j-1}, a_j]) = \frac{1}{n\ (a_j - a_{j-1})} \sum_{i=1}^{n} I(a_{j-1} < x_i \leq a_j) \tag{5}$$



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Average of uniform densities

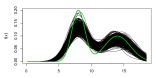
Let us come back to (4) and collect properly the terms, using a more compact notation:

$$\hat{f}_X(x) = \frac{1}{n} \sum_{i=1}^{n} U_h(x - x_i) \tag{6}$$

where:

$$U_h(z) = \begin{cases} \frac{1}{2h} & |z| \leq h \\ 0 & |z| > h \end{cases}$$

$U_h(z)$ is the density of a *uniform distribution* in the range $\{-h, h\}$. To simplify future generalizations, it is better working with densities of standardized variables.



Average of standardized uniform densities

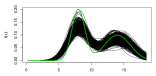
Using a standardized uniform variable , i.e. defined in the interval (-1, 1):

$$\hat{f}_X(x) = \frac{1}{n\ h} \sum_{i=1}^{n} W\left(\frac{x - x_i}{h}\right) \tag{7}$$

where:

$$W(z) = \begin{cases} \frac{1}{2} & |z| \leq 1 \\ 0 & |z| > 1 \end{cases}$$

This form leads us to future generalizations, because $f(U)$ is discontinuous at the ends and this always leads to the **steps** in the estimate of $f(x)$ by (7)



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

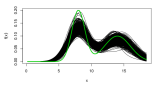
Examples

R code

examples $\Rightarrow$

code R

./panel_kernel2013.R



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Density estimation as combination as uniform densities

Example of the previous method with two different values of h

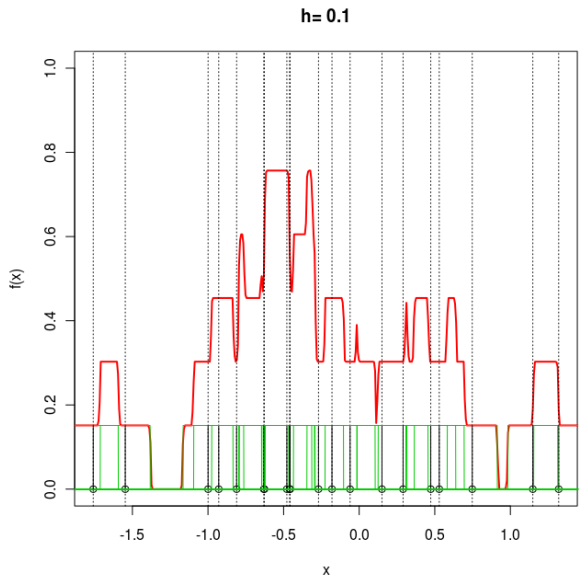


Figure: step density estimates $n = 20$

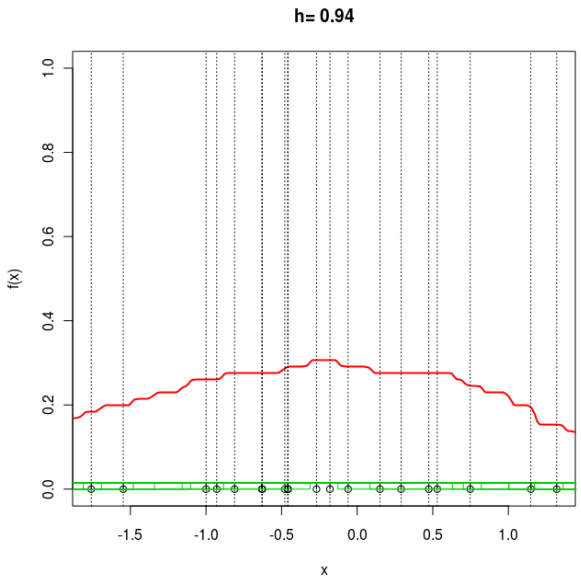
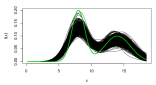


Figure: step density estimates $n = 20$



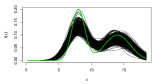
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Origin of kernel estimators

To eliminate the discontinuity, we can replace the uniform density with another generic density function $K(\cdot)$
The general form is:

$$\hat{f}_X(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - x_i}{h}\right) \tag{8}$$

- the function $K(\cdot)$ is called the kernel
- h is the width, or window, or bandwidth (or *smoothing parameter*)
- $\frac{1}{k}K(\frac{x-x_i}{h})$ is the weight of each observation in the determination of $\hat{f}(x)$.

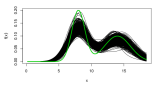


dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Construction of the kernel estimator

Basically, every point *spreads* its influence according to $K(\cdot)$ and h.

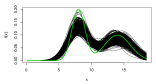
- As regards $K(\cdot)$, we consider at the moment proper densities, symmetrical around zero and standardized.
- The most obvious example for $K(\cdot)$ is the density of a standard normal distribution:
- **even if not 100% excellent, this choice gives results useful in many applications, and briefly it is very confortable**
- Obviously observations x_i closer to x are more influent, large values of h attenuate the influence of the nearest observations and spread their influence over a wider range.



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Construction of the kernel estimator

- Small values of h give weights concentrated in the immediate proximity of each x_i and thus provide less overall density estimates vary from smooth to x. The influence of each x_i is limited to a small range.
- If $K(\cdot)$ is a standard normal kernel , the density out of range $\{x_i - 4h, x_i + 4h\}$ is virtually zero, which means that the observations far from x more than 4 times the bandwidth h will not influence the estimate of $f(x)$ (actually even distances of over twice the bandwidth make the influence of an observation very small)



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

density estimation with normal kernel

Density with normal kernel with two different values of h

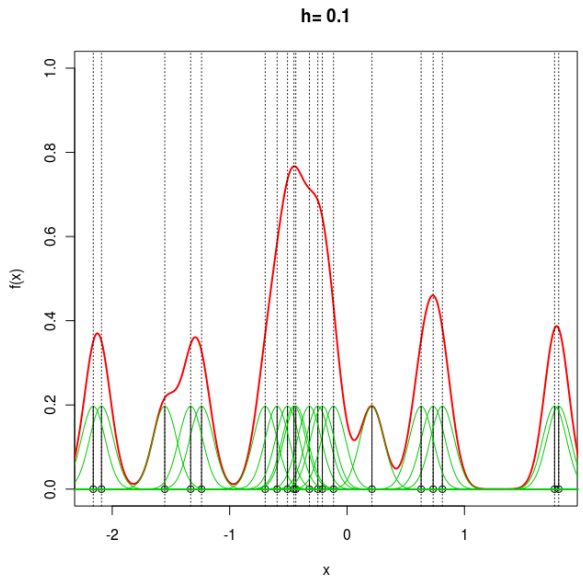


Figure: normal kernel density estimate, $n = 20$

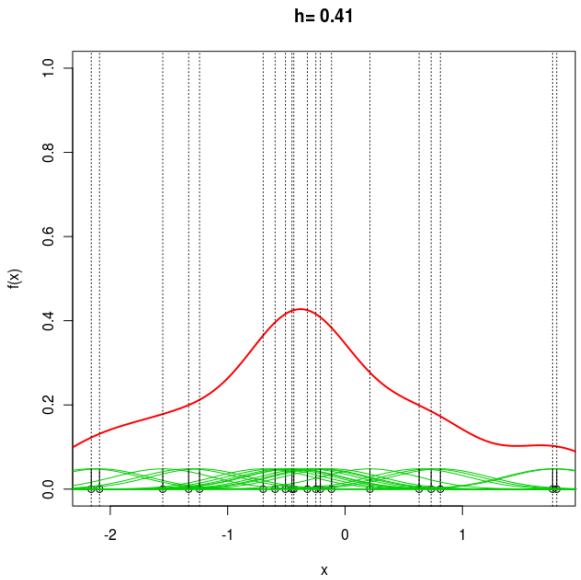
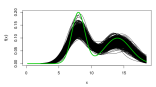


Figure: normal kernel density estimate, $n = 20$



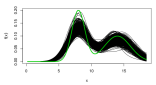
examples

examples with different h
 and different K
 graphics
 dynamic examples with R

and run draw1 ()

\url{./panel_kernel2013.R}

RIVEDERE



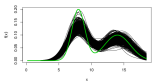
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Types of kernel functions

$K(x)$ determines the shape of the curves while the bandwidth h determines the width. The kernel estimator therefore depends on two elements:

1. the kernel K, (**its choice has little influence on the results** as we will see later)
2. the bandwidth h, (**it influences substantially the results: acting on h we range from a density with n modes (or peaks) to a flat one**)

We will see these properties theoretically and practically



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Estimation of density with different kernels

Density with 4 different kernels and two different values of h

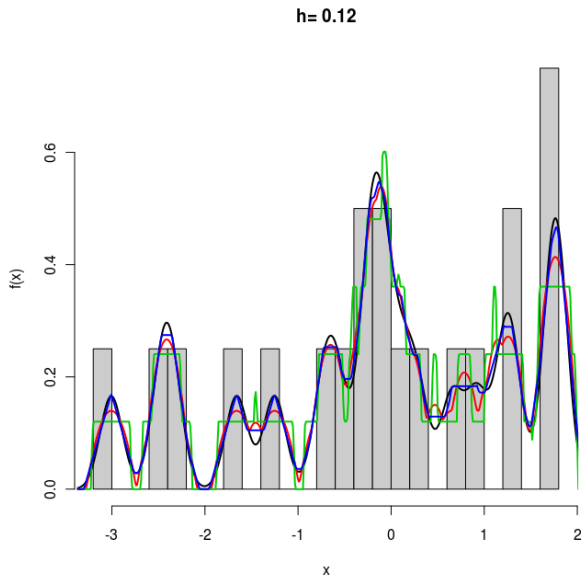


Figure: 4 different kernels

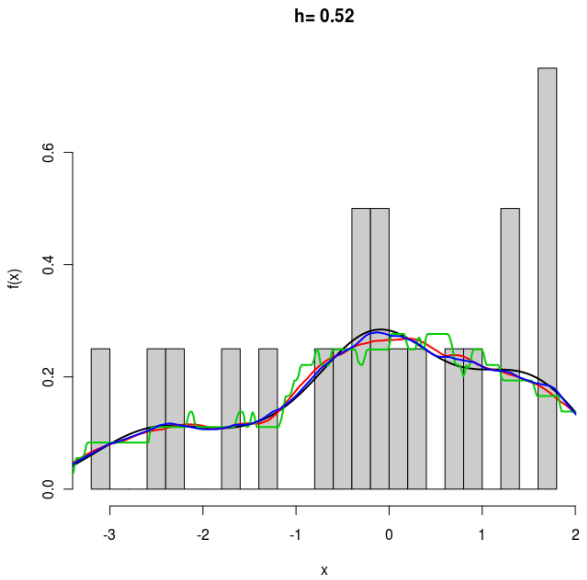
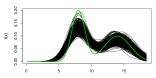


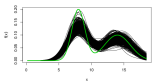
Figure: 4 different kernels



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Types of kernel functions

code examples in R and
`./panel_kernel2013.R`
these properties will be proved analytically and exemplified with procedures in R
h is denoted in several ways:
bandwidth,
window,
smoothing parameter,
scale parameter of the kernel function,
etc..

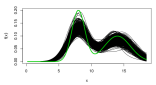


dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Common kernel functions

- Epanechnikov $\left(K = \begin{array}{ll} \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}t^2\right) & -\sqrt{5} \leq t \leq \sqrt{5} \\ 0 & \text{otherwise} \end{array} \right)$

  $eff(K) = 1.$
- Triangular $(K = 1 - |t|$ if $|t| < 1, 0$ otherwise $)$: $eff(K) \approx 0.9859.$
- Gaussian $\left(K = \frac{1}{\sqrt{2\pi}} e^{-(1/2)t^2}\right)$: $eff(K) \approx 0.9512.$
- Rectangular $\left(K = \frac{1}{2}$ if $|t| < 1, 0$ otherwise $\right)$: $eff(K) \approx 0.9295.$
- All efficiency values are not very different from 1, even for the rectangular kernel. So the choice of the kernel can not be based on the MISE,



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Common kernel functions

the MISE will be discussed later, at the moment it is sufficient to say that it is a global measure of the behavior of an estimator of a density

MISE = Integrated Mean Square Error

Of course, other less statistical considerations hold, such as the degree of differentiability required and the computational effort.

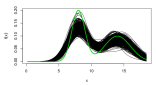
Example (code R)

`./panel_kernel2013.R`

codice R interattivo

R

`panel_kernel2013.R`



dSEAS dipartimento di scienze economiche aziendali e statistiche department of economics business and statistics

Density estimation with different kernel functions

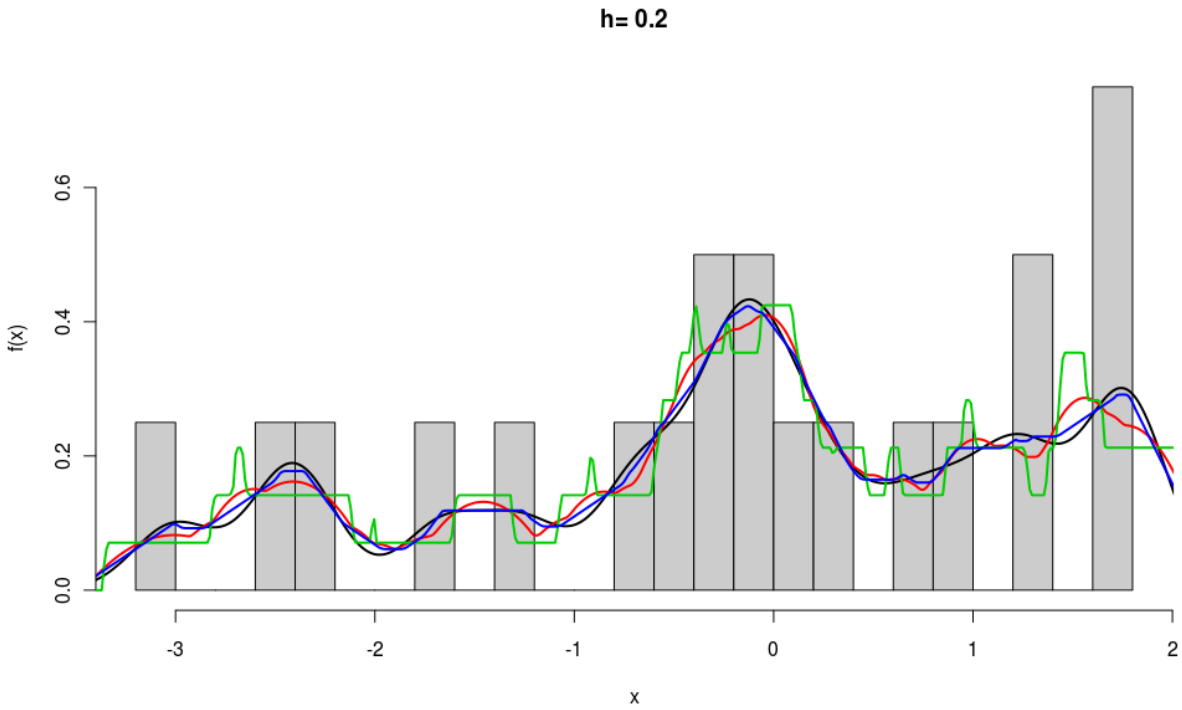
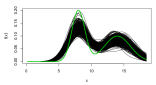


Figure: 4 different kernel estimators $n = 20$



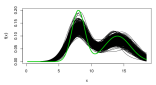
Meaning of the kernel estimator of a density

estimator of $f(x)$ based on the incremental ratio

In the estimator of $f(x)$ based on the incremental ratio, the contribution of each point x_i is 0 or 1
according to the fact that the distance from x is greater or less than h

kernel estimator of $f(x)$

In the kernel estimator of $f(x)$, the contribution of each point x_i varies according to a function: $\frac{1}{h}\ \mathrm{K}\ \left(\frac{x-x_i}{h}\right)$



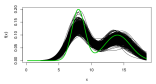
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Analytical characteristics of the kernel function

The kernel function must satisfy the following conditions:

- $\int_R K(x)\ dx = 1$
- $\int_R x\ K(x)\ dx = 0$
- $\int_R x^2\ K(x)\ dx = k_2 \neq 0 \qquad (< \infty)$
- (standardized functions are used, so that $k_2 = 1$)
- $K(x) \geq 0 \qquad \forall x;$

Therefore, the kernel estimator $\hat{f}_h(x)$ is a density of probability, such that $\int_R \hat{f}_h(x)dx = 1$
($K(x)$ should also be symmetrical **but it is not the only possible choice ...**).

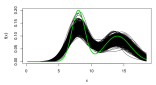


Variable width kernels

Let h_{j,n_p} the radius of the circle centered at x_j that contains other n_p points; a definition of a variable kernel can be obtained from:

$$\hat{f}_{h_j}(x) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h_{j,n_p}} K\left(\frac{x - X_j}{h_{j,n_p}} \right)$$

- in this case, the observed points in regions with sparse data have kernel flatter (or smoothed).
- The result depends on n_p.
- The kernel estimator is still a probability density.
- More detailed with respect to the kernel h fixed, even if we use a smaller h.



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

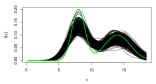
Technical justifications of kernel estimators

- Approximation of an incremental ratio
- Average of the density contributions of each point
- Approximation problem or smoothing of a histogram
- as a convolution of densities

It can shown that the histogram $\hat{f}_h(x)$ is a consistent estimator of $f(x)$, ie

$$h \to 0, nh \to \infty \Rightarrow MSE(f_h(x)) \to 0.$$

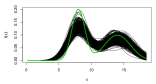
we will prove similar properties for kernel estimators.



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Statistical properties of kernel estimators

1. how to measure the goodness of an estimator? If possible, we will try to exploit the classic definitions used in a parametric framework.
2. ISE and MISE
3. simulations
4. asymptotic behavior
5. choice of h?
   - effect of h
   - effect of $K(\cdot)$
   - **cross validation or other techniques**

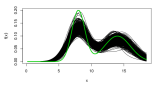


dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Choice of $K(\cdot)$

The choice of K is based on considerations relative to:

1. the efficiency of the estimator $\hat{f}_h(x)$;
   but we will see that the choice of $K(\cdot)$ *is not crucial*
2. the degree of regularity that we expect $\hat{f}_h(x)$ (i.e. if we choose the uniform kernel, it will be discontinuous!);
3. the computational effort required
   (less and less relevant over time)
4. analytical convenience
   Some asymptotic and analytical results are obtained more easily if $K(\cdot)$ *is normal, and this is even more true in the multivariate case*

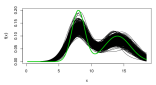


dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

The choice of h

1. if $h \to 0$, $\hat{f}_h(x)$ tends to a sum of quantities $(nh)^{-1}K[(x - x_i)/h]$ which are high in correspondence of x_i and small elsewhere. So the estimated density with $\hat{f}_h(x)$ will be irregular and rough with a peak at each observation;
2. if $h \to \infty$, $\hat{f}_h(x)$ tends to a sum of quantities $(nh)^{-1}K[(x - x_i)/h]$ which are small and flat, and then the curve estimated will be smooth with a tendency to smooth the spurious peaks

In summary: small values of h give a local high weight to each observation (strong irregularities), while high values of h give a low weight to the individual observations (regular $\hat{f}_h(x)$).



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

MISE computation through simulations

Samples of size 100

A simulated sampling distribution: explain the elements of the figure (we will make during the lesson other simulation experiments)

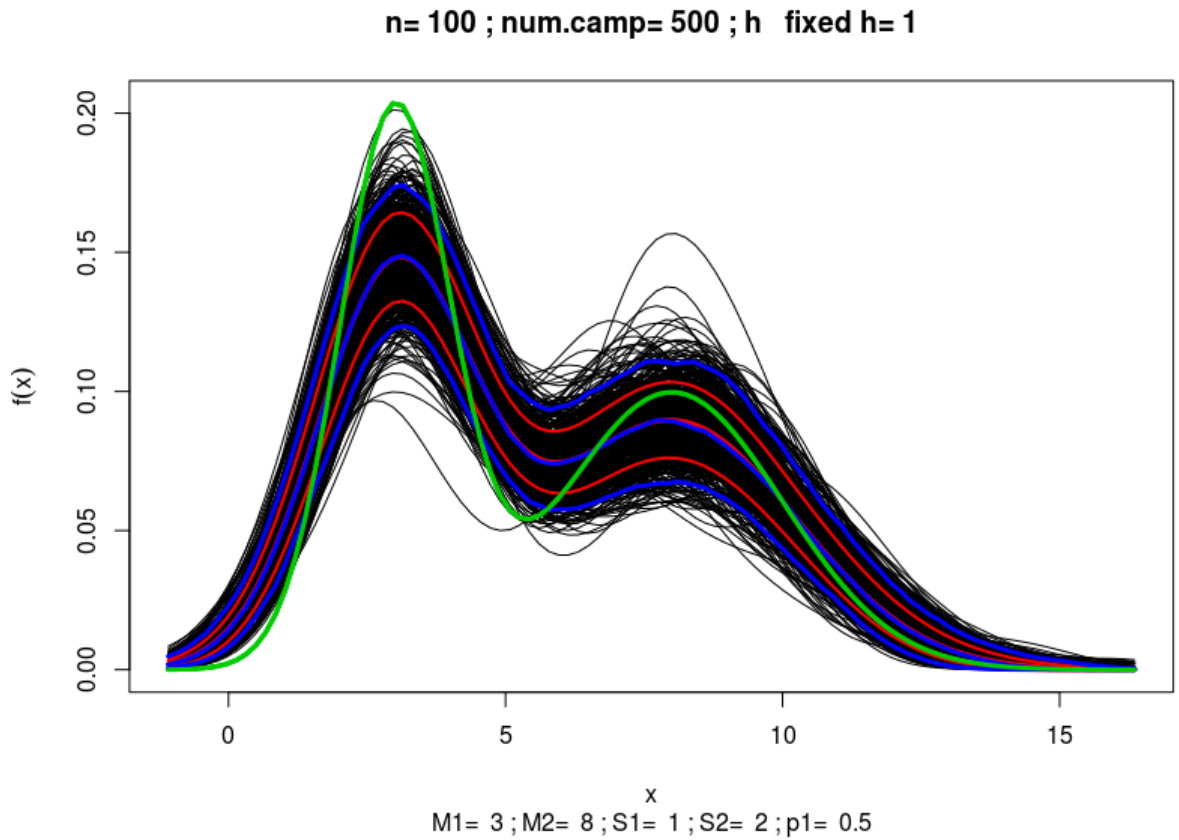


Figure: Simulations from a mixture of two normal distributions

Simulations

Example (R code for simulations)

./codiceNP2013simul1.R

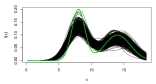
./simulazioni_kernel2012functions.R

codice R

R

```
codiceNP2013simul1.R
source("simulazioni_kernel2012functions.R")
```

EXPLAIN THE ELEMENTS OF THE FIGURE



Measurements of the properties of a nonparametric estimator

Let $X_1, \cdots, X_n$ iid v.a. with unknown density $f(\cdot)$.

- **Bias**: $B[\hat{f}_h(x)] = E_F[\hat{f}_h(x) - f(x)]$
- **Variance** $V[\hat{f}_h(x)] = E_F[\hat{f}_h(x) - E[\hat{f}_h(x)]]^2$

Trade-off between bias and variance

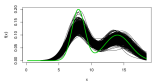
As we shall see, there is an inevitable contradiction between the goal of reducing both distortion and the variance as a function of h.
as in any estimation problem!

$$\Downarrow$$

The choice of h is a **trade-off between bias and variance !** (as already seen in estimators based on probabilities on intervals)
The old problem of statistical inference

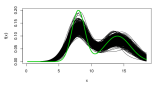
Properties of the estimators -1

- Initially we consider **local properties** for a fixed value x_0, as if $f(x_0)$ be an unknown parameter to be estimated.
- The main point is that we are essentially extending to **nonparametric problems** methods of evaluation of estimators **designed for parametric problems**
- However the importance of the density estimators lies in their ability to provide information on $f(x)$ in an **exploratory phase of data analysis**



Explorative tools

- Another fundamental aspect is the *graphical aspect* that may have the representation of a density estimated by the kernel, with a bandwidth large enough be smooth, but small enough to highlight possible multimodalities
- Given the great number of graphical interactive tools present now, like those used in this brief course, we can think that practically we can use all of them.



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Recall on Mean Square Error

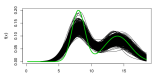
The mean square error of a generic point estimator $\hat{\theta}$, is a general measure of the quality of an estimator:
That is, its capability to give information about the true parameter θ. It is defined as the expected value (on the sampling distribution of $\hat{\theta}$) of the squared difference between estimator and true values of the parameter

MSE=*Variance* + *Bias*^{2}

$$MSE[\hat{\theta}] = \mathrm{E}\left[(\hat{\theta} - \theta)^2\right] =$$

$$= \mathrm{E}\left[\left(\hat{\theta} - \mathrm{E}\left[\hat{\theta}\right]\right)^2\right] + \left\{\mathrm{E}\left[\hat{\theta} - \theta\right]\right\}^2 = Var[\hat{\theta}] + (Bias[\hat{\theta}])^2$$

$$Var[\hat{\theta}] = \mathrm{E}\left[\left(\hat{\theta} - \mathrm{E}\left[\hat{\theta}\right]\right)^2\right] \; Bias[\hat{\theta}] = \mathrm{E}\left[\hat{\theta} - \theta\right]$$



dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Mean Square Error

The mean square error of $\hat{f}(x)$ at a fixed point x_0, reflects the trade-off between the two components (variance and bias):

`Let us use` x_0 `instead of the usual` x`, to highlight that we are interested in local properties of the estimator, i.e. the property at a particular point` x_0

Here $f(x_0)$ plays the role of θ, and $\hat{f}_h(x_0)$ the role of $\hat{\theta}$

$MSE[\hat{f}_h(x_0)]$

$$MSE[\hat{f}_h(x_0)] = \mathrm{E}\left[(\hat{f}_h(x_0) - f(x_0))^2\right] =$$

$$= Var[\hat{f}_h(x_0)] + \left\{\mathrm{E}\left[\hat{f}_h(x_0) - f(x_0)\right]\right\}^2 = Var[\hat{f}_h(x_0)] + (Bias[\hat{f}_h(x_0)])^2$$

$$Var[\hat{f}_h(x_0)] = \mathrm{E}\left[\left\{\hat{f}_h(x_0) - \mathrm{E}\left[\hat{f}_h(x_0)\right]\right\}^2\right]$$

$$Bias[\hat{f}_h(x_0)] = \mathrm{E}\left[\hat{f}_h(x_0)\right] - f(x_0)$$

Mean Square Error

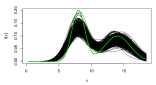
We have

$$Var[\hat{f}_h(x_0)] = \mathrm{E}\left[\left\{\hat{f}_h(x_0) - \mathrm{E}\left[\hat{f}_h(x_0)^2\right]\right\}\right]$$

$$BIAS[\hat{f}_h(x_0)] = \mathrm{E}\left[\hat{f}_h(x_0)\right] - f(x_0)$$

The **minimization of the MSE** with respect to h (if we **knew the true** $f()$) is a trade-off between two characteristics:

- *oversmoothing* (if we choose high values of h to reduce the variance) and
- *undersmoothing* (if we choose small h to reduce the bias).

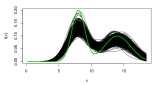


Mean Integrated Square Error (MISE)

- MSE measures the accuracy of the estimator $\hat{f}_h(x_0)$ of the density $f()$ at a particular point x_0. It is a **local measure**
- A global measure of goodness of fit, the **MISE** , can be obtained by integrating the value of **MSE** over the whole range of x:

$$MISE(\hat{f}_h) = \int_{-\infty}^{+\infty} MSE(\hat{f}_h(x))dx$$

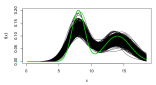
- Under certain assumptions on $f(\cdot)$ and $K(\cdot)$ we will try to choose, for the construction of a kernel estimator, the value of h that minimizes the MISE.



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Integrated Mean Square Error (MISE)

- This measure is not always optimal, (and not necessarily the integral exists or is finite) and has the disadvantage of measuring squared absolute distances and not relative distances: an error of 0.01 in the estimate of $f(x)$ should have a different importance according to the fact that $f(x)$ is equal to 0.3 or 0.02!
- The MISE as a measure of the overall behavior of $\hat{f}_h(x)$ is not optimal but it is computable, or at least approximated, in many standard situations.
- For analytical convenience we take as a measure of the overall behavior of the estimator of a density



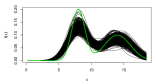
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

MISE: operators M and I

It should be stressed **the difference between the operators M and I used in the definition of the MISE**

- M (mean): expected value of the squared error (relative to the random distribution of X, and for given $f(\cdot)$, or computed by simulation)
- I (Integrated): integration with respect to x in the domain of X

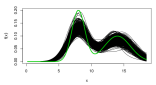
The two operations should not be confused. It is very important to keep in mind and understand their difference.



MISE: operators M and I

M and I of MISE

- Operator M in the MSE is relative to the sampling distribution of $\hat{f}_h(x_0)$, x_0 is fixed, the expected value is based on the probability distribution of the random variables $X_1, X_2, ..., X_n$
- The operator I instead is simply an integration with respect to a real variable ($MSE(x)$ is integrated with respect to x), we could approximate this with a sum on a finite set of x values which cover well its range of variation.



Simulations

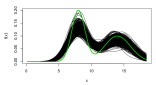
Example (R code for simulations)

codice R

R

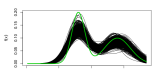
```
codiceNP2013simul1.R
source("simulazioni_kernel2012functions.R")
```

- Examples in R on the difference between the operators M and I used in the definition of the MISE through simulations
- R code
- EXPLAIN THE ELEMENTS OF THE FIGURE



dSEAS dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

```
source("simulazioni_kernel2012functions.R")

## 1: esempi di simulazione da una mistura di due norm
## su ogni campione di ampiezza n viene stimata una de
## e finestra fissa (in questo esempio)
system.time(f<-simul.normalmix2(n=100,ncamp=5,hmet="fi

## spiegazione dettagliata della simulazione
system.time(f<-simul.normalmix2(n=100,ncamp=5,hmet="fi



## altri valori di h
system.time(f<-simul.normalmix2(n=100,ncamp=5,hmet="fi



system.time(f<-simul.normalmix2(n=100,ncamp=5,hmet="fi
```



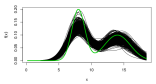
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Simulations

Simulations

To show basic behaviour we use code in
`simulazioni_kernel2012functions.R`

1. Fix the true density (green line)
2. Draw a sample of size n (not shown)
3. Estimate a kernel density with a fixed h (the black lines)
4. repeat steps 2 and 3 n_s times (500 in these examples)
5. Compute summary statistics (average , median, quartiles,...) for each point of the x axis
6. Draw summary statistics (blue and red lines)



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

MISE computation through simulations: Explanation of elements

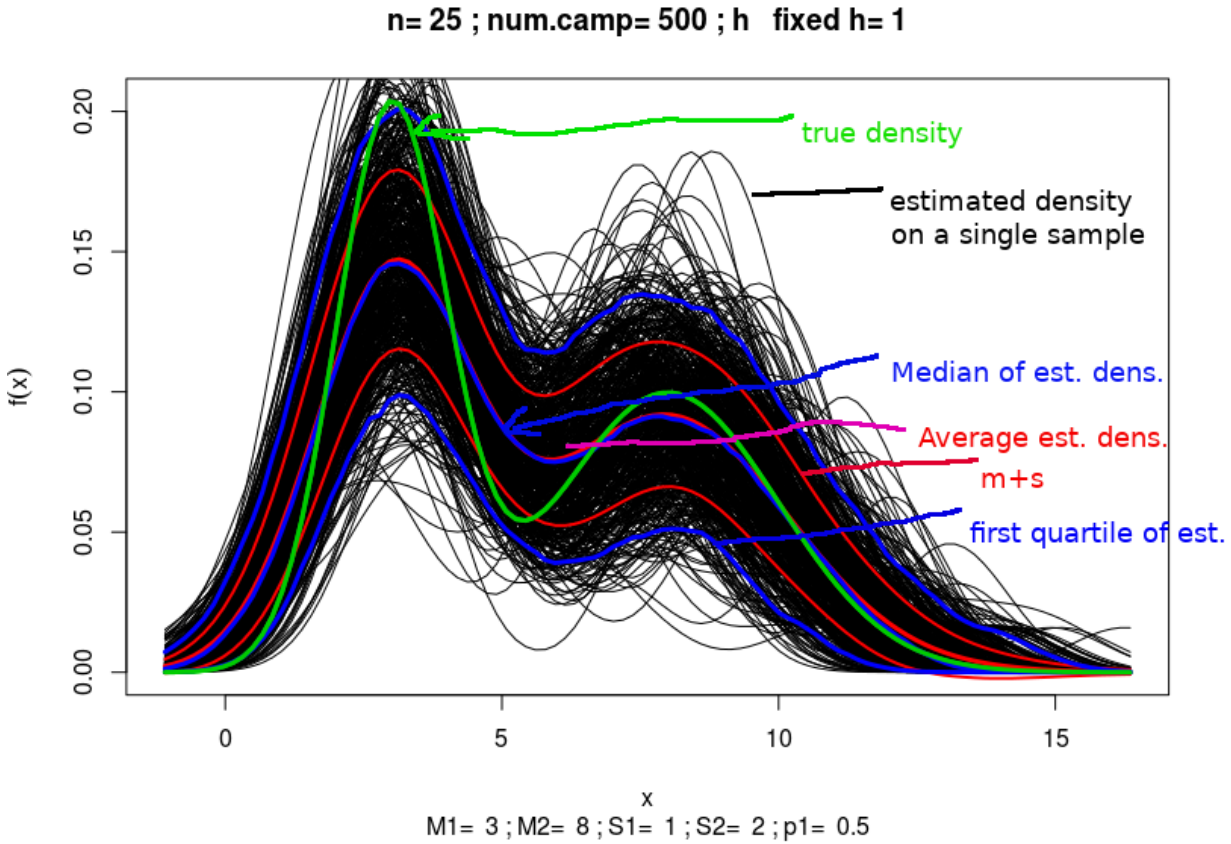
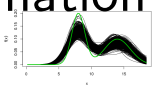


Figure: Simulations from a mixture of two normal distributions: Explanation of elements



MISE computation through simulations

Samples of size 25

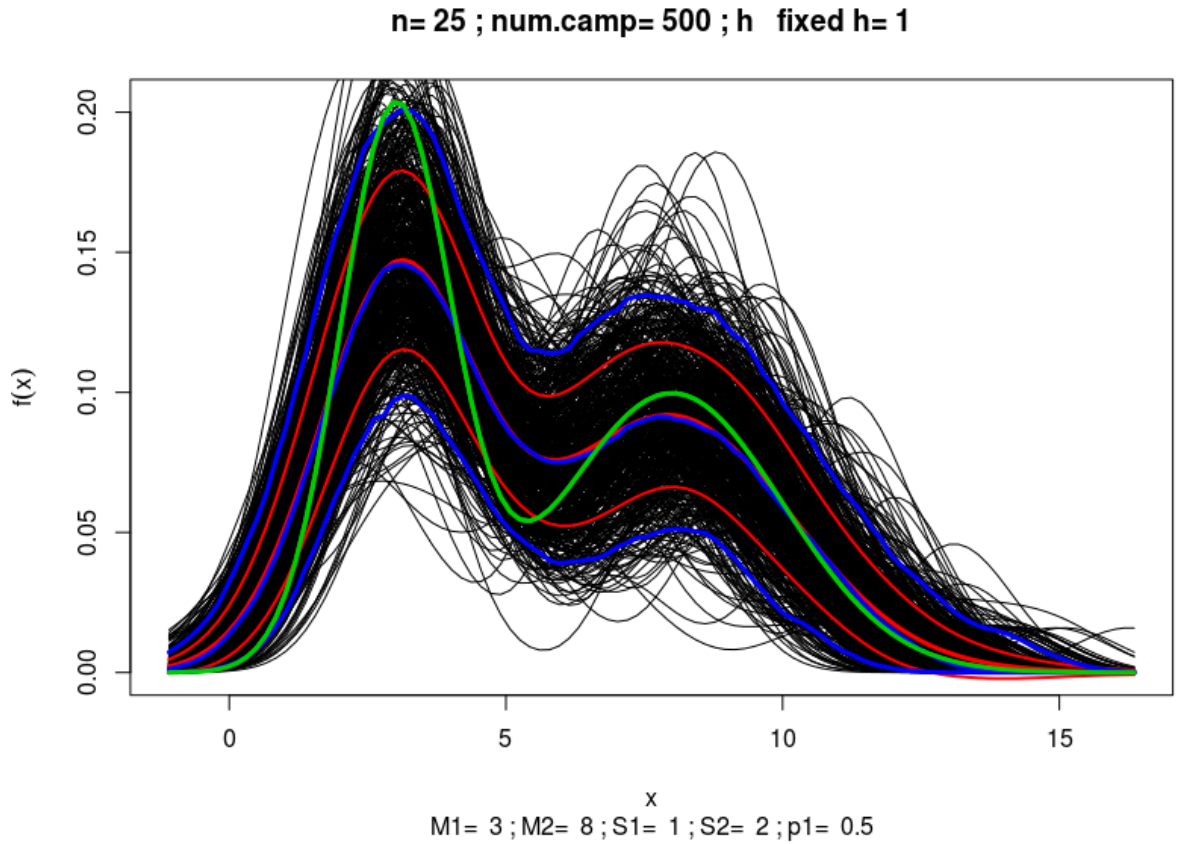


Figure: Simulations from a mixture of two normal distributions

MISE computation through simulations

Samples of size 100

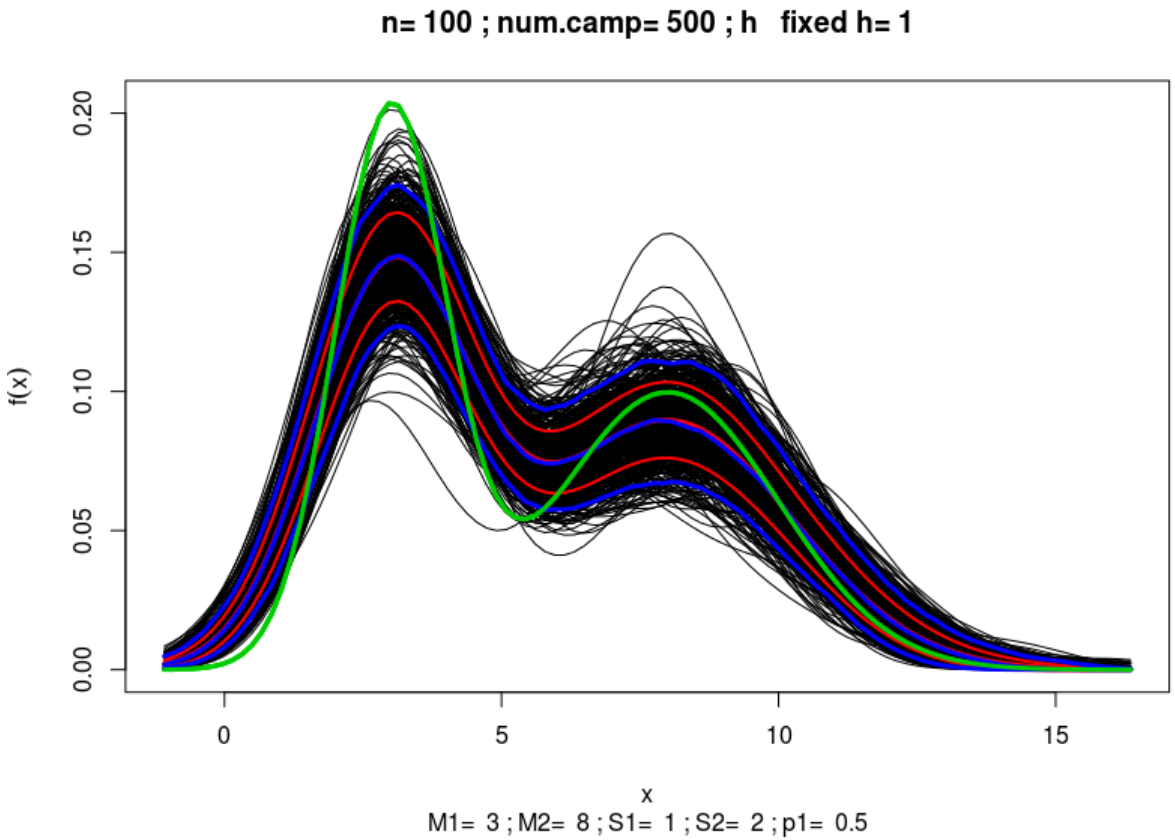


Figure: Simulations from a mixture of two normal distributions

MISE computation through simulations

Samples of size 400

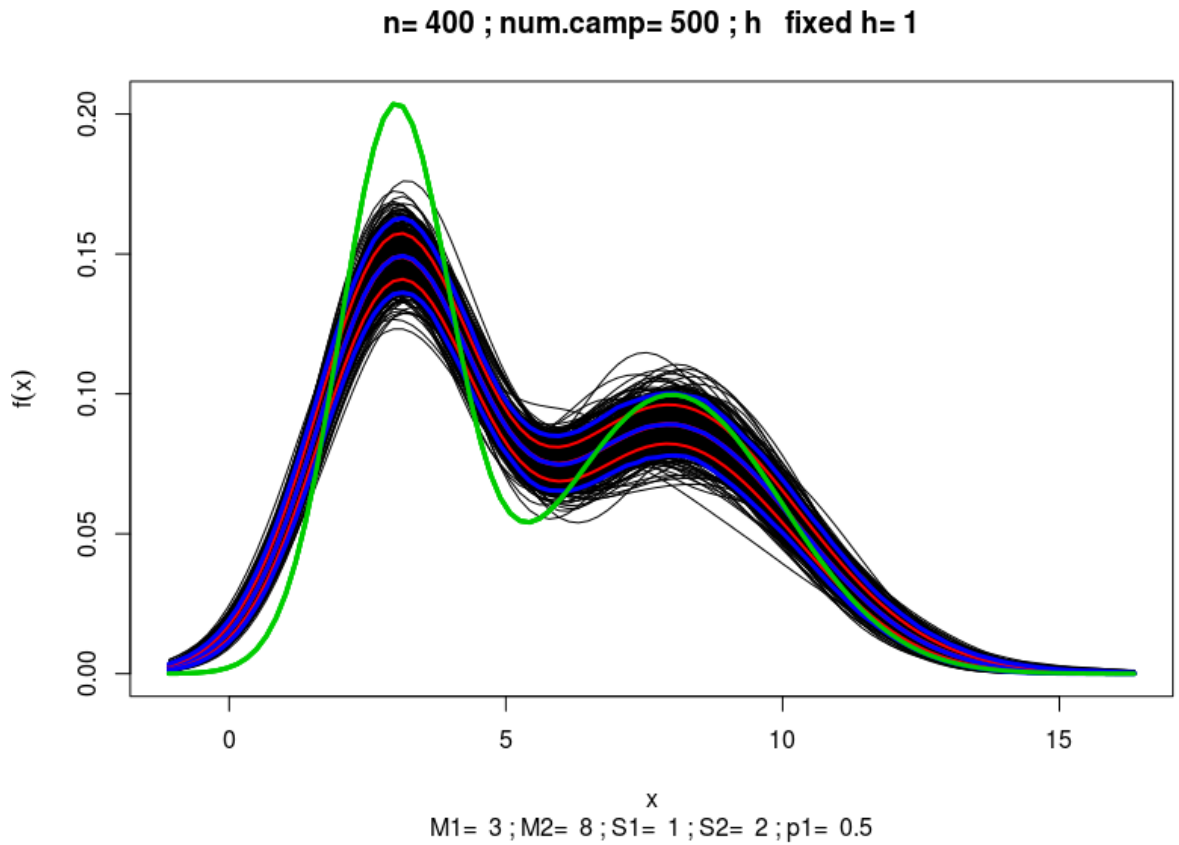


Figure: Simulations from a mixture of two normal distributions

MISE computation through simulations

Samples of size 1600

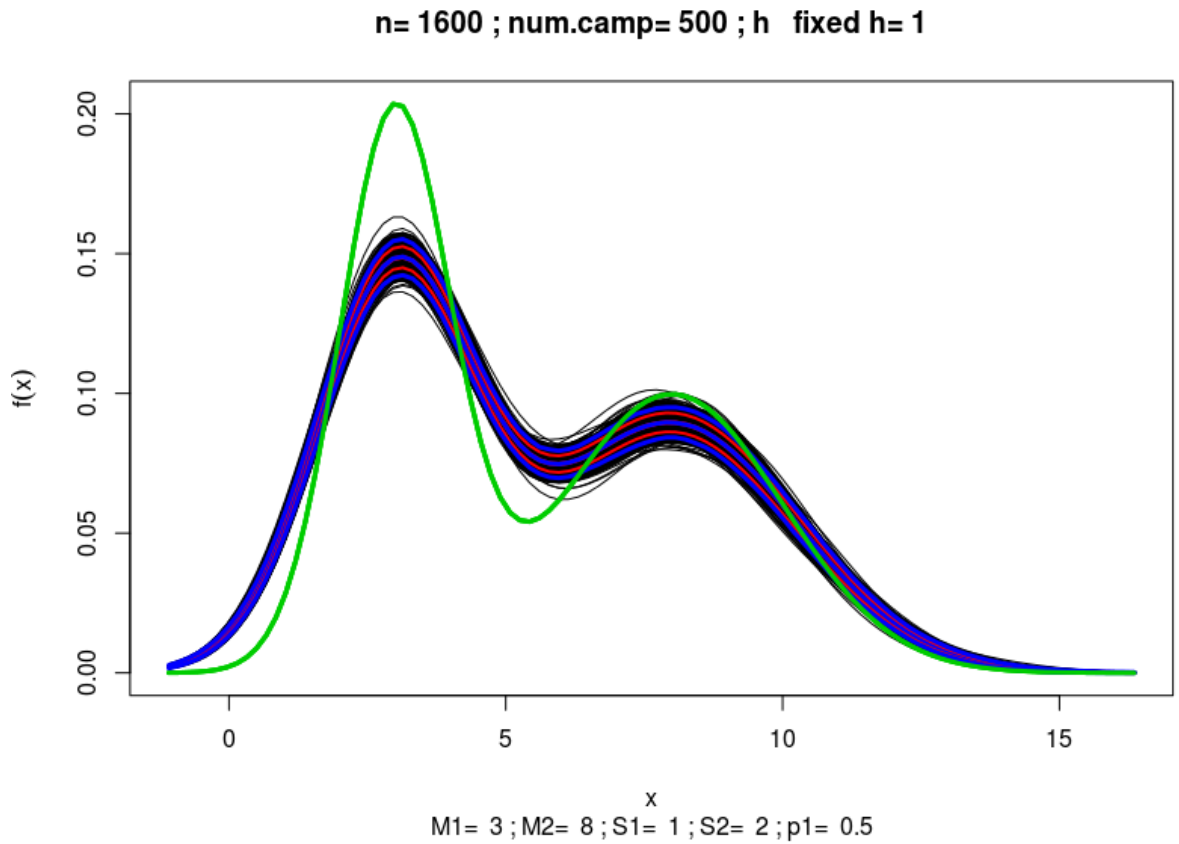


Figure: Simulations from a mixture of two normal distributions

Basic asymptotic results ($n \to \infty$)

To obtain basic asymptotic results (that is, as $n \to \infty$), it is necessary to assume some dependence of h by n and therefore in this section the symbol $h(n)$ will be used $h(n)$ should be a decreasing function. The key results are:

Asymptotic Bias and Variance

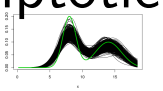
$$\text{Asymptotic Bias}(\hat{f}_{h(n)}(x)) = \frac{1}{2} h(n)^2 \; f''(x)$$

in simulations, bias seems small near inflection points

$$\text{Asymptotic Variance}(\hat{f}_{h(n)}(x)) = \frac{f(x)}{nh(n)} \int_{\Re} K(t)^2 dt$$

in simulations, variance seems higher near maximum points

Combining the two expressions we obtain the expression of asymptotic mean square error (AMSE $= AV + AB^2$)



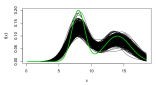
dSEAS dipartimento di scienze economiche aziendali e statistiche department of economics business and statistics

Asymptotic results ($n \to \infty$)

slides repeated later!

Some important considerations are deduced:

- $h(n)$ has a direct effect on bias but an inverse one on variance
- Bias is a function of $h(n)$: if we want it to be asymptotically equal to zero, the bandwidth $h(n)$ must tend to zero as $n \to \infty$ (and this seems quite reasonable)
- Instead the variance is an inverse function of $h(n)$: If we want that this variance is asymptotically equal to zero, it is necessary that $\frac{1}{nh(n)}$ tends to zero.
- and then $nh(n)$ must diverge.



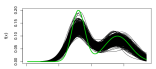
dSEAS dipartimento di scienze economiche aziendali e statistiche department of economics business and statistics

asymptotic results

- Combining the two previous requirements, $h(n)$ must tend to zero as $n \to \infty$, to delete the bias, *but slowly* , because to delete also the asymptotic variance we must have:

$$\lim_{n\to\infty} \frac{1}{nh(n)} = 0 \qquad \text{and so:} \qquad \lim_{n\to\infty} nh(n) = \infty$$

  and then n^{-1} *must be an infinitesimal of higher order than* $h(n)$
- dependence (local) from $f(x)$ and $f''(x)$
- **the asymptotic behavior of $\hat{f}(x)$, in terms of order of magnitude, is determined by $h(n)$ and not by the choice of $K(\cdot)$**



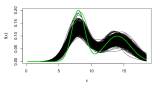
dSEAS dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

asymptotic results

A value of $h(n)$ that meets the previous requirements and that minimizes the asymptotic mean squared error is:

$$h(n)_{opt} = n^{-1/5} \left(A(K) \frac{f(x)}{f''(x)} \right)^{1/5}$$

It depends on x (and therefore we will minimize an integrated form of the mean square error)
It must be noted that the order of magnitude is $n^{-1/5}$, in this case the two components of the AMSE are both of order of $n^{-4/5}$ (indeed $AV = 4AB^2$)



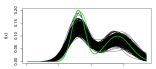
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

On bias and variance

Let us consider, in the standard expression of the kernel estimator, x fixed, while the x_i are n determinations of random variables Y_i (to avoid confusion with the x fixed) independent and equally distributed according to a density distribution $f(y)$ (in the domain $\mathcal{D}$)
We have:

$$\mathrm{E}\left[\hat{f}_h(x)\right] = \int_{\mathcal{D}} h^{-1} K\left(\frac{x-y}{h}\right) f(y)dy, \tag{9}$$

$$n\ \mathrm{V}\left[\hat{f}_h(x)\right] = \int_{\mathcal{D}} h^{-2} K\left(\frac{x-y}{h}\right)^2 f(y)dy - \left[\int_{\mathcal{D}} h^{-1} K\left(\frac{x-y}{h}\right) f(y)dy\right]^2. \tag{10}$$



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

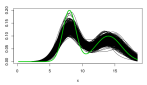
Bias and variance

The bias of $\hat{f}_h(x)$ depends on h (knowing anyway that h depends on n!).

$$B(x) = \mathrm{E}\left[\hat{f}_h(x)\right] - f(x) \; = \; \int_{\mathcal{D}} h^{-1} K\left(\frac{x-y}{h}\right) f(y)dy - f(x)$$

Consider the change of variable $y = x - ht$, and then $|dy| = h|dt|$ then:

$$B(x) = \int_{\mathcal{D}*} K(t)f(x-ht)dt - f(x) \; = \; \int_{\mathcal{D}*} K(t)(f(x-ht) - f(x))dt$$



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

asymptotic results

The Taylor series expansion is:

$$f(x-ht) = f(x) - htf'(x) + \frac{1}{2}h^2t^2f''(x) + or[(ht)^2]$$

So with the assumptions on $K(\cdot)$:

$$B(x) = -hf'(x)\int_{\mathcal{D}*} tK(t)dt + \frac{1}{2}h^2f''(x)\int_{\mathcal{D}*} t^2K(t)dt + \dots$$

$$= \frac{1}{2}h^2f''(x)k_2 + o(h^2)$$

and the **integrated** squared bias is:

Asymptotic integrated squared Bias

$$\int_{\mathcal{D}} B(x)^2dx \approx \frac{1}{4}h^4k_2^2\int_{\mathcal{D}} f''(x)^2dx$$

depends on h and the curvature of $f''(x)$!

Asymptotic results

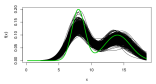
For the variance, from (9) and (10):

$$\mathrm{V}\left[\hat{f}_h(x)\right] = n^{-1}\int_{\mathcal{D}} h^{-2}\left\{K\left(\frac{x-y^2}{h}\right)\right\} f(y)dy$$

$$-n^{-1}(f(x)+B(x))^2$$

Using the Taylor series approximation as before, (and omitting the terms in n^{-1} and h) we have:

proof



dSEAS dipartimento di scienze economiche aziendali e statistiche department of economics business and statistics

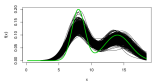
proof (???)

$$\text{var}\left[\hat{f}_h(x)\right] \approx n^{-1}h^{-1}f(x)\int_{\mathcal{D}*} K(t)^2 dt$$

and then (integrating with respect to x)

$$\int_{\mathcal{D}} \text{V}\left[\hat{f}_h(x)dx\right] \approx \frac{1}{n\ h}\int_{\mathcal{D}*} K^2(t)dt$$

Therefore you can see the trade-off between bias and variance in the choice of h.



dSEAS dipartimento di scienze economiche aziendali e statistiche department of economics business and statistics

Asymptotic results

In order to minimize the MISE(h) with respect to h (or better, its asymptotic expression AMISE):

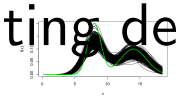
$$\text{AMISE}(h) = \frac{1}{4}h^4 k_2^2 \int_{\mathcal{D}} f''(x)^2 dx + n^{-1}h^{-1} \int_{\mathcal{D}*} K(t)^2 dt$$

differentiating with respect to h and equating to zero ($\frac{\partial AMISE(h)}{\partial h} = 0$),we have, :

$$h_{opt} = k_2^{-2/5} \left\{ \int_{\mathcal{D}*} K(t)^2 dt \right\}^{1/5} \left\{ \int_{\mathcal{D}} f''(x)^2 dx \right\}^{-1/5} n^{-1/5}$$

We observe that:

- the ideal value of h converges to zero with increasing n but with a very slow rate;
- since $\int_{\mathcal{D}} f''(x)^2 dx$ measures the quickness of the fluctuation of the density $f(x)$, smaller values of h will produce more fluctuating density estimates.



Minimization of MISE (asymptotically)

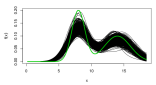
Substituting the value of h_{opt} to the expression of the MISE in correspondence of the optimal value of h, we have:

$$\text{AMISE}(h_{opt}) = \frac{5}{4} C(K) \left\{ \int_{\mathcal{D}} f''(x) dx \right\}^{1/5} n^{-4/5}$$

with

$$C(K) = k_2^{2/5} \left\{ \int_{\mathcal{D}*} K(t)^2 dt \right\}^{4/5}$$

Therefore $K(\cdot)$ should be chosen in order to return a small value of $C(K)$, because in this case it would have a small MISE (if we were able to choose h correctly!).



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Order of infinitesimals in the expression of AMISE

Order of infinitesimal of h

$$h(n)_{opt} = O\left(n^{-1/5}\right)$$

$h(n)$ decreases, as n diverges, but slowly

Substituting this value in AMISE we have:

AMISE (Asymptotic MISE)

$$AMISE[\hat{f}(x); h(n)] = O\left(n^{-4/5}\right)$$

In the regular parametric case we have:

$$V[\hat{f}(x); \hat{\theta}] = O\left(n^{-1}\right)$$

for example for the classical sample mean: $V(M) = \frac{\sigma^2}{n}$

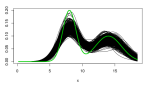
asymptotic results : optimal $K(\cdot)$

We need to minimize the value of $C(K)$ (or $\int K(t)^2 dt$) under the constraints $\int_{\mathcal{D}} K(t)dt = 1$ and $\int_{\mathcal{D}} t^2 K(t)dt = 1$ (assuming $k_2 = 1$). It is shown that the solution to this problem is given by:

$$K_e(t) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}t^2\right) & -\sqrt{5} \leq t \leq \sqrt{5} \\ 0 & \text{otherwise} \end{cases}.$$

$K_e(t)$ is known as the Epanechnikov kernel .
We can talk about efficiency of each symmetric kernel with respect to $K_e()$:

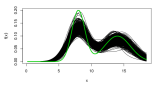
$$eff(K) = \left\{ \frac{C(K_e)}{C(K)} \right\}^{5/4}$$



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Rules to choose h

- Exploiting previous relationships we can get some *rules* (or better some *guidelines*) for the choice of h , for a given set of observed data.
- It is clear that the properties of the estimator depend (as already seen) on the characteristics of the true density function $f(x)$ in the whole range of x,
- So it is *unlikely* that we can talk about *optimal* rules in absolute sense



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Silverman Rule

The Silverman rule

Silverman rule (assuming that $f(x)$ is normal)

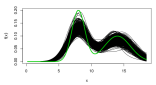
$$h = \left(\frac{4}{3n}\right)^{\frac{1}{5}} \sigma x$$

(in R σ_X is estimated with a robust estimator to avoid over-smoothing due to some observations far away from the mass of data)

The rule of Silverman

The use of Silverman rule, in absence of other informations, gives a satisfactory approximation as an initial estimate of $f(x)$

very useful in applications



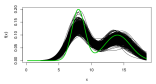
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

iterative techniques

- Since, as we have seen , the value of h that minimizes **integrated** *mean square error* is asymptotically given by:

$$h_{opt} = n^{-1/5} \left(A(K) n \int_{\mathcal{D}} f''(x) dx \right)^{1/5}$$

we can iteratively estimate h by giving an initial estimate of f and then using it to evaluate the curvature of $f(\cdot)$ in the integral above , and replacing iteratively .



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

And the likelihood (to estimate h) ?

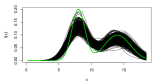
- Likelihood ?
- why not?
- Let us try to maximize the likelihood ...

$$L(h) = \prod \hat{f}(x_i, h))$$

construction of the likelihood

$$L(h; x_1, \ldots, x_n) = \prod_{i=1}^{n} \hat{f}(x_i; h) = \prod_{i=1}^{n} \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h} K(\frac{x_i - x_j}{h})$$

`recall codes in R`



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Likelihood with two values of h

Density with two values of h (see the properties of $\hat{f}(x, h)$ when $h \rightarrow 0$ and according to $x = x_i$ or not)

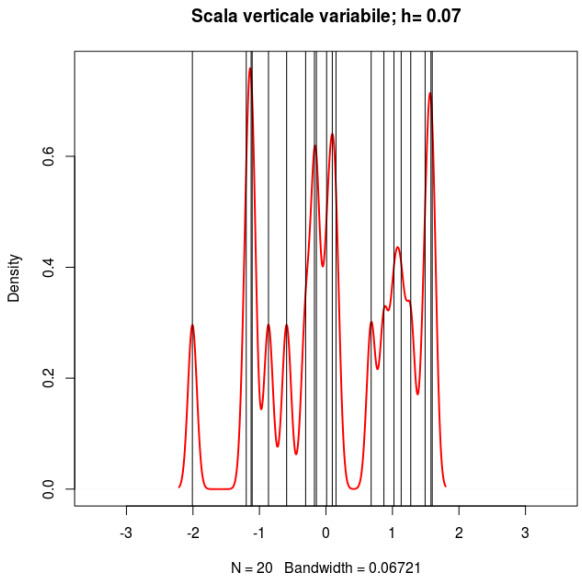


Figure: Likelihood computed on a sample

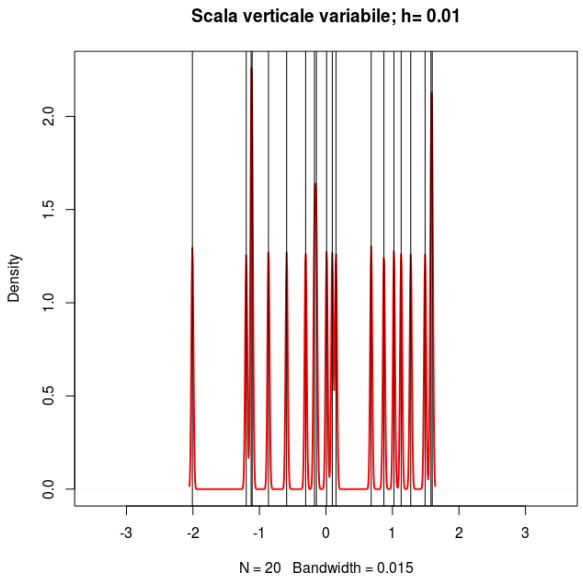
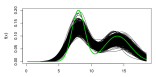


Figure: Likelihood computed on a sample



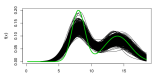
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Likelihood ?

You would get an estimator of zero width that focuses the whole mass on the observed data !
(on the other hand , *only* according to the observed data , and without parametric models, the most *probable* are the same data !)

$$L(h; x_1, \ldots, x_n) = \prod_{i=1}^{n} \hat{f}(x_i; h)) = \prod_{i=1}^{n} \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h} K(\frac{x_i - x_j}{h})$$

In the likelihood the problem is when $i = j$ (these are the terms that lead to divergence)
what can we do? **try to remove them ....**
and define a new criterion. We can therefore maximize a CV likelihood (Cross Validation)



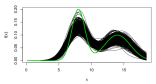
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Cross Validation

We define $\hat{f}_{-i}(x_i)$ as an estimator of $f(x_i)$ based on $n-1$ sample values obtained by excluding x_i

$$\hat{f}_{-i}(x_i; h) = \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} \frac{1}{h} K\left(\frac{x_i - x_j}{h}\right)$$

$\hat{f}_{-i}(x_i; h)$: an estimate of density at x_i **obtained without using x_i!!!**



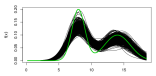
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Likelihood Cross Validation

- We build a likelihood function based on the values $\hat{f}_{-i}(x_i, h)$ defined:

$$L(h, x_1, \ldots, x_n) = \prod_{i=1}^{n} \hat{f}_{-i}(x_i, h) = \prod_{i=1}^{n} \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} \frac{1}{h} K\left(\frac{x_i - x_j}{h}\right)$$

- **We deleted the terms that lead to the divergence of the likelihood**
- Since one can not directly maximize the likelihood (because we would get a degenerated estimator that focuses the entire mass on the observed data) we can maximize a CV likelihood (Cross Validation).



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Cross validation density estimation

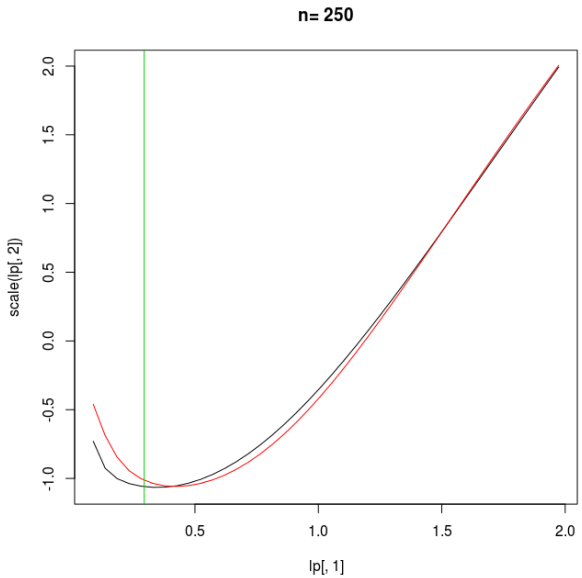


Figure: CV function with two methods

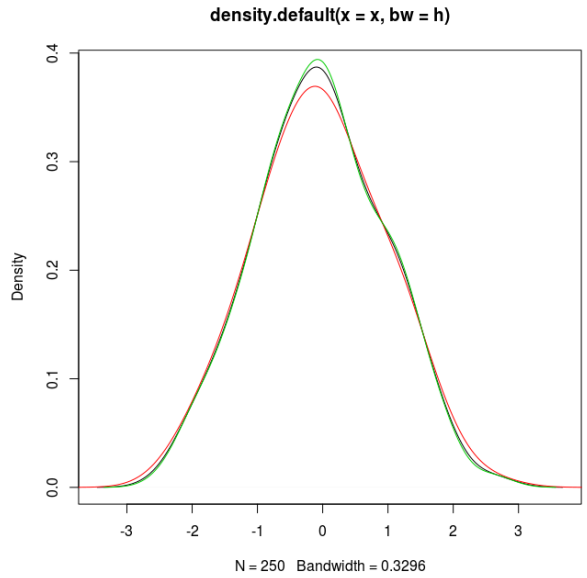
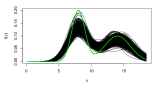


Figure: Estimated densities

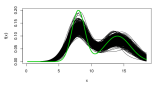


dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Cross Validation

In compact form , and using logarithms, we determine h such that:

$$\hat{h} : \max_{h} \sum_{i=1}^{n} log(\hat{f}_{-i}(x_i, h))$$



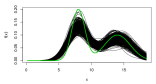
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Maximum Likelihood Cross Validation

There are two forms of cross-validation : **maximum likelihood CV** and **least- squares CV** .
In the estimation of the integrated squared error:

$$\int_{\mathcal{D}} \left\{ \hat{f}(x) - f(x) \right\}^2 dx$$

some quantities involving $\hat{f}(x)$ are estimated by using the estimate $\hat{f}_{-i}(x)$ non-parametric estimate obtained deleting the i- th observation



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Least squares CV

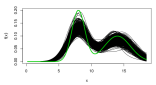
The least squares method measures the distance between f and $\hat{f}$ by the **Integrated Squared Error (ISE)** .

$$d_I(h) = \int (\hat{f}_h - f)^2(x)dx$$

$$= \int \hat{f}_h^2(x)dx - 2\int (\hat{f}_h f)(x)dx + \int f^2(x)dx$$

Note that in this formula

- $\int \hat{f}_h^2(x)dx$ is calculated from the data
- $\int f^2(x)dx$ does not depend on h
- $\int (\hat{f}_h f)(x)dx$ must be estimated from the data.



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Cross Validation

Subtracting the constant term , the minimization of the ISE corresponds to minimization of the quantity

$$d_I(h) - \int f^2(x)dx = \int \hat{f}_h^2(x)dx - 2\int (\hat{f}_h f)(x)dx$$

We can express the last term as an expected value

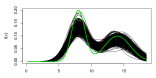
$$\int (\hat{f}_h f)(x)dx = E_X[\hat{f}_h(x)]$$

We can estimate this term:

$$E_X[\hat{f}_h(X)] = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{-i}(x_i; h)$$

Therefore, using this estimate we can obtain a *good* value of h minimizing:

$$\textbf{CV (h)} = \int \hat{f}_h^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(x_i, h)$$

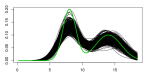


dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

with respect to h

Cross Validation property

- Recall of asymptotic properties of the of Least Squares Cross Validation method (ISE (CV) tends ISE with optimal h)
- Setting the CV likelihood from a hypothetical additional Y observation (*example on the board*)
- matches the Kullback - Leibler distance
- Recall the asymptotic properties of the Likelihood Cross Validation method (not many asymptotic properties , sensitivity to outliers , eg with $|X_1 - X_2| = R$ and then $\hat{f}_1(X_1) = 0$ if $h < R$ and then ...)
- We need stronger conditions on $f(x)$ and $K(\cdot)$ (support and behavior in the tails)



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

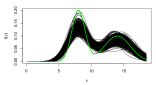
Estimation of a multivariate density function

Multivariate density estimation

- The problem substantially does not change **but only at a first sight!!** , compared to the univariate case .
- In fact we will soon see that the similarities are only technical, but there are **some substantial differences and complications**
- Actually it does not only concern the case of the of density or intensity estimation, but generally, the problem of multivariate data analysis

Special features of the multivariate case

- Definition of the concept of distance between points in $\mathcal{R}_d$
- Filling a region of $\mathcal{R}_d$, in particular when d is not small



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Density of points (uniform)

R code from: dimensions.R ./dimensions.R

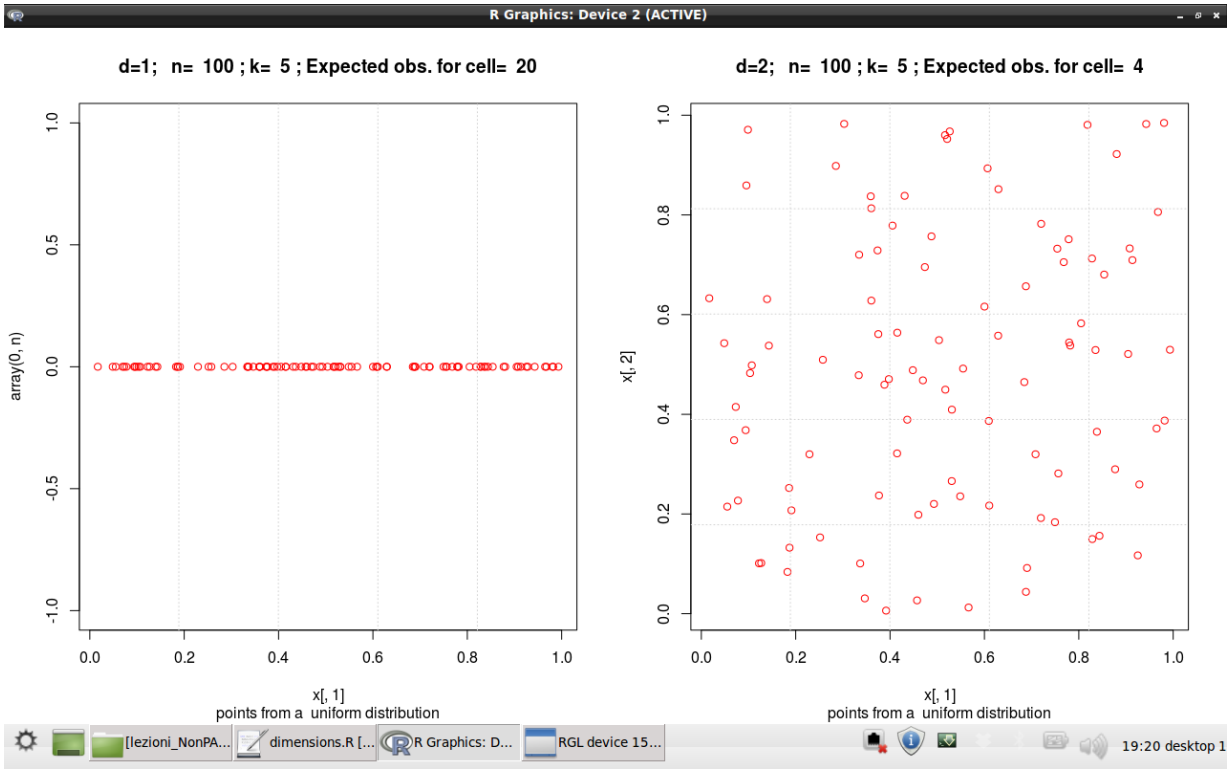
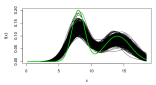


Figure: Sample of size 100 from a uniform distribution, d=1 and d=2



Density of points (uniform)

`R code from: dimensions.R`

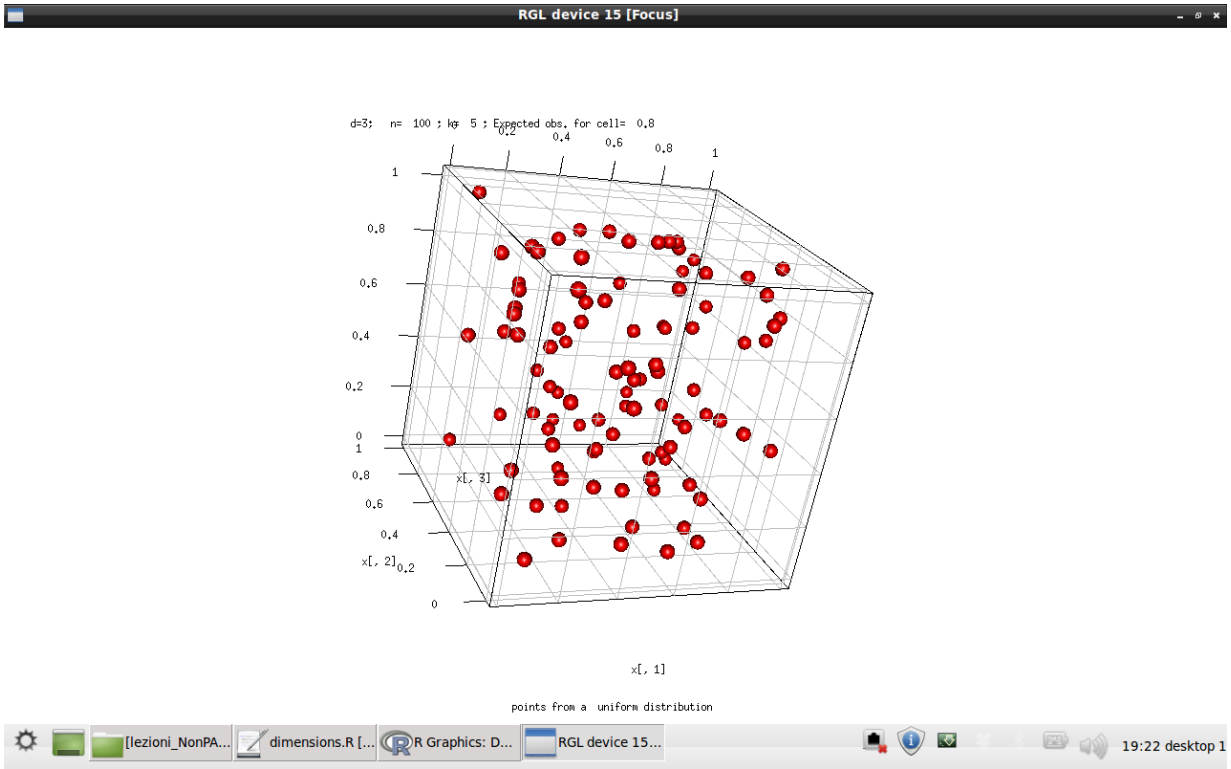
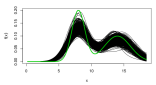


Figure: Sample of size 100 from a trivariate uniform distribution, d=3



Density of points (normal)

`R code from: dimensions.R`

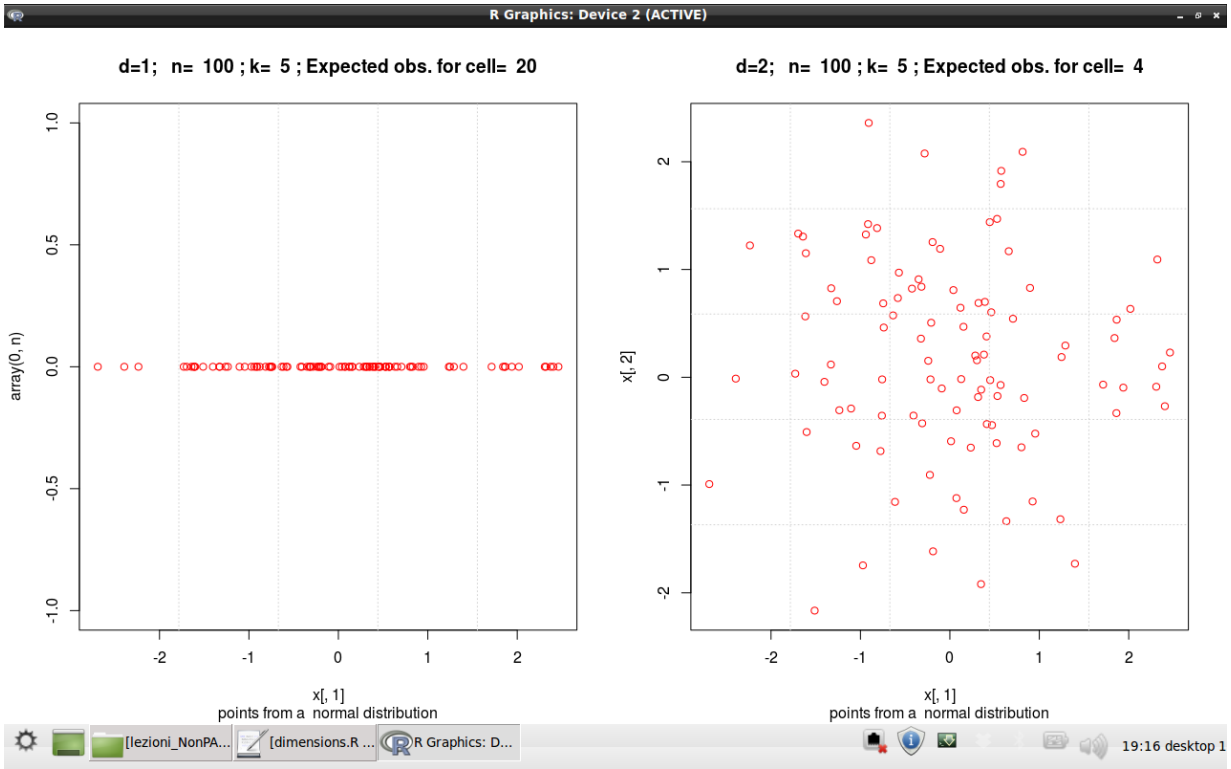
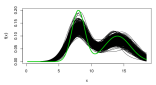


Figure: Sample of size 100 from a normal distribution, d=1 and d=2



Density of points (normal)

`R code from: dimensions.R`

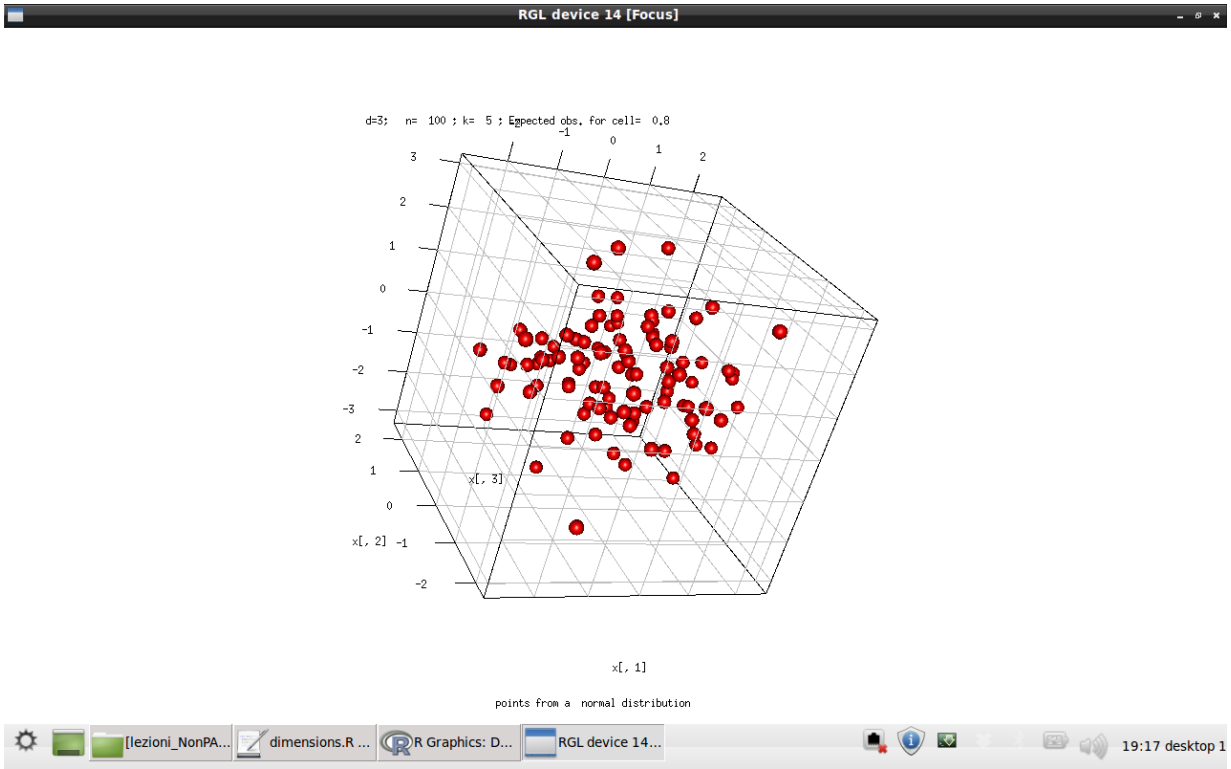
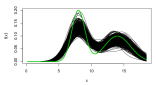


Figure: Sample of size 100 from a trivariate normal distribution, d=3



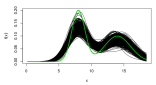
Density of points (normal)

`R code from: dimensions.R`

Some new concept

New concepts passing fromm 1 to d dimensions:

- Sparsity
- Curse of Dimensionality
- Fifferent definitions of distance



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

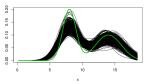
Estimation of a multivariate density function

We start with the bivariate case , generalizing , as long as possible, the approach used in the univariate case .
The definition of a bivariate density function can be obtained starting from the distribution function , which is given by:

$$F_{X_1,X_2}(x_1,x_2) = \text{Prob}\left\{X_1 \leq x_1 \cap X_2 \leq x_2\right\}$$

if $F(x)$ is differentiable then

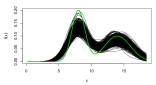
$$f_{X_1,X_2}(x_1,x_2) = \frac{\partial^2 F_{X_1,X_2}(x_1,x_2)}{\partial x_1 \partial x_2}$$



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Difficulties in the definition of $\hat{f}(x_1, x_2)$

- There are the same conceptual difficulties in the definition of an estimator for a density function of a bivariate variable as in the case of simple variables. Namely the conceptual difference between the theoretical density function and the distribution function for a random variable (simple or multiple) .
- While the distribution function of a $d-$variate random variable is always defined as the probability that $\mathbf{X}$ fall in a certain intersection of open intervals (and the equivalent empirical is easy to construct) , the density is always defined as the derivative (of order d) of that function.
- Once again, this difficulty concerns with the difficulty of defining the probability for a continuous variable; this probability is defined for regions and not for points.



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

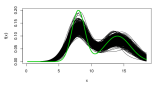
bivariate density estimation : approximations

- Let's try again to approximate the bivariate density at a point; let us return to the approximation of the probability in a small region around (x_1, x_2):
-

$$f(x_1, x_2) \approx$$

$$\frac{\text{Prob}\left\{x_1 - \frac{\delta_1}{2} < X_1 \leq x_1 + \frac{\delta_1}{2} \bigcap x_2 - \frac{\delta_2}{2} < X_2 \leq x_2 + \frac{\delta_2}{2}\right\}}{\delta_1 \delta_2}$$

The situation is altered by the fact that we have regions and not intervals in $\mathcal{R}_2$: and it will be even worse in $\mathcal{R}_d$

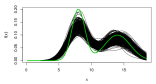


Approximation of the incremental ratio

Again a first approoximation is given by the histogram

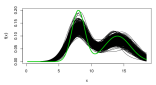
which provides discontinuous estimate of the density

Furthermore we have to divide a region into rectangles (not a straight line into segments)



approximation of the incremental ratio as average of contributions

In the univariate case every point x_i contributes to the estimate of f if it is far from x less than h, and we write $\hat{f}(x)$ as the arithmetic mean of the contribution of each point .
Now every point $\mathbf{x}_i$ (bivariate case , but even d - variate) contributes to estimate whether each component (x_{ij}) is far from x_j less than $h_j (j = 1, 2, \ldots, d)$.
But the concept of distance can have a wide definition in d dimensions.



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

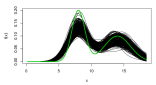
Approximation of the as average of contributions

bivariate case

approximation of the incremental ratio as average of the contributions of the individual observations

$$\hat{f}(x_1, x_2) = \frac{1}{4nh_1h_2} \sum_{i=1}^{n} I\left(x_1 - h_1 < x_{i1} \leq x_1 + h_1 \bigcap x_2 - h_2 < x_{i2} \leq x_2 + h_2\right)$$

This is an intuitive extension of the approximation of the incremental ratio to the $d-$variate case :
$I()$ is the indicator function.



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Average of uniform densities

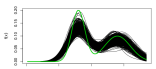
We can express all as a function of a bivariate standardized uniform density (and then $d-$ variate) in a region bounded by the intervals (-1 , 1), collecting the terms appropriately :

$$\hat{f}_{X_1,X_2}(x_1,x_2) = \frac{1}{h_1 h_2 \ n} \sum_{i=1}^{n} W\left(\frac{x_1 - x_{i1}}{h_1}, \frac{x_2 - x_{i2}}{h_2}\right) \qquad (11)$$

where we define:

$$W(z) = \begin{cases} \frac{1}{2^2} & (|z_1| \leq 1) \cap (|z_2| \leq 1) \\ 0 & \text{elsewhere} \end{cases}$$

From here we can achieve new generalizations of kernel estimators that avoid the steps in the estimate of $\hat{f}(x_1, x_2)$ obtained from (11)



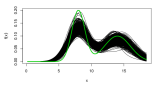
bivariate kernel estimator

To obtain a kernel estimator of a bivariate density we can replace the uniform density in (11) with another generic bivariate density function $K(\cdot,\cdot)$

Generalization of univariate kernel

$$\hat{f}_{X_1,X_2}(x_1,x_2) = \frac{1}{h_1 h_2\ n}\sum_{i=1}^{n} K\left(\frac{x_1 - x_{i1}}{h_1}, \frac{x_2 - x_{i2}}{h_2}\right) \tag{12}$$

- the function $K(\cdot,\cdot)$ is a bivariate kernel function, in standardized variables (zero mean and unit variance , or at least constant)
- h_1 and h_2 are two bandwidths, or smoothing parameters (one for each dimension)
- $\frac{1}{h_1}\frac{1}{h_2}K\left(\frac{x_1-x_{i1}}{h_1}, \frac{x_2-x_{i2}}{h_2}\right)$ is the weight of each observation in the computation of $\hat{f}(x_1,x_2)$.



dSEAS dipartimento di scienze economiche aziendali e statistiche department of economics business and statistics

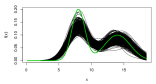
Bivariate kernel estimator

Each point *spreads its influence* as a function of $K(\cdot,\cdot)$ and (h_1, h_2)

- $K(\cdot,\cdot)$ is a bivariate density function centered on zero and with standardized components **(but not necessarily independent!)**
- If **the kernel function has independent components** $K_1(\cdot)$ **and** $K_2(\cdot)$, we have:

$$\hat{f}_{X_1,X_2}(x_1,x_2) = \frac{1}{h_1 h_2\ n}\sum_{i=1}^{n} K_1\left(\frac{x_1 - x_{i1}}{h_1}\right) K_2\left(\frac{x_2 - x_{i2}}{h_2}\right) \qquad (13)$$

The most frequent example for $K_j(\cdot)(j = 1, 2)$ is the density of a standard normal distribution.



Two estimates (with a bivariate normal kernel independent components) on the same sample , with different h

`./panel_kernel_biv2.R`

`interactive execution with slider`

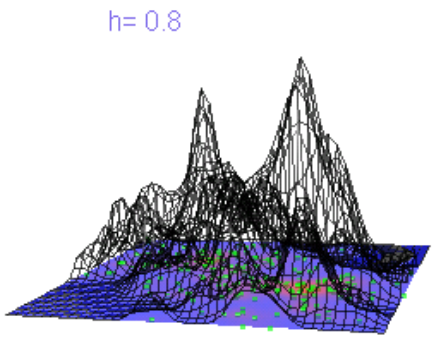


Figure: Sample of size 200

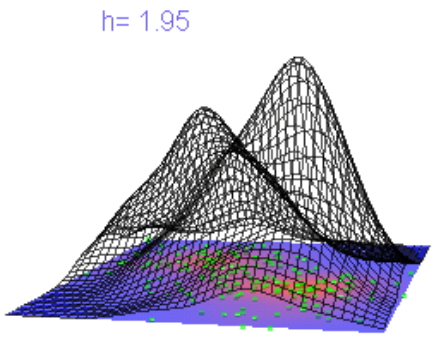
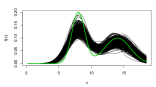


Figure: Sample of size 200



Density plot (with a normal bivariate kernel with independent components), with different h

`interactive execution with slider and rotations`

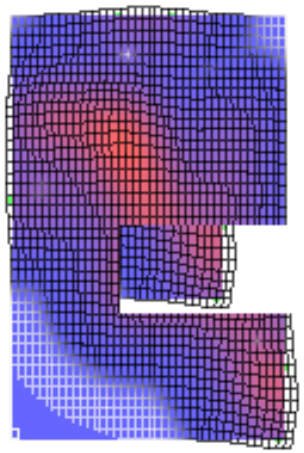


Figure: Sample of size 200

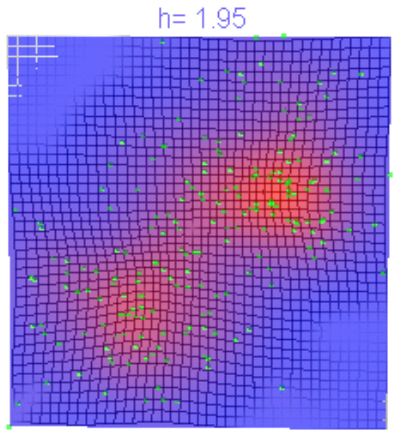
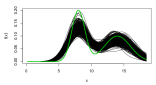
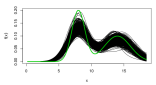


Figure: Sample of size 200



bivariate kernel estimator and distance of the points in $\mathcal{R}_2$

- Each observed point $\mathbf{x}_i$ *contributes to the estimated density* in the region that surrounds its influence as a function of $K(\cdot,\cdot)$ and (h_1, h_2)
- For each coordinate, observations $\mathbf{x}_i$ closer to $\mathbf{x}$ weight more. Larger h smooth the weight of closer observation and spread the influence of each observed x_i over a wider range.
- But what do we mean by **close** in multivariate context ?
- In the univariate case , in $\mathcal{R}_1$ the distance is unquestionably measured by the segment $x - x_1$. **but how to measure it in $\mathcal{R}_2$ and , in general, in $\mathcal{R}_d$?**



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Bivariate kernel estimator and distance of points in $\mathcal{R}_2$

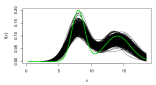
- Everything seems simple and obvious turning to the two-dimensional case , but already in the two previous examples it is not.
- Let us limit for now to a bivariate kernel with independent components and consider the bivariate uniform kernel on rectangular domains:

$$\hat{f}_{X_1,X_2}(x_1,x_2) = \frac{1}{4h_1h_2\ n}\sum_{i=1}^{n} \mathsf{I}\left(\left|\frac{x_1 - x_{i1}}{h_1}\right| \leq 1\right) \bigcap \left(\left|\frac{x_2 - x_{i2}}{h_2}\right| \leq 1\right) \quad (14)$$

and a bivariate normal kernel (independent components) :

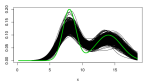
$$\hat{f}_{X_1,X_2}(x_1,x_2) = \frac{1}{h_1h_2\ n}\sum_{i=1}^{n} \phi\left(\frac{x_1 - x_{i1}}{h_1}\right) \phi\left(\frac{x_2 - x_{i2}}{h_2}\right) \quad (15)$$

where $\phi(\cdot)$ is the density of a standard normal



Bivariate kernel estimator and distance of points in $\mathcal{R}_2$

- It 's clear that in the uniform kernel (14) points $\mathbf{x}_i$ give the same contribution on a rectangle area
- But in a normal kernel with independent components (15) points $\mathbf{x}_i$ gives the same contribution on an ellipse (centered at $\mathbf{x}_i$ and with axes proportional to the h_i and *parallel* to the Cartesian axes)
- Similar considerations can be done in the multivariate case in $\mathcal{R}_d$ (with hyper- parallelepipeds and ellipsoids)



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Two very different examples of bivariate pattern of points

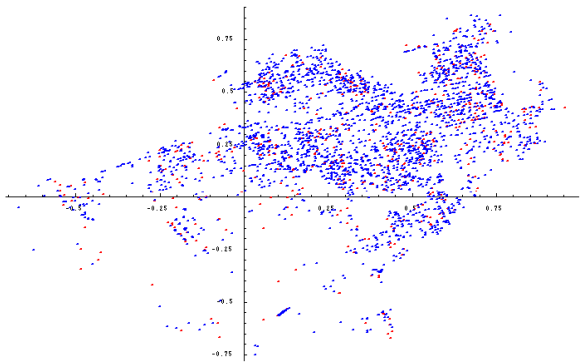


Figure: Spatial distribution of cases and controls

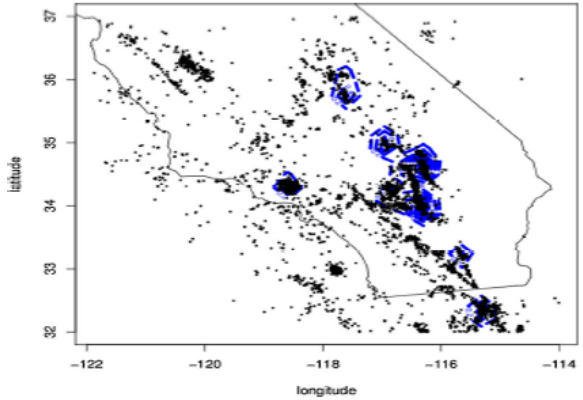
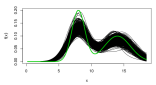
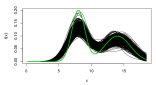


Figure: Spatial distribution of seismic events (California)



Bivariate kernel estimator with correlated components

- **If we leave the case of kernel with independent components, we could think about a bivariate normal kernel with correlated components**
- **In this case, x_i gives the same contribution to the points of the ellipse centered at x_i but with axes not parallel to the Cartesian axes** . In the latter ' case, the ellipses have eccentricity determined by ρ



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

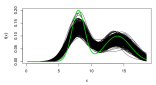
Bivariate kernel estimator with correlated components

- Density estimation based on a bivariate normal kernel (components with correlation ρ) :

$$\hat{f}_{X_1,X_2}(x_1,x_2) = \frac{1}{h_1 h_2\ n} \sum_{i=1}^{n} \phi_{\text{biv}}\left(\frac{x_1 - x_{i1}}{h_1}, \frac{x_2 - x_{i2}}{h_2}, \rho\right) \tag{16}$$

$$\phi_{\text{biv}}(z_1,z_2,\rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[z_1^2 - 2\rho z_1 z_2 + z_2^2\right]}$$

is the density of a bivariate normal with correlated standardized components.

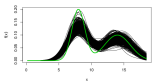


Multivariate Kernels

In general, if $\mathbf{K}(\cdot)$ is a Multivariate kernel function (components with zero mean and standardized dependent or independent), in compact form the $d-$variate kernel estimator is given by:

multivariate kernel estimator

$$\hat{f}_{\mathbf{X}}(\mathbf{x}) = \frac{1}{n\sqrt{|\mathbf{H}|}} \sum_{i=1}^{n} \mathbf{K}(\mathbf{x} - \mathbf{x}_i, \mathbf{H}) \tag{17}$$



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

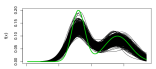
multivariate normal Kernel

A natural and comfortable choice for the kernel $\mathbf{K}(\cdot)$ is a *multivariate normal density function*

$$\mathbf{K}(\mathbf{x}) = \|\mathbf{\Sigma}\|^{-\frac{1}{2}}(2\pi)^{-d/2}\exp\left\{-\frac{1}{2}[\mathbf{x}-\boldsymbol{\mu}]^{\mathrm{T}}\mathbf{\Sigma}^{-1}[\mathbf{x}-\boldsymbol{\mu}]\right\} \qquad (18)$$

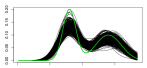
where μ is a d component vector of mean values
$\mathbf{\Sigma}$ is a $d \times d$ variance covariance matrix

For brevity of notation $\mathbf{H}$ is a matrix of smoothing parameters ($\mathbf{\Sigma}$ if we use a multivariate normal kernel), which generally includes on the main diagonal the terms related to the parameters of scale and off-diagonal terms that measure the correlations (it will be a matrix of variances and covariances for the multivariate normal kernel, or correlation for normal multivariate with standardized components)



Kernel multivariate independent components

- A multivariate kernel can be **with independent components** (isotropic kernel)
- The contribution of each point (with standardized coordinates) $\frac{x_j - x_{ij}}{h_j}$ is constant for points $\mathbf{x}$ of hypersphere centered on $\mathbf{x}_i$
- and in ellipsoids with axes parallel to the coordinate axes if the coordinates are not standardized .
- We can express the estimator with kernel with independent components in relation to the univariate kernel functions (with d scale parameters h_j)



multivariate Kernel with independent components

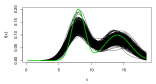
$$\hat{f}_{\mathbf{X}}(\mathbf{x}) = \frac{1}{n\prod_{j=1}^{d} h_j} \sum_{i=1}^{n} \prod_{j=1}^{d} K\left(\frac{x_j - x_{ij}}{h_j}\right) \tag{19}$$

$$\mathbf{x}^{\mathrm{T}} = \{x_1, x_2, \ldots, x_j, \ldots, x_d\}$$

given n $d-$variate observations $\mathbf{x}_i$, $i = 1, 2, \ldots, n$

$$\mathbf{x}_i^{\mathrm{T}} = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{id}\}$$

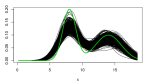
- evidently (19) is a special case of (17) when the **H** matrix is diagonal .
  (and with the notation adopted here $h_j = \sqrt{h_{jj}}$)



dSEAS dipartimento di scienze economiche aziendali e statistiche department of economics business and statistics

Multivariate Kernel with correlated components

- If the symmetric matrix **H** is not diagonal , we have a kernel estimator with correlated components.
- The contribution of each point is constant for the points **x** of an ellipsoid with axes not parallel to the coordinate axes .
- It is sometimes referred to as *anisotropic kernel* but it is not a general terminology
- In general we can also fix some non-diagonal components equal to zero , and estimating the other .



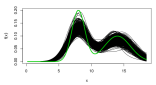
dSEAS dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Multivariate kernel with correlated components : example

`example from my paper`

anisotropic space-time kernel (anisotropic in the *xyz* space) versus an isotropic space-time kernel :

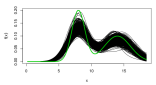
$$\mathbf{h} = \begin{pmatrix} \mathbf{h_x} & \mathbf{h_{xy}} & \mathbf{h_{xz}} & \mathbf{0} \\ \mathbf{h_{xy}} & \mathbf{h_y} & \mathbf{h_{yz}} & \mathbf{0} \\ \mathbf{h_{xz}} & \mathbf{h_{yz}} & \mathbf{h_z} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{h_t} \end{pmatrix} \text{ or } \mathbf{h} = \begin{pmatrix} \mathbf{h_x} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{h_y} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{h_z} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{h_t} \end{pmatrix}$$



dSEAS dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

kernel multivariate variables bands (just a hint)

- The concept of variable bands , or **Adaptive kernel** now must be understood in a multivariate sense , and concerns the entire **H** matrix.
- A direct extension of the univariate concept regards the possibility of working with bands of varying widths for each coordinate
- however, you may think you have correlation structures varying as a function of sampling points:
- $$\hat{f}_{\mathbf{X}}(\mathbf{x}) = \frac{1}{n\sqrt{|\mathbf{H}|}} \sum_{i=1}^{n} \mathbf{K}(\mathbf{x} - \mathbf{x}_i, \mathbf{H}_i)$$
- A very effective adaptive method is the **nearest neighbor**



dSEAS dipartimento di scienze economiche aziendali e statistiche department of economics business and statistics

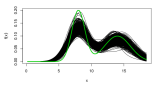
Dimensionality

So far we have seen examples with small values of d (2,3,4). What happens when d is high ?

Increase in the number of dimensions

It might seem that increasing the number of dimensions d everything proceed in the same way , but ...

example



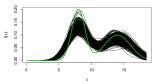
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

The Curse of dimensionality

The Curse of dimensionality

- As the number of dimensions d increases, and using equal width intervals , the number of cells that divide the region Ω_d of $\mathcal{R}^d$ grows exponentially with d .
- And the set of n points fills less and less significant areas of the region d -dimensional

example



examples

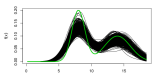
In a multivariate normal distribution, the quadratic form in the numerator follows a χ^2 distribution with d degrees of freedom

$$\frac{f((X))}{f((0))} \sim \exp(-\frac{1}{2}\chi^2_d)$$

considerations on the density and the fill volume of observation

```
round ( pchisq ( 1,1:10 ), 3)
[1] 0.683 0.393 0.199 0.090 0.037 0.014 0.005 0.002 0.001 0.0
```

For $d = 1, 2, \ldots, 10$ `pchisq (1, d )` represents the probability that a $d-$variate normal observation has a norm less than unity (part of a hyper-sphere of unit radius)



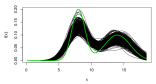
examples

```
round ( exp (- qchisq ( .5,1:10 ) / 2 ), 3)
[1] 0.797 0.500 0.306 0.187 0.114 0.069 0.042 0.025 0.015 0.0
```

For $d = 1, 2, \ldots, 10$ `exp(- qchisq (.5 , d) / 2 )` represents the relative density of a *median* quadratic form
For example, with 5 variables ($d = 5$) 50 per cent of the observations on average should have a density of 11% if compared to the maximum theoretical density !

Density in d - dimensions

Increasing the number of dimensions, this median density decreases dramatically



Hypersphere Volume

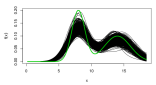
density of the hyper - sphere
if d is equal to the volume of hypersphere is:

$$Vol(S_d) = \frac{\pi^{d/2} r^d}{d!}$$

while the volume of the circumscribed hypercube :

$$Vol(C_D) = (2r)^d$$

Clearly their ratio tends to diverge from zero to d !



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Asymptotic results for $d-$variates kernels

optimum $h_d(n)$

$$h_d(n)_{opt} = O\left(n^{-\frac{1}{d+4}}\right)$$

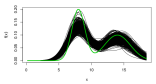
$h_d(n)$ decreases as n diverges, but slower for larger values for d

AMISE

$$AMISE[\hat{f}_d(x); h(n)] = O\left(n^{-\frac{4}{d+4}}\right)$$

While in the parametric case we have

$$AMISE[\hat{f}_d(x); \hat{\theta}] = O\left(n^{-1}\right)$$



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics