COS 513: Nonparametric Estimation

Lecturer: Philippe Rigollet Scribe: Lanhui Wang

Dec 9, 2010

1 What is Nonparametric Estimation?

1.1 Parametric Estimation: estimate parameters in $\mathbb{R}^k$

Examples:

1. $X_{1:n}$ i.i.d $\sim N_d(\mu,\Sigma)$ Estimating μ or Σ is a parametric problem.
2. $Y_i = X_i^T\beta + \sigma\epsilon_i$Esimating β or σ is a parametric problem.

1.2 Noparametric Esimation: estimate infinite dimensional parameters, typically functions.

Examples:

1. $X_{1:n}(X_i \in \mathbb{R}^k$ for $i = 1 : n)$ i.i.d $\sim$probability density function p. Estimating $p : \mathbb{R}^k \to$ R is a nonparametric problem.
2. $Y_i = f(X_i) + \sigma\epsilon_i$Esimating $f : \mathbb{R}^k \to$ R is a nonparametric problem.

2 Nonparametric Density Estimation

2.1 The histogram

$X_{1:n}$i.i.d with density $p(\bullet)$ on [0,1]. Let $B_{1:m}$be a partition of [0,1] called bins. Assume that all bins are intervals with length h. Estimate p by a constant value on each bin: for $x \in B_j$,

$$p(x) \quad \simeq \quad \frac{1}{|B_j|}\int_{B_j} p(x)dx \simeq \frac{1}{h}\mathbb{P}(X \in B_j) = \frac{1}{h}\mathbb{E}[\mathbb{I}(X \in B_j)]$$

To estimate, $\mathbb{E} \rightsquigarrow \frac{1}{n}\sum_{i=1}^n$, so here

$$\hat{p}(x) = \frac{1}{nh}\sum_{i=1}^n \mathbb{I}(X_i \in B_j) = \frac{n_j}{nh},\ x \in B_j$$

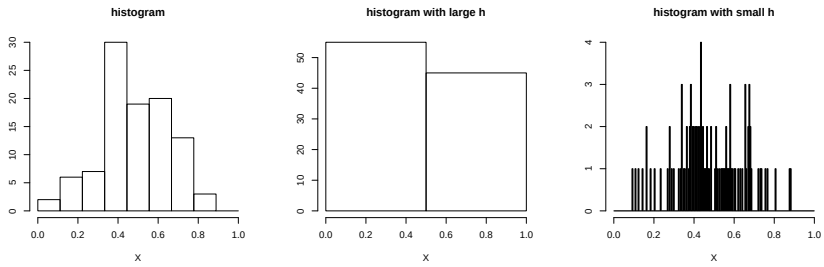


Figure 1:

where $n_j$is the number of ovservations in bin B_j. And full piecewise constant estimator is

$$\hat{p}_n(x) = \sum_{j=1}^{m} \frac{n_j}{nh} \mathbb{I}(x \in B_j)$$

called HISTOGRAM.

2.2 Kernel density estimator

Let $F(\bullet)$be the cdf of the sample, then

$$F(x) = \mathbb{P}(X \leq x) = \mathbb{E}[\mathbb{I}(X \leq x)]$$

To estimate, $\mathbb{E} \rightsquigarrow \frac{1}{n} \sum_{i=1}^{n}$, so we obtain the estimator

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq x)$$

called Empirical CDF.
According to Glivenko-Cantelli's theorem

$$\sup_x |\hat{F}_n(x) - F(x)| \to 0, n \to \infty, \text{ a.s.}$$

From F to p:

$$p(x) = \frac{dF(x)}{dx} \simeq \frac{F(x + h/2) - F(x - h/2)}{h}$$

for small $h > 0$.
And this yields estimator (Rosenblatt, 1956)

$$\hat{p}_n(x) = \frac{\hat{F}_n(x + h/2) - \hat{F}_n(x - h/2)}{h}$$

and its generic form

$$\hat{p}_n(x) = \frac{1}{nh}\sum_{i=1}^{n}\mathbb{I}(x - h/2 < X_i \leq x + h/2) = \frac{1}{nh}\sum_{i=1}^{n}\mathbb{I}(|\frac{X_i - x}{h}| \leq \frac{1}{2})$$

We have

$$\hat{p}_n(x) = \frac{1}{nh}\sum_{i=1}^{n}K(\frac{X_i - x}{h})$$

where

$$K(u) = \mathbb{I}(-1/2 < u \leq 1/2)$$

called rectangular kernel.

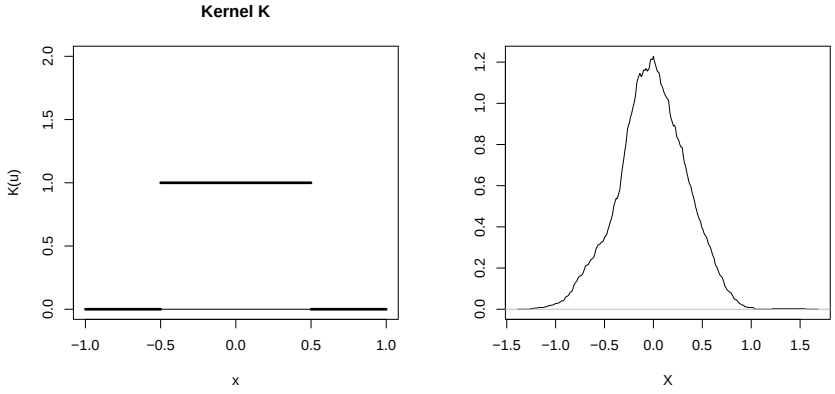


Figure 2:

Properties of K:$\int K = 1 \rightsquigarrow \int \hat{p}_n = 1$.
And according to Rosenblatt's theorem

$$\hat{p}_n(x) \to p(x), \forall x \text{ in probability}$$

This rectangular kernel can be generalized to other kernels:

Kernel density estimator (Parzen, 1962) We call kernel density estimator (or Parzen-Rosenblatt estimator) a nonparametric estimator of the form

$$\hat{p}_n(x) = \frac{1}{nh}\sum_{i=1}^{n}K(\frac{X_i - x}{h})$$

where $h > 0$ is called BANDWIDTH and $K : \mathbb{R} \to \mathbb{R}$ is any function such that

$$\int |K| < \infty, \int K = 1$$

and K is called KERNEL.
(h, K) are the two parameters of this esimator.

Examples of kernels:

- Rectangular: $K(u) = \mathbb{I}(|u| \leq 1/2)$ (figure 2)
- Triangular: $K(u) = (1 - |u|)_+$ (figure 3)
- Epanechnikov: $K(u) = \frac{3}{4}(1 - u^2)_+$ (figure 4)
- Gaussian: $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2})$ (figure 5)

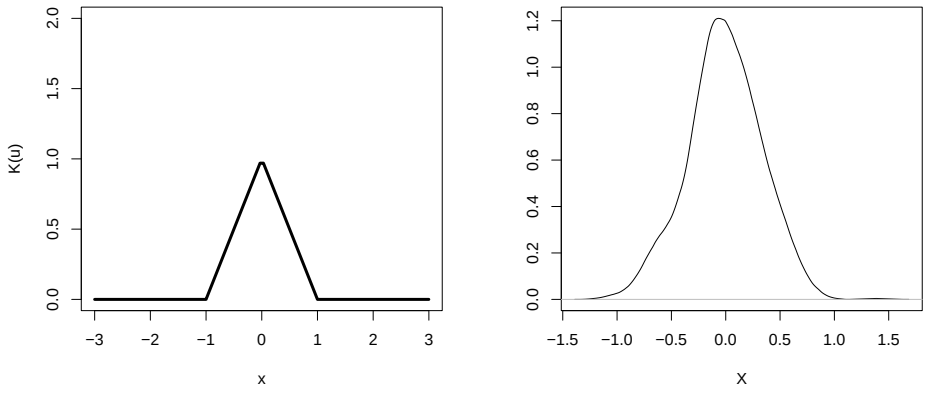


Figure 3:

Effect of bandwith The following figures are histogram with same and kernel density estimators with different h.

Gaussian kernel plot of $z \mapsto \frac{1}{h}K(\frac{z}{h})$

For $z = X_i - x = 1$

for exmaple: for $h = 1, 1/2,$ and $1/3$, $\frac{1}{h}K(\frac{X_i - x}{h}) = \frac{1}{h}K(\frac{1}{h}) = 0.24,\ 0.11,$ and 0.013 respectively. (See the following figures for $h = 1, 1/2, 1/3$)

Desirable properties of a kernel We typically want (but not necessary): $K(u) \geq 0$, and $K(u) = K(-u)$

See Silverman (1986): Density estimation for statistics and data analysis.

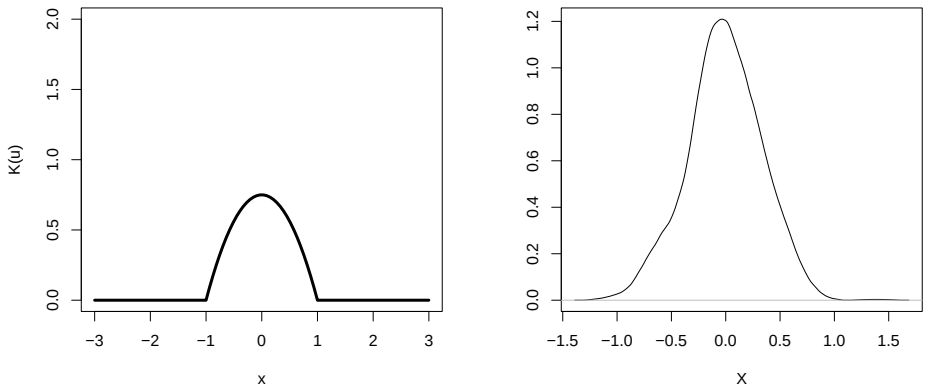


Figure 4:

Multivariate kernel density estimator Each observation $X_i = (X_{i1}, \ldots, X_{id}) \in \mathbb{R}^d$, $i = 1, \ldots, n$ with density $p(\bullet)$ on $\mathbb{R}^d$.

Product kernel $\mathbb{K} : \mathbb{R}^d \mapsto \mathbb{R}$,

$$\mathbb{K}(u) = \prod_{j=1}^{d} K(u_j), \ u = (u_1, \ldots, u_d)$$

where K is a kernel on $\mathbb{R}$.

Multivariate kernel density estimator

$$\hat{p}_n(x) = \frac{1}{nh^d} \sum_{i=1}^{n} \mathbb{K}(\frac{X_i - x}{h})$$

Curse of dimensionality Assume $n = 1000$, $d = 10$ and the rectangular kernel $K(u) = \mathbb{I}(|u| \leq .5)$, then the multivariate kernel is

$$\mathbb{K}(u) = \prod_{j=1}^{d} K(u_j) = \mathbb{I}(u \in [-0.5, 0.5]^d)$$

Say we want to estimate $p(0)$. How many points (in average) fall in the window when X has uniform distribution on $[0, 1]^d$?

We get

$$\mathbb{E}[\sum_{i=1}^{n} \mathbb{I}(X_i \in \text{window })] = n\mathbb{P}(X \in \text{window }) = nh^d$$

Assume that $h = 0.5$ (which is already very large), we obtain:

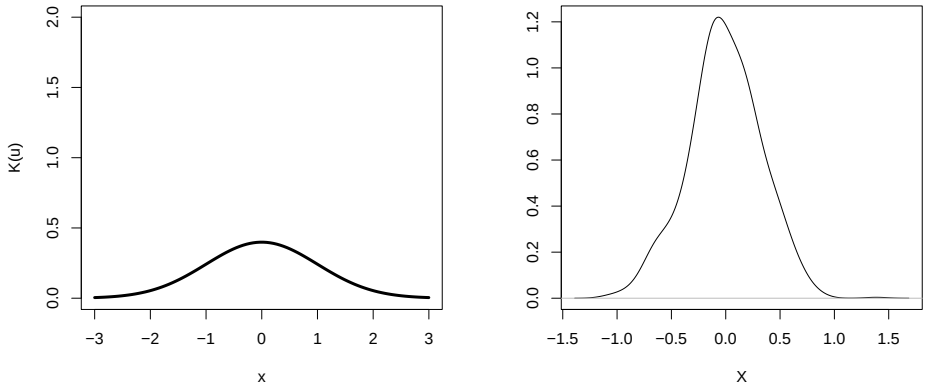


Figure 5:

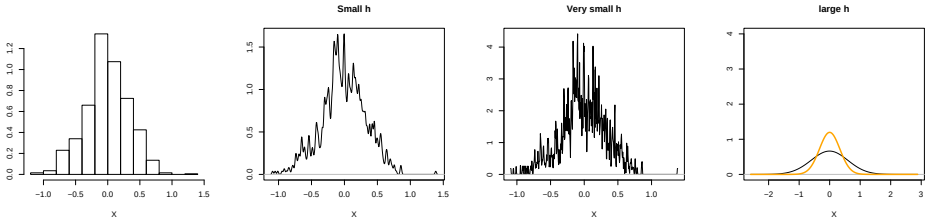


Figure 6:

$$\mathbb{E}\# = 1000 \bullet 2^{-10} = \frac{1000}{1024} < 1$$

not even one in average.

3 Nonparametric regression estimation

3.1 Conditional density

Let (X, Y) be such that

$$f(x) = \mathbb{E}[Y|X = x]$$

observations: n i.i.d copies of $(X, Y) : (X_1, Y_1), \ldots, (X_n, Y_n)$
$X_1, \ldots, X_n$ are called design or effects.
Two types: random design vs. fixed design (typically $X_i = i/n$).

Figure 7:

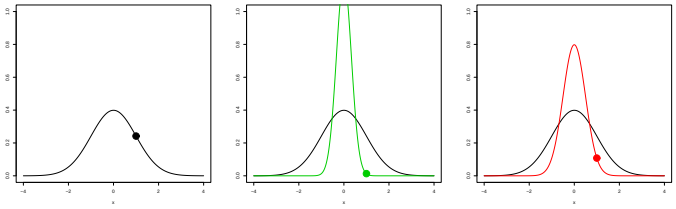
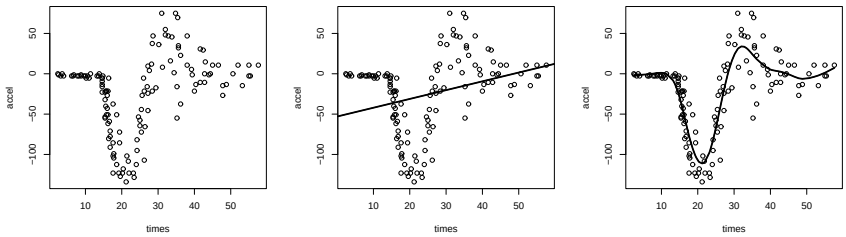


Figure 8:



Example: motorcycle data A sensor is placed on the helmet of a motorcyclist. We measure the acceleration at different times of a crash. The second and the third figure are fitting linear regression by least squares and nonparamteric estimation respectively.

Assume random design and that (X,Y)has joint density $p(x,y)$.

Regression function

$$f(x) = \mathbb{E}[Y|X=x] = \int yp(y|x)dy = \frac{\int yp(x,y)dy}{\int p(x,y)dy}$$

3.2 Nadaraya-Watson estimator

We can use bivariate kernel density estimator

$$\hat{p}_n(x,y) = \frac{1}{nh^2}\sum_{i=1}^{n} K(\frac{X_i - x}{h})K(\frac{Y_i - y}{h})$$

Replae $p(x,y)$ by $\hat{p}_n(x,y)$, we obtain

$$\hat{f}_n(x) = \frac{\int y\hat{p}_n(x,y)dy}{\int \hat{p}_n(x,y)dy}$$

Nadaraya-Watson estimator (proposed independently by Nadaraya (1964) and Watson (1964))

$$\hat{f}_n^{NW}(x) = \frac{\sum_{i=1}^n Y_i K(\frac{X_i - x}{h})}{\sum_{i=1}^n K(\frac{X_i - x}{h})}$$

if denominator $\neq 0$ and 0 otherwise.

Assume that $\int K = 1$ and $\int uK(u)du = 0$. then

$$\hat{f}_n^{NW}(x) = \hat{f}_n(x) = \frac{\int y\hat{p}_n(x,y)dy}{\int \hat{p}_n(x,y)dy}$$

3.3 Local polynomial estimator

Constant estimator Assume first that we have the following idea: find the best constant approximation to the data, which corresponds to fitting the best line with slope=0. We solve least squares: $\min_c \frac{1}{n}\sum_{i=1}^n (Y_i - c)^2$ and obtain c.

Locally constant estimator It is a bad idea to assume that accel is constant over time. But locally it is approximately true because the function is smooth.

For a given point x, to estimate $f(x)$, we downweight the X_i's that are far from x using a kernel

Local criterion

$$\min_c \frac{1}{n}\sum_{i=1}^n (Y_i - c)^2 K(\frac{X_i - x}{h})$$

$\hat{f}(x) = c(x)$ is the locally constant estimator.

Remark: $h \to 0$: all points are far away. $h \to \infty$: all points are at the same distance (back to constant estimator).

Fact: the locally constant estimator is the Nadaraya-Watson estimator.

Locally linear estimator We localize the least squares criterion for linear regression:

$$\min_{a,b} \frac{1}{n}\sum_{i=1}^n (Y_i - aX_i - b)^2 K(\frac{X_i - x}{h})$$

$\hat{f}(x) = b(x)$ is the locally linear estimator.

Locally polynomial estimator Even further, given $p \geq 0$:

$$\min_{a,b} \frac{1}{n}\sum_{i=1}^n (Y_i - (a_0 + a_1 X_i + \ldots + a_p X_i^p))^2 K(\frac{X_i - x}{h})$$

$\hat{f}(x) = a_0(x)$ is the locally polynomial estimator.

Local vs Global methods The above methods are local in the sense that for a given x, you can compute $\hat{f}(x)$. To be opposed to global methods that compute $\hat{f}(x)$ for all x at once. Each method has its + and -.

3.4 Projection methods

Basis of functions Assume that the design is random with uniform distribution.

Just like in linear algebra, there are bases of vectors, there exists bases of functions. They are typically infinite (infinite dimention).

Assume that $\varphi_1, \varphi_2, \ldots, \varphi_n$is such a bsis. In particular, we can consider the representation of f in this basis:

$$f(x) = \sum_{j=1}^{\infty} \alpha_j \varphi_j(x)$$

Note that the α_j's do not depend on x (global).

Projection estimator Given the basis, it is equivalent to estimate the α_j's :

$$\hat{f}(x) = \sum_{j=1}^{N} \hat{\alpha}_j \varphi_j(x)$$

We basically estimate $\alpha_j$by 0 for $j \geq N+1$. This makes sense if we assume that $\alpha_j \searrow 0$ fast. This is the case for smooth functions if the basis is well chosen.

Such bases include: Fourier (trigonometric) basis, wavelet bases, Polynomial bases (Legendre, Hermitte, ...).

Regardless of the basis (Fourier or not), the α_j's are called Fourier coefficients.

How do we compute the estimators $\hat{\alpha}_j$? By the definition of the Fourier coefficient, we have

$$\alpha_j = \int f(x)\varphi_j(x)dx = \mathbb{E}[Y\varphi(X)]$$

this is because the design has uniform distribution.

To estimate, $\mathbb{E} \rightsquigarrow \frac{1}{n}\sum_{i=1}^{n}$, so we obtain

$$\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i \varphi(X_i)$$

The parameter N plays the same role as h and has to be chosen carefully: large N leads to undersmoothing and small N leads to oversmoothing.