Introduction to statistical learning

3.5 Nonparametric methods

V. Lefieux

June 2018

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Table of contents

Smoothing kernel methods
- For density estimation
  - KDE
  - Bandwidth and kernel choice
  - Multivariate case
- For regression
  - Nadaraya-Watson estimator
  - Local polynomials

Nearest neighbors

Table of contents

Smoothing kernel methods

For density estimation

KDE

Bandwidth and kernel choice

Multivariate case

For regression

Nadaraya-Watson estimator

Local polynomials

Nearest neighbors

Table of contents

Smoothing kernel methods

For density estimation

KDE

Bandwidth and kernel choice

Multivariate case

For regression

Nadaraya-Watson estimator

Local polynomials

Nearest neighbors

Goal

Estimate the density f of a random variable:

- Based on an i.i.d sample $(X_1, \ldots, X_n)$.
- Without parametric hypothesis.

Reference: (Silverman, 1986)

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

A first idea: the histogram I

Let $(X_1, \ldots, X_n)$ be an i.i.d sample from $X \in \mathbb{R}$.

Assume that X has a probability density f bounded support $([a, b[)$.

We divide the support into a series of k small intervals $[\alpha_i, \alpha_{i+1}[$, $i \in \{1, \ldots, k\}$, with $\alpha_1 = a$ and $\alpha_{k+1} = b$.

Histogram estimator is a step function:

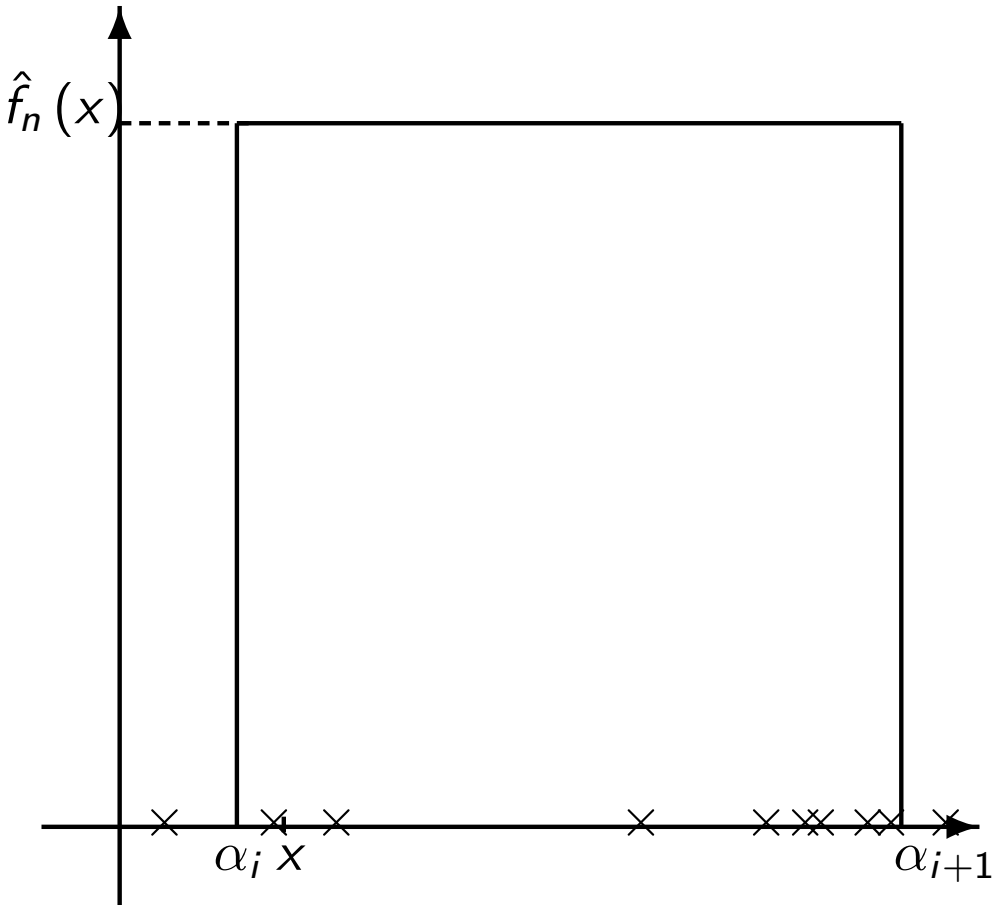
$$\forall x \in [a, b[: \widehat{f}_n(x) = \sum_{i=1}^{k} \frac{f_i}{\alpha_{i+1} - \alpha_i} \mathbb{1}_{[\alpha_i, \alpha_{i+1}[}(x)$$

where:

$$f_i = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{[\alpha_i, \alpha_{i+1}[}(X_j) \ .$$

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

A first idea: the histogram II



Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

A first idea: the histogram III

Usually we consider equal intervals:

$$\forall i \in \{1, \ldots, k\} : \alpha_{i+1} - \alpha_i = \delta_n \ .$$

We choose δ_n such as:

$$\delta_n \xrightarrow{n \to +\infty} 0 \ ,$$
$$n\delta_n \xrightarrow{n \to +\infty} +\infty \ .$$

Mean Square Error MSE $= \mathbb{E}\left[\left(\widehat{f}_n(x) - f(x)\right)^2\right]$ is of $n^{-\frac{2}{3}}$ order for an optimal choice $\delta_n \sim n^{-\frac{1}{3}}$.

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

From histogram to KDE I

We can consider a "moving" histogram: the value of the density estimator at x is equal to the height of the rectangle in the histogram, for the interval $[x - h_n, x + h_n]$:

$$\widehat{f}_n(x) = \frac{1}{2nh_n} \sum_{i=1}^{n} \mathbf{1}_{[x-h_n, x+h_n]}(X_i).$$

We can notice that:

$$x - h_n \leq X_i \leq x + h_n \quad \Longleftrightarrow \quad -1 \leq \frac{x - X_i}{h_n} \leq 1$$

so:

$$\widehat{f}_n(x) = \frac{1}{2nh_n} \sum_{i=1}^{n} \mathbf{1}_{[-1,1]}\left(\frac{x - X_i}{h_n}\right).$$

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

From histogram to KDE II

The moving histogram can be rewriten:

$$\widehat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)$$

where

$$K(x) = \frac{1}{2} \mathbf{1}_{[-1,1]}(x) \ .$$

Note that considering $[-1, 1]$ or $[-1, 1[$ is equivalent.

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Kernel Density Estimator (KDE)

We can improve the "moving histogram" estimation considering another K function (kernel).

Kernel Density Estimator (KDE) (in the univariate case) is defined by:

$$\widehat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)$$

where h_n is the bandwidth and K is a Parzen-Rosenblatt kernel:

- K is bounded,
- $\int_{-\infty}^{+\infty} K(x)\,\mathrm{d}x = 1$,
- $xK(x) \xrightarrow{x \to +\infty} 0$.

Generally we use symmetric kernels:

$$\forall x \in \mathbb{R} : K(-x) = K(x) \ .$$

Note that the kernel isn't necessary nonnegative.

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Kernel order

Consider for a function f and $k \in \mathbb{N}$:

$$m_k(f) = \int_{-\infty}^{+\infty} x^k f(x) \, \mathrm{d}x.$$

The order of a kernel K is:

$$r = \min \{k \in \mathbb{N} : m_k(K) > 0\} .$$

Note that $r \geq 2$ for symmetric kernels.

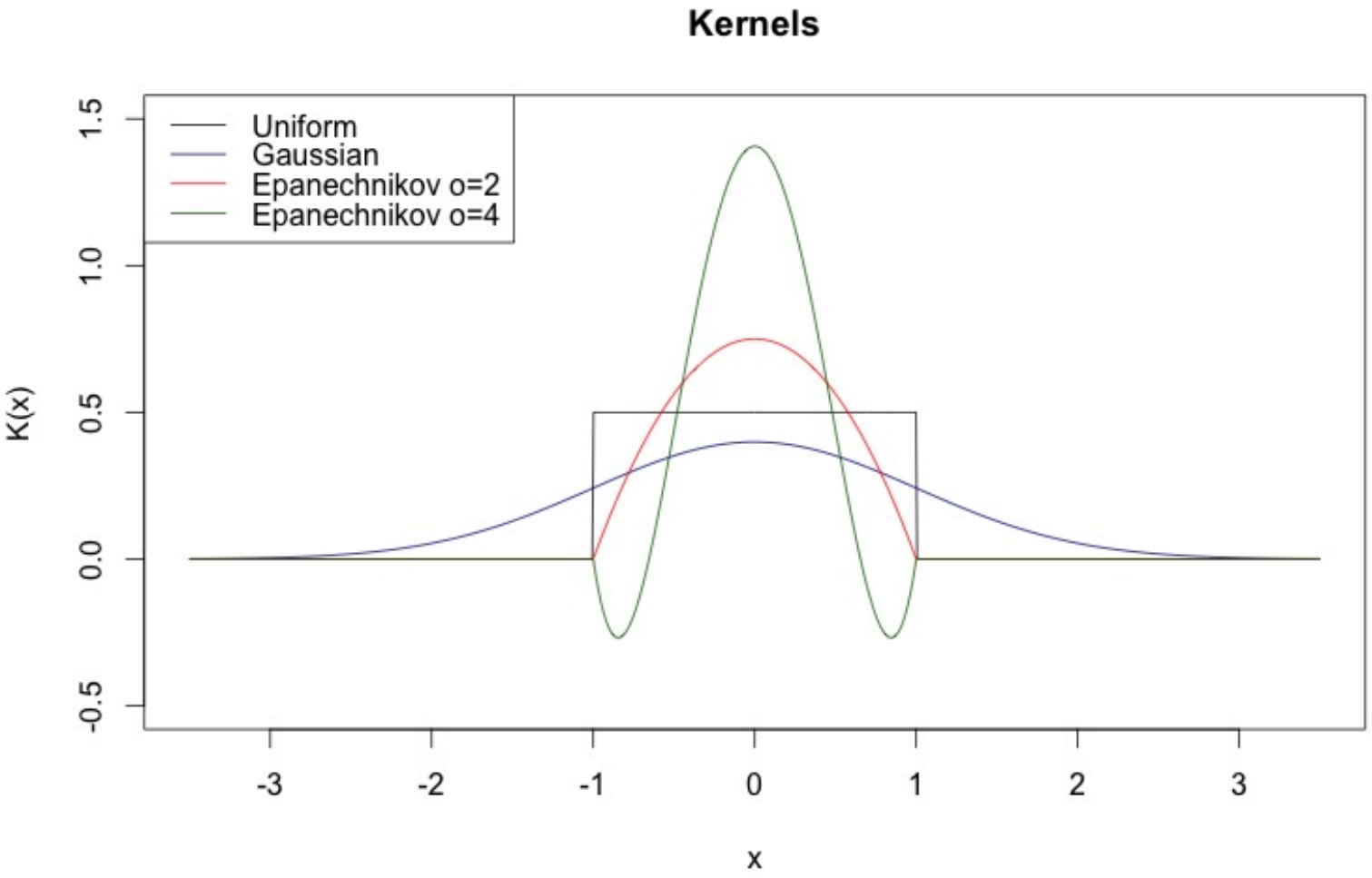
Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Classic kernels I

Uniform	$K(x) = \frac{1}{2}\mathbb{1}_{[-1,1]}(x)$
Gaussian	$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$
Epanechnikov (o=2)	$K(x) = \frac{3}{4}\left(1 - x^2\right)\mathbb{1}_{[-1,1]}(x)$
Epanechnikov (o=4)	$K(x) = \frac{15}{8}\,\frac{3}{4}\left(1 - \frac{7}{3}x^2\right)\left(1 - x^2\right)\mathbb{1}_{[-1,1]}(x)$

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Classic kernels II



Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

KDE properties

Consider for a function f:

$$R(f) = \int_{-\infty}^{+\infty} f^2(x)\,\mathrm{d}x\ .$$

We have:

$$\int_{-\infty}^{+\infty} \widehat{f}_n(x)\,\mathrm{d}x = 1\ .$$

$$\begin{aligned}
\mathrm{Bias}\left(\widehat{f}_n(x)\right) &= \mathbb{E}\left(\widehat{f}_n(x)\right) - f(x) \\
&= \frac{1}{r!} f^{(r)}(x)\, h_n^r m_r(K) + o\left(h_n^r\right)\ . \\
\mathrm{Var}\left(\widehat{f}_n(x)\right) &= \frac{f(x)\,R(K)}{nh_n} + o\left(\frac{1}{n}\right)\ .
\end{aligned}$$

The bias expression is obtained (with a Taylor approximation) for a kernel of order r.

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Remarks

- As small the bandwidth is, as small is the bias.
- For a second-order kernel, the bias increases with the square of the bandwidth.
- With higher (than 2) order kernels, bias is of lower than the order obtained from second-order kernels.
  Higher-order kernels are bias-reducing kernels.
- Variance tends to 0 if $nh_n \xrightarrow{n \to +\infty} 0$.

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Mean Square Error

The mean square error of the KDE is:

$$\begin{aligned}
\mathrm{MSE}\left(\widehat{f}_n(x)\right) &= \mathbb{E}\left[\left(\widehat{f}_n(x) - f(x)\right)^2\right] \\
&= \left[\mathrm{Bias}\left(\widehat{f}_n(x)\right)\right]^2 + \mathrm{Var}\left(\widehat{f}_n(x)\right) \\
&= \left(\frac{1}{r!} f^{(r)}(x)\, h_n^r m_r(K)\right)^2 + \frac{f(x)\, R(K)}{n h_n} \\
&\quad + o\left(h_n^r + \frac{1}{n}\right) .
\end{aligned}$$

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Asymptotic Mean Square Error

The AMSE, and the following AMISE, are useful to obtain the optimal bandwidth.

$$\mathrm{AMSE}\left(\widehat{f}_n(x)\right) = \left(\frac{1}{r!} f^{(r)}(x)\, h_n^r m_r(K)\right)^2 + \frac{f(x)\, R(K)}{n h_n} .$$

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Asymptotic Mean Integrated Square Error

$$\begin{aligned}\mathrm{AMISE}\left(\widehat{f_n}\right) &= \int_{-\infty}^{+\infty} \mathrm{AMSE}\left(\widehat{f_n}(x)\right) \mathrm{d}x \\ &= \left(\frac{1}{r!} h_n^r m_r(K)\right)^2 R\left(f^{(r)}\right) + \frac{R(K)}{nh_n} .\end{aligned}$$

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Optimal bandwidth

After differentiation of the AMISE, we obtain the following optimal values:

$$h_{opt} = \left[\left(R\left(f^{(r)}\right)\right)^{\frac{-1}{2r+1}} \left(\frac{(r!)^2 R(K)}{2r.m_r^2(K)} \right)^{\frac{1}{2r+1}} \right] n^{-\frac{1}{2r+1}} ,$$

$$\text{AMISE}_{opt}\left(\widehat{f}_n\right) = (2r+1) \left[\frac{R\left(f^{(r)}\right) m_r^2(K) (R(K))^{2r}}{(r!)^2 (2r)^{2r}} \right]^{\frac{1}{2r+1}} n^{-\frac{2r}{2r+1}} .$$

Considering a second-order kernel, we obtain $h_{opt} \sim n^{-\frac{1}{5}}$ and $\text{AMISE}_{opt}\left(\widehat{f}_n\right) \sim n^{-\frac{4}{5}}$, which is less than the classic n^{-1} rate for parametric methods.

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Rule of thumb

The optimal bandwidth depends on $R\left(f^{(r)}\right)$ which is unknown.The idea is to replace this unknown quantity by $R\left(g^{(r)}\right)$ where $g^{(r)}$ is a plausible parametric density.

We can consider for example $\mathcal{N}\left(\overline{x_n}, s^2\right)$ where $(\overline{x}, s)$ are respectively the sample mean and standard deviation. The bandwidth choice would be optimal if the density is gaussian, and close to the optimal one if the density is close to a gaussian one. Whatever it could be a good initialization point in optimal bandwidth searching.

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Kernel choice

Considering the kernel order r, and the optimal bandwidth, the AMISE is minimized with $m_r(K)(R(K))^r$.

Epanechnikov (1969) has minimized $R(K)$ under constraint $m_r(K) = 1$ (only the shape is important), and obtained a scaled version of the following one:

$$K(x) = \frac{3}{4}\left(1 - x^2\right) \mathbb{1}_{[-1,1]}(x) .$$

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Cross validation I

Consider the ISE (Integrated Squared Error):

$$\mathrm{ISE}\left(\widehat{f}\right) = \int \left(\widehat{f}(x) - f(x)\right)^2 \mathrm{d}x \ .$$

We can rewrite:

$$\mathrm{ISE}\left(\widehat{f}\right) = \int \widehat{f}^2(x)\ \mathrm{d}x - 2\int \widehat{f}(x) f(x)\ \mathrm{d}x + \int f^2(x)\ \mathrm{d}x \ .$$

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Cross validation II

1. $\int f^2(x)\,\mathrm{d}x$ doesn't depend on h_n.
2. For a bandwidth h_n, $\int \widehat{f}^2(x)\,\mathrm{d}x$ can be computed.
3. We could abusively consider that $\int \widehat{f}(x) f(x)\,\mathrm{d}x$ is equivalent to $\int \widehat{f}(x) f(x)\,\mathrm{d}x = \mathbb{E}\left[\widehat{f}(X)\right]$, indeed $\widehat{f}$ depends on the sample.
   To correct this, we use a "leave-one out estimator" and estimate $\int \widehat{f}(x) f(x)\,\mathrm{d}x$ by:

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{f}_{-i}(X_i)$$

where

$$\widehat{f}_{-i}(x) = \frac{1}{(n-1)h}\sum_{j=1, j\neq i}^{n} K\left(\frac{x - X_j}{h_n}\right).$$

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Cross validation III

The cross validation bandwidth is obtained by minimizing:

$$\mathrm{CV}\left(h_n\right) = \int \widehat{f}^2\left(x\right)\,\mathrm{d}x - \frac{2}{n(n-1)h_n}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} K\left(\frac{X_i - X_j}{h_n}\right).$$

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

KDE in the multivariate case I

Let $(X_1, \ldots, X_n)$ be an i.i.d sample from $X \in \mathbb{R}^p$, $p \in \mathbb{N}^*$. Kernel Density Estimator (KDE) (in the multivariate case) is defined by:

$$\widehat{f}_n(x) = \frac{1}{nh_{1,n} \ldots h_{p,n}} \sum_{i=1}^{n} K_p \left(\frac{x_1 - X_{1,i}}{h_{1,n}}, \ldots, \frac{x_p - X_{p,i}}{h_{p,n}} \right)$$

where $x = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$, $H = (h_{1,n}, \ldots, h_{p,n})^\top \in \mathbb{R}^p$ is the bandwidth and K_p is a multivariate Parzen-Rosenblatt kernel:

- K_p is bounded on $\mathbb{R}^p$.
- $\int_{\mathbb{R}^p} K(x) \, \mathrm{d}x = 1$.
- $\|x\|^p K_p(x) \xrightarrow{\|x\| \to +\infty} 0$ where $\|x\|$ is the euclidean norm.

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

KDE in the multivariate case II

Usually we choose the same bandwidth $H = (h_n, \ldots, h_n) \in \mathbb{R}^p$ and we rewrite:

$$\widehat{f}_n(x) = \frac{1}{nh_n^p} \sum_{i=1}^{n} K_p\left(\frac{x - X_i}{h_n}\right)$$

with $X_i = (X_{1,i}, \ldots, X_{p,i})$.
Generally we use a product kernel (we consider this case in the following):

$$K_p(x_1, \ldots, x_p) = \prod_{j=1}^{p} K(x_j) \ .$$

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Bias, variance and AMISE

We still consider for a function f on $\mathbb{R}^p$:

$$R(f) = \int_{\mathbb{R}^p} f^2(x)\,\mathrm{d}x\ ,$$
$$m_k(f) = \int_{\mathbb{R}^p} x^k f(x)\,\mathrm{d}x\ .$$

We have:

$$\mathrm{Bias}\left(\widehat{f}_n(x)\right) = \frac{1}{r!} m_r(K)\, h_n^r \bigtriangledown^r f(x) + o\left(h_n^r\right)\ ,$$
$$\mathrm{Var}\left(\widehat{f}_n(x)\right) = \frac{f(x)\, R^p(K)}{n h_n^p} + o\left(\frac{1}{n}\right)$$

where $\bigtriangledown^r f(x) = \sum_{i=1}^p \frac{\partial^r}{\partial x_i^r} f(x_1, \ldots, x_p)$.
We obtain:

$$\mathrm{AMISE}\left(\widehat{f}_n\right) = \left(\frac{1}{r!} h_n^r m_r(K)\right)^2 R\left(\bigtriangledown^r f\right) + \frac{R^p(K)}{n h_n^p}\ .$$

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Remarks

- $h_{opt} \sim n^{-\frac{1}{2r+p}}$ and $\text{AMISE}_{opt}\left(\widehat{f}_n\right) \sim n^{-\frac{2r}{2r+p}}$.
- As small the bandwidth is, as small is the bias.
- Variance tends to 0 if $nh_n^p \xrightarrow{n \to +\infty} 0$.
- The AMISE rate decreases with the dimension, this is called the curse of dimensionality.
- From a practical point of view, we choose the optimal bandwidth by cross validation.

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Table of contents

Smoothing kernel methods

For density estimation

KDE

Bandwidth and kernel choice

Multivariate case

For regression

Nadaraya-Watson estimator

Local polynomials

Nearest neighbors

Goal

Let $((X_1, Y_1), \ldots, (X_n, Y_n))$ be an i.i.d sample from $(X, Y) \in \mathbb{R} \times \mathbb{R}$.
Consider the regression between

$$Y = m(X) + \varepsilon$$

where ε is a white noise.
The goal is to estimate the link function m without parametric hypothesis.

References: Härdle et al. (2004),Fan and Gijbels (1996)

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

From KDE to Nadaraya-Watson estimator I

The following link function:

$$
\begin{aligned}
m(x) &= \mathbb{E}(Y \mid X = x) \\
&= \int_{-\infty}^{+\infty} y f_{Y|X=x}(y)\, \mathrm{d}y \\
&= \frac{1}{f_X(x)} \int_{-\infty}^{+\infty} y f_{X,Y}(x,y)\, \mathrm{d}y
\end{aligned}
$$

minimizes the quadratic cost function $\mathbb{E}\left(\varepsilon^2\right)$.

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

From KDE to Nadaraya-Watson estimator II

Using the 2 following KDE (product kernel for X and Y, symetric kernel K, same bandwidth h_n):

$$\widehat{f}_X(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) ,$$

$$\widehat{f}_{X,Y}(x,y) = \frac{1}{nh_n^2} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) K\left(\frac{y - Y_i}{h_n}\right) ,$$

we can consider the following estimator:

$$\widehat{m}_{nw}(x) = \frac{1}{\widehat{f}_X(x)} \int_{-\infty}^{+\infty} y \widehat{f}_{X,Y}(x,y) \, \mathrm{d}y .$$

After calculations we obtain:

$$\widehat{m}_{nw}(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)} .$$

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Nadaraya-Watson estimator I

Let $((X_1, Y_1), \ldots, (X_n, Y_n))$ be an i.i.d sample from $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$.
Consider the regression between

$$Y = m(X) + \varepsilon$$

where ε is a white noise.
The Nadaraya-Watson estimator of the link function is:

$$\widehat{m}_{nw}(x) = \frac{\sum_{i=1}^{n} Y_i K_p\left(\frac{x - X_i}{h_n}\right)}{\sum_{i=1}^{n} K_p\left(\frac{x - X_i}{h_n}\right)} .$$

Generally we use a product kernel, and it's possible to consider different bandwidths for covariates.

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Nadaraya-Watson estimator II

We can rewrite:

$$\widehat{m}_{nw}(x) = \sum_{i=1}^{n} \omega_i Y_i$$

where:

$$\omega_i = \frac{K_p\left(\frac{x-X_i}{h_n}\right)}{\sum_{j=1}^{n} K_p\left(\frac{x-X_j}{h_n}\right)} .$$

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Cross validation

We can choose the optimal bandwidth by a cross validation criterium. We can estimate the:

$$\mathrm{MISE}\,(m) = \mathbb{E}\left[(\widehat{m}_{nw}(X) - m(X))^2\right]$$

by:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{m}_{nw}\left(X_i\right) - m\left(X_i\right)\right)^2 \ .$$

In fact we consider:

$$\mathrm{CV}\left(\widehat{m}_{nw}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{m}_{nw,-i}\left(X_i\right) - Y_i\right)^2$$

where:

$$\widehat{m}_{nw,-i}\left(x\right) = \frac{\sum_{j=1,j\neq i}^{n} Y_j K\left(\frac{x-X_j}{h_n}\right)}{\sum_{j=1,j\neq i}^{n} K\left(\frac{x-X_j}{h_n}\right)} \ .$$

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Local polynomials (univariate case)

For Nadaraya-Watson estimator, continuity criterion on m allows to write:

$$x \approx X_i \Rightarrow m(x) \approx m(X_i) \ .$$

Thus the folllowing definition of $\widehat{m}$:

$$\widehat{m}(x) = \underset{m(x)=a}{\arg\min} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) \ [Y_i - m(X_i)]^2 \ .$$

If m is d-differentiable, we can write:

$$x \backsimeq X_i \Longrightarrow m(x) \backsimeq \sum_{j=0}^{d} [\gamma_j(x)] \ (x - X_i)^j$$

with $\gamma_j(x) = \frac{m^{(j)}(x)}{j!}$.

We can write in this case:

$$\widehat{m}(x) = \underset{m(x)=\sum_{j=0}^{d} a_j (x-X_i)^j}{\arg\min} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) \ [Y_i - m(X_i)]^2 \ .$$

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Table of contents

Smoothing kernel methods
- For density estimation
  - KDE
  - Bandwidth and kernel choice
  - Multivariate case
- For regression
  - Nadaraya-Watson estimator
  - Local polynomials

Nearest neighbors

k-nearest neighbors

Let $(X_1, \ldots, X_n)$ be a sample in $\mathbb{R}^p$ and consider $x \in \mathbb{R}^p$.
We compute distances between x and the sample:

$$\forall i \in \{1, \ldots, n\} : D_i = d\left(x, X_i\right)$$

where d is a distance (for example: euclidean, Manhattan).
We reorder these distance:

$$D_{(1)} \leq \ldots \leq D_{(n)} \ .$$

The k-nearest neighbors of x (k-nn for k-nearest neighbours) are the k points associated to $D_{(1)} \leq \ldots \leq D_{(k)}$: $X_{(1)}\left(x\right), \ldots, X_{(k)}\left(x\right)$ (in cas of equality, it's possible to use a random selection).
Based on k-nearest neighbors, it's possible to estimate a probability density, build a classification or a regression.

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Regression

Let $((X_1, Y_1), \ldots, (X_n, Y_n))$ be an i.i.d sample of $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$.
Consider the following regression:

$$Y = m(X) + \varepsilon$$

where ε is a white noise.
The regression with k-nearest neighbors is:

$$\widehat{m}_{knn}(x) = \frac{1}{k} \sum_{i=1}^{n} \mathbb{1}_{d(x, X_i) \leq D_k} Y_i \ .$$

It's possible to smooth this link function.

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

Choice of k

We find heuristically with cross validation the optimal number of nearest neighbors, based on an error criterion (RMSE, MAPE...).

Bigger is k, smoother is the regression estimation.

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References

References I

Epanechnikov, V. A. (1969). Nonparametric estimation of a multidimensional probability density. *Theory of Probability and Applications*, 14(1):156–161.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. Monographs on Statistics & Applied Probability (66). Chapman & Hall.

Härdle, W., Müller, M., Sperlich, S., and Werwatz, A. (2004). *Nonparametric and semiparametric models*. Springer.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning. Data Mining, inference, and prediction*. Springer Series in Statistics. Springer, 2 edition.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Monographs on Statistics & Applied Probability (26). Chapman & Hall.

Smoothing kernel methods
For density estimation
KDE
Bandwidth and kernel choice
Multivariate case
For regression
Nadaraya-Watson estimator
Local polynomials
Nearest neighbors
References