

## Lec 10: Introduction to Nonparametric Statistics


---

Yanju Han

Nov. 21, 2023

---

---



Nonparametric model:  $y \sim p_\theta$ ,  $\theta$ : infinite-dimensional, typically a function  
(typically written as  $y \sim p_f$ )

Canonical examples:

Regression: given  $(x_1, y_1), \dots, (x_n, y_n) \sim P_{X,Y}$ , estimate  
the regression function  
$$f(x) := \mathbb{E}[Y | X=x]$$

Density estimation: given  $x_1, \dots, x_n \sim f$  with an unknown density  $f$ ,  
estimate  $f$ .

Other examples: in causal inference, interested in:

- causal function:  $c(x) = \text{Cov}(Y, W | X=x)$

- conditional/heterogeneous ATE (CATE):

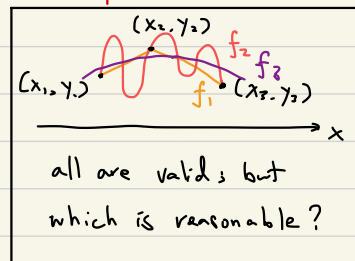
$$\tau(x) = \mathbb{E}[Y | W=1, X=x] - \mathbb{E}[Y | W=0, X=x]$$

Features of nonparametric models:

- model size > sample size, assumptions are necessary to prevent  
overfitting (typically smoothness or shape)

- MLE not well-defined / non-unique / hard  
to find

- explicit bias-variance tradeoff!



## Nonparametric regression

### A simple binning estimator

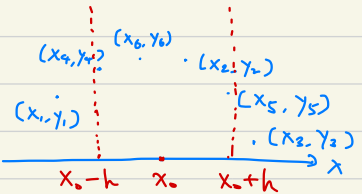
Assumption:  $x_1, \dots, x_n \in [0, 1]$  nonrandom (fixed design)

Target: given  $x_0 \in [0, 1]$ , estimate  $f(x_0) = \mathbb{E}[Y|X=x_0]$

A simple estimator:

bandwidth  
parameter

$$\hat{f}_h(x_0) = \frac{\sum_{i: |x_i - x_0| \leq h} Y_i}{\#\{i: |x_i - x_0| \leq h\}}$$



(  $h \rightarrow 0$ : overfit to the closest data point;  
 $h \rightarrow \infty$ : underfit to the sample average of  $y$  )

Analysis: assume that  $f$  is  $L$ -Lipshitz, i.e.  $|f'(x)| \leq L \forall x$

(or equivalently,  $|f(x) - f(y)| \leq L|x - y| \forall x, y$ )

also, assume that  $\text{Var}(Y|X=x) \leq \sigma_0^2$  for all  $x$

Variance of  $\hat{f}_h(x_0)$ :

$$\begin{aligned} \text{Var}(\hat{f}_h(x_0)) &= \frac{\text{Var}(\sum_{i: |x_i - x_0| \leq h} Y_i)}{(\#\{i: |x_i - x_0| \leq h\})^2} \quad (\text{non-random } \{x_i\}) \\ &= \frac{\sum_{i: |x_i - x_0| \leq h} \text{Var}(Y_i)}{(\#\{i: |x_i - x_0| \leq h\})^2} \quad (\text{independence}) \\ &\leq \frac{\sigma_0^2}{\#\{i: |x_i - x_0| \leq h\}} \end{aligned}$$

If  $\{x_i\}$  are evenly spaced in  $[0, 1]$ , then  $\text{Var}(\hat{f}_h(x_0)) = O(\frac{\sigma_0^2}{nh})$ .

Bias of  $\hat{f}_h(x_0)$ ,

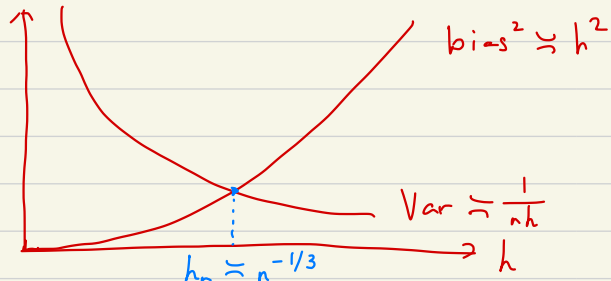
$$\begin{aligned}
 |\text{Bias}(\hat{f}_h(x_0))| &= |\mathbb{E}[\hat{f}_h(x_0)] - f(x_0)| \\
 &= \left| \frac{\sum_{i: |x_i - x_0| \leq h} f(x_i)}{\#\{i: |x_i - x_0| \leq h\}} - f(x_0) \right| \quad (\mathbb{E}[y_i] = f(x_i)) \\
 &= \left| \frac{\sum_{i: |x_i - x_0| \leq h} (f(x_i) - f(x_0))}{\#\{i: |x_i - x_0| \leq h\}} \right| \\
 &\leq \frac{\sum_{i: |x_i - x_0| \leq h} |f(x_i) - f(x_0)|}{\#\{i: |x_i - x_0| \leq h\}} \quad (\text{triangle inequality}) \\
 &\leq \frac{\sum_{i: |x_i - x_0| \leq h} L |x_i - x_0|}{\#\{i: |x_i - x_0| \leq h\}} \quad (\text{Lipshitz condition}) \\
 &\leq Lh.
 \end{aligned}$$

Final mean-squared error (MSE)

$$\begin{aligned}
 \text{MSE}(\hat{f}_h(x_0)) &= \mathbb{E}(\hat{f}_h(x_0) - f(x_0))^2 \\
 &= \text{Bias}(\hat{f}_h(x_0))^2 + \text{Var}(\hat{f}_h(x_0)) \\
 &= O(L^2 h^2 + \frac{\sigma_0^2}{nh})
 \end{aligned}$$

$$\begin{aligned}
 \text{Optimal choice of } h: h &= h_n = \left( \frac{\sigma_0^2}{nL^2} \right)^{1/2} \\
 \text{Optimal MSE} &= O(L^{3/2} \sigma_0^{4/3} n^{-2/3})
 \end{aligned}$$

Bias-variance  
tradeoff:



## Generalization: Nadaraya-Watson estimator

Kernel: a (non-negative) function  $K: \mathbb{R}^d \rightarrow \mathbb{R}$  s.t.  $\int_{\mathbb{R}^d} K(x) dx = 1$ .

Rescaled kernel: for  $h > 0$ , let  $K_h(x) = \frac{1}{h^d} K(\frac{x}{h})$

Examples: rectangle/box kernel:  $K(x) = 1(\|x\|_\infty \leq \frac{1}{2})$

Gaussian kernel:  $K(x) = (\frac{1}{2\pi})^{d/2} \exp(-\frac{1}{2}\|x\|_2^2)$

Property:  $\int_{\mathbb{R}^d} K_h(x) dx = \int_{\mathbb{R}^d} K_h(hz) h^d dz = \int_{\mathbb{R}^d} K(z) dz = 1, \forall h > 0$ .

### Kernel-regression estimator / Nadaraya-Watson estimator

$$\hat{f}_h(x_0) = \frac{\sum_{i=1}^n K_h(x_0 - x_i) y_i}{\sum_{i=1}^n K_h(x_0 - x_i)}$$

- previous estimator corresponds to  $K$  being the box kernel;
- an equivalent expression of  $\hat{f}_h(x_0)$  is

$$\hat{f}_h(x_0) = \sum_{i=1}^n w_i(x_0) y_i, \quad \text{with} \quad w_i(x_0) = \frac{K_h(x_0 - x_i)}{\sum_{i=1}^n K_h(x_0 - x_i)}$$

being the "weight" of  $x_i$  for  $x_0$ .

Analysis. Assume that: 1)  $|K(x)| \leq B$  for all  $x$ ;  
2)  $K(x) = 0$  for all  $|x| \geq M$ .

$$\begin{aligned} \text{Var}(\hat{f}_h(x_0)) &\leq \frac{\sum_{i=1}^n K_h^2(x_0 - x_i) \sigma_0^2}{\left(\sum_{i=1}^n K_h(x_0 - x_i)\right)^2} \leq \frac{\frac{B}{h^d} \sigma_0^2}{\sum_{i=1}^n K_h(x_0 - x_i)} \\ &= \frac{B \sigma_0^2}{n h^d} \cdot \frac{1}{\underbrace{\frac{1}{n h^d} \sum_{i=1}^n K\left(\frac{x_0 - x_i}{h}\right)}_{\text{expected to be a constant}}} \end{aligned}$$

$$\begin{aligned}
|\text{Bias}(\hat{f}_h(x_0))| &= \left| \frac{\sum_{i=1}^n K_h(x_0 - x_i) f(x_i)}{\sum_{i=1}^n K_h(x_0 - x_i)} - f(x_0) \right| \\
&\leq \frac{\sum_{i=1}^n K_h(x_0 - x_i) |f(x_i) - f(x_0)|}{\sum_{i=1}^n K_h(x_0 - x_i)} \quad \begin{array}{l} \text{blue arrow: } K_h(x_0 - x_i) = 0 \text{ if } |x_0 - x_i| \geq hM \\ \text{red arrow: } \leq L |x_i - x_0| \end{array} \\
&\leq \frac{\sum_{i=1}^n K_h(x_0 - x_i) \cdot L h M}{\sum_{i=1}^n K_h(x_0 - x_i)} = L M h
\end{aligned}$$

$$\text{MSE}(\hat{f}_h(x_0)) = O(L^2 M^2 h^2 + \frac{B \sigma^2}{n h^{\frac{2}{d}}}) = O(n^{-\frac{2}{2+d}})$$

$\uparrow$   
 optimal bandwidth  $h = h_n \asymp n^{-\frac{1}{2+d}}$

Capturing higher smoothness of  $f$ : next lecture (local polynomials / splines)

Density estimation: estimate  $f$  from  $x_1, \dots, x_n \sim f$ .

Kernels are still useful: let  $K$  be a kernel with  $\int_{\mathbb{R}^d} x K(x) dx = 0$ .

Kernel density estimator (KDE):

$$\hat{f}_h(x_0) = \frac{1}{n} \sum_{i=1}^n K_h(x_0 - X_i)$$

- Intuition: when  $K$  is the box kernel,

$$\begin{aligned}
\hat{f}_h(x_0) &= \frac{\#\{i: x_i \text{ lies in the box centered at } x_0 \text{ of edgelenhth } h\}}{n h^d} \\
&\approx \frac{f(x_0) \cdot n h^d}{n h^d} = f(x_0)
\end{aligned}$$

Analysis: assume that  $\|f''\|_\infty \leq L$ ,  $\int x^2 K(x) dx < \infty$ ,  $\int K^2(x) dx < \infty$ ,  
 and  $d = 1$ .

$$\begin{aligned}
\text{Var}(\hat{f}_h(x_0)) &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(K_h(x_0 - X_i)) \\
&\leq \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}[K_h(x_0 - X_i)^2] \\
&= \frac{1}{nh^2} \int K\left(\frac{x_0 - x}{h}\right)^2 dx \\
&= \frac{1}{nh} \int K^2(z) dz \quad (x = x_0 - hz)
\end{aligned}$$

$$\begin{aligned}
|\text{Bias}(\hat{f}_h(x_0))| &= |\mathbb{E}[K_h(x_0 - X_1)] - f(x_0)| \\
&= \left| \underbrace{\int \frac{1}{h} K\left(\frac{x_0 - x}{h}\right) f(x) dx}_{\text{convolution of } f \text{ and } K_h} - f(x_0) \right| \\
&= \left| \int \frac{1}{h} K\left(\frac{x_0 - x}{h}\right) (f(x) - f(x_0)) dx \right| \quad (\int K_h(x) dx = 1) \\
&= \left| \int \frac{1}{h} K\left(\frac{x_0 - x}{h}\right) (f(x) - f(x_0) - f'(x_0)(x - x_0)) dx \right| \\
&\quad \left( \int \frac{1}{h} K\left(\frac{x_0 - x}{h}\right) (x_0 - x) dx = h \cdot \int z K(z) dz = 0 \right) \\
&\leq \int \frac{1}{h} K\left(\frac{x_0 - x}{h}\right) \cdot \frac{L}{2} (x_0 - x)^2 dx \\
&= \frac{Lh^2}{2} \int z^2 K(z) dz \quad (x = x_0 - hz)
\end{aligned}$$

$$\text{MSE}(\hat{f}_h(x_0)) = O(L^2 h^4 + \frac{1}{nh}) \underset{\text{optimal bandwidth } h = h_n \sim n^{-1/5}}{=} O(n^{-4/5})$$

View Nadaraya-Watson as KDE:

$$\begin{aligned}\mathbb{E}[Y|X=x] &= \frac{\int y f(x, y) dy}{\int f(x, y) dy} \approx \frac{\int y \cdot \frac{1}{n} \sum_{i=1}^n K_h(x-X_i) K_h(y-Y_i) dy}{\int \frac{1}{n} \sum_{i=1}^n K_h(x-X_i) K_h(y-Y_i) dy} \\ &= \frac{\sum_{i=1}^n K_h(x-X_i) Y_i}{\sum_{i=1}^n K_h(x-X_i)}\end{aligned}$$

## Nearest-neighbor density estimator

Define  $r_i = \|X_i - x_0\|_2$  as the distance between  $X_i$  and  $x_0$ .

For  $k=1, 2, \dots, n$ , let  $r_{(k)}$  be the  $k$ -th smallest element of  $(r_1, \dots, r_n)$   
( $k$ -th nearest neighbor)

$$\hat{f}_k(x_0) = \frac{k/n}{\underset{\substack{\uparrow \\ \text{volume of } d\text{-dim ball} \\ \text{of radius } r_{(k)}}}{\text{Vol}_d(r_{(k)})}} = \frac{k/n}{\frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)} r_{(k)}^d}$$

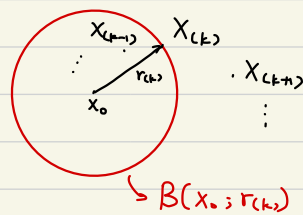
Intuition:  $f(x_0) \cdot \text{Vol}_d(r_{(k)})$

$$\approx \int_{B(x_0, r_{(k)})} f(x) dx$$

= actual prob. of  $X \in B(x_0, r_{(k)})$

$\approx$  empirical prob. of  $X \in B(x_0, r_{(k)})$

$$= \frac{k}{n}.$$



Rigorous claim:  $\int_{B(x_0, r_{(k)})} f(x) dx \sim \text{Beta}(k, n-k+1)$

Pf. LHS =  $k$ -th smallest element of  $\{Z_i := \int_{B(x_0, \|X_i - x_0\|_2)} f(x) dx\}_{i=1}^n$   
where  $P(Z_1 \leq t) = P(\|X_1 - x_0\|_2 \leq g^{-1}(t))$  ( $g(r) := \int_{B(x_0, r)} f(x) dx$ )  
 $= g(g^{-1}(t)) = t \Rightarrow Z_i \sim \text{Unif}[0, 1] \quad \square$



Analysis (Optional) Assume that  $\|f'\|_\infty \leq L$  &  $c \leq f(x) \leq C$  for all  $x \in \text{supp}(f)$ .

$$\begin{aligned} \text{Step I. } & \left| \int_{B(x_0; r_{ck})} f(x) dx - f(x_0) \cdot \text{Vol}_d(r_{ck}) \right| \\ &= \left| \int_{B(x_0; r_{ck})} (f(x) - f(x_0)) dx \right| \\ &= \left| \int_{B(x_0; r_{ck})} [f(x) - f(x_0) - f'(x_0)(x - x_0)] dx \right| \\ &\leq \int_{B(x_0; r_{ck})} \frac{L}{2} \|x - x_0\|_2^2 dx \leq \frac{L}{2} r_{ck}^2 \text{Vol}_d(r_{ck}) \end{aligned}$$

$$\begin{aligned} \text{Step II. } & \mathbb{E} \left| \int_{B(x_0; r_{ck})} f(x) dx - \hat{f}_k(x_0) \cdot \text{Vol}_d(r_{ck}) \right|^2 \\ &= \mathbb{E} \left| \text{Beta}(k, n+1-k) - \frac{k}{n} \right|^2 = O\left(\frac{k}{n^2}\right). \end{aligned}$$

Step III. Since  $f(x) \asymp 1$  everywhere,

$$\begin{aligned} \frac{k}{n} &\stackrel{\text{w.p.}}{\approx} \int_{B(x_0; r_{ck})} f(x) dx \asymp \text{Vol}_d(r_{ck}) \\ \Rightarrow \text{Vol}_d(r_{ck}) &\asymp \frac{k}{n} \Rightarrow r_{ck} \asymp \left(\frac{k}{n}\right)^{1/d}. \end{aligned}$$

$$\begin{aligned} \text{Conclusion. } & f(x_0) \text{Vol}_d(r_{ck}) \stackrel{O(r_{ck}^d \text{Vol}_d(r_{ck}))}{\approx} \int_{B(x_0; r_{ck})} f(x) dx \\ & \stackrel{O(\sqrt{k}/n)}{\approx} \hat{f}_k(x_0) \cdot \text{Vol}_d(r_{ck}) \end{aligned}$$

$$\Rightarrow \text{MSE}(\hat{f}_k(x_0)) = O\left(r_{ck}^4 + \frac{k/n^2}{\text{Vol}_d(r_{ck})^2}\right)$$

$$= O\left(\left(\frac{k}{n}\right)^{4/d} + \frac{1}{k}\right) \underset{\substack{\uparrow \\ k = k_n \asymp n^{\frac{4}{4+d}}}}{=} O\left(n^{-\frac{4}{4+d}}\right)$$

(matching the KDE result for  $d=1$ )