Chapter 2.

References:

↳ Chap. 24 vdV

↳ Sec. 6.3 All of Stats.

- We observe $X_1, \ldots, X_n \overset{iid}{\sim} Q$

- Let $F(x) = Q\{X \le x\}$ denote the CDF
  $f(x)$ denote the density.

- <u>Our goal</u>: Estimate f at a point x_0.

- Given that $f(x_0) = \frac{d}{dx} F(x)\big|_{x=x_0}$ we might consider estimating $f(x_0)$ w/ a plug-estimator

empirical CDF

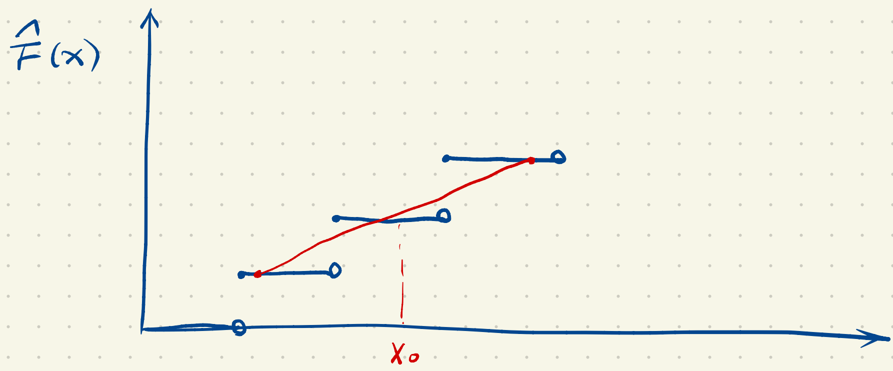
$$\hat{f}(x_0) = \frac{d}{dx} \hat{F}(x)\Big|_{x=x_0}$$

- One way to estimate F is via the empirical CDF.

(1) $$\hat{F}(x_0) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \le x_0\}$$

- We know that this is a "good" estimator of CDF.
  ↳ E.g.: $\sqrt{n}\left[\hat{F}(x_0) - F(x_0)\right] \rightsquigarrow N(0, \sigma^2)$

- <u>But</u>: Estimating $f(x_0)$ via (1) turns out to be a very bad idea



A lot of things will be 0.

A better one will lead us to KDE.

- Another option uses that

$$f(x_0) = \lim_{h \to 0} \frac{F(x_0+h) - F(x_0-h)}{2h}$$

and so, when h is small

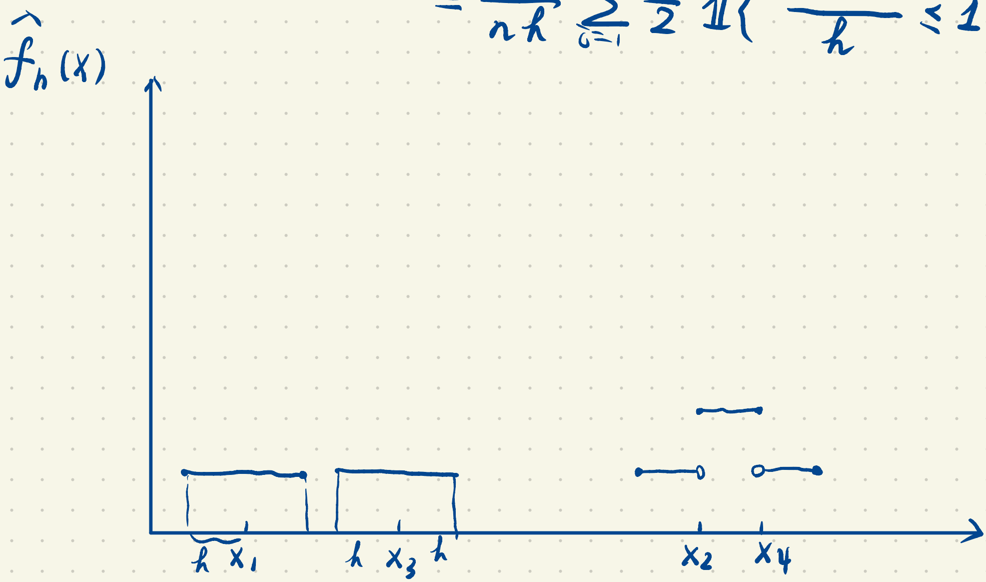
$$f(x_0) \approx \frac{F(x_0+h) - F(x_0-h)}{2h}$$

- This suggests an estimator

$$\hat{f}_h(x_0) := \frac{\hat{F}(x_0+h) - \hat{F}(x_0-h)}{2h}$$

$$= \frac{1}{2nh} \sum_{i=1}^{n} \mathbb{1}\{x_0 - h < X_i \le x_0 + h\}$$

$$\overset{(a.s.)}{=} \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} \mathbb{1}\left\{ \frac{|X_i - x_0|}{h} \le 1 \right\} \star$$



<u>Note</u>: This estimate of f is not smooth.

<u>Question</u>: Can we define a smoother estimate of f?

<u>Ans</u>: Yes!

- Let $K: \mathbb{R} \to \mathbb{R}$ be a kernel, that is, a function satisfying $\int K(u)\,du = 1$.

Def: An s-th order kernel is a kernel K that satisfies

$$\int u^r K(u)\,du = 0 \quad \text{for } r = 1, 2, \dots, s-1$$

$$\left| \int u^s K(u)\,du \right| < \infty$$

↳ If K is symmetric about zero, $[K(u) = K(-u)]$ then K is always at least a 2nd order kernel $(s=2)$

↳ Using higher order kernels $(s>2)$ can lead to estimators with lower bias.

General form of the KDE

For $h>0$,

$$\hat{f}_h(x_0) := \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} K_h\left(\frac{X_i - x_0}{h}\right)$$

where $K_h(u) = \frac{1}{h} K\left(\frac{u}{h}\right)$

Example of Kernels:

1) Uniform : $K(u) = \frac{1}{2} \mathbb{1}\{|u| \le 1\}$ bounded support ☆

2) Epanechnikov : $K(u) = \frac{3}{4}(1-u^2)\mathbb{1}\{|u| \le 1\}$ bounded support

3) Gaussian : $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-u^2/2\right)$

Note: All of these kernels are 2nd-order.

Nice fact: If K is non-negative, then for any $h > 0$, $\hat{f}_h$ is a PDF.

$$\int \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) dx$$

$$= \frac{1}{n} \sum \int \frac{1}{h} K\left(\frac{X_i - x_0}{h}\right) dx \qquad \left(u = \frac{X_i - x}{h}\right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \int K(u)\, du = 1.$$

Studying $\hat{f}_h(x_0)$:

We now study the performance of $\hat{f}_h(x_0)$ as an estimator of $f(x_0)$.

We'll quantify the performance in terms of the MSE :

$$\mathbb{E}\left[\{\hat{f}_h(x_0) - f(x_0)\}^2\right] = \overbrace{\left\{\mathbb{E}\left[\hat{f}_h(x_0)\right] - f(x_0)\right\}^2}^{\text{bias}^2}$$
$$+ \underbrace{\mathrm{var}\left[\hat{f}_h(x_0)\right]}_{\text{variance}}$$

Here, we'll suppose that f belongs to (β, L) Hölder class w/ $\beta = 2$, and $L > 0$.

↳ Note: All of the calculations we do still go through if the restriction of f to a nbhd. of x_0 is (β, L)-Hölder.

Recall: Saying that f is $(2, L)$ Hölder means that
$$|f'(x_1) - f'(x_2)| \le L|x_1 - x_2| \quad \forall x_1, x_2$$

We focus on the case where the kernel K is:

- ↳ bounded
- ↳ non-negative
- ↳ 2nd order
- ↳ bounded support

We're going to see that choosing a small bandwidth $h > 0$ yields low bias and high variance and vice versa.

Bias of KDE

Recall: $\text{Bias} = \mathbb{E}[\hat{f}_h(x_0)] - f(x_0)$

$$\mathbb{E}[\hat{f}_h(x_0)] = \frac{1}{nh}\sum_{i=1}^{n} \mathbb{E}\left[K\left(\frac{X_i - x_0}{h}\right)\right]$$

$$(\text{iid}) \quad = \frac{1}{h}\mathbb{E}\left[K\left(\frac{X_i - x_0}{h}\right)\right]$$

$$\left(u = \frac{x_1 - x_0}{h}\right)$$

$$= \frac{1}{h}\int K\left(\frac{x_1 - x_0}{h}\right) f(x_1)\, dx_1$$

$$= \int K(u)\, f(x_0 + uh)\, du$$

Recalling $\int K(u)\, du = 1$, and so

$$\text{Bias} = \mathbb{E}[\hat{f}_h(x_0)] - f(x_0)$$

smooth!

$$= \int K(u) \underbrace{\left[f(x_0 + uh) - f(x_0)\right]}_{\star} du$$

By the mean-value Thm., we know there exists

$\tilde{x}_{uh}$ such that

$$\star = uh f'(\tilde{x}_{uh})$$

Hence

$$\text{Bias} = \int K(u)\, uh\, f'(\tilde{x}_{uh})\, du$$

$$= \underbrace{\int K(u)\, uh\, f'(x_0)\, du} + \int K(u)\, uh\left[f'(\tilde{x}_{uh}) - f'(x_0)\right] du$$

2nd order

$$= h f'(x_0) \int K(u)\, u\, du$$

$$= 0$$

Hence

$$|\text{Bias}| = \left| \int K(u)\, u h \left[f'(\widetilde{x_{uh}}) - f'(x_0) \right] du \right|$$

Jensen $\leq h \int K(u)\, |u|\, | f'(\widetilde{x_{uh}}) - f'(x_0) |\, du$

Hölder (Lipschitz) $\leq L h \int K(u)\, |u|\, | \widetilde{x_{uh}} - x_0 |\, du$

$|\widetilde{x_{uh}} - x_0|$ $\leq L h^2 \underbrace{\int K(u)\, u^2\, du}_{\sigma_K^2}$

$$= L \sigma_K^2 h^2$$

Hence, $\boxed{\text{Bias}^2 \leq L^2 \sigma_K^4 h^4}$

Variance of the KDE

$$\text{Var}(\hat{f}_h(x_0)) = \text{Var}\left[\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h} \right) \right]$$

(independence) $= \frac{1}{n^2 h^2} \sum_{i=1}^{n} \text{var}\left[K\left(\frac{x_i - x_0}{h} \right) \right]$

(identical) $= \frac{1}{nh^2} \text{var}\left[K\left(\frac{x_1 - x_0}{h} \right) \right]$

$\leq \frac{1}{nh^2} \mathbb{E}\left[K\left(\frac{x_1 - x_0}{h} \right)^2 \right]$ ← 2nd moment

$= \frac{1}{nh^2} \int K\left(\frac{x_1 - x_0}{h} \right)^2 f(x_1)\, dx_1$ $\qquad u = \frac{x_1 - x_0}{h}$

$$= \frac{1}{nh} \underbrace{\int K(u)^2 f(x_0 + uh)\, du}_{\star\star} \qquad (2)$$

We'll study ☆☆ in what follows.

To do this, we'll make use of two facts:

1) f is Hölder continuous $\Rightarrow$ continuous.

2) K has bounded support.

Let $k_1 = \inf\{u : K(u) > 0\}$, $k_2 = \sup\{u : K(u) > 0\}$.

We have that

$$
\begin{aligned}
☆☆ &= \int K(u)^2 f(x_0 + uh)\, du \\
&= \int_{k_1}^{k_2} K(u)^2 f(x_0 + uh)\, du \\
&\leq \left[\sup_{u \in [k_1, k_2]} f(x_0 + uh) \right] \int_{k_1}^{k_2} K(u)^2\, du
\end{aligned}
$$

If $h \leq 1$, then this shows that

$$
\leq \underbrace{\left[\sup_{t \in [k_1, k_2]} f(x_0 + t) \right] \int_{k_1}^{k_2} K(u)^2\, du}_{=: \tilde{C}}
$$

Hence, we have shown that (by plugging in)

$$
\mathrm{Var}(\hat{f}_h(x_0)) \leq \frac{\tilde{C}}{nh}
$$

Plugging in our bound on the $\boxed{\text{bias}^2}$ and $\boxed{\text{variance}}$, we find that MSE

$$MSE \le L^2 \sigma_K^4 h^4 + \frac{\tilde{c}}{nh}$$

By setting

$$L^2 \sigma_K^4 h^4 = \frac{\tilde{c}}{nh}$$

$$\Longrightarrow \quad h = c n^{-1/5}$$

$$\therefore \quad MSE \le O(n^{-4/5})$$

Generalization

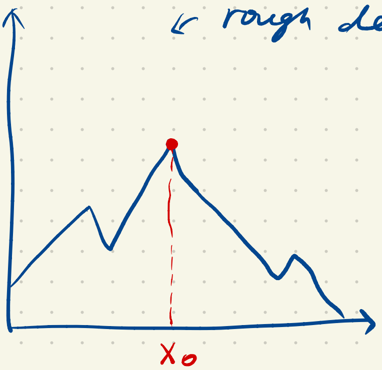
1) 1-dimensional setting w/ different amounts of smoothness

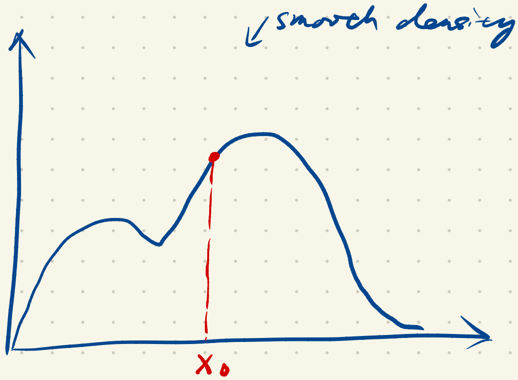
If f belongs to a (β, L) Hölder class, then similar arguments show that

$$MSE = O\left(n^{-\frac{2\beta}{2\beta+1}}\right)$$

if a kernel has sufficiently high order.

We saw this is class w/ $\beta = 2$.





sufficiently high order kernel leads to very smooth density which makes x_0 easier to be estimated.

2) d-dimensional probs.

Suppose X is d-dimensional and we want to estimate $f(x_0)$ at a fixed $x_0 \in \mathbb{R}^d$.

A KDE in this setting takes the form

$$\hat{f}_h(x_0) = \frac{1}{nh^d} \sum_{i=1}^{n} \prod_{j=1}^{d} K\left(\frac{X_{ij} - x_{0j}}{h}\right)$$

If f is β-times differentiable and all partial derivatives up to order β is bounded, then

$$MSE = O\left(n^{-\frac{2\beta}{2\beta+d}}\right)$$

Note: The dimension d appears in the denominator of exponent.

d	MSE UB
1	$n^{-0.8}$
2	$n^{-0.67}$
4	$n^{-0.5}$
10	$n^{-0.29}$

One way of thinking about the exponent on n:

If $MSE \asymp n^{-\alpha}$, then, when n is large, to half MSE, you need to collect about $2^{1/\alpha}$ more data.