Non-parametric density estimation

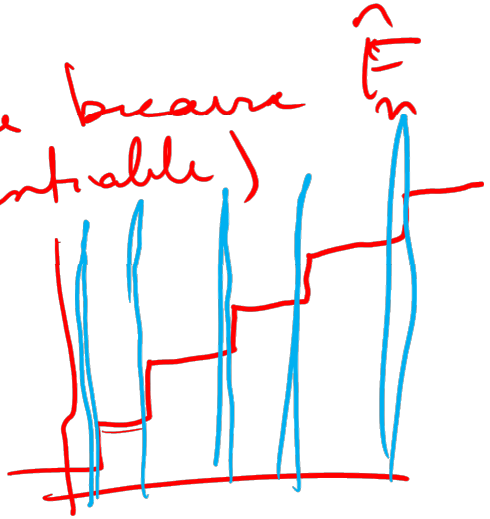
Reading material: https://faculty.washington.edu/yenchic/18W_425/Lec6_hist_KDE.pdf

Motivation

- $D = \{X_1, X_2, \ldots, X_n\}$ sampled i.i.d from an unknown f(X)
- Estimate $\hat{f}(x)$ without committing on a specific parametric form of f(X)
- Why not use empirical CDF?

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

$$\hat{f}_n(x) = \frac{\partial}{\partial x}\hat{F}_n(x)$$ (Not very usable because $\hat{F}_n$ is not differentiable)

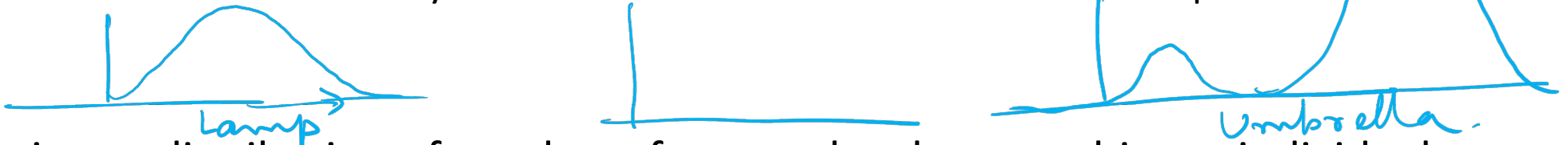


  - Too inefficient to maintain. Need to store entire data
  - Too jerky. Non-zero density at observed points, zero elsewhere.

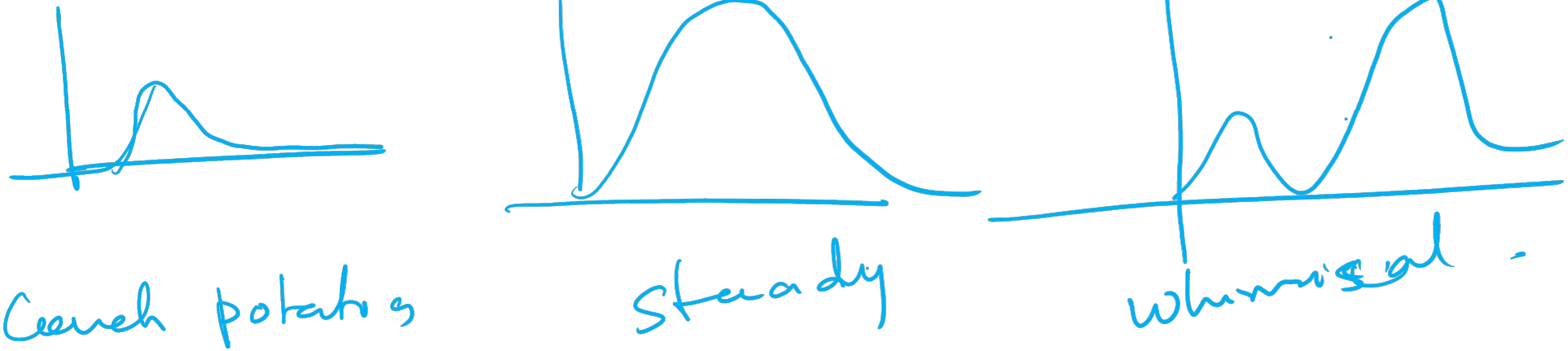
- Density estimation: assume some form of smoothness of f(X)

Real-life examples where parametric form is hard.

- Estimate density for weekly sales of each product in an online store.
  - Difficult to know exactly the best fit of distribution of sales for each product.

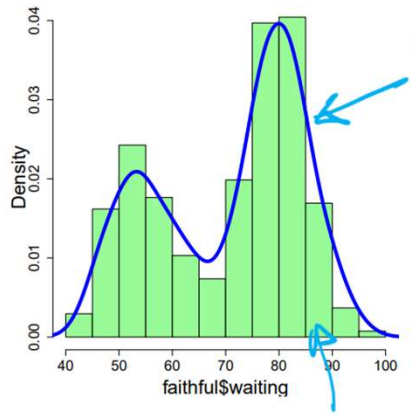


- Estimate distribution of number of steps taken by any arbitrary individual on a phone.



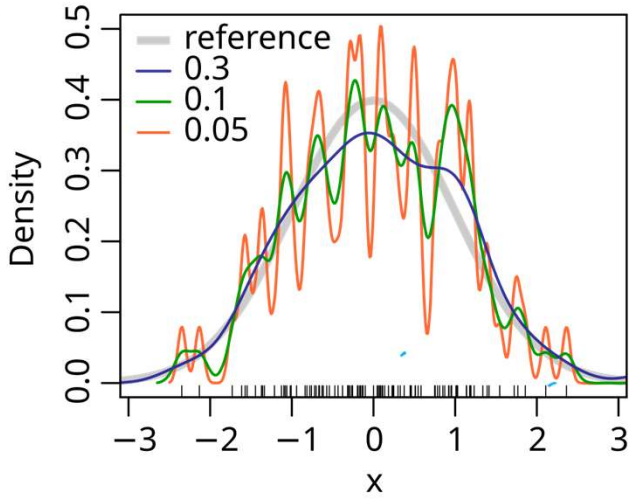
Density estimation methods

- Histogram



True density

- Kernel density

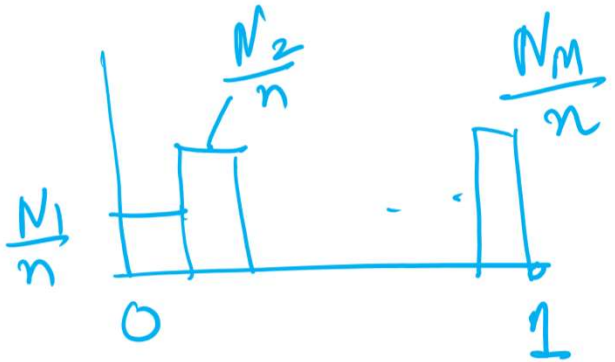


Histogram

- Let $X \in [0,1]$.
- To estimate $\hat{f}_n(x)$ using D: $\{x_1, x_2 \; - \; - \; x_n\}$
- Partition the [0,1] range into M equal sized bins.

$$B_1 = \left[0, \frac{1}{M}\right), B_2 = \left[\frac{1}{M}, \frac{2}{M}\right), \cdots, B_{M-1} = \left[\frac{M-2}{M}, \frac{M-1}{M}\right), B_M = \left[\frac{M-1}{M}, 1\right].$$

$\frac{1}{M}$



- Estimated density

$x \in B_j \Rightarrow \frac{j-1}{M} \le x \le \frac{j}{M}$

N_j = # of x_i in D in B_j

$$\hat{f}_n(x) = \text{(Fraction of instances in D in } B_j\text{)/Width of bin}$$

$$= \frac{\sum_{i=1}^{n} I(x_i \in B_j)\} N_j}{n \, (Y_M)}$$

20.2 Example. Figure 20.3 shows three different histograms based on $n = 1,266$ data points from an astronomical sky survey. Each data point represents the distance from us to a galaxy. The galaxies lie on a "pencilbeam" pointing directly from the Earth out into space. Because of the finite speed of light, looking at galaxies farther and farther away corresponds to looking back in time. Choosing the right number of bins involves finding a good tradeoff between bias and variance. We shall see later that the top left histogram has too few bins resulting in oversmoothing and too much bias. The bottom left histogram has too many bins resulting in undersmoothing and too few bins. The top right histogram is just right. The histogram reveals the presence of clusters of galaxies. Seeing how the size and number of galaxy clusters varies with time, helps cosmologists understand the evolution of the universe. ■

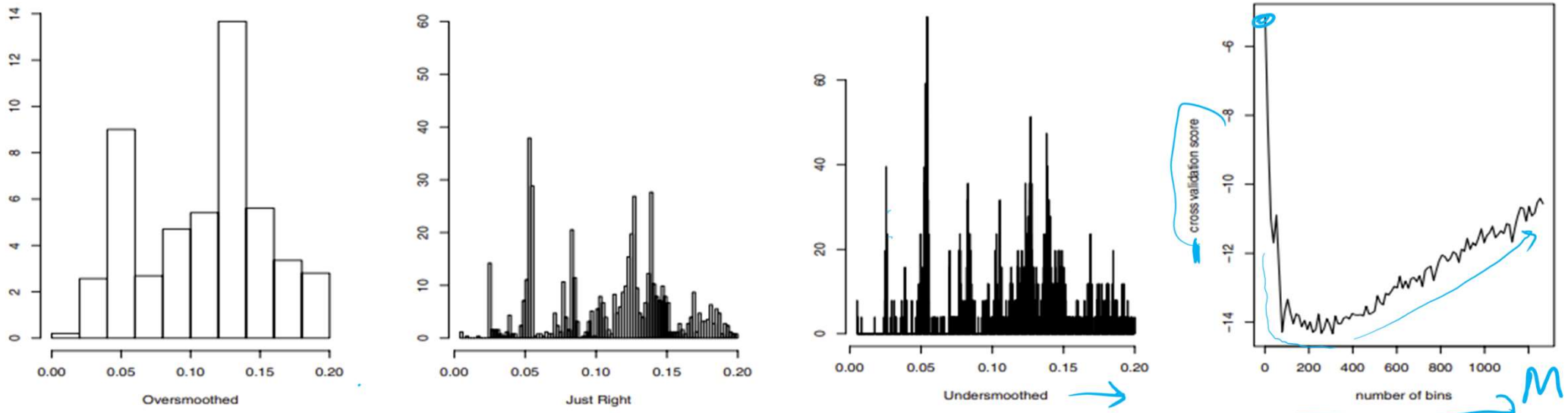


FIGURE 20.3. Three versions of a histogram for the astronomy data. The top left histogram has too few bins. The bottom left histogram has too many bins. The top right histogram is just right. The lower, right plot shows the estimated risk versus the number of bins.

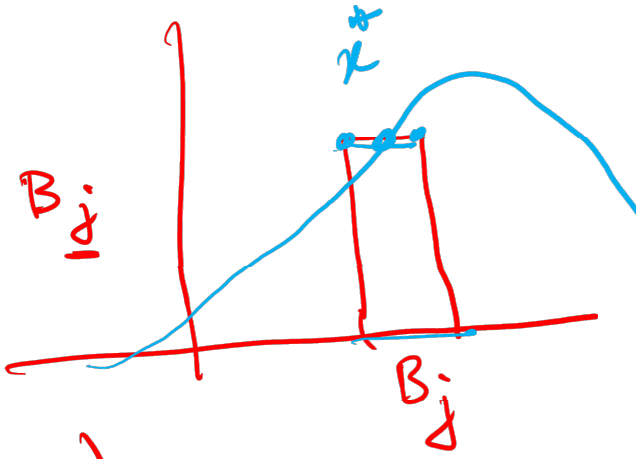
$n = 1266$

$X \sim$ distance of an observed galaxy from earth

Bias, Variance, Risk

$E[\hat{f}_n(x)] - f(x) \equiv$ Bias.

$E[\hat{f}_n(x)] = E_D\left[\frac{M}{n}\sum_{i=1}^{n} I(X_i \in B_j)\right]$ if $x \in B_j$



$= \frac{M}{n} \cdot n \, E_X(I(X \in B_j))$

$= \frac{M}{n} \cdot n \int_{x \in B_j} f(x)dx = M\,P(X \in B_j)$

$= \left[M\left[F\left(\frac{j}{M}\right) - F\left(\frac{(j-1)}{M}\right)\right]\right]$

$= \frac{F\left(\frac{j}{M}\right) - F\left(\frac{j-1}{M}\right)}{\left(\frac{j}{M} - \frac{j-1}{M}\right)} = F'(x^*) = f(x^*)$

$x^* \in B_j$

$F(x)$ is differentiable

$$P(X \in B_j) = \frac{f(x^*)}{M}$$

Variance of $\hat{f}_n(x)$

$$\text{Var}(\hat{f}_n(x)) = \frac{M^2}{n^2} \text{Var}\left(\sum_{i=1}^{n} I(x_i \in B_j)\right) \quad \text{if} \quad x \in B_j.$$

$$= \frac{M^2}{n} \text{Var}_x\left(I(X \in B_j)\right)$$

$$= \frac{M^2}{n} P(X \in B_j)\left(1 - P(X \in B_j)\right)$$

$$= \frac{M^2}{n}\left[\frac{f(x^{(2)})}{M}\right]\left[1 - \frac{f(x^{(2)})}{M}\right]$$

$$= \frac{M f(x^{(2)})}{n} - \frac{f(x^{(2)})^2}{n}$$

Variance increases as $\underline{M}$ is increased

$M \equiv$ # of bins:

Risk $(\hat{f}_n(x)) = \text{Bias}^2 + \text{Var}$

$$[f(x^*) - f(x)]^2 + \frac{M f(x^*)}{n} - \frac{f(x^*)^2}{n} \qquad x^* \in B_j.$$

$$= f'(\tilde{x})(x^* - x) + \frac{M}{n} f(x^*) - \frac{f(x^*)^2}{n}.$$

MV thm:

$$\leq |f'(\tilde{x})| \left\{\frac{1}{M}\right\} + \frac{M}{n} \frac{f(x^*)}{n} - \frac{f(x^*)}{n}.$$

UB on bias — Variance

Bias drops with increasing M vs Variance which increases.

Kernel Density Estimation

- One of the most convenient and accurate ways to estimate any density.
- Given data sample

$$D = \{x_1, x_2 - - x_n\}.$$

- Assume a special function called a kernel that puts a density mass around each training point.

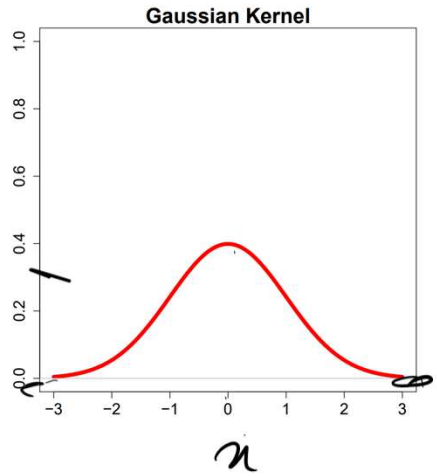
1. $K(x)$ is symmetric.
2. $\int K(x)dx = 1.$
3. $\lim_{x\to-\infty} K(x) = \lim_{x\to+\infty} K(x) = 0.$

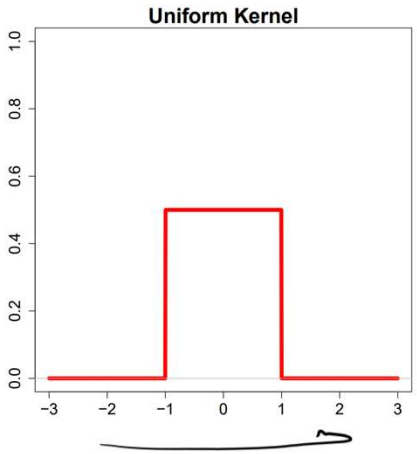
- Examples of kernels

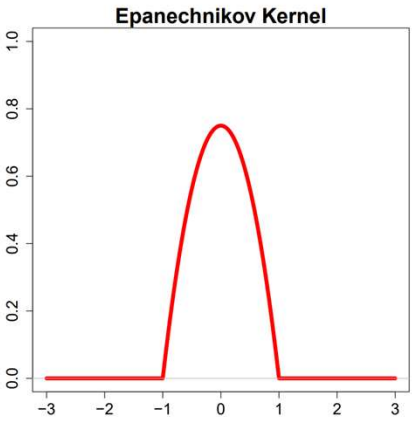
$$\text{Gaussian} \quad K(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}},$$

$$\text{Uniform} \quad K(x) = \frac{1}{2} I(-1 \leq x \leq 1),$$

$$\text{Epanechnikov} \quad K(x) = \frac{3}{4} \cdot \max\{1 - x^2, 0\}.$$





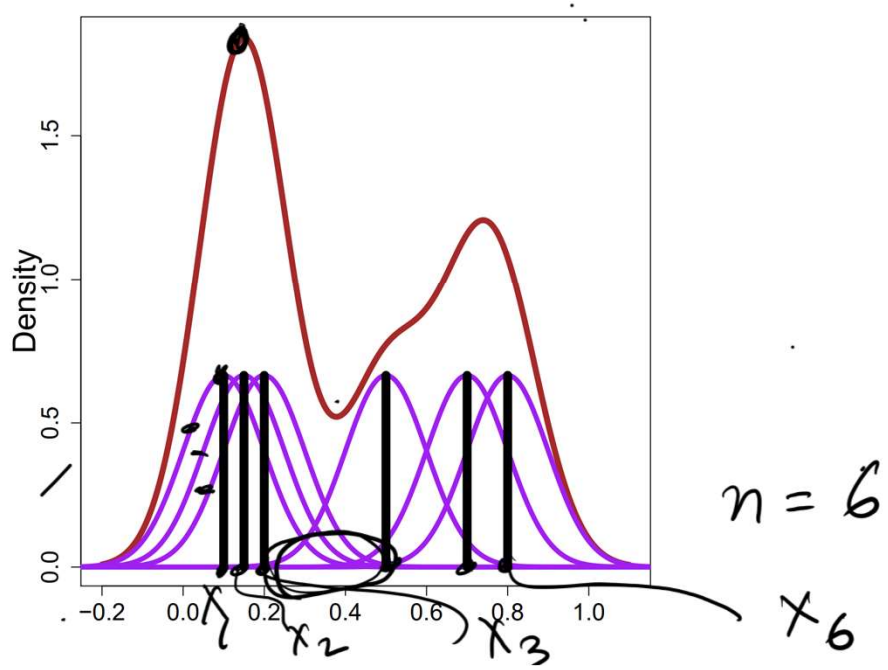


Kernel density estimator

$h = \frac{1}{M}$

20.12 Definition. *Given a kernel K and a positive number h, called the* **bandwidth**, *the* **kernel density estimator** *is defined to be*

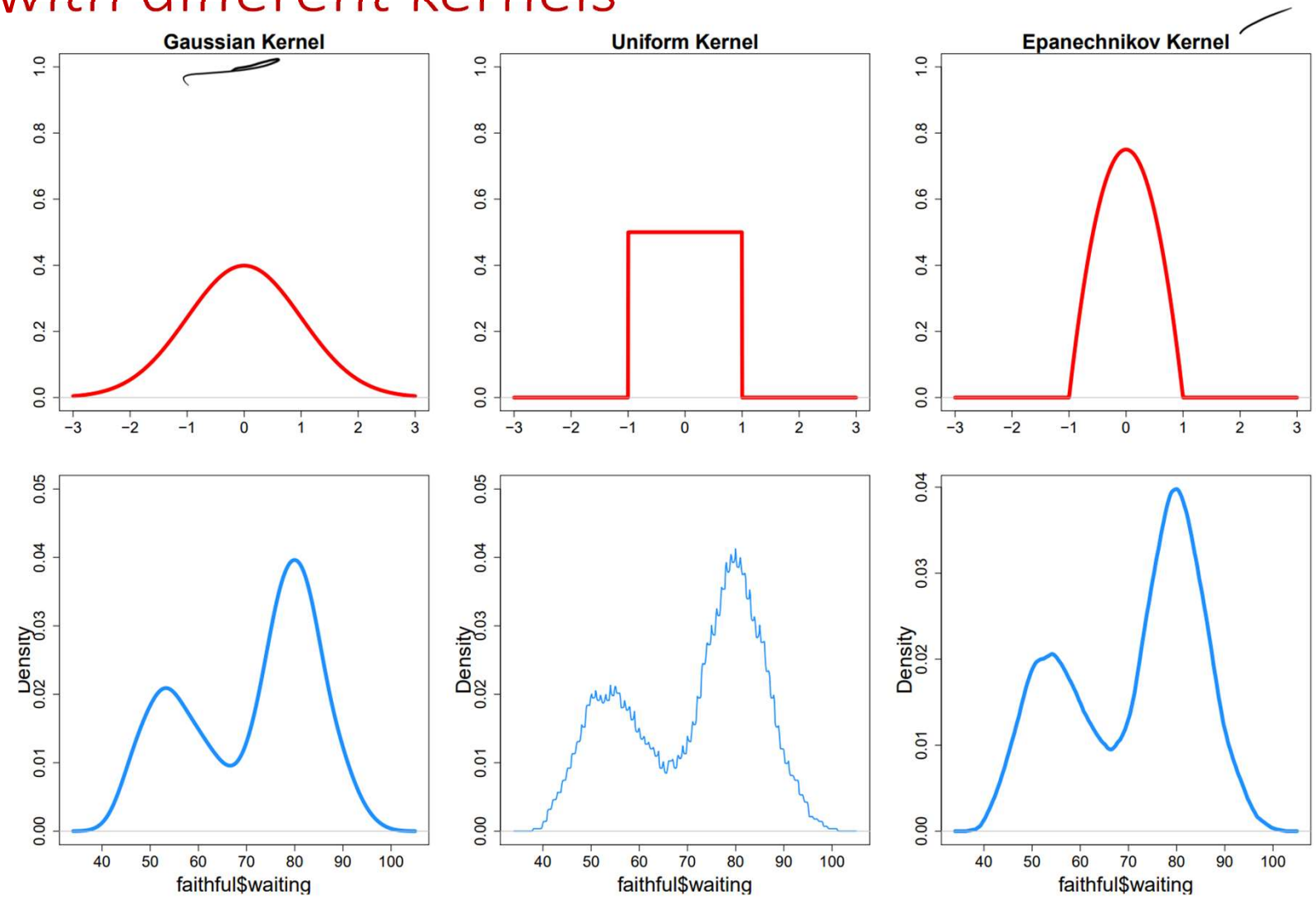
$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right). \tag{20.21}$$



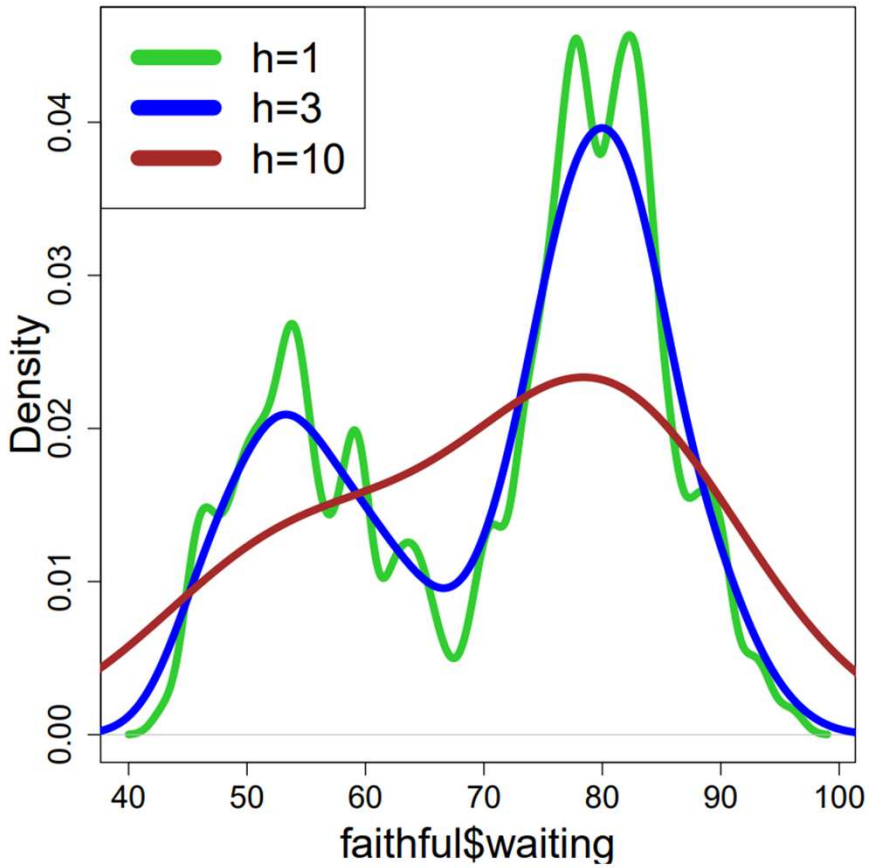
$n = 6$

X_1 X_2 X_3 X_6

KDE with different kernels



Effect of kernel width



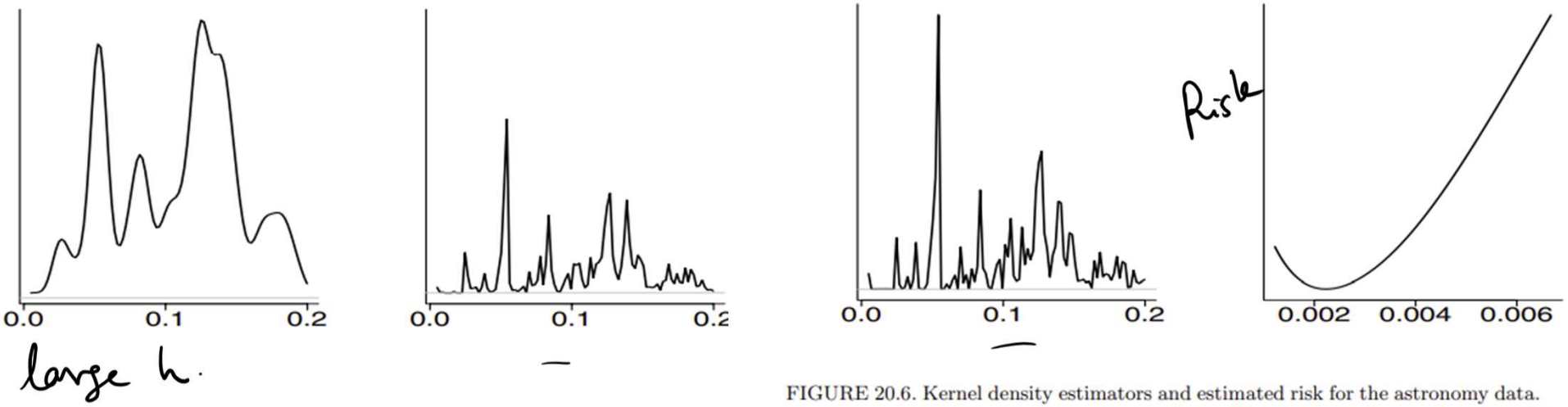


FIGURE 20.6. Kernel density estimators and estimated risk for the astronomy data. Top left: oversmoothed. Top right: just right (bandwidth chosen by cross-validation). Bottom left: undersmoothed. Bottom right: cross-validation curve as a function of bandwidth h. The bandwidth was chosen to be the value of h where the curve is a minimum.

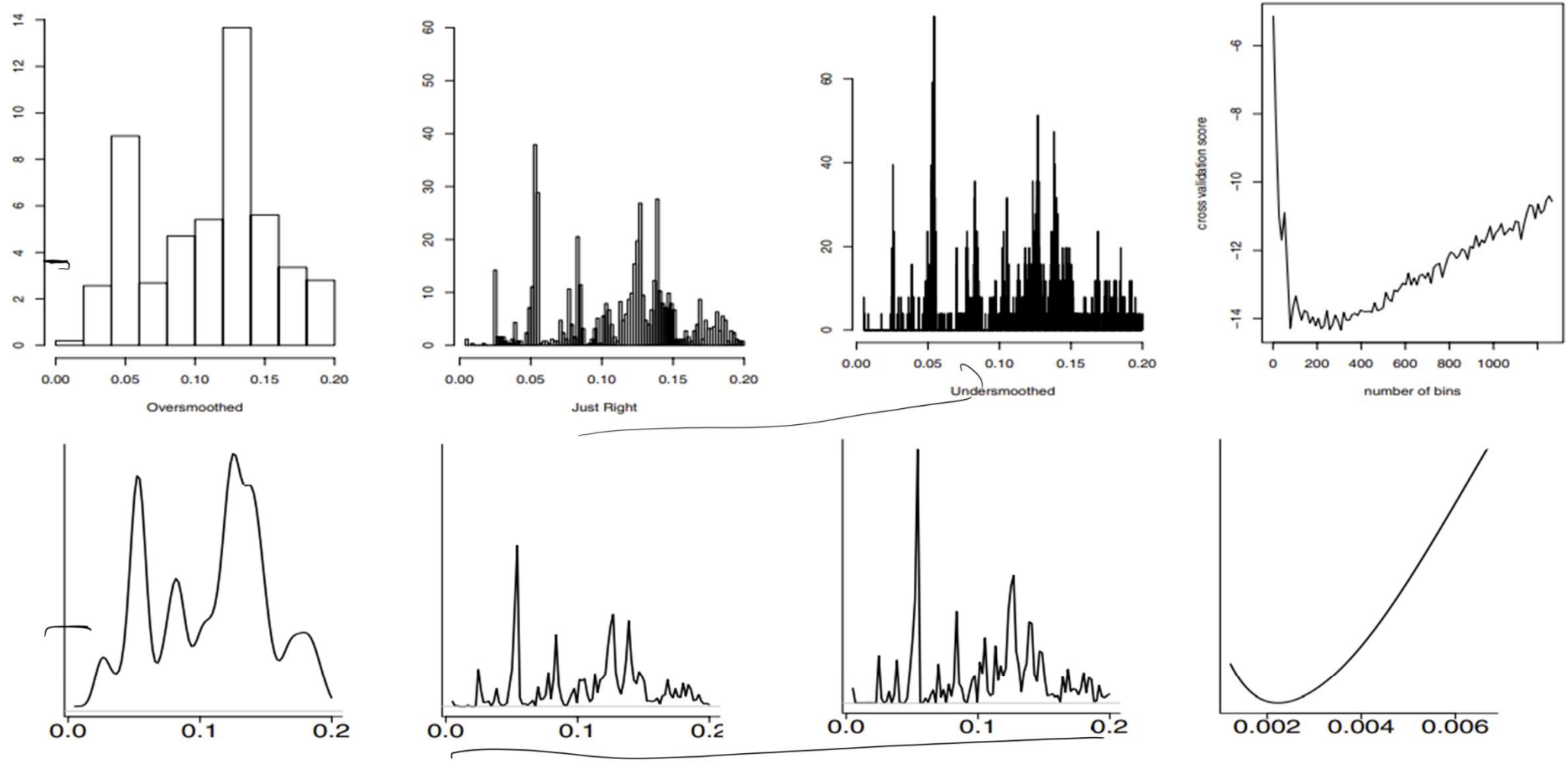


FIGURE 20.6. Kernel density estimators and estimated risk for the astronomy data. Top left: oversmoothed. Top right: just right (bandwidth chosen by cross-validation). Bottom left: undersmoothed. Bottom right: cross-validation curve as a function of bandwidth h. The bandwidth was chosen to be the value of h where the curve is a minimum.

Demo

https://colab.research.google.com/github/fbeilstein/machine_learning/blob/master/lecture_15_kernel_density_estimation.ipynb#scrollTo=pGU9KIkxe-FY