Lecture 11:

Nonparametric Curve Estimation

Nonparametric Curve Estimation

We'll discuss nonparametric estimation of PDF-s and regression functions referred to as **curve estimation** or **smoothing**.

We saw how to estimate a CDF F, without making any assumptions about it. If one needs to estimate a PDF $f(x)$ or a regression function $r(x) = \mathbb{E}(Y|X = x)$, things are different – we can't estimate them well without assuming their **smoothness**.

The Bias-Variance Tradeoff

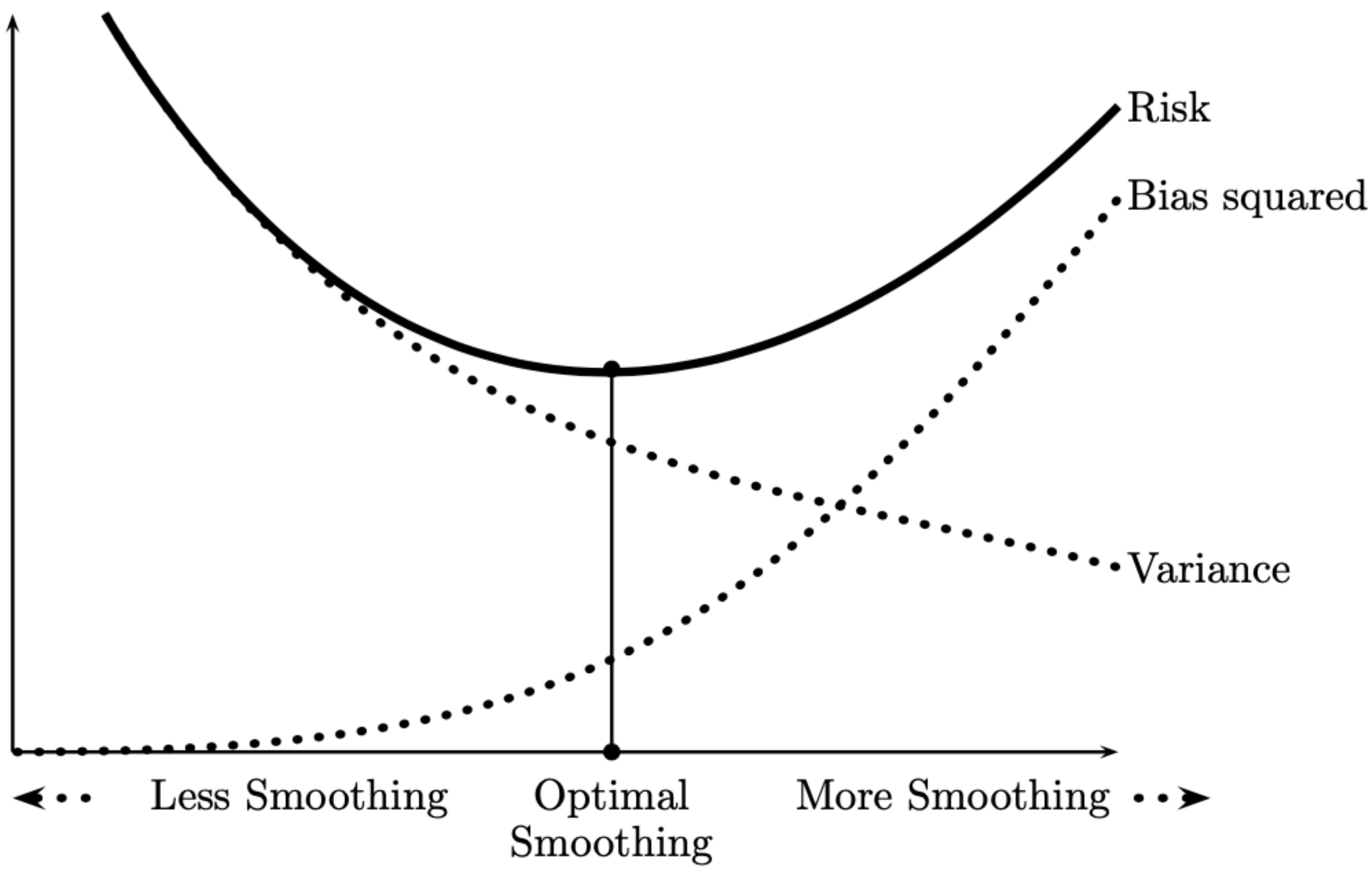
The Bias-Variance Tradeoff

Let g denote an unknown function (density / regression). Let $\widehat{g}_n$ denote its estimator ($\widehat{g}_n(x)$ is a random function, since it depends on the data). We typically use **integrated square error (ISE)** as a loss $L(g, \widehat{g}_n) = \int \left(g(u) - \widehat{g}_n(u)\right)^2 du$. The **risk** or **mean integrated squared error (MISE)** is expectation of it, $R(g, \widehat{g}) = \mathbb{E}\left(L(g, \widehat{g})\right)$.

Lemma: Risk can be written as $R(g, \widehat{g}_n) = \int b^2(x)\, dx + \int v(x)\, dx$

where $b(x) = \mathbb{E}(\widehat{g}_n(x)) - g(x)$ is the **bias** of $\widehat{g}_n(x)$ at x, and $v(x) = \mathbb{V}\left(\widehat{g}_n(x)\right) = \mathbb{E}\left[\left(\widehat{g}_n(x) - \mathbb{E}\,\widehat{g}_n(x)\right)^2\right]$ is the **variance**.

The Bias-Variance Tradeoff



To summarise, RISK $=$ BIAS^2 $+$ VARIANCE. When data is oversmoothed, bias is large and variance is small. When data is undersmoothed, it’s the opposite. This is called the **bias-variance tradeoff** – minimizing risk is balancing these two.

Histograms

Histograms

To speak about KDE-s, first recall histograms.

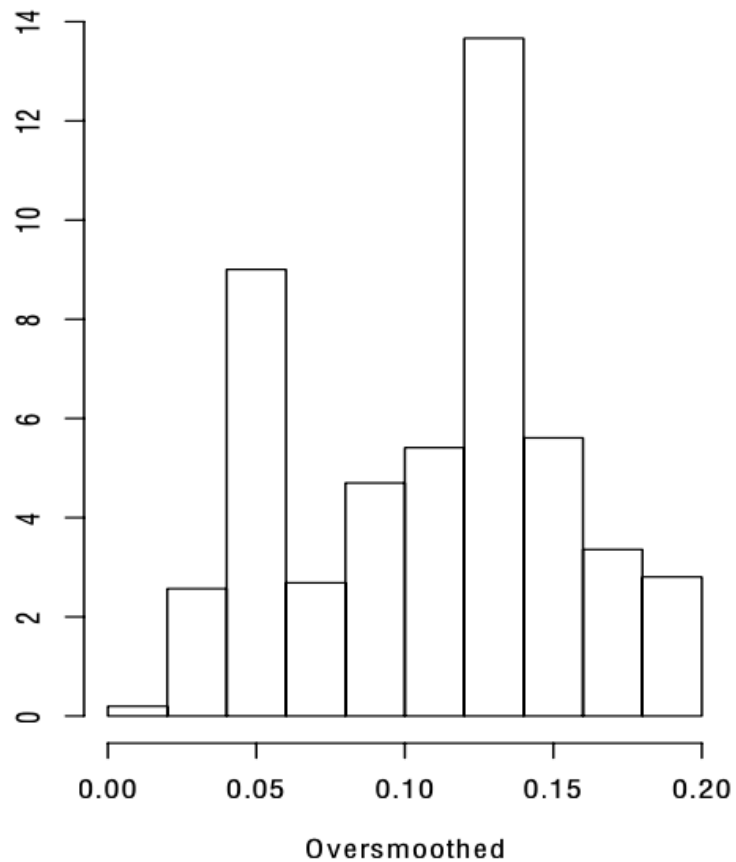
Let $X_1, \ldots, X_n$ be IID on $[0,1]$ with density f. Restricting to $[0,1]$ is not crucial, we can always rescale to this interval. Let m be the integer number of **bins**

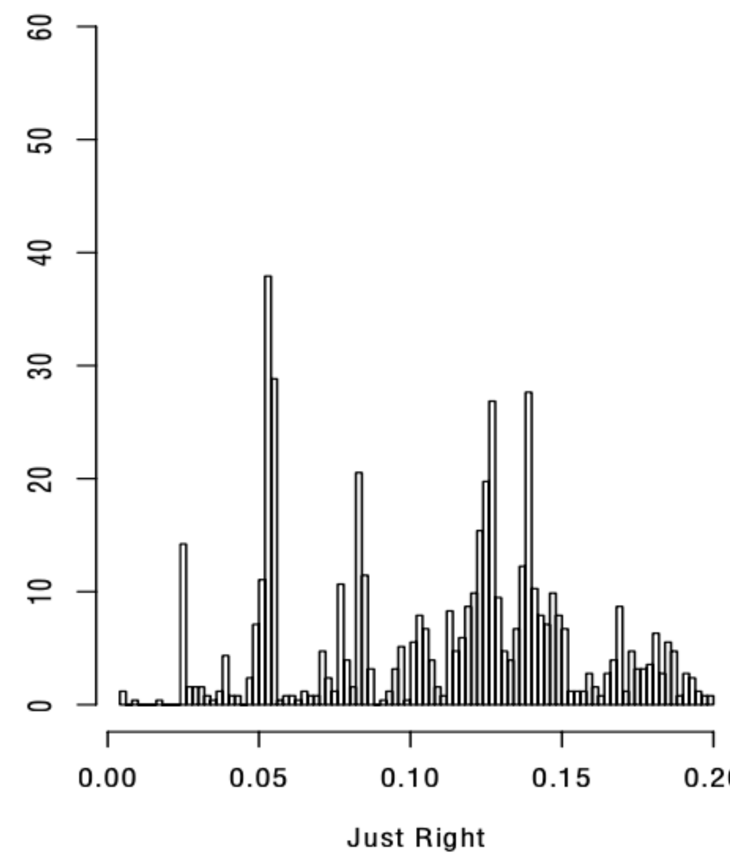
$$B_1 = \left[0, \frac{1}{m}\right), B_2 = \left[\frac{1}{m}, \frac{2}{m}\right), \ldots, B_m = \left[\frac{m-1}{m}, 1\right]$$

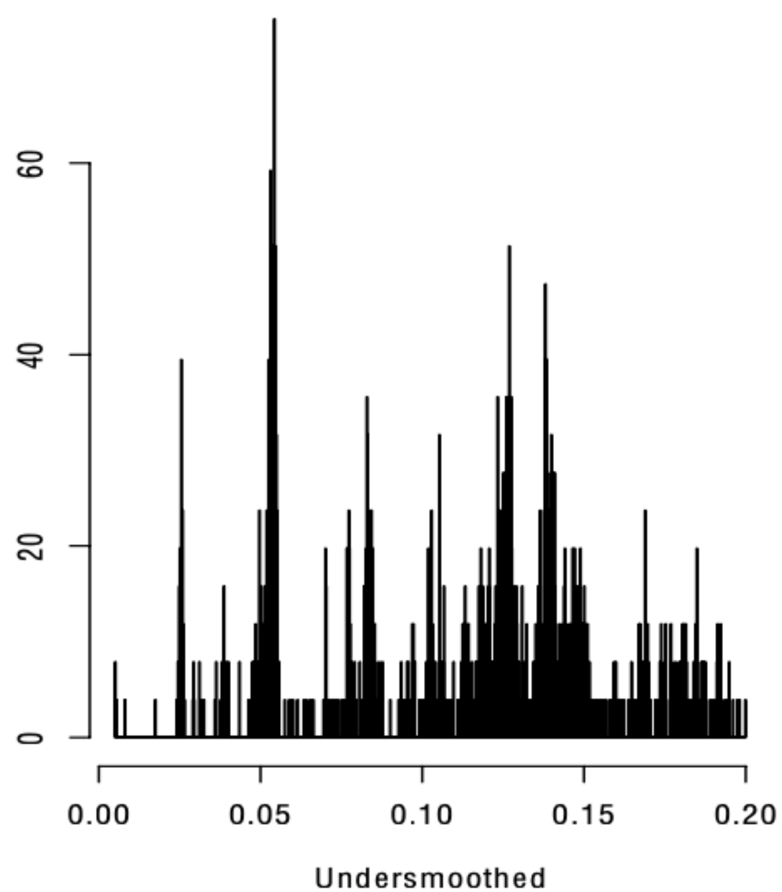
The **binwidth** $h = 1/m$, and ν_j denoting the # of observations in B_j, let $\widehat{p}_j = \nu_j/n$ and $p_j = \int_{B_j} f(u)\,du$. The **histogram estimator** is

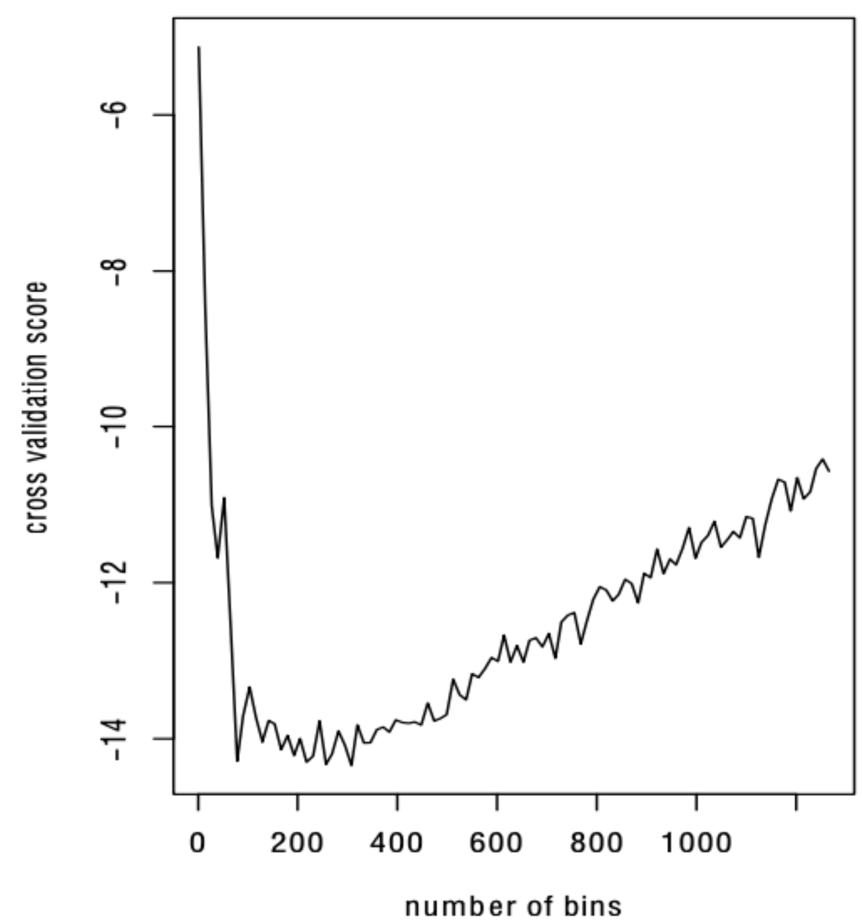
$\hat{f}_n(x) = \sum_{j=1}^{n} \frac{\widehat{p}_j}{h} I(x \in B_j)$. For small h, $\mathbb{E}\hat{f}_n(x) \approx f(x)$.

Histograms









What's this?

Histograms

Theorem: For fixed x and m, let B_j be the bin containing x, then

$$\mathbb{E}(\hat{f}_n(x)) = \frac{p_j}{h} \quad \text{and} \quad \mathbb{V}(\hat{f}_n(x)) = \frac{p_j(1-p_j)}{nh^2}$$

Doing Taylor expansion, one can also prove:

Theorem: Assuming $\displaystyle\int \left(f'(u)\right)^2 du < \infty$, one has

$$R(\hat{f}_n, f) \approx \frac{h^2}{12}\int \left(f'(u)\right)^2 du + \frac{1}{nh}, \text{ which is minimized by value}$$

$$h^* = \frac{1}{n^{1/3}}\left(\frac{6}{\int (f'(u))^2\, du}\right)^{1/3}, \text{ with which } R(\hat{f}_n, f) \approx \frac{C}{n^{2/3}}$$

Histograms

So with an optimally chosen bin-width, the MISE decreases at rate $n^{-2/3}$. Most parametric estimators converge at rate n^{-1}. This slower rate is the price we pay for being nonparametric. Formula for h^* is interesting, but not practical, since depends on unkownn f.

Recall the loss $L(h) = \int \hat{f}_n^2(x)\,dx - 2\int \hat{f}_n(x) f(x)\,dx + \int f^2(x)\,dx$,

where last term doesn't depend on h. So we minimise first two.

Definition: The **cross-validation estimator of risk** is $\hat{J}(h) = \int \left(\hat{f}_n(x)\right)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{(-i)}\left(X_i\right)$ where $\hat{f}_{(-i)}$ is the histogram with i-th observation removed.

Histograms

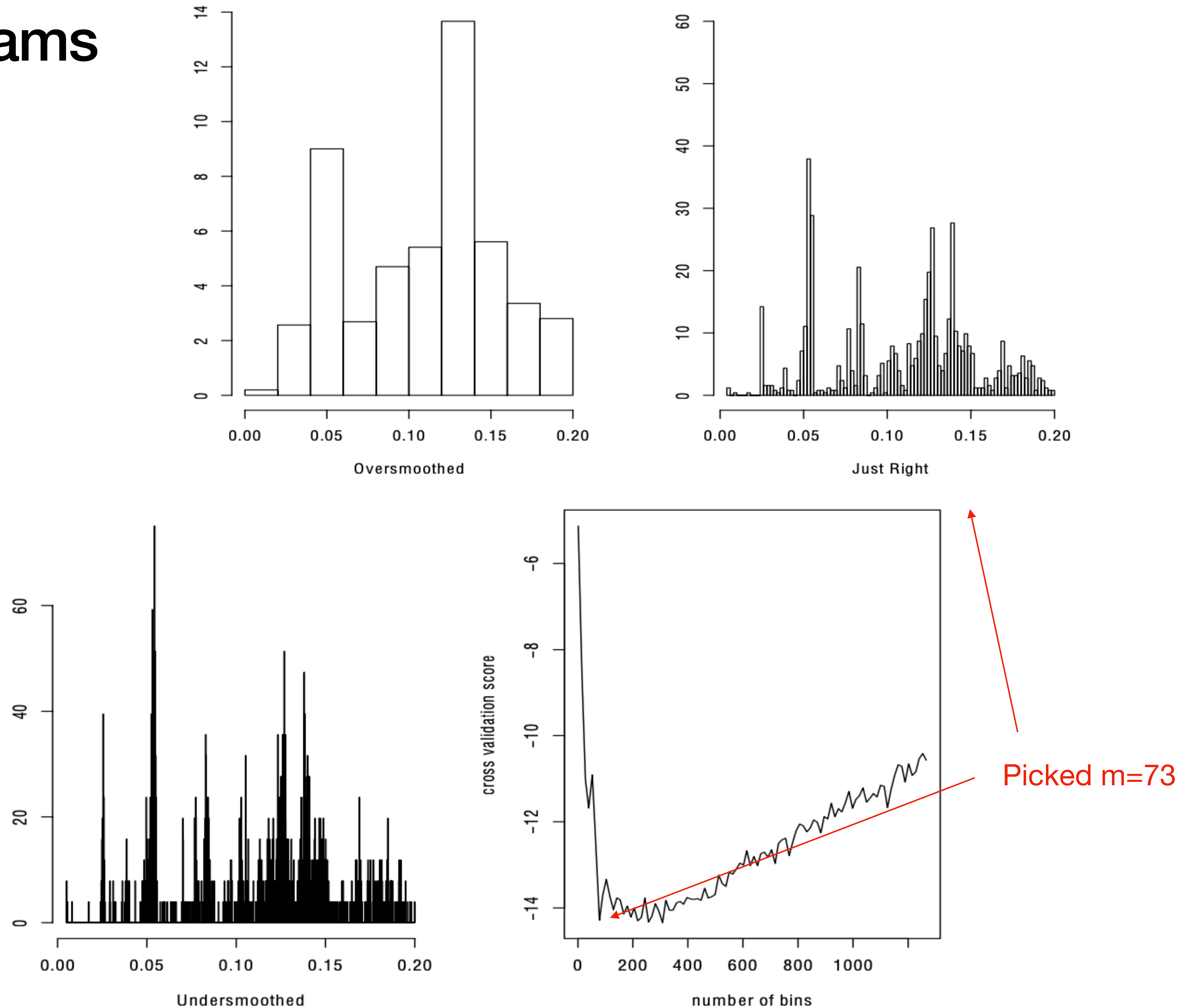
Theorem: That estimator is nearly unbiased: $\mathbb{E}(\widehat{J}(x)) \approx \mathbb{E}(J(x))$.

With those $\hat{f}_{(-i)}$-s, we have to recompute the histogram n times – for all values of h. This is not very practical, and there's a shortcut:

Theorem: $\widehat{J}(h) = \dfrac{2}{(n-1)\,h} - \dfrac{n+1}{n-1} \sum_{j=1}^{m} \widehat{p}_j^{\,2}$

On our exemplar plot, the minimum of cross-validation estimator is quite flat. The "optimal" histogram was constructed using $m = 73$ bins.

Histograms



Oversmoothed
Just Right
Undersmoothed
cross validation score
number of bins
Picked m=73

Histograms

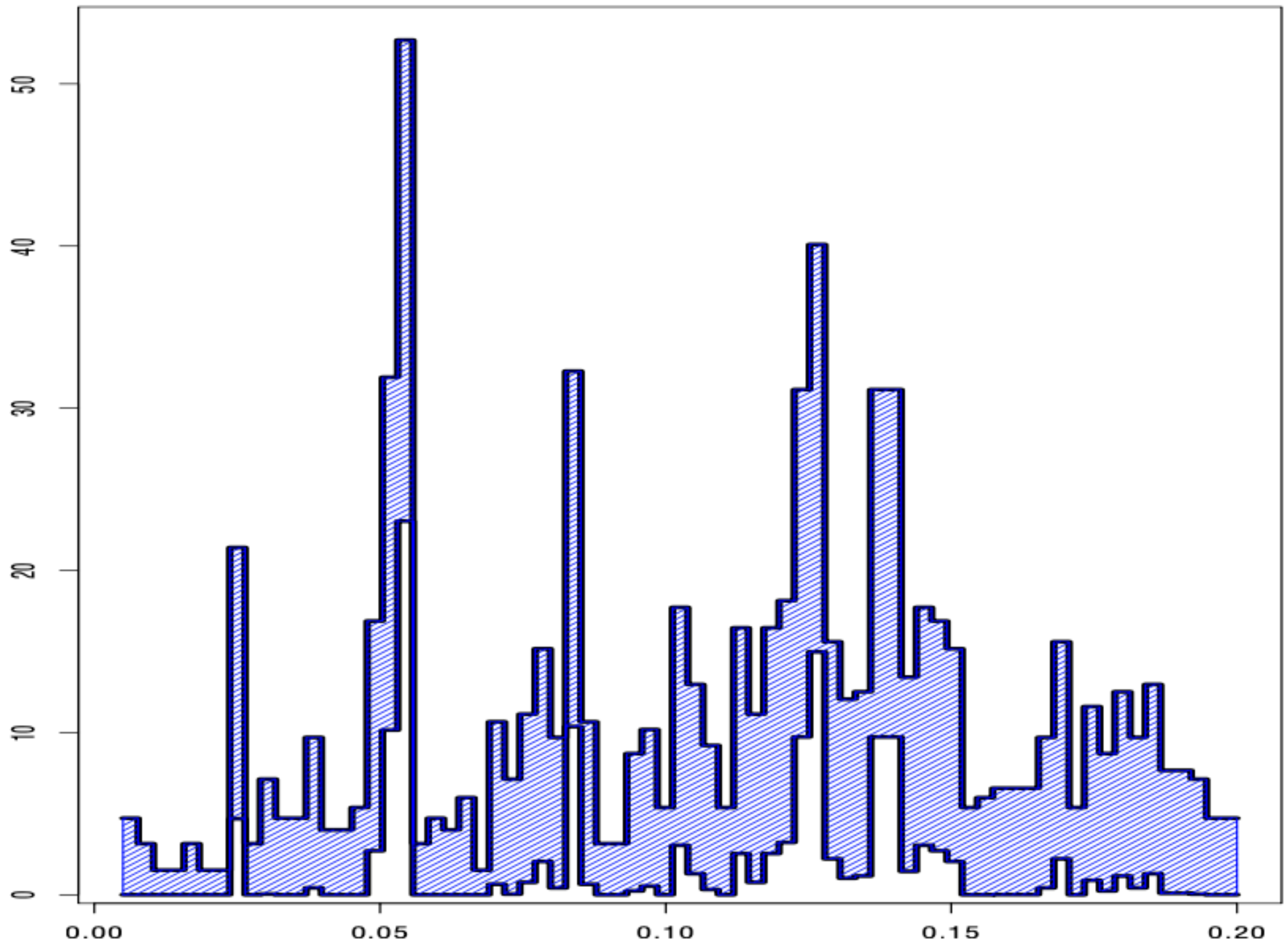
How about confidence intervals?

Theorem: The $(1-\alpha)$-confidence interval for a histogram is (assuming $m \to \infty$ and $m(n)\log(n)/n \to 0$ as $n \to \infty$) from

$$\ell_n(x) = \left(\max\left\{\sqrt{\hat{f}_n(x)} - c, 0\right\}\right)^2$$ – lower bound, to

$$u_n(x) = \left(\sqrt{\hat{f}_n(x)} + c\right)^2$$ – upper bound.

Histograms



95%-confidence interval for our data

Kernel Density Estimators (KDEs)

Kernel Density Estimators

Histograms are discontinuous. **Kernel density estimators** are smoother and they converge faster to the true density.

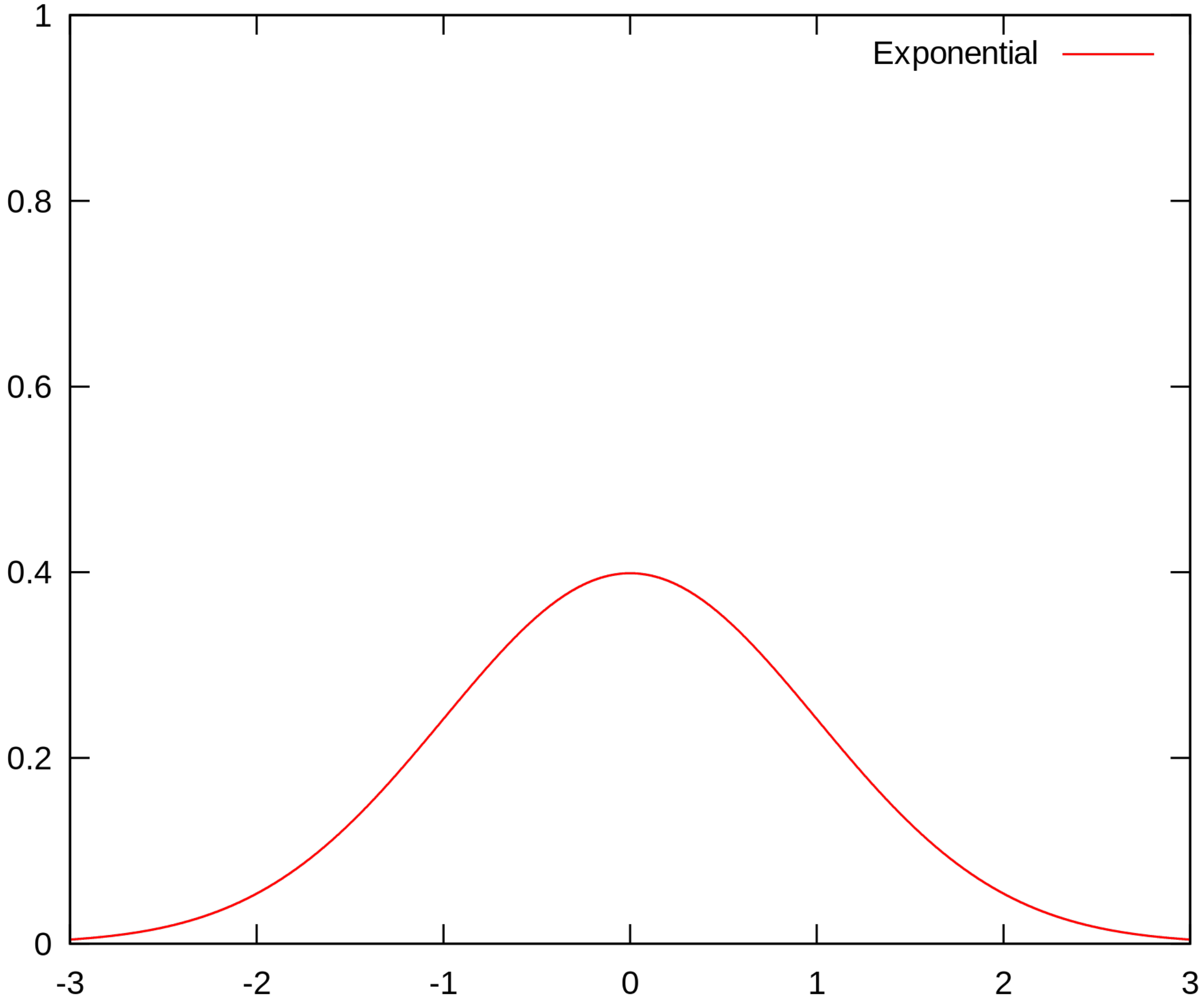
A **kernel** is any smooth function K such that 1) $K(x) \geq 0$, 2) $\displaystyle\int K(x)\,dx = 1$, 3) $\displaystyle\int xK(x)\,dx = 0$ 4) $\sigma_K^2 = \displaystyle\int x^2 K(x)\,dx > 0$.

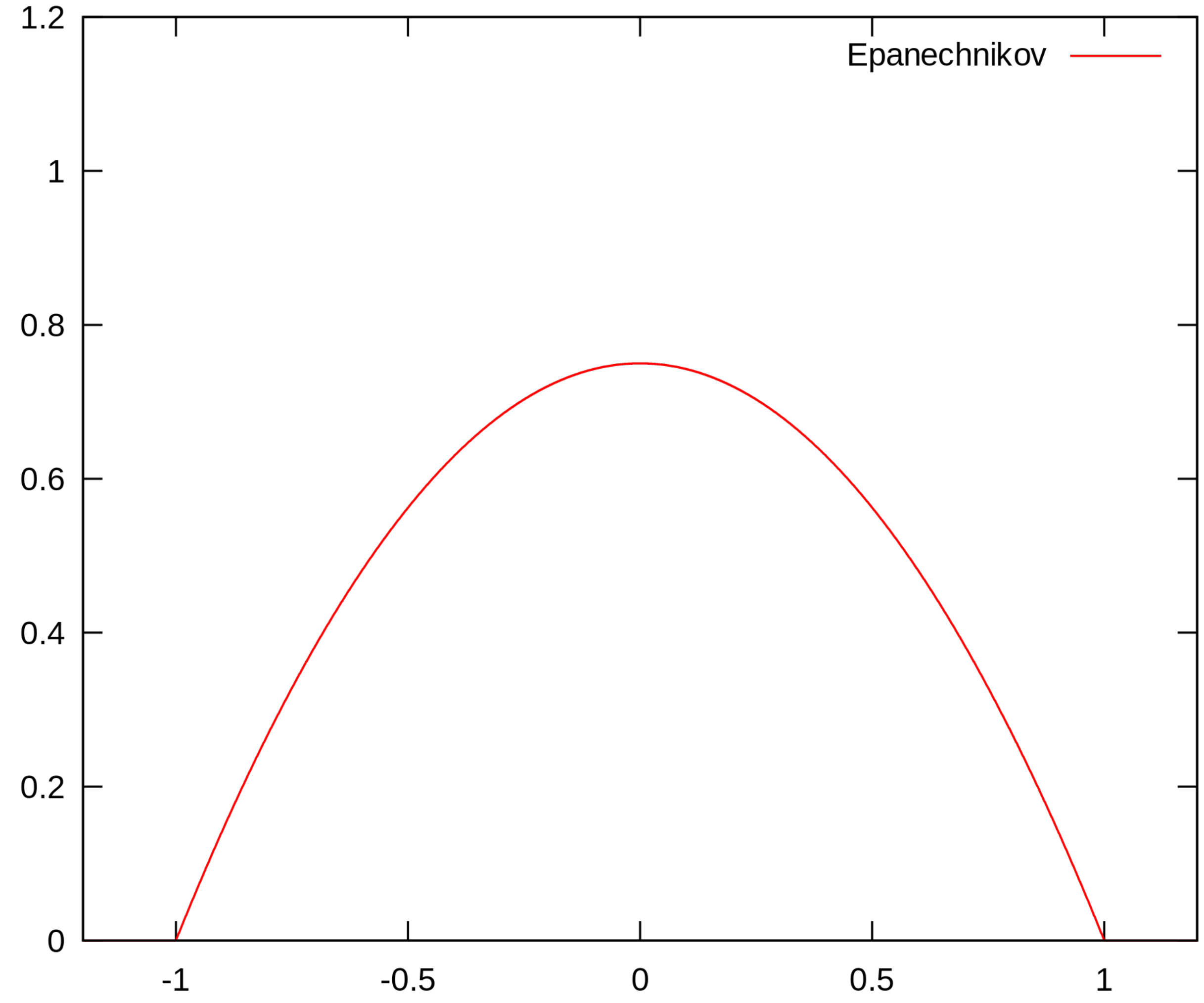
Two important kernels are Gaussian (normal) $K(x) = (2\pi)^{-1/2} e^{-x^2/2}$ and **Epanechnikov**

$$K(x) = \begin{cases} \frac{3}{4}(1 - x^2/5)/\sqrt{5}, & |x| < \sqrt{5} \\ 0 & \text{otherwise} \end{cases}$$

(there are many others, see Kernel page on Wiki)

Kernel Density Estimators





In some sense, Epanechikov kernel is the best finite kernel

Kernel Density Estimators

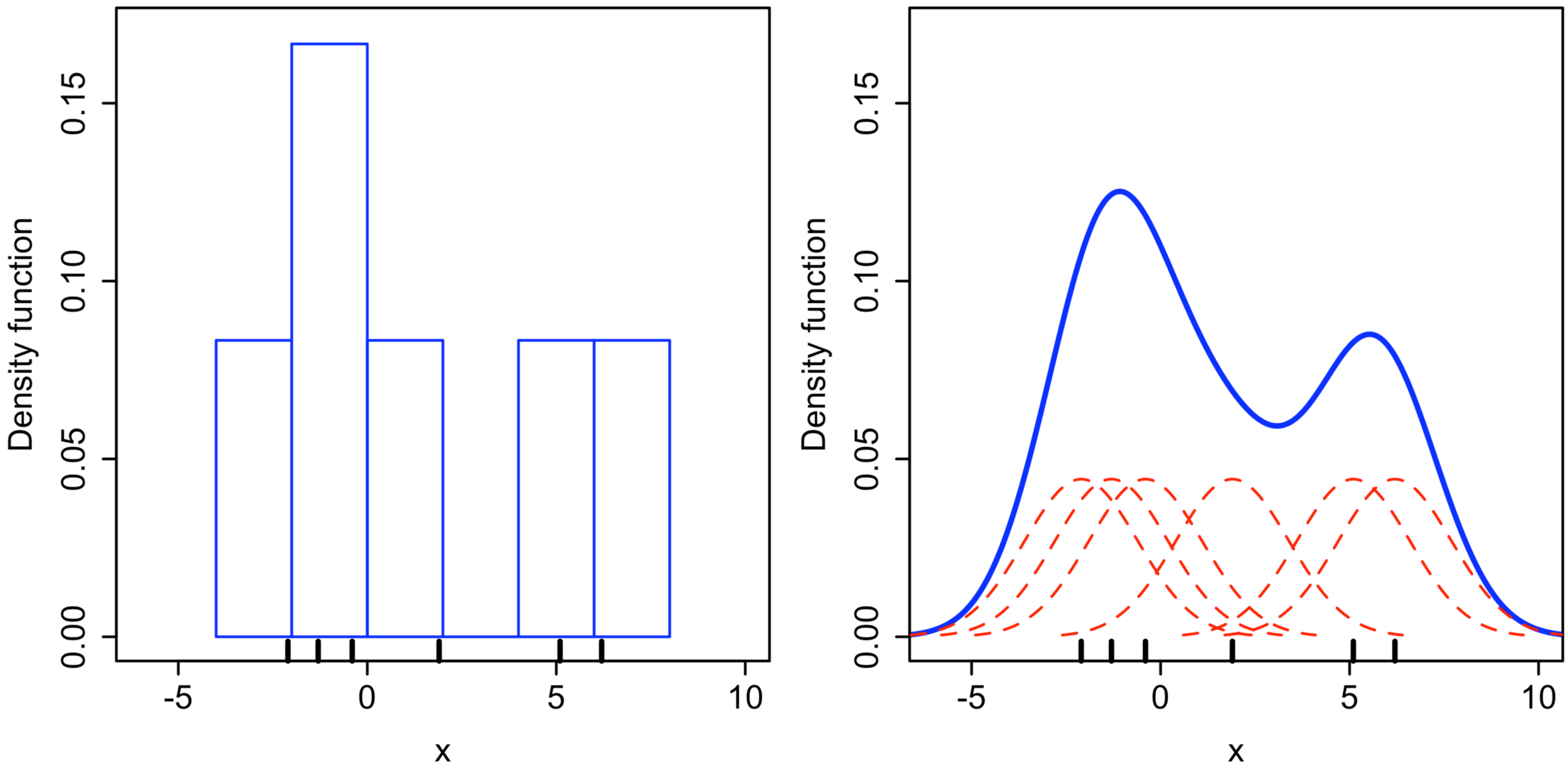
Definition: Given a kernel K and a positive number h, called the **bandwidth**, the **kernel density estimator** is

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

So it puts a smooth lump of mass $1/n$ at each data point X_i. Bandwidth parameter h controls the amount of smoothing – when $h \to 0$, we get infinitely-high spikes of zero width at each data point. When $h \to \infty$, we get uniform density.

Choosing K is arguably not as important as properly choosing h.

Kernel Density Estimators



Kernel Density Estimators

Theorem: Under some weak assumptions on f and K,

$R(f, \hat{f}_n) \approx \frac{1}{4}\sigma_K^4 h^4 \int \left(f''(x)\right)^2 + \frac{\int K^2(x)dx}{nh}$ where

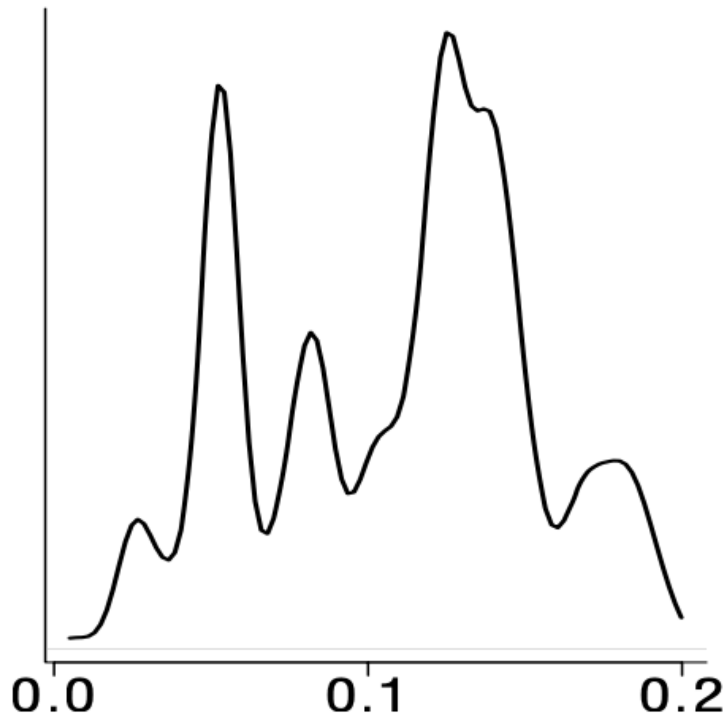
$\sigma_K^2 = \int x^2 K(x)\, dx$. Optimal bandwidth $h^* = \frac{c_1^{-2/5} c_2^{1/5} c_3^{-1/5}}{n^{1/5}}$ where

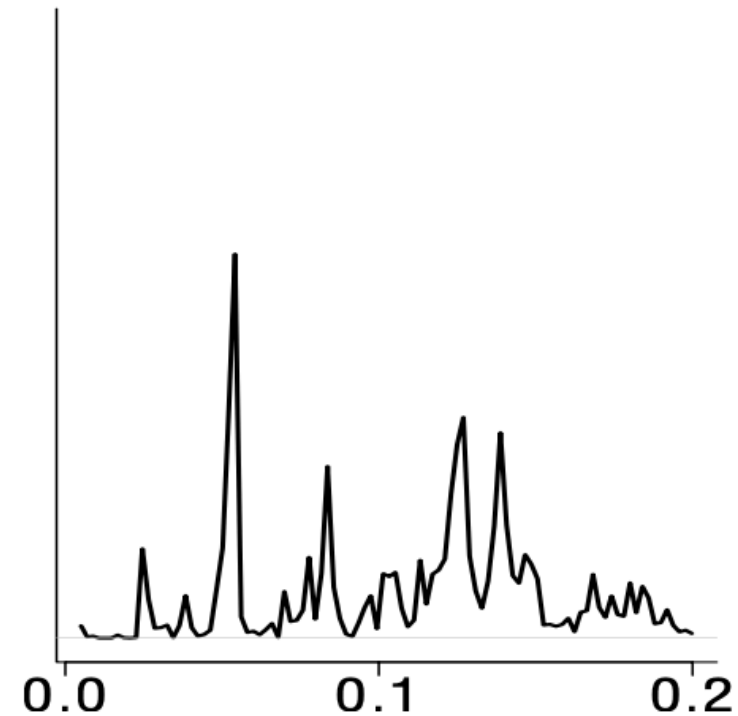
$c_1 = \int x^2 K(x)\, dx$, $c_2 = \int K^2(x)\, dx$ and $c_3 = \int (f''(x))^2\, dx$. With

this choice of bandwidth, $R(f, \hat{f}_n) \approx \frac{c_4}{n^{4/5}}$ with some $c_4 > 0$.

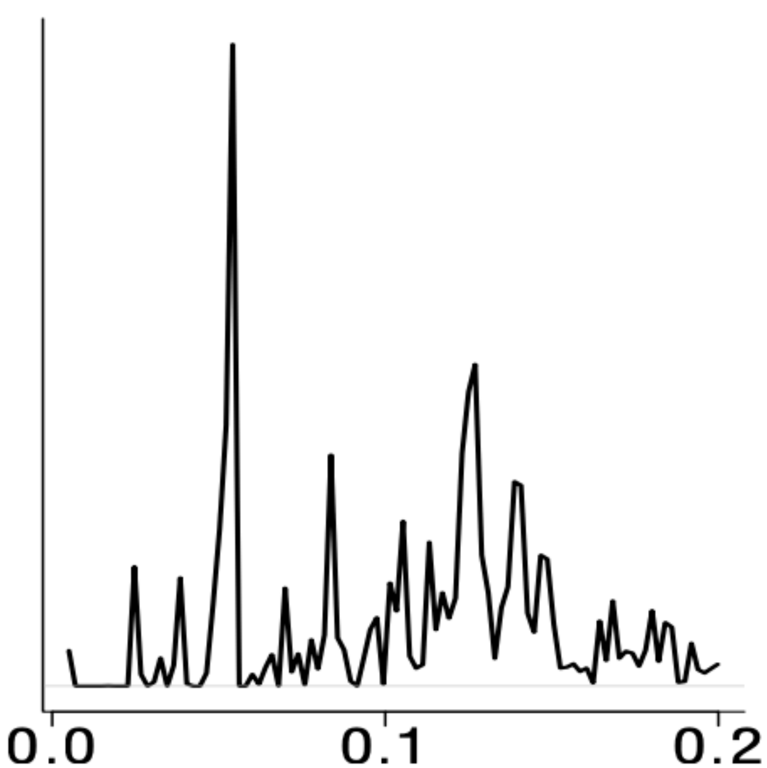
So KDEs converge at rate $n^{-4/5}$, while histograms – at slower $n^{-2/3}$. It can be shown, with weak assumptions, that **no** nonparametric estimator converges faster than $n^{-4/5}$!

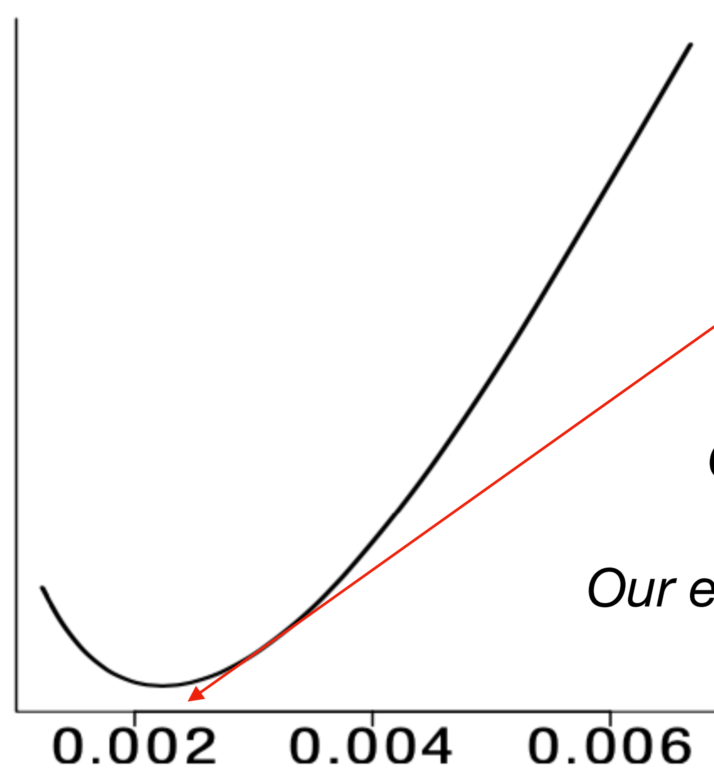
Kernel Density Estimators





Just right h





Optimal-bandwidth KDE may seem "wiggly", but that's ok! Our eyes are not the best judges here :)

Kernel Density Estimators

As with histograms, optimally choosing bandwidth = minimizing the risk.

Theorem: For any $h > 0$, $\mathbb{E}(\widehat{J}(h)) = \mathbb{E}(J(h))$. Also,

$$\widehat{J}(h) \approx \frac{1}{hn^2} \sum_i \sum_j K^* \left(\frac{X_i - X_j}{h} \right) + \frac{2}{nh} K(0) \text{ where}$$

$$K^*(x) = K^{(2)}(x) - 2K(x) \text{ and } K^{(2)}(z) = \int K(z-y)K(y)\, dy.$$

Particularly, if $K(x) \equiv \mathcal{N}(0,1)$, then $K^{(2)}(x) \equiv \mathcal{N}(0,2)$.

FFT helps computing this!

Kernel Density Estimators

A remarkable (Stone's) **Theorem:** Suppose f is bounded, and $\hat{f}_h$ is the KDE with bandwidth h, h_n being optimal h from cross-validation.

Then $$\frac{\int \left(f(x) - \hat{f}_{h_n}(x)\right)^2 dx}{\inf_h \int \left(f(x) - \hat{f}_h(x)\right)^2 dx} \xrightarrow{P} 1$$

Kernel Density Estimators

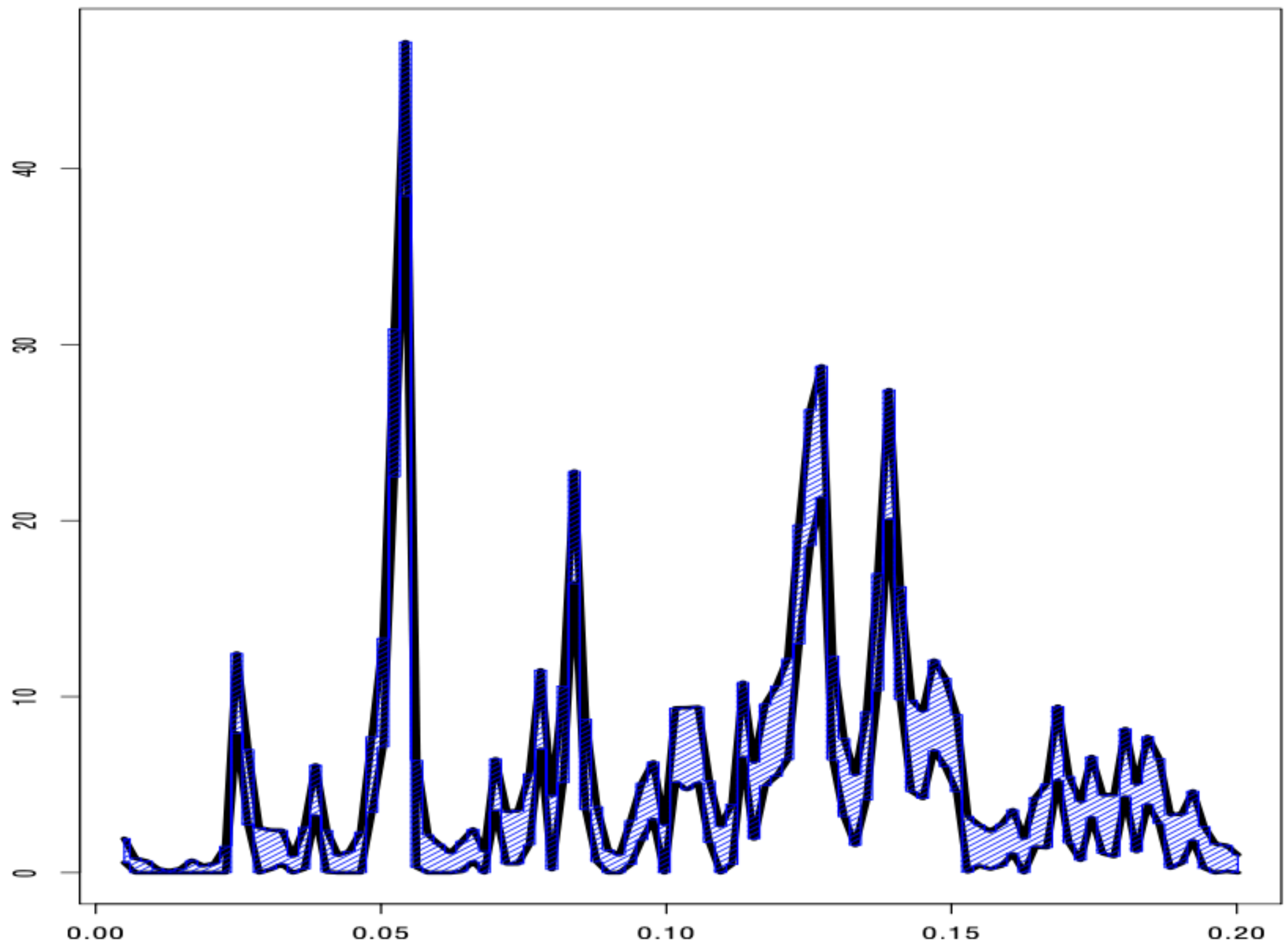
Confidence intervals again!

$$\ell_n(x) = \hat{f}_n(x) - q\,\mathrm{se}(x), \quad u_n(x) = \hat{f}_n(x) + q\,\mathrm{se}(x)$$

$$\mathrm{se}(x) = \frac{s(x)}{\sqrt{n}}, \quad s^2(x) = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(x) - \overline{Y}_n(x))^2,$$

$$Y_i(x) = \frac{1}{h}K\left(\frac{x - X_i}{h}\right), \quad q = \Phi^{-1}\left(\frac{1 + (1-\alpha)^{1/m}}{2}\right)$$

Kernel Density Estimators



95%-confidence interval for our data

Kernel Density Estimators

Curse of dimensionality: KDEs can be generalised to arbitrary dimension (all you need is a multivariate version of the kernel).

Optimal bandwidth would be $h \sim n^{-1/(4+d)}$, the risk would be $\sim n^{-4/(4+d)}$ – quickly increases with dimension.

Consider the following (Silverman, 1986) table: sample size required to ensure RMSE less than 0.1 at 0 (density is multivar. normal) with optimal bandwidth selected:

Dimension	Sample Size
1	4
2	19
3	67
4	223
5	768
6	2790
7	10,700
8	43,700
9	187,000
10	842,000