

CSE 6740 Lecture 4

How Do I Learn Any Density? (Nonparametric Estimation)

Alexander Gray

agray@cc.gatech.edu

Georgia Institute of Technology

Today

1. Nonparametric estimation (*What if I don't want to specify a simple parametric form?*)
2. Kernel density estimation (*How can I estimate a density nonparametrically?*)

Nonparametric Estimation

What if I don't want to specify a simple parametric form?

Nonparametric Estimation

What exactly do we mean by “nonparametric”? Example of a nonparametric model class, called a *Sobolev space*:

$$\mathcal{F} = \left\{ f : \int (f''(x))^2 dx < \infty \right\} \quad (1)$$

“Nonparametric” doesn’t mean there are no parameters. There is typically a local “model”. It refers to model classes, like the one above, which aren’t parametric (having finite number of parameters). We sometimes say such a class is *distribution-free*.

Nonparametric Estimation

A *nonparametric method* is one for which we can pretend the model class is actually such a class, as far as its asymptotic properties.

In other words, it is a method for which one can show something like consistency with respect to a very general class of distributions (we want to say “any distribution” but this is of course never quite true).

Examples of Nonparametric Methods

Some examples of popular nonparametric methods:

- Histogram, kernel density estimation (density estimation)
- Splines, wavelet regression (regression)
- Kernel discriminant analysis, nearest neighbor, support vector machines (classification)

Histogram

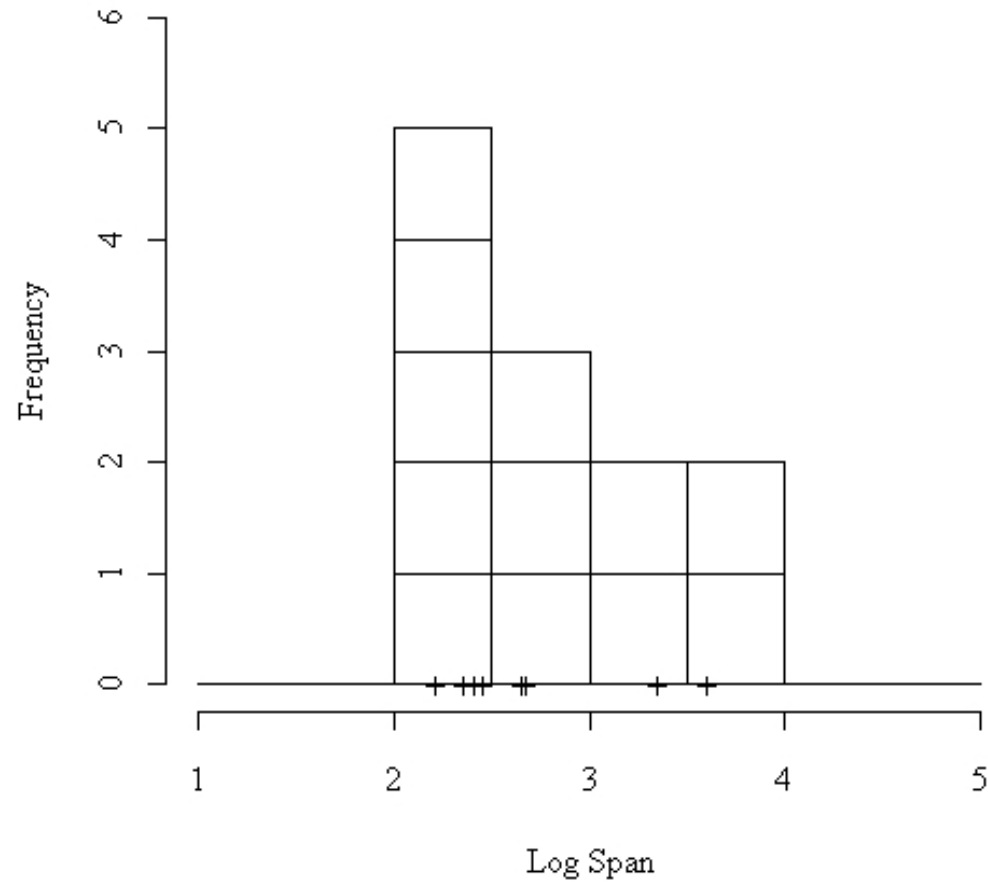
Perhaps the simplest nonparametric density estimator is the *histogram*:

$$\hat{f}_N(x) = \sum_{j=1}^m \frac{\hat{p}_j}{h} I(x \in B_j) \quad (2)$$

where $h = 1/m$ is the *binwidth*, Y_j is the number of observations in bins $B_1 = [0, \frac{1}{m})$, $B_2 = [\frac{1}{m}, \frac{2}{m})$, \dots , $\hat{p}_j = Y_j/N$, and $p_j = \int_{B_j} f(u)du$.

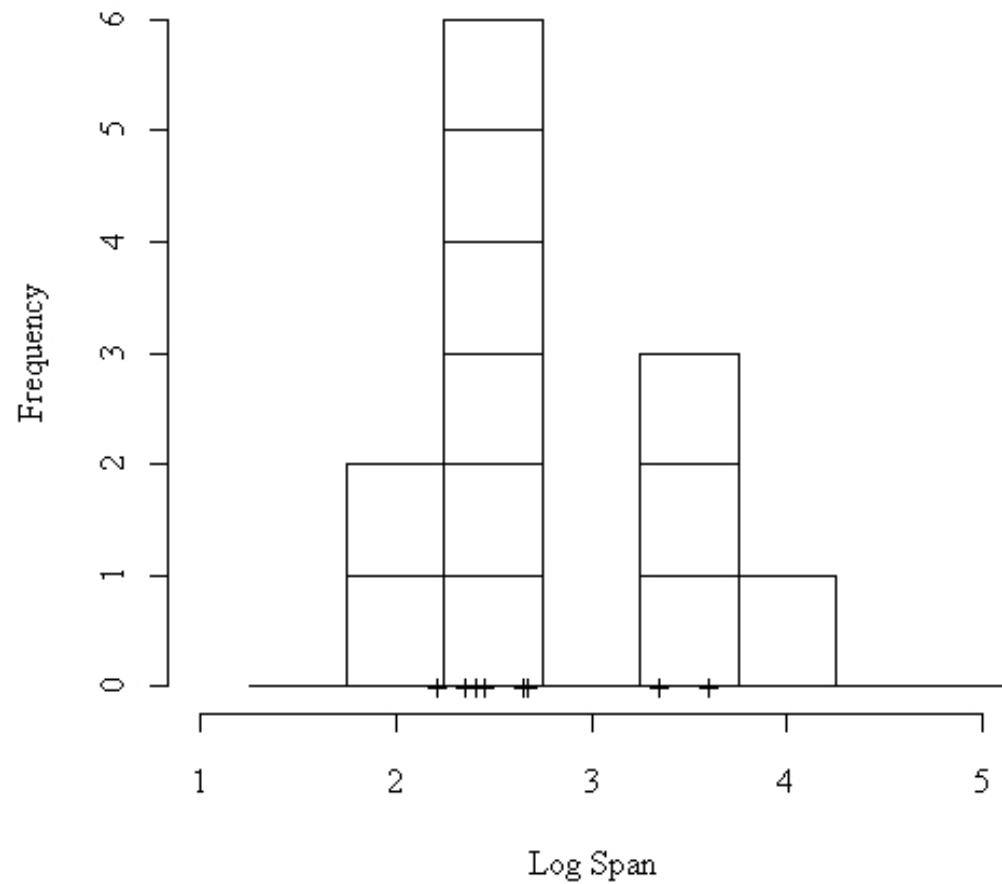
Histogram

**Histogram with breaks at n.0 and n.5
binwidth=0.5**



Histogram

**Histogram with breaks at n.25 and n.75
binwidth=0.5**



Histogram

Note a few things. First, the placement of the bins (*i.e.* shifting a bit to the left or right) can make a significant qualitative difference. Second, the density estimate is not smooth.

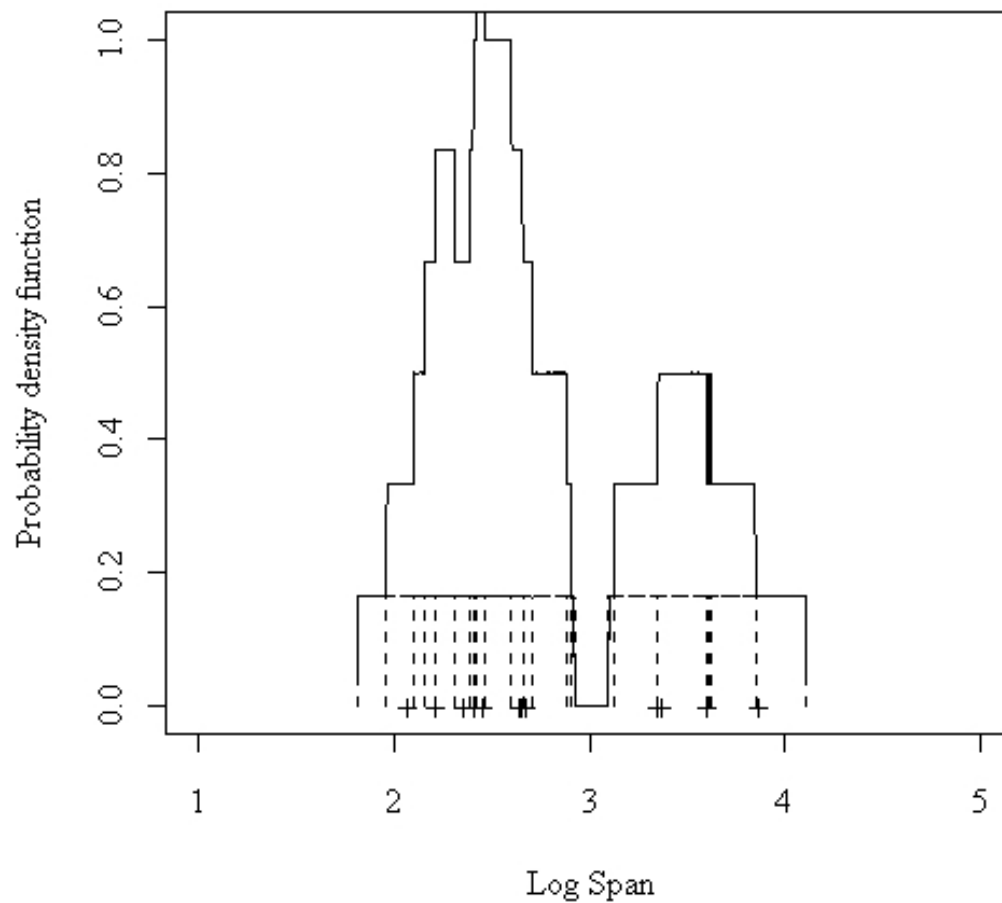
Nonetheless, we can show that $\mathbb{E}(\hat{f}_N(x)) \approx f(x)$, under certain conditions. Remarkable, but we can do better.

Kernel Density Estimation

How can I estimate a density nonparametrically?

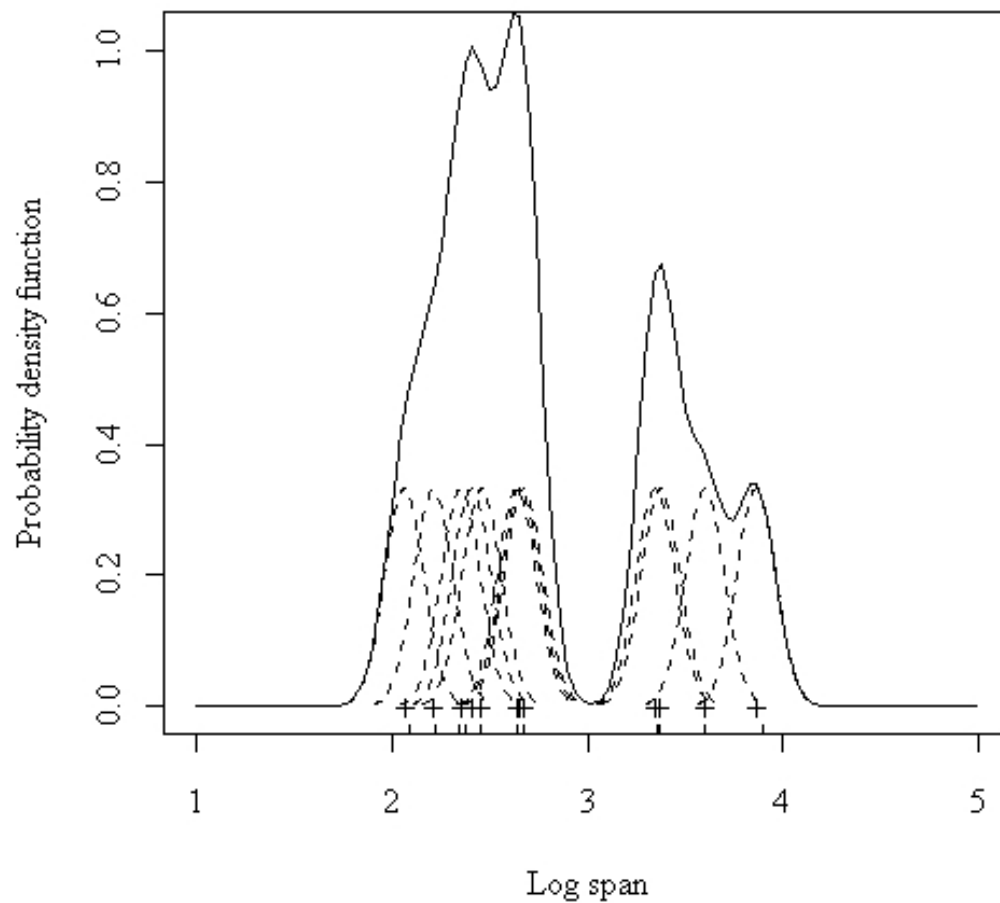
Kernel Density Estimator

'Histogram' with blocks centred over data points



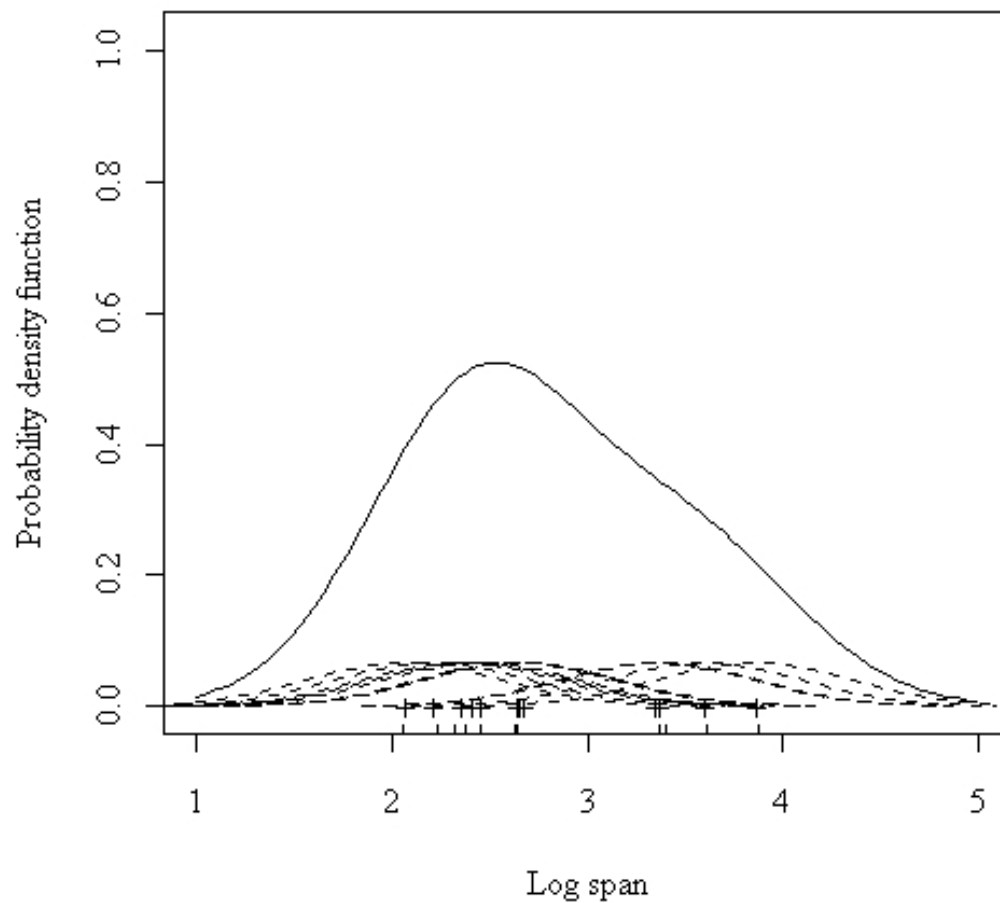
Kernel Density Estimator

Undersmoothed



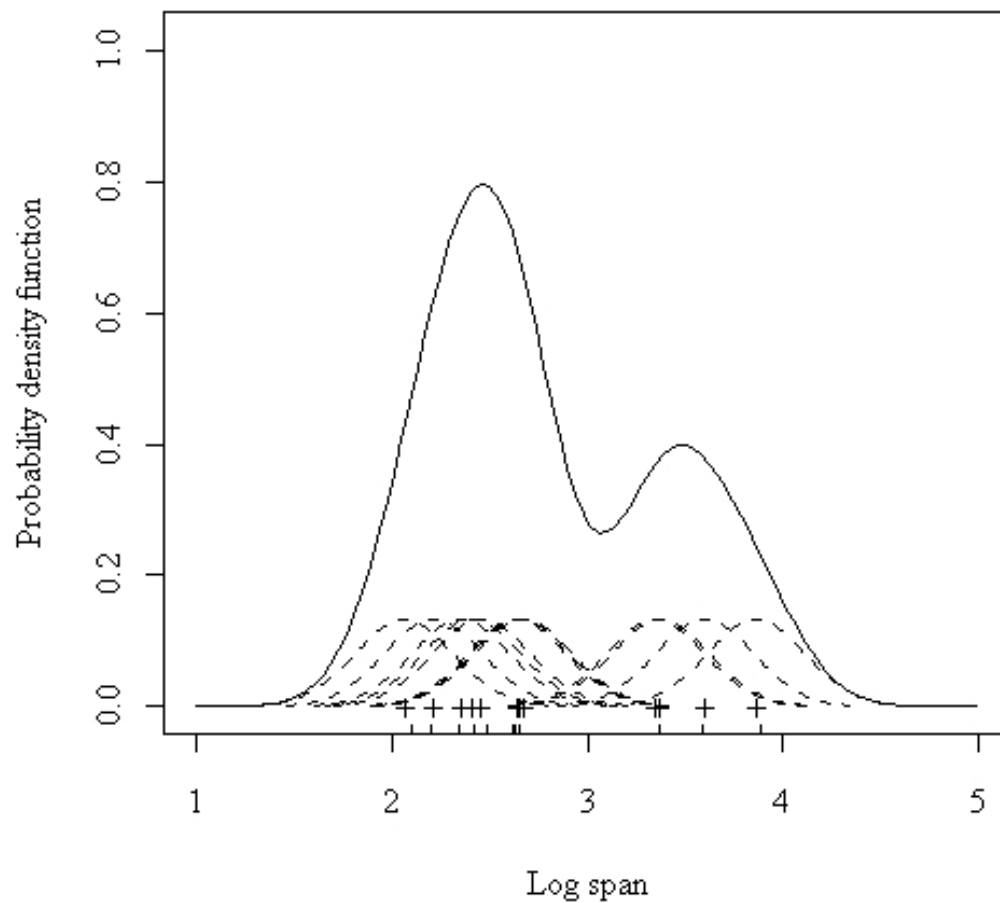
Kernel Density Estimator

Oversmoothed



Kernel Density Estimator

Optimally smoothed



Histogram Versus KDE

By centering the blocks on each data point, and generalizing to a smooth kernel function from the block (which we'll now call the *rectangular kernel*), we have a more satisfying density estimator, called the *kernel density estimator* (KDE). We also saw both histograms and KDE *have a parameter*, the kernel width, and that its proper choice is critical.

We will be able to show asymptotically that KDE is a better estimator of the density than the histogram, and be able to specify a procedure for choosing the optimal kernel width in KDE (as well as the optimal bin width for a histogram).

Kernel Density Estimator

The kernel density estimator is defined as

$$\hat{f}_N(x) = \frac{1}{N} \sum_{i=1}^N \frac{1}{h} K\left(\frac{x - X_i}{h}\right), \quad (3)$$

where the *kernel function* is any smooth function K such that $K(u) \geq 0$, $\int K(u)du = 1$, $\int uK(u)du = 0$, and $\sigma_K^2 = \int u^2 K(u)du > 0$, and its parameter h is called the *bandwidth*.

An example is the Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$.

Kernel Density Estimator

The parts of KDE:

Task: density estimation

Model class: Sobolev space

Loss: L_2 error

Optimizer: exhaustive or gradient descent

Generalization mechanism: cross-validation

Evaluation algorithm: *generalized N -body algorithm*

Note that we have changed from the likelihood, which we used in our parametric example, mixtures of K Gaussians, to L_2 error. We will return to the reason for this. Also note the need for a fast evaluation algorithm, which we will discuss in a later lecture.

L_2 : MSE, MISE

Suppose $\hat{f}_N(x)$ is an estimate of a function $f(x)$. The *squared error* (or L_2) loss is

$$L(f(x), \hat{f}_N(x)) = (f(x) - \hat{f}_N(x))^2. \quad (4)$$

The average of any loss is called the *risk* or in this case the *mean squared error*

$$\text{MSE} = R(f(x), \hat{f}_N(x)) = \mathbb{E}L(f(x), \hat{f}_N(x)). \quad (5)$$

To summarize the risk over all values of x , we use the *integrated risk* or in this case the *mean integrated squared error* (MISE):

$$\text{MISE} = R(f, \hat{f}_N) = \int R(f(x), \hat{f}_N(x)) dx. \quad (6)$$

L_2 : Bias-Variance Tradeoff

For L_2 loss we have a convenient decomposition (dropping reference to x and N for the moment):

$$\begin{aligned} & \mathbb{E} \left(f - \hat{f} \right)^2 \\ = & \mathbb{E} \left(f - \mathbb{E}\hat{f} + \mathbb{E}\hat{f} - \hat{f} \right)^2 \\ = & \mathbb{E} \left(\left(f - \mathbb{E}\hat{f} \right) + \left(\mathbb{E}\hat{f} - \hat{f} \right) \right)^2 \\ = & \mathbb{E} \left(f - \mathbb{E}\hat{f} \right)^2 + \mathbb{E} \left(\mathbb{E}\hat{f} - \hat{f} \right)^2 + 2\mathbb{E} \left(\left(\mathbb{E}\hat{f} - \hat{f} \right) \left(f - \mathbb{E}\hat{f} \right) \right) \quad (1) \\ = & \text{bias}^2(\hat{f}) + \mathbb{V}(\hat{f}) + 2 \left(\mathbb{E} \left(f\mathbb{E}\hat{f} \right) - \mathbb{E} \left(\mathbb{E}\hat{f}^2 \right) - \mathbb{E} f \hat{f} + \mathbb{E} \left(\hat{f}\mathbb{E}\hat{f} \right) \right) \\ = & \text{bias}^2(\hat{f}) + \mathbb{V}(\hat{f}) + 2 \left(f\mathbb{E}\hat{f} - \mathbb{E}\hat{f}^2 - f\mathbb{E}\hat{f} + \mathbb{E}\hat{f}^2 \right) \quad (1) \\ = & \text{bias}^2(\hat{f}) + \mathbb{V}(\hat{f}) \quad (1) \end{aligned}$$

L_2 : Bias-Variance Tradeoff

...by just using a “completing the square” trick and the properties of expectation. Back to our original notation, we have that

$$R(f(x), \hat{f}_N(x)) = \text{bias}^2(\hat{f}_N(x)) + \mathbb{V}(\hat{f}_N(x)) \quad (14)$$

where

$$\text{bias}(\hat{f}_N(x)) = \mathbb{E}(\hat{f}_N(x)) - f(x). \quad (15)$$

KDE Consistency

Assume that f is continuous at x and that $h_N \rightarrow 0$ and $Nh_N \rightarrow \infty$ as $N \rightarrow \infty$. Then

$$\hat{f}_N(x) \xrightarrow{p} f(x) \quad \text{as } N \rightarrow \infty. \quad (16)$$

Note that the bandwidth must shrink as we get more data, but not go to zero as rapidly as $1/N$, *i.e.* the expected number of points falling in the interval $x \pm h_N$ must tend to infinity, however slowly, as N tends to infinity.

A stronger notion, *uniform consistency*, can also be shown, under some conditions that are only slightly stronger:

$$\sup_x \left| \hat{f}_N(x) - f(x) \right| \xrightarrow{p} 0 \quad \text{as } N \rightarrow \infty. \quad (17)$$

KDE Risk

Now, we're going to come up with a detailed expression for the risk of a kernel density estimator. This will let us conclude some important things.

Let $R(x) = \mathbb{E}(\hat{f}_N(x) - f(x))^2$ be the risk at a point x and $R = \int R(x)dx$ denote the integrated risk. Assume that f'' is absolutely continuous and that $\int (f''(x))^2 dx < \infty$.

Also recall our assumptions about K . We'll write $K_h(x, X) = \frac{1}{h}K((x - X)/h)$.

KDE Risk

For our estimator $\hat{f}_N(x) = \frac{1}{N} K_h(x, X)$ we have

$$\mathbb{E}(\hat{f}_N(x)) = \mathbb{E}(K_h(x, X)) \quad (18)$$

$$= \int \frac{1}{h} K\left(\frac{x-t}{h}\right) f(t) dt \quad (19)$$

$$= \int K(u) f(x-hu) du \quad (20)$$

$$= \int K(u) \left[f(x) - hu f'(x) + \frac{1}{2} h^2 u^2 f''(x) + \dots \right] du \quad (21)$$

$$= f(x) + \frac{1}{2} h^2 f''(x) \int u^2 K(u) du + \dots \quad (22)$$

since $\int K(x) dx = 1$ and $\int x K(x) dx = 0$.

KDE Risk

Then the bias is

$$\mathbb{E}(\hat{f}_N(x)) - f(x) = \frac{1}{2}\sigma_K^2 h_N^2 f''(x) + O(h_N^4). \quad (23)$$

By a similar calculation, the variance is

$$\mathbb{V}(\hat{f}_N(x)) = \frac{f(x) \int K^2(x) dx}{N h_N} + O\left(\frac{1}{N}\right). \quad (24)$$

Putting them together we have

$$R = \frac{1}{4}\sigma_K^4 h_N^4 (f''(x))^2 + \frac{f(x) \int K^2(x) dx}{N h_N} + O\left(\frac{1}{N}\right) + O(h_N^4). \quad (25)$$

KDE Optimal Bandwidth

If we differentiate the risk with respect to h and set it equal to 0, we see that the asymptotically optimal bandwidth is

$$h^* = \left[\left(\frac{\int K(x)^2 dx}{\left(\int x^2 K(x) dx \right)^2 \int (f''(x))^2 dx} \right) \frac{1}{N} \right]^{1/5}. \quad (26)$$

So the best bandwidth decreases at rate $N^{-1/5}$.

Effectively balances bias and variance.

KDE Convergence Rate

Now if we plug h^* into the risk, we see that if the optimal bandwidth is used then $R = O(N^{-4/5})$.

It can be shown that histograms converge at rate $O(N^{-2/3})$.

It turns out there does not exist a density estimator that converges faster than $O(N^{-4/5})$.

Convergence Rate Lower Bound

Let \mathcal{F} be the set of all PDF's and let $f^{(m)}$ denote the m^{th} derivative of f . Define

$$\mathcal{F}_m(c) = \left\{ f \in \mathcal{F} : \int |f^{(m)}(x)|^2 dx \leq c^2 \right\}. \quad (27)$$

For any estimator \hat{f}_N ,

$$\sup_{f \in \mathcal{F}_m(c)} \mathbb{E}_f \int (\hat{f}_N(x) - f(x))^2 dx \geq b \left(\frac{1}{N} \right)^{2m/(2m+1)} \quad (28)$$

where $b > 0$ is a universal constant that depends only on m and c . Plugging in $m = 2$ yields $O(N^{-4/5})$.

KDE Optimal Kernel Function

We can also see what kernel function K minimizes the risk. If we do this we obtain this kernel:

$$K(x) = \frac{3}{4} (1 - x^2) I(|x| \leq 1). \quad (29)$$

This is called the *Epanechnikov kernel*.

Thus the Gaussian kernel is not the optimal kernel.

In practice the choice of bandwidth is much more important than the choice of kernel function.

Main Things You Should Know

- What it means to be nonparametric
- What kernel density estimation (KDE) and its parts are
- What asymptotic properties we can show about KDE
- What the bias-variance tradeoff is

Sample Final Questions

1. (T/F) A histogram has no parameters.
2. (T/F) Cross-validation does not apply to density estimation, since density estimation is an unsupervised task.
3. (T/F) The Gaussian kernel is the asymptotically optimal kernel function.