# STAT 460/560 Class 13: Non-parametric curve estimation

Ben Bloem-Reddy

**Reading:** Ch. 20, [Was04]. (I'm also referring to [Was06] for more detailed exposition of certain results; the corresponding sections there are easy to find.)

We'll study nonparametric curve estimation today. Although it can be used for many things, we'll encounter it in the two most widely used settings. The basic idea is that there is an unknown function $f(x)$ that we want to estimate from data. In density estimation, $f$ is the PDF of $F$, the unknown distribution of our data. In regression, $f$ is the conditional expectation $\mathbb{E}(Y|X = x)$, to be estimated from $(X, Y)$ pairs.

We'll estimate $f$ with $\hat{f}$, and it's important to keep in mind that in general, $\hat{f}$ is a function of the same type as $f$, *and* it is a function of our data; hence it is random. We'll often use the subscript $n$, as in $\hat{f}_n$, to remind us that it depends on a random sample of size $n$.

We'll measure performance with the **integrated squared error** (ISE) loss, defined as

$$L(f, \hat{f}) = \int (f(u) - \hat{f}(u))^2 du \,, \tag{13.1}$$

with corresponding risk,

$$R(f, \hat{f}) = \mathbb{E}[L(f, \hat{f})] \,, \tag{13.2}$$

where the expectation is taken with respect to the data used for estimation. As long we can interchange integrals, so that $R(f, \hat{f}) = \int R_u \, du$, with $R_u := \mathbb{E}[(f(u) - \hat{f}(u))^2]$ for each $u$,

$$R(f, \hat{f}) = \int R_u du = \int b^2(u) du + \int v(u) du \,, \tag{13.3}$$

where $b(u) = \mathbb{E}[\hat{f}(u)] - f(u)$ is the bias function, and $v(u) = \mathbb{E}[(\hat{f}(u) - \mathbb{E}[\hat{f}(u)])^2]$ is the variance function. So the bias-variance trade-off we've previously encountered is alive and well in this setting. We'll see that in nonparametric curve estimation, the trade-off is controlled primarily by the smoothness of the functions we use to estimate the curve: too much smoothing results in high bias and low variance; not enough smoothing results in low bias and high variance. As such, most of this class is about how to find the right balance.

### 1. Kernel density estimation

Kernel density estimators are linear combinations of "bumps" centered on our observations. Typically, the bumps are smooth—with the amount of smoothness controlled by a parameter—so that the result is a smoothed version of our observations. For our purposes, a **kernel** is any smooth function $K$ such that $K(x) \geq 0$, $\int K(x)dx = 1$, $\int xK(x)dx = 0$, and $0 < \sigma_K^2 := \int x^2 K(x)dx < \infty$. For simplicity, we'll use the Gaussian kernel, which is just the PDF of the standard Gaussian PDF, $K(x) = (2\pi)^{-1/2} e^{-x^2/2}$.

The **kernel density estimator** (KDE) with **bandwidth** $h$ is

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right) \,. \tag{13.4}$$

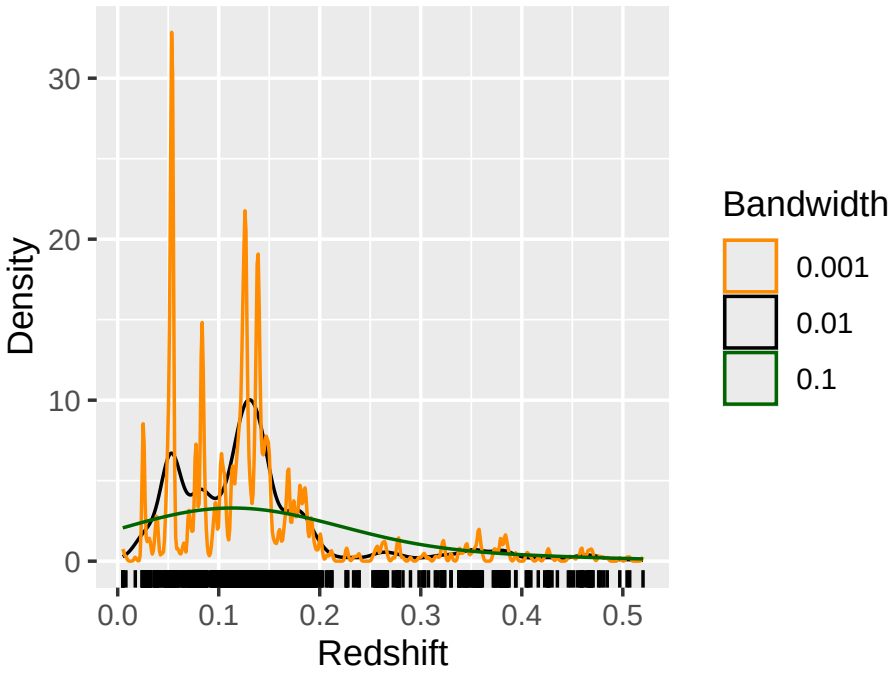


Figure 1: Kernel density estimator of the redshift in the data in Exercise 13.1, using the Gaussian kernel and various bandwidths.

**Exercise 13.1.** Load the Sloan Digital Sky Survey data with:

```
library(readr)
galaxy <-
  read_table("https://www.stat.cmu.edu/~larry/all-of-nonpar/=data/galaxy.dat",
                  col_names = c('ra', 'declination', 'rs'))
```

Write a function that takes as input arguments a point $x$, a vector of observations, and a bandwidth, and returns the KDE at $x$. Plot the KDE for a few different values of $h$. What is the effect of increasing $h$? Of decreasing $h$?

The KDE using various bandwidths is shown in Fig. 1. Clearly, the choice of $h$ is important to how the resulting KDE appears. It also controls the bias-variance trade-off, and hence the theoretical properties of the KDE. In general, $h$ will be set at different values for different $n$, so we denote it by $h_n$.

**Theorem 13.1.** *Under regularity conditions on $f$,*[1] *the risk of the KDE using the Gaussian kernel is*

$$R(f, \hat{f}_n) = \frac{1}{4}\sigma_K^4 h_n^4 \int (f''(x))^2 dx + \frac{1}{nh}\int K^2(x)dx + O(n^{-1}) + O(h_n^6) \; . \tag{13.5}$$

For a sequence $x_1, x_2, \ldots$, the notation $x_n = O(a_n)$ means there is some finite $N$ such that $|x_n/a_n| \leq M < \infty$ for all $n > N$. So in the statement of the theorem, the error of approximating $R(f, \hat{f}_n)$ by the first two terms decays like $M/n + M'h_n^6$ as $n \to \infty$ and $h_n \to 0$.

*Proof.* Write $K_h(x, X) = K((x - X)/h)/h$, so that $\hat{f}_n(x) = \sum_{i=1}^n K_h(x, X_i)/n$. Thus (show this!),

$$\mathbb{E}[\hat{f}_n(x)] = \mathbb{E}[K_h(x, X)] \quad \text{and} \quad \text{Var}[\hat{f}_n(x)] = \text{Var}[K_h(x, X)]/n \; . \tag{13.6}$$

[1] Essentially, continuity/smoothness conditions on the second and third derivatives of $f$; see Theorem 6.28 of [Was06].

We'll analyze the bias function via Taylor expansion, with (note that there are typos in the corresponding proof in [Was04]; they are corrected in [Was06])

$$\begin{aligned}
\mathbb{E}[K_h(x,X)] &= \int \frac{1}{h} K((x-t)/h) f(t) dt && (13.7)\\
&= \int K(u) f(x-hu) du && (13.8)\\
&= \int K(u) \left( f(x) - huf'(x) + \frac{1}{2} h^2 u^2 f''(x) + \cdots \right) du && (13.9)\\
&= f(x) + \frac{1}{2} h^2 \sigma_K^2 f''(x) + O(h^4) \,, && (13.10)
\end{aligned}$$

where on the last line we use the facts that $\int K(u)du = 1$ and that all odd moments of the Gaussian distribution are zero. Hence, the bias function is

$$b_n(x) = \mathbb{E}[\hat{f}_n(x)] - f(x) = \frac{1}{2} h_n^2 \sigma_K^2 f''(x) + O(h_n^4) \,. \tag{13.11}$$

Similarly, the variance is

$$v_n(x) = \frac{f(x) \int K^2(u) du}{n h_n} + O(n^{-1}) \,. \tag{13.12}$$

Squaring the bias and integrating over $x$ yields (13.5) (show this!). □

**Activity 13.1.** Show that squaring the bias and integrating over $x$ yields (13.5). Show that the asymptotically optimal bandwidth is

$$h_n^* = \left( \frac{\int (K(x))^2 dx}{n \sigma_K^4 \int (f''(x))^2 dx} \right)^{1/5} \,, \tag{13.13}$$

and that using $h_n^*$ results in a risk of

$$R^*(f, \hat{f}_n) = O(n^{-4/5}) \,. \tag{13.14}$$

**Solution:** Squaring the bias yields

$$b_n(x)^2 = \frac{1}{4} h_n^4 \sigma_K^4 (f''(x))^2 + O(h_n^6) \,.$$

Summing with $v_n(x)$ and integrating over $x$ yields (13.5). To find the asympntotically optimal bandwidth, we can differentiate the risk with respect to $h_n$ and focus only on the leading order terms when solving for $h_n^*$. Doing so yields $h_n^*$ above. Substituting back into the risk and defining $\lambda := \int (f''(x))^2 dx$ and $\bar{K}^2 := \int K(x)^2 dx$, we get

$$\begin{aligned}
R^*(f, \hat{f}_n) &= \frac{1}{4} \sigma_K^4 \lambda \left( \frac{\bar{K}^2}{n \sigma_K^4 \lambda} \right)^{4/5} + \frac{\bar{K}^2}{n} \left( \frac{n \sigma_K^4 \lambda}{\bar{K}^2} \right)^{1/5} + O(n^{-1}) \\
&= \frac{5}{4} \left( \frac{\bar{K}^2 \sigma_K}{n} \right)^{4/5} \lambda^{1/5} + O(n^{-1}) \,.
\end{aligned}$$

This decreases at rate $n^{-4/5}$. Unfortunately, we can't calculate $\lambda$ because it depends on the unknown density.

The rate of convergence of the asymptotically optimal KDE is $n^{-4/5}$. This is slower than the typical parametric rate of $n^{-1}$, but faster than the rate of $n^{-2/3}$ that is achieved by estimating $f$ with histograms

(as in Ch. 20.2 of [Was04]). Qualitatively, this is typical of nonparametric estimators; we pay a price for not making parametric assumptions, but using smooth functions to estimate smooth functions works better than using non-smooth functions to estimate smooth functions.

In practice, $h$ is selected by cross-validation, which in this case can be computed (up to a constant) by

$$\hat{J}(\hat{f}) = \int \hat{f}(x)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i) \; , \tag{13.15}$$

where $\hat{f}_{-i}$ denotes the KDE obtained by excluding $X_i$ from the data. This is an unbiased estimator, and has a convenient approximation that can be computed efficiently (see [Was04; Was06]).

**2. Nonparametric regression**

Recall that the aim of regression is to estimate

$$r(x) = \mathbb{E}(Y|X = x) \; . \tag{13.16}$$

There are many nonparametric regression estimators, most of which take the form of a weighted average of the $Y_i$'s, where the weights are formed from nearby observations to the corresponding $X_i$'s. The **Nadaraya–Watson** kernel estimator (NWKE) is

$$\hat{r}(x) = \sum_{i=1}^{n} w_i(x) Y_i \; , \quad \text{with} \quad w_i(x) = \frac{K_h(x, x_i)}{\sum_{j=1}^{n} K_h(x, x_j)} \; . \tag{13.17}$$

We can view $w_i(x)$ as the KDE of the conditional PDF $f(y|x)$ at $y = y_i$, and the overall NWKE as the plug-in estimator of

$$\mathbb{E}(Y|X = x) = \int y f(y|x) dy \; . \tag{13.18}$$

**Activity 13.2.** Show that the NWKE is equal to

$$\int y \hat{f}(y|x) dy = \frac{\int y \hat{f}(x, y) dy}{\hat{f}(x)} \; ,$$

where $\hat{f}(x, y)$ and $\hat{f}(x)$ are the KDEs, assuming that the kernel on $(x, y)$ is $K(x, y) = K(x)K(y)$.

**Solution:** The KDE for $f(x, y)$ is

$$\hat{f}(x, y) = \frac{1}{n}\sum_{i=1}^{n} K_h(x, X_i) K_h(y, Y_i) \; .$$

Using that, we have

$$\int y \hat{f}(x, y) \; dy = \frac{1}{n}\sum_{i=1}^{n} K_h(x, X_i) \int y K_h(y, Y_i) \; dy \; .$$

So we just need to compute the integral

$$\begin{aligned}\int yK_h(y, Y_i)\ dy &= \int \frac{1}{h} yK((y - Y_i)/h)\ dy \\ &= \int \frac{1}{h}(hu + Y_i)K(u)h\ du \\ &= Y_i \int K(u)\ du + h \int uK(u)\ du \\ &= Y_i\ ,\end{aligned}$$

where the last equality follows from the kernel properties $\int K(u)du = 1$ and $\int uK(u)du = 0$.

Putting this together, along with the KDE for $f(x)$, we get

$$\hat{r}(x) = \frac{\int y\hat{f}(x, y)\ dy}{\hat{f}(x)} = \sum_{i=1}^{n} \frac{K_h(x, X_i)Y_i}{\sum_{j=1}^{n} K_h(x, X_j)}\ .$$

The analysis of risk here is much more involved, but the asymptotically optimal sequence of bandwidths is $h^* = O(n^{-1/5})$, under which the risk converges at $O(n^{-4/5})$, as with KDE.

# References


[Was04] L. Wasserman. *All of Statistics*. Springer New York, 2004.

[Was06] L. Wasserman. *All of Nonparametric Statistics*. Springer New York, 2006.