

## 45 Density Estimation

A problem closely related to nonparametric regression, is nonparametric density estimation. Let

$$X_1, \dots, X_n \sim f$$

where  $f$  is some probability density. We want to estimate  $f$ .

**45.1 Example (Bart Simpson).** The top left plot in Figure 48 shows the density

$$f(x) = \frac{1}{2}\phi(x; 0, 1) + \frac{1}{10} \sum_{j=0}^4 \phi(x; (j/2) - 1, 1/10) \quad (183)$$

where  $\phi(x; \mu, \sigma)$  denotes a Normal density with mean  $\mu$  and standard deviation  $\sigma$ . Based on 1000 draws from  $f$ , I computed a kernel density estimator, described later. The top right plot is based on a small bandwidth  $h$  which leads to undersmoothing. The bottom right plot is based on a large bandwidth  $h$  which leads to oversmoothing. The bottom left plot is based on a bandwidth  $h$  which was chosen to minimize estimated risk. This leads to a much more reasonable density estimate. ■

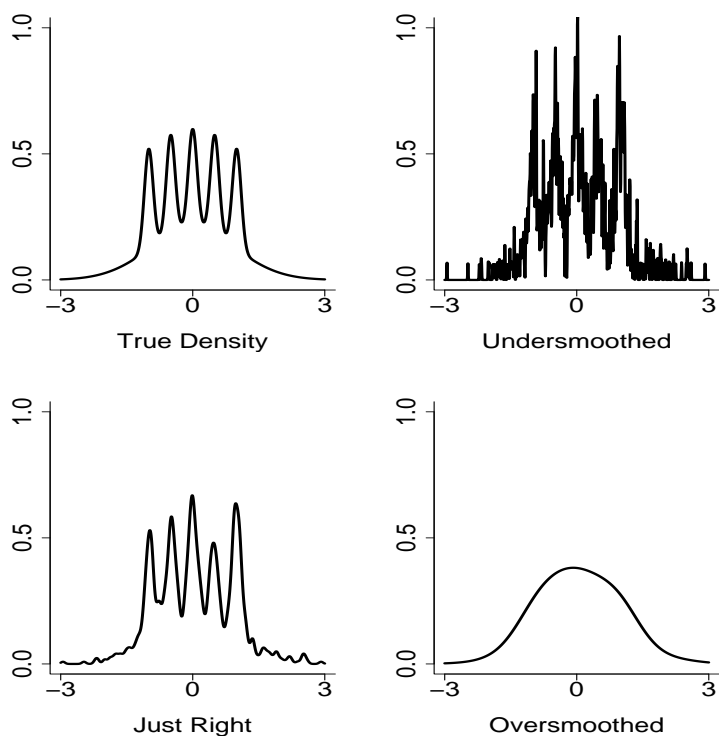


Figure 48: The Bart Simpson density from Example 45.1. Top left: true density. The other plots are kernel estimators based on  $n = 1000$  draws. Bottom left: bandwidth  $h = 0.05$  chosen by leave-one-out cross-validation. Top right: bandwidth  $h/10$ . Bottom right: bandwidth  $10h$ .

We will evaluate the quality of an estimator  $\hat{f}_n$  with the risk, or integrated mean squared error,  $R = \mathbb{E}(L)$  where

$$L = \int (\hat{f}_n(x) - f(x))^2 dx$$

is the integrated squared error loss function. The estimators will depend on some smoothing parameter  $h$  and we will choose  $h$  to minimize an estimate of the risk. The usual method for estimating risk is **leave-one-out cross-validation**. The details are different for density estimation than for regression. In the regression case, the cross-validation score was defined as  $\sum_{i=1}^n (Y_i - \hat{r}_{(-i)}(X_i))^2$  but in density estimation, there is no response variable  $Y$ . Instead, we proceed as follows.

The loss function, which we now write as a function of  $h$ , (since  $\hat{f}_n$  will depend on some smoothing parameter  $h$ ) is

$$\begin{aligned} L(h) &= \int (\hat{f}_n(x) - f(x))^2 dx \\ &= \int \hat{f}_n^2(x) dx - 2 \int \hat{f}_n(x)f(x)dx + \int f^2(x) dx. \end{aligned}$$

The last term does not depend on  $h$  so minimizing the loss is equivalent to minimizing the expected value of

$$J(h) = \int \hat{f}_n^2(x) dx - 2 \int \hat{f}_n(x)f(x)dx. \quad (184)$$

We shall refer to  $\mathbb{E}(J(h))$  as the risk, although it differs from the true risk by the constant term  $\int f^2(x) dx$ .

**45.2 Definition.** *The cross-validation estimator of risk is*

$$\hat{J}(h) = \int \left( \hat{f}_n(x) \right)^2 dx - \frac{2}{n} \sum_{i=1}^n \hat{f}_{(-i)}(X_i) \quad (185)$$

where  $\hat{f}_{(-i)}$  is the density estimator obtained after removing the  $i^{\text{th}}$  observation. We refer to  $\hat{J}(h)$  as the cross-validation score or estimated risk.

Perhaps the simplest nonparametric density estimator is the histogram. Suppose  $f$  has its support on some interval which, without loss of generality, we take to be  $[0, 1]$ . Let  $m$  be an integer and define **bins**

$$B_1 = \left[0, \frac{1}{m}\right), \quad B_2 = \left[\frac{1}{m}, \frac{2}{m}\right), \quad \dots, \quad B_m = \left[\frac{m-1}{m}, 1\right]. \quad (186)$$

Define the **binwidth**  $h = 1/m$ , let  $Y_j$  be the number of observations in  $B_j$ , let  $\hat{p}_j = Y_j/n$  and let  $p_j = \int_{B_j} f(u)du$ .

The **histogram estimator** is defined by

$$\hat{f}_n(x) = \sum_{j=1}^m \frac{\hat{p}_j}{h} I(x \in B_j). \quad (187)$$

To understand the motivation for this estimator, note that, for  $x \in B_j$  and  $h$  small,

$$\mathbb{E}(\hat{f}_n(x)) = \frac{\mathbb{E}(\hat{p}_j)}{h} = \frac{p_j}{h} = \frac{\int_{B_j} f(u)du}{h} \approx \frac{f(x)h}{h} = f(x).$$

**45.3 Example.** Figure 49 shows three different histograms based on  $n = 1,266$  data points from an astronomical sky survey. Each data point represents a “redshift,” roughly speaking, the distance from us to a galaxy. 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 many bins resulting in oversmoothing and too much bias. The bottom left histogram has too few bins resulting in undersmoothing. The top right histogram is based on 308 bins (chosen by cross-validation). The histogram reveals the presence of clusters of galaxies. ■

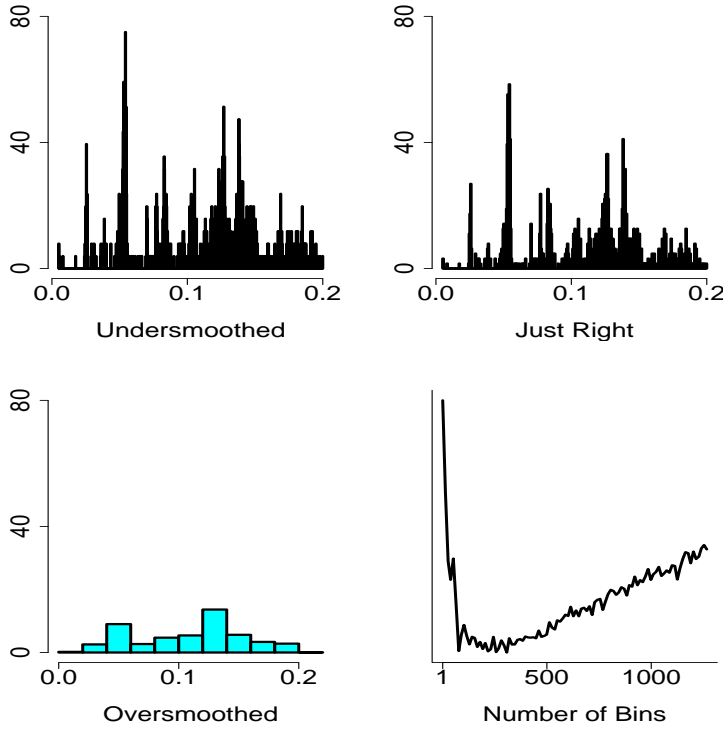


Figure 49: Three versions of a histogram for the astronomy data. The top left histogram has too many bins. The bottom left histogram has too few bins. The top right histogram uses 308 bins (chosen by cross-validation). The lower right plot shows the estimated risk versus the number of bins.

Consider fixed  $x$  and fixed  $m$ , and let  $B_j$  be the bin containing  $x$ . Then,

$$\mathbb{E}(\hat{f}_n(x)) = \frac{p_j}{h} \quad \text{and} \quad \mathbb{V}(\hat{f}_n(x)) = \frac{p_j(1-p_j)}{nh^2}. \quad (188)$$

The risk satisfies

$$R(\hat{f}_n, f) \approx \frac{h^2}{12} \int (f'(u))^2 du + \frac{1}{nh}. \quad (189)$$

The value  $h^*$  that minimizes (189) is

$$h^* = \frac{1}{n^{1/3}} \left( \frac{6}{\int (f'(u))^2 du} \right)^{1/3}. \quad (190)$$

With this choice of binwidth,

$$R(\hat{f}_n, f) \sim \frac{C}{n^{2/3}}. \quad (191)$$

We see that with an optimally chosen binwidth, the risk decreases to 0 at rate  $n^{-2/3}$ . We will see shortly that kernel estimators converge at the faster rate  $n^{-4/5}$ .

**45.4 Theorem.** *The following identity holds:*

$$\hat{J}(h) = \frac{2}{h(n-1)} - \frac{n+1}{h(n-1)} \sum_{j=1}^m \hat{p}_j^2. \quad (192)$$

**45.5 Example.** We used cross-validation in the astronomy example. We find that  $m = 308$  is an approximate minimizer. The histogram in the top right plot in Figure 49 was constructed using  $m = 308$  bins. The bottom right plot shows the estimated risk, or more precisely,  $\hat{J}$ , plotted versus the number of bins. ■

Histograms are not smooth. Now we discuss kernel density estimators which are smoother and which converge to the true density faster.

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

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x - X_i}{h}\right). \quad (193)$$

This amounts to placing a smoothed out lump of mass of size  $1/n$  over each data point  $X_i$ ; see Figure 50.

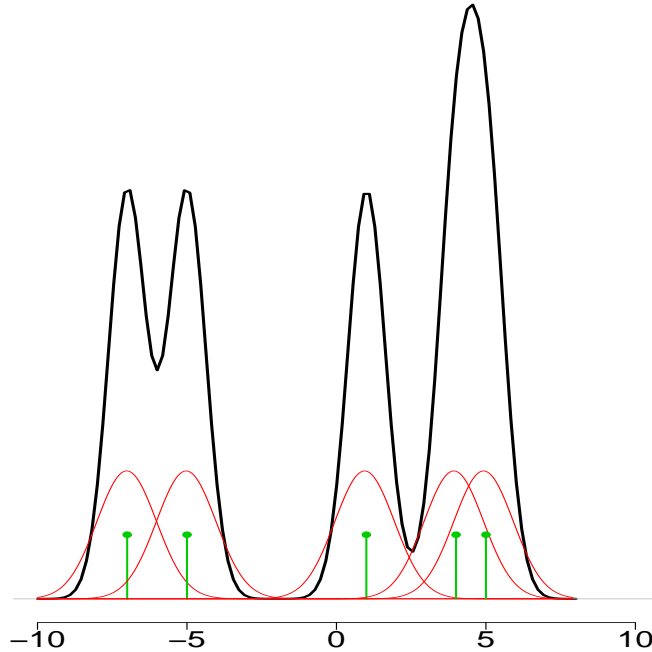


Figure 50: A kernel density estimator  $\hat{f}_n$ . At each point  $x$ ,  $\hat{f}_n(x)$  is the average of the kernels centered over the data points  $X_i$ . The data points are indicated by short vertical bars. The kernels are not drawn to scale.

In R use: `kernel(x, bw=h)` where  $h$  is the bandwidth.

As with kernel regression, the choice of kernel  $K$  is not crucial, but the choice of bandwidth  $h$  is important. Figure 51 shows density estimates with several different bandwidths. Look also at Figure 48. We see how sensitive the estimate  $\hat{f}_n$  is to the choice of  $h$ . Small bandwidths give very rough estimates while larger bandwidths give smoother estimates. In general we will let the bandwidth depend on the sample size so we write  $h_n$ . Here are some properties of  $\hat{f}_n$ .

The risk is

$$R \approx \frac{1}{4} \sigma_K^4 h_n^4 \int (f''(x))^2 dx + \frac{\int K^2(x) dx}{nh} \quad (194)$$

where  $\sigma_K^2 = \int x^2 K(x) dx$ .

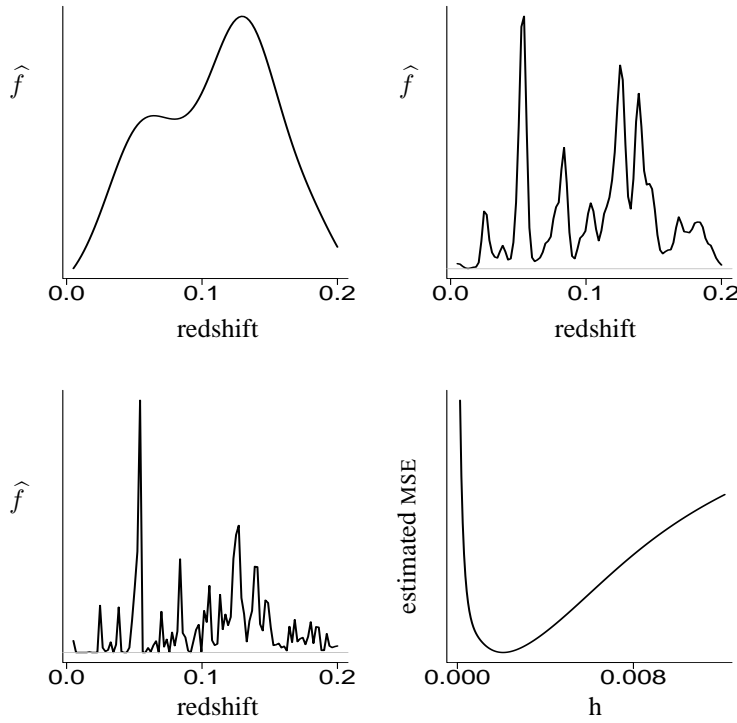


Figure 51: Kernel density estimators and estimated risk for the astronomy data. Top left: oversmoothed. Top right: just right (bandwidth chosen by cross-validation). Bottom left: undersmoothed. Bottom right: cross-validation curve as a function of bandwidth  $h$ . The bandwidth was chosen to be the value of  $h$  where the curve is a minimum.

If we differentiate (194) with respect to  $h$  and set it equal to 0, we see that the asymptotically optimal bandwidth is

$$h_* = \left( \frac{c_2}{c_1^2 A(f)n} \right)^{1/5} \quad (195)$$

where  $c_1 = \int x^2 K(x) dx$ ,  $c_2 = \int K(x)^2 dx$  and  $A(f) = \int (f''(x))^2 dx$ . This is informative because it tells us that the best bandwidth decreases at rate  $n^{-1/5}$ . Plugging  $h_*$  into (194), we see that if the optimal bandwidth is used then  $R = O(n^{-4/5})$ . As we saw, histograms converge at rate  $O(n^{-2/3})$  showing that kernel estimators are superior in rate to histograms.

In practice, the bandwidth can be chosen by cross-validation but first we describe another method which is sometimes used when  $f$  is thought to be very smooth. Specifically, we compute  $h_*$  from (195) under the idealized assumption that  $f$  is Normal. This yields  $h_* = 1.06\sigma n^{-1/5}$ . Usually,  $\sigma$  is estimated by  $\min\{s, Q/1.34\}$  where  $s$  is the sample standard deviation and  $Q$  is the interquartile range.<sup>4</sup> This choice of  $h_*$  works well if the true density is very smooth and is called the **Normal reference rule**.

<sup>4</sup>Recall that the interquartile range is the 75th percentile minus the 25th percentile. The reason for dividing by 1.34 is that  $Q/1.34$  is a consistent estimate of  $\sigma$  if the data are from a  $N(\mu, \sigma^2)$ .

### The Normal Reference Rule

For smooth densities and a Normal kernel, use the bandwidth

$$h_n = \frac{1.06 \hat{\sigma}}{n^{1/5}}$$

where

$$\hat{\sigma} = \min \left\{ s, \frac{Q}{1.34} \right\}.$$

Since we don't want to necessarily assume that  $f$  is very smooth, it is usually better to estimate  $h$  using cross-validation. Recall that the cross-validation score is

$$\hat{J}(h) = \int \hat{f}^2(x) dx - \frac{2}{n} \sum_{i=1}^n \hat{f}_{-i}(X_i) \quad (196)$$

where  $\hat{f}_{-i}$  denotes the kernel estimator obtained by omitting  $X_i$ .

**R code.** use the `bw.ucv` function to do cross-validation:

```
h = bw.ucv(x)
plot(density(x, bw=h))
```

The bandwidth for the density estimator in the upper right panel of Figure 51 is based on cross-validation. In this case it worked well but of course there are lots of examples where there are problems. Do not assume that, if the estimator  $\hat{f}$  is wiggly, then cross-validation has let you down. The eye is not a good judge of risk.

Constructing confidence bands for kernel density estimators is similar to regression. Note that  $\hat{f}_n(x)$  is just a sample average:  $\hat{f}_n(x) = n^{-1} \sum_{i=1}^n Z_i(x)$  where

$$Z_i(x) = \frac{1}{h} K \left( \frac{x - X_i}{h} \right).$$

So the standard error is  $\text{se}(x) = s(x)/\sqrt{n}$  where  $s(x)$  is the standard deviation of the  $Z_i(x)$ 's:

$$s(x) = \sqrt{\frac{1}{n} \sum_{i=1}^n (Z_i(x) - \hat{f}_n(x))^2}. \quad (197)$$

Then we use

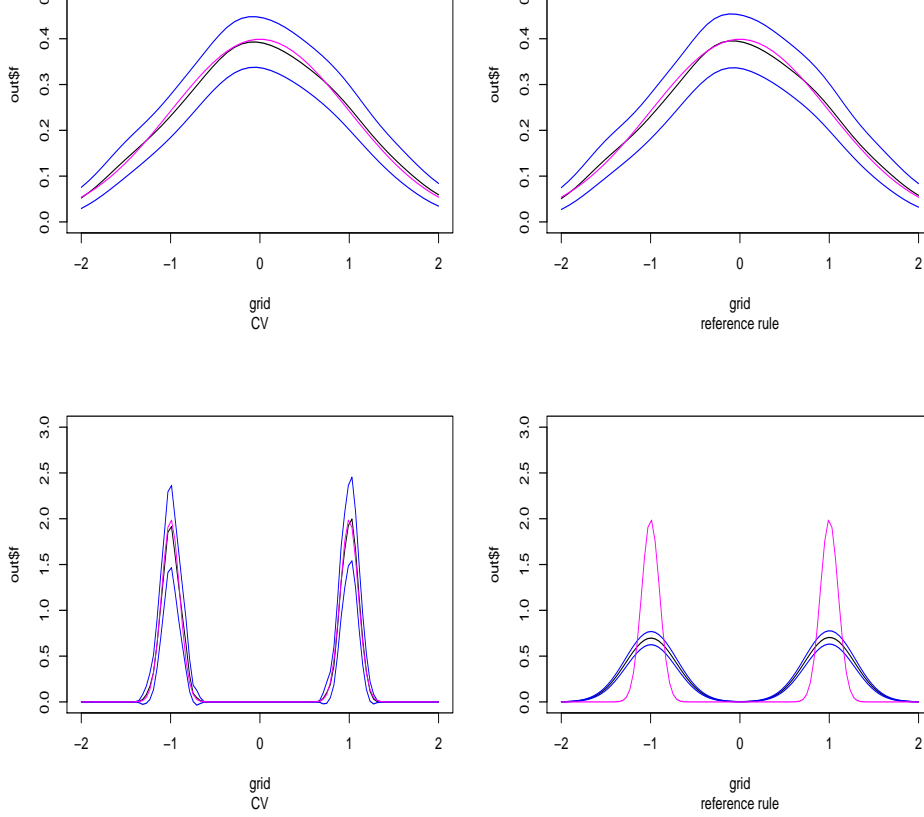
$$\hat{f}_n(x) \pm z_{\alpha/(2n)} \text{se}(x).$$

**45.7 Example.** Figure 45 shows two examples. The first is data from  $N(0,1)$  and second from  $(1/2)N(-1, .1) + (1/2)N(1, .1)$ . In both cases,  $n = 1000$ . We show the estimates using cross-validation and the Normal reference rule together with bands. The true curve is also shown. That's the curve outside the bands in the last plot.

Suppose now that the data are  $d$ -dimensional so that  $X_i = (X_{i1}, \dots, X_{id})$ . The kernel estimator can easily be generalized to  $d$  dimensions. Most often, we use the product kernel

$$\hat{f}_n(x) = \frac{1}{nh_1 \cdots h_d} \sum_{i=1}^n \left\{ \prod_{j=1}^d K \left( \frac{x_j - X_{ij}}{h_j} \right) \right\}. \quad (198)$$

To further simplify, we can rescale the variables to have the same variance and then use only one bandwidth.



A LINK BETWEEN REGRESSION AND DENSITY ESTIMATION. Consider regression again. Recall that

$$r(x) = \mathbb{E}(Y|X = x) = \int yf(y|x)dy = \frac{\int yf(x, y)dy}{f(x)} \quad (199)$$

$$= \frac{\int yf(x, y)}{\int f(x, y)dx}. \quad (200)$$

Suppose we compute a bivariate kernel density estimator

$$\hat{f}(x, y) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_1} K\left(\frac{x - X_i}{h_1}\right) \frac{1}{h_2} K\left(\frac{y - Y_i}{h_2}\right) \quad (201)$$

and we insert this into (200). Assuming that  $\int uK(u)du = 0$ , we see that

$$\int y \frac{1}{h_2} K\left(\frac{y - Y_i}{h_2}\right) dy = \int (h_2u + Y_i)K(u)du \quad (202)$$

$$= h_2 \int uK(u)du + Y_i \int K(u)du \quad (203)$$

$$= Y_i. \quad (204)$$

Hence,

$$\int y \hat{f}(x, y) dy = \frac{1}{n} \sum_{i=1}^n \int y \frac{1}{h_1} K\left(\frac{x - X_i}{h_1}\right) \frac{1}{h_2} K\left(\frac{y - Y_i}{h_2}\right) dy \quad (205)$$

$$= \frac{1}{n} \sum_{i=1}^n \frac{1}{h_1} K\left(\frac{x - X_i}{h_1}\right) \int y \frac{1}{h_2} K\left(\frac{y - Y_i}{h_2}\right) dy \quad (206)$$

$$= \frac{1}{n} \sum_{i=1}^n Y_i \frac{1}{h_1} K \left( \frac{x - X_i}{h_1} \right). \quad (207)$$

Also,

$$\int \widehat{f}(x, y) dy = \frac{1}{n} \sum_{i=1}^n \int \frac{1}{h_1} K \left( \frac{x - X_i}{h_1} \right) \frac{1}{h_2} K \left( \frac{y - Y_i}{h_2} \right) dy \quad (208)$$

$$= \frac{1}{n} \sum_{i=1}^n \frac{1}{h_1} K \left( \frac{x - X_i}{h_1} \right). \quad (209)$$

Therefore,

$$\widehat{r}(x) = \frac{\int y \widehat{f}(x, y) dy}{\int \widehat{f}(x, y) dy} \quad (210)$$

$$= \frac{\frac{1}{n} \sum_{i=1}^n Y_i \frac{1}{h_1} K \left( \frac{x - X_i}{h_1} \right)}{\frac{1}{n} \sum_{i=1}^n \frac{1}{h_1} K \left( \frac{x - X_i}{h_1} \right)} \quad (211)$$

$$= \frac{\sum_{i=1}^n Y_i K \left( \frac{x - X_i}{h_1} \right)}{\sum_{i=1}^n K \left( \frac{x - X_i}{h_1} \right)} \quad (212)$$

which is the kernel regression estimator. In other words, the kernel regression estimator can be derived from kernel density estimation.