20.6. BIAS, VARIANCE AND ALIASING

20.6.1 Bias vs. Variance

Recall from Section 19.2.5 that for an estimator $\hat{\theta}$ of a population quantity $\theta$ we have that an overall measure of the accuracy of the estimator is

$$E[(\hat{\theta} - \theta)^2] = bias(\hat{\theta})^2 + Var(\hat{\theta})$$ (20.10)

In many cases there is a tradeoff between the bias and variance components here. We can have a smaller bias, for instance, but at the expense of increased variance. This is certainly the case with nonparametric density estimation.

As an illustration, suppose the true population density is $f_R(t) = 4t^3$ for $t$ in $(0,1)$, 0 elsewhere. Let’s use (20.6):

$$\hat{f}_R(t) = \frac{\#(t-h, t+h)}{2hn}$$ (20.11)
What is the bias? The numerator has a binomial distribution with \( n \) trials and success probability

\[
p = P(t - h < R < t + h) = \int_{t-h}^{t+h} 4u^3 \, du = (t + h)^4 - (t - h)^4 = 8t^3h + 8th^3 \quad (20.12)
\]

By the binomial property, the numerator of (20.11) has expected value \( np \), and thus

\[
E[\hat{f}_R(t)] = \frac{np}{2nh} = 4t^3 + 4th^2 \quad (20.13)
\]

Subtracting \( f_R(t) \), we have

\[
\text{bias}[\hat{f}_R(t)] = 4th^2 \quad (20.14)
\]

\[\text{Note in the calculation here that it doesn’t matter whether we write } \leq t + h \text{ or } < t + h, \text{ since } R \text{ is a continuous random variable.}\]
So, the smaller we set $h$, the smaller the bias, consistent with intuition.

How about the variance? Again using the binomial property, the variance of the numerator of (20.11) is $np(1-p)$, so that

$$\text{Var}[\hat{f}_R(t)] = \frac{np(1-p)}{(2nh)^2} = \frac{np}{2nh} \cdot \frac{1-p}{2nh} = (4t^3 + 4th^2) \cdot \frac{1-p}{2nh}$$

(20.15)

This matches intuition too: On the one hand, for fixed $h$, the larger $n$ is, the smaller the variance of our estimator—i.e. larger samples are better, as expected. On the other hand, the smaller we set $h$, the larger the variance, because with small $h$ there just won’t be many $R_i$ falling into our interval $(t-h,t+h)$.

So, you can really see the bias-variance tradeoff here, in terms of what value we choose for $h$\footnote{You might ask about finding the $h$ to minimize (20.10). This would not make sense in our present context, in which we are simply assuming a known density in order to explore the bias and variance issues here. In practice, of}
20.6.2 Aliasing

There is another issue here to recognize: The integration in (20.12) tacitly assumed that \( t - h > 0 \) and \( t + h < 1 \). But suppose we are interested in \( f_R(1) \). Then the upper limit in the integral in (20.12) will be 1, not \( t+h \), which will approximately reduce the value of the integral by a factor of 2.

This results in strong bias near the endpoints of the support.\(^3\) Let’s illustrate this with the same density explored in our last section.

Using the general method in Section 5.7 for generating random numbers from a specified distribution, we have that this function will generate \( n \) random numbers from the density \( 4t^3 \) on (0,1):

---

\(^3\) Recall that this term was defined in Section 5.4.4.
f <- function(n) runif(n)^0.25

So, let’s generate some numbers and plot the density estimate:

> plot(density(f(1000)))

The result is shown in Figure 20.5. Sure enough, the estimated density drops after about $t = 0.9$, instead of continuing to rise.

### 20.7 Nearest-Neighbor Methods

Consider (20.6) again. We count data points that are within a fixed distance from $t$; the number of such points will be random. With the nearest-neighbor approach, it’s just the opposite: Now the number will be fixed, while the maximum distance from $t$ will be random.

Specifically, at any point $t$ we find the $k$ nearest $R_i$ to $t$, where $k$ is chosen by the analyst just like $h$ is selected in the kernel case. (And the method is usually referred to as the $k$-Nearest Neighbor method, kNN.) The estimate is now
CHAPTER 20. HISTOGRAMS AND BEYOND: NONPARAMETRIC DENSITY ESTIMATION

Figure 20.6: Empirical cdf, toy example

\[
\hat{f}_R(t) = \frac{k/n}{2 \max_i |R_i - t|} \tag{20.16}
\]

\[
= \frac{k}{2n \max_i |R_i - t|} \tag{20.17}
\]

\[
= \frac{k}{2n \max_i |R_i - t|} \tag{20.18}
\]

20.8 Estimating a cdf

Let’s introduce the notion of the empirical distribution function (ecdf), based on a sample \(X_1, \ldots, X_n\). It is a sample estimate of a cdf, defined to be the proportion of \(X_i\) that are below \(t\) in the sample. Graphically, \(F_X\) is a step function, with jumps at the values of the \(X_i\).

As a small example, say \(n = 4\) and our data are 4.8, 1.2, 2.2 and 6.1. We can plot the empirical cdf by calling R’s \text{ecdf()}\ function:

\[
> \text{plot(ecdf(x))}
\]

The graph is in Figure [20.6]. (In \texttt{ggplot2}, the function \text{stat_ecdf()}\ is similar.)