Elements of Density Estimation
Math 6070, Spring 2006

Davar Khoshnevisan
University of Utah

March 2006

Contents

1 Introduction

The basic problem in density estimation is this: Suppose $X_1, \ldots, X_n$ is an independent sample from a density function f that is unknown. In many cases, f is unknown only because it depends on unknown parameter(s). In such cases, we proceed by using methods that are discussed in Math 5080–5090. For example, if $X_1, \ldots, X_n \sim N(\mu\,, \sigma^2)$, then the density is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Then, we estimate f by

$$\hat{f}(x) = \frac{1}{\hat{\sigma}\sqrt{2\pi}} \exp\left(-\frac{(x-\hat{\mu})^2}{2\hat{\sigma}^2}\right),$$

where $\hat{\mu} = \bar{X}_n$ and $\hat{\sigma}^2 = (1/n)\sum_{i=1}^n (X_i - \bar{X}_n)^2$ are the usual (maximum likelihood) estimates of mean and variance.

Here, we are studying the more interesting case that f is generally unknown. In this more general case, there are several different approaches to density estimation. Later on we shall concentrate our efforts on the socalled "kernel density estimators." But for now, let us begin with a discussion of the most commonly-used, quick-and-dirty approach: The histogram.

1.1 The Histogram

A standard histogram of data $X_1, \ldots, X_n$ starts with agreeing on a point x_0—called the *origin*—and a positive number h—called *bandwidth*. Then, we define *bins* B_j for all integers $j = 0, \pm 1, \pm 2, \ldots$ as follows:

$$B_j := [x_0 + jh, x_0 + (j+1)h].$$

The ensuing *histogram* is the plot of the density estimator,

$$\hat{f}(x) := \frac{1}{nh} \sum_{j=1}^n \mathbf{I}\{X_j \text{ is in the same bin as } x\}.$$

Note that for all $x \in B_k$, $\hat{f}(x)$ is equal to $(1/h)$ times the fraction of the data that falls in bin k. The bandwidth h is a "smoothing parameter." As h is increased, the plot of $\hat{f}$ becomes "smoother," and conversely as h is decreased, $\hat{f}$ starts to look "rougher." Fine-tuning h is generally something that one does manually. This is a skill that is honed by being thoughtful and after some experimentation.

Warnings:

1. Generally the graph of $\hat{f}$ is also very sensitive to our choice of x_0.
2. The resulting picture/histogram is jagged by design. More often than not, density estimation is needed to decide on the "shape" of f. In such cases, it is more helpful to have a "smooth" function estimator.
3. There are estimators of f that have better mathematical properties than the histogram.

Example 1 Consider the following hypothetical data set:

1, 1, 2, 3, 4, 4, 4, 2, 1.5, 1.4, 2.3, 4.8.

Here, $n = 12$. Suppose we set $x_0 := 0$ and $h := 1.5$. Then, the bins of interest are

$$[0, 1.5),\ [1.5, 3),\ [3, 4.5),\ [4.5, 6).$$

Therefore,

$$\hat{f}(x) = \frac{1}{18} \times \begin{cases} 3 & \text{if } 0 \le x < 1.5, \\ 4 & \text{if } 1.5 \le x < 3, \\ 4, & \text{if } 3 \le x < 4.5, \\ 1, & \text{if } 4.5 \le x < 6, \end{cases}$$

$$= \begin{cases} 1/6 & \text{if } 0 \le x < 1.5, \\ 2/9 & \text{if } 1.5 \le x < 3, \\ 2/9, & \text{if } 3 \le x < 4.5, \\ 1/18, & \text{if } 4.5 \le x < 6. \end{cases}$$

In order to see how changing x_0 can change the picture consider instead $x_0 = 1$. Then,

$$\hat{f}(x) = \frac{1}{18} \times \begin{cases} 4 & \text{if } 1 \le x < 2.5, \\ 4, & \text{if } 2.5 \le x < 4, \\ 1, & \text{if } 4 \le x < 5.5. \end{cases}$$

The preceding example showcases the problem with the choice of the origin: By changing x_0 even a little bit we can change the entire shape of $\hat{f}$. Nevertheless, the histogram can be a useful (i.e., fast) starting-point for the data analyst. For instance, in `R`, you first type the expression "`X = c(1,1,2,3,4,4,4,2,1.5,1.4,2.3,4.8)`" to get X to denote the data vector of the previous example. Then, you type "`hist(X)`" to produce Figure 1. The `R` command `hist` has several parameters that you can use to fine-tune your histogram plotting. For instance, `hist(X,breaks=6)` produces Figure 2. [Figure 1 can be produced also with `hist(X,breaks=3)`.]

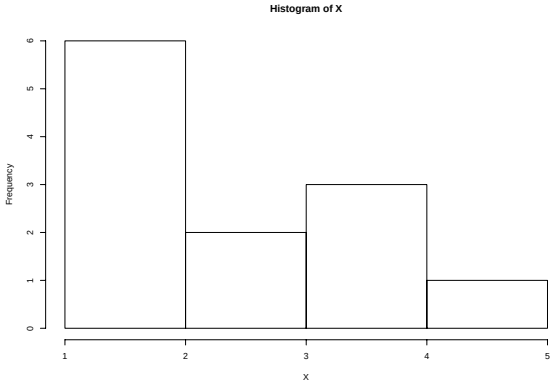

Figure 1: Histogram of the data of Example 1.
Three breaks (automatic).

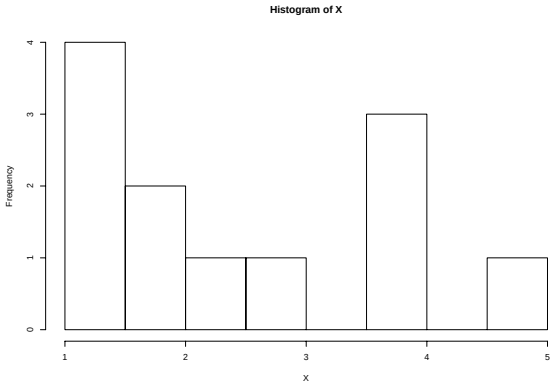

Figure 2: Histogram of the data of Example 1.
Six breaks (manual).

1.2 The Kernel Density Estimator

Kernel density estimators are a smoother substitute for histograms. We start with a heuristic argument: If h is a small number, and if f is continuous at x, then

$$f(x) \approx \frac{1}{2h} \mathrm{P}\{x-h < X < x+h\}.$$

Here, $X \sim f$, of course. On the other hand, by the law of large numbers, if n is large then

$$\mathrm{P}\{x-h < X < x+h\} \approx \frac{1}{n} \sum_{j=1}^{n} \mathbf{I}\left\{x-h < X_j < x+h\right\},$$

in probability. So we can consider the density estimator

$$\begin{aligned}
\hat{f}(x) &:= \frac{1}{2nh} \sum_{j=1}^{n} \mathbf{I}\left\{x-h < X_j < x+h\right\}, \\
&= \frac{1}{n} \sum_{j=1}^{n} \frac{\mathbf{I}\{|X_j - x| \le h\}}{2h} \\
&= \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h} w\left(\frac{x-X_j}{h}\right),
\end{aligned}$$

where w is the "kernel,"

$$w(x) := \frac{1}{2}\mathbf{I}\{|x| \le 1\}.$$

This definition of $\hat{f}(x)$ yields a variant of the histogram. In order to obtain a smoother estimator, note that if h is small then $w_h(x) := (1/h)w((x-X_j)/h)$ is approximately a "delta function at X_j" That is: (1) w_h is highly peaked at X_j, and (2) the area under w_h is fixed to be one. So our strategy is to replace the role of w by a smoother function so that a smoother delta function is obtained.

So now consider a "kernel" K. It is a function such that $K(x) \ge 0$ and $\int_{-\infty}^{\infty} K(x)\,dx = 1$. Then, define

$$\hat{f}(x) := \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x-X_j}{h}\right).$$

The parameter h is used to tune the estimator. It is alternatively called the *window width*, the *bandwidth*, and/or the *smoothing parameter*. Roughly speaking, the kernel desnity estimator puts a smooth but concentrated "bump function" over each observation, and then averages over the bumps.

1.3 The Nearest-Neighborhood Density Estimator

Let us choose and fix some integer k with the property that $k \ll n$ [usually, $k \approx \sqrt{n}$.] Then, define $\rho_1(t) \le \rho_2(t) \le \cdots \le \rho_n(t)$ to be the ordered distances from t to the sample $X_1, \ldots, X_n$.[1] Then, we can consider

$$\hat{f}(x) = \frac{k-1}{2n\rho_k(x)}. \tag{1}$$

This is called the *nearest-neighbor density estimator* (also known as the "NN density estimator.") In order to see why it is sensible first note that if f is continuous at x and r is sufficiently small, then

$$\mathrm{E}\left[\sum_{j=1}^{n} \mathbf{I}\left\{x-r < X_j < x+r\right\}\right] = n\mathrm{P}\left\{x-r < X_1 < x+r\right\} \approx 2rnf(x).$$

Therefore, by the law of large numbers, if n is large then one might expect that

$$\sum_{j=1}^{n} \mathbf{I}\left\{x-r < X_j < x+r\right\} \approx 2rnf(x),$$

in probability. Thus, one might expect that for n large, the following has high probability:

$$\sum_{j=1}^{n} \mathbf{I}\left\{x-\rho_k(x) < X_j < x+\rho_k(x)\right\} \approx 2\rho_k(x)nf(x).$$

[1] For instance, if $X_1 = 1, X_2 = 0, X_3 = 2$, then $\rho_1(0.6) = 0.4$, $\rho_2(0.6) = 0.6$, and $\rho_3(t) = 1.4$.

[This is not obvious because $\rho_k(x)$ is a random variable. But remember that we are merely developing a heuristic argument here.] Because

$$\sum_{j=1}^{n} \mathbf{I}\{x-\rho_k(x) \le X_j \le x+\rho_k(x)\} = k-1,$$

this leads us to (1).

NN density estimators have some well-known setbacks. Here are two:

1. $\hat{f}$ is not smooth. Typically, this problem is addressed by using instead

$$\hat{f}(x) = \frac{1}{n\rho_k(x)} \sum_{j=1}^{n} K\left(\frac{x-X_j}{\rho_k(x)}\right).$$

   This one performs somewhere between the NN-estimator and the kernel estimator.

2. $\hat{f}$ is a better estimator of f "locally." For instance, this is a better method if we are interested in the values/shape of f near a point x. Indeed, we can check easily that

$$\int_{-\infty}^{\infty} \hat{f}(x)\,dx = \frac{k-1}{2n} \int_{-\infty}^{\infty} \frac{dx}{\rho_k(x)} = \infty.$$

   The reason is that $\rho_k(x) \sim |x|$ as $|x| \to \infty$. Therefore, the NN estimator is not itself a density function.

1.4 Variable Kernel Density Estimation

Let K be a nice kernel, and choose and fix a positive integer k. Define $\delta_{j,k}$ to be the distance between X_j and the kth nearest point in $\{X_1,\dots,X_n\} \setminus \{X_k\}$. Formally speaking,

$$\delta_{j,k} := \min_{\ell \ne j} |X_j - X_\ell|.$$

Then we consider the *variable kernel density estimator*,

$$\hat{f}(x) := \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h\delta_{j,k}} K\left(\frac{x-X_j}{h\delta_{j,k}}\right).$$

The "window width" h determines the degree of "smoothing," and k determines how strongly the window width responds to "local detail."

1.5 The Orthogonal Series Method

Suppose f is a density on $[0,\infty)$. Define

$$\begin{aligned}
\phi_0(1) &:= 1\\
\phi_1(x) &:= \sqrt{2}\cos(2\pi x),\\
\phi_2(x) &:= \sqrt{2}\sin(2\pi x),\\
&\vdots\\
\phi_{2j-1}(x) &:= \sqrt{2}\cos(2\pi jx),\\
\phi_{2j}(x) &:= \sqrt{2}\sin(2\pi jx),
\end{aligned}$$

for $j \ge 1$. Then, the theory of Fourier series tells us that

$$f(x) \sim \sum_{j=0}^{\infty} f_j \phi_j(x),$$

where

$$f_j := \int_0^1 f(x)\phi_j(x)\,dx,$$

and "$f \sim \sum_{j=1}^{\infty} f_j\phi_j$" means that the infinite sum converges in $\mathscr{L}^2(\mathbf{R})$ to f. That is,

$$\lim_{N\to\infty} \int_0^{\infty} \left| f(x) - \sum_{j=0}^{N} f_j(x)\phi_j(x) \right|^2 dx = 0.$$

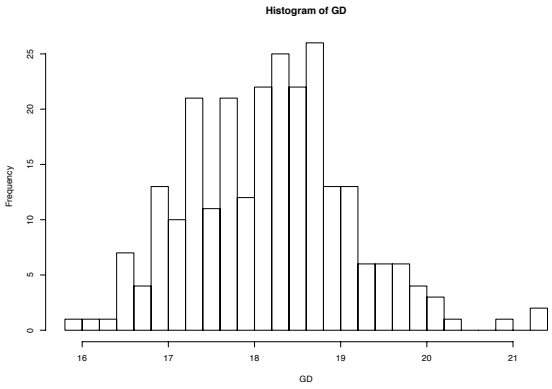

Figure 3: Histogram of the variable "GD".
Thirty breaks.

Now suppose $X \sim f$. Then, f_j is nothing but $\mathrm{E}[\phi_j(X)]$, and we are led to the law-of-large-numbers estimator

$$\hat{f}_j := \frac{1}{n}\sum_{\ell=1}^{n} \phi_j(X_\ell).$$

Therefore, we are led to the estimator

$$\hat{f}(x) := \sum_{j=0}^{N} \hat{f}_j \phi_j(x),$$

where M is a large, pre-determined, and fixed constant. This estimator has a serious setback: In general, $\hat{f}(x)$ need not be ≥ 0!

1.6 Maximum Penalized Likelihood Estimation

Define the "likelihood" of g to be

$$\mathscr{L}(g) := \mathscr{L}(g \,|\, X_1, \ldots, X_n) := \prod_{j=1}^{n} g(X_j).$$

Then we can try to find g that maximizes $\mathscr{L}(g)$. Unfortunately, this is doomed to fail. Indeed, let $\hat{f}(x)$ denote the histogram with origin $x_0 = 0$ and bandwidth $h > 0$. Then it is evident that $\hat{f}(X_i) \geq (nh)^{-1}$, whence it follows that $\prod_{j=1}^{n} \hat{f}(X_i) \geq (nh)^{-n}$. Consequently, $\max_g \mathscr{L}(g) \geq (nh)^{-n}$ for all $h > 0$. Let $h \to 0$ to find that $\max_g \mathscr{L}(g) = \infty$.

Although the preceding attempt failed, it is not without its merits. The reason that our first attempt failed was that we are maximizing $\mathscr{L}(g)$ over too many functions g. Therefore, we can restrict the class of g's over which the maximization is taken. For instance, consider the "penalized log-likelihood,"

$$\ell(g) := \sum_{j=1}^{n} \ln g(X_j) - \lambda F(g),$$

where $\lambda > 0$ is a smoothing parameter and $F(g)$ measures the roughness of g (say!). An example to have in mind is $F(g) := \int_{-\infty}^{\infty} (g''(x))^2\, dx$. Then, we can try and find g that solves the maximization problem, $\max_{g \in W^{1,1}} \ell(g)$, where $W^{1,1}$ denotes the class of all functions g such that $\int_{-\infty}^{\infty} (g(x))^2\, dx < \infty$ and $\int_{-\infty}^{\infty} (g''(x))^2\, dx < \infty$.

The statistic $\sum_{j=1}^{n} \ln g(X_j)$ corresponds to the goodness of fit; $F(g)$ to smoothness; and λ to how much of each (goodness of fit versus smoothness) we wish to opt for. The major setback of this method is that it is technically (and computationally) *very* hard and intensive.

2 Kernel Density Estimation in One Dimension

Recall that $X_1, \ldots, X_n$ are i.i.d. with density function f. We choose and fix a probability density function K and a binwidth h, and then define our kernel density estimate as

$$\hat{f}(x) := \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right), \qquad -\infty < x < \infty.$$

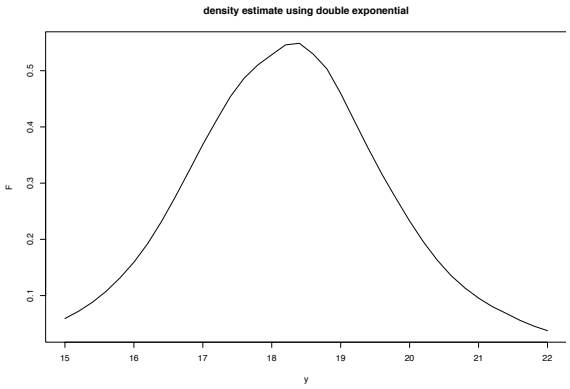

Figure 4: Kernel density estimate using DE ($h = 0.5$).

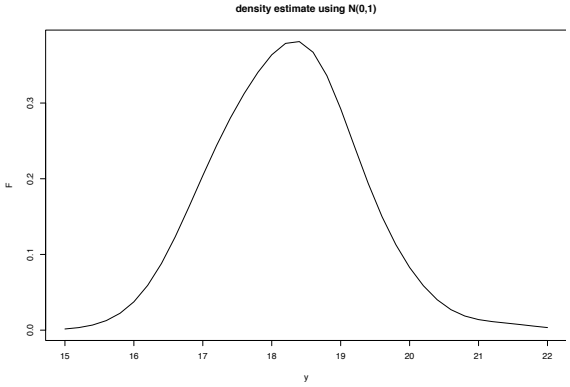

Figure 5: Kernel density estimate using $N(0,1)$ ($h = 0.5$).

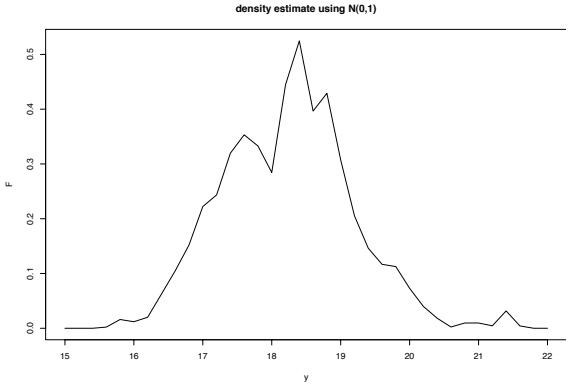

Figure 6: Kernel density estimate using $N(0,1)$ ($h = 0.1$).

Before we start our analysis, let us see how kernel density estimators looks for a certain data set whose variable I call "GD." In order to have a reasonable starting point, I have drawn up the histogram of the data. This appears in Figure 3. The number of breaks was 30. This number was obtained after a little experimentation.

Figures 4, 5, and 6 depict three different kernel density estimates of the unknown density f. They are all based on the same dataset.

1. Figure 4 shows the kernel density estimator of "GD" with bandwidth $h := 0.5$ and $K :=$ the double-exponential density; i.e., $K(x) = \frac{1}{2}e^{-|x|}$. The density K is plotted in Figure 7.

2. Figure 5 shows the kernel density estimator for the same bandwidth ($h = 0.5$), but now $K := (2\pi)^{-1/2}\exp(-x^2/2)$ is the $N(0,1)$ density. The density K is plotted in Figure 8 for the purposes of comparison.

3. Figure 6 shows the kernel density estimator for the smaller bandwidth $h = 0.1$, but still K is still the $N(0,1)$ density.

Before we analyse kernel density estimators in some depth, let us try and understand the general notion of "smoothing," which translates to the mathematical "convolution." In actual practice, you raise h in order to obtain a smoother kernel density estimator; you lower h to obtain a rougher one. Figures 5 and 6 show this principle for the variable "GD."

2.1 Convolutions

If f and g are two non-negative functions on $\mathbf{R}$, then their *convolution* is defined as

$$(f*g)(x) := \int_{-\infty}^{\infty} f(y)g(x-y)\,dy,$$

provided that the integral exists, of course. A change of variables shows that $f*g = g*f$, so that convolution is a symmetric operation. You have seen convolutions in undergraduate probability already: If X and Y are independent random variables with respective densities f and g, then $X+Y$ is a continuous random variable also, and its density is exactly $f*g$.

Quite generally, if f and g are probability densities then so is $f*g$. Indeed, $(f*g)(x) \geq 0$ and

$$\int_{-\infty}^{\infty}(f*g)(x)\,dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(y)g(x-y)\,dy\,dx = \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} g(x-y)\,dx\right) f(y)\,dy = 1,$$

after a change of the order of integration.

Quite generally, convolution is a "smoothing operation." One way to make this precise is this: Suppose f and g are probability densities; g is continuously differentiable with a bounded derivative. Then, $f*g$ is also differentiable and

$$(f*g)'(x) = \int_{-\infty}^{\infty} f(y)g'(x-y)\,dx.$$

The continuity and boundedness of g' ensure that we can differentiate under the integral sign. Similar remarks apply to the higher derivatives of $f*g$, etc.

In other words, if we start with a generic density function f and a smooth one g, then $f*g$ is in general not less smooth than g. By symmetry, it follows that $f*g$ is at least as smooth as the smoother one of f and g.

2.2 Approximation to the Identity

Let K be a real-valued function on $\mathbf{R}$ such that $K(x) \geq 0$ for all $x \in \mathbf{R}$, and $\int_{-\infty}^{\infty} K(x)\,dx = 1$. That is, K is a density function itself. But it is one that we choose according to taste, experience, etc. Define for all $h > 0$,

$$K_h(x) := \frac{1}{h}K\left(\frac{x}{h}\right).$$

For example, if K is the standard-normal density, then K_h is the $N(0,h)$ density. In this case, K_h concentrates more and more around 0 as $h \downarrow 0$. This property is valid more generally, e.g., if K "looks" like a normal, Cauchy, etc.

Recall that K is a density function. This implies that K_h is a density also. Indeed, $K_h(x) \geq 0$, and

$$\int_{-\infty}^{\infty} K_h(x)\,dx = \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{x}{h}\right)dx = \int_{-\infty}^{\infty} K(y)\,dy = 1,$$

after a change of variables. The collection $\{K_h\}_{h>0}$ of functions is sometimes called an *approximation to the identity*. The following justifies this terminology.

Theorem 2 *Let f be a density function. Suppose that either:*

1. *f is bounded; i.e., there exists B such that $|f(x)| \le B$ for all x; or*
2. *K vanishes at infinity; i.e., $\lim_{|z|\to\infty} K(z) = 0$.*

Then, whenever f is continuous in an open neighborhood of $x \in \mathbf{R}$,

$$\lim_{h\to 0}(K_h * f)(x) = f(x).$$

In many applications, our kernel K is infinitely differentiable and vanishes at infinity. The preceding then proves that f can be approximated, at all its "continuity points," by an infinitely-differentiable function.

Proof of Theorem 2: Because K_h is a density function, we have $f(x) = \int_{-\infty}^{\infty} f(x)K_h(y)\,dy$ for all $x \in \mathbf{R}$. Therefore,

$$\begin{aligned} f(x) - (f * K_h)(x) &= \int_{-\infty}^{\infty} K_h(y) f(x)\,dy - \int_{-\infty}^{\infty} K_h(y) f(x-y)\,dy \\ &= \int_{-\infty}^{\infty} K_h(y)\Big[f(x) - f(x-y)\Big]\,dy. \end{aligned}$$

We apply the triangle inequality for integrals to find that

$$\begin{aligned} |f(x) - (f * K_h)(x)|\,dx &\le \int_{-\infty}^{\infty} K_h(y)\Big|f(x) - f(x-y)\Big|\,dy \\ &= \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{y}{h}\right)\Big|f(x) - f(x-y)\Big|\,dy \\ &= \int_{-\infty}^{\infty} K(z)\Big|f(x) - f(x-zh)\Big|\,dz. \end{aligned}$$

Fix $\varepsilon > 0$, and choose $\delta > 0$ such that $|f(x) - f(y)| \le \varepsilon$ whenever $|y - x| \le \delta$. We split the last integral up in two pieces:

$$\begin{aligned} &\int_{-\infty}^{\infty} K(z)\Big|f(x) - f(x-zh)\Big|\,dz \\ &= \int_{|zh|>\delta} K(z)\Big|f(x) - f(x-zh)\Big|\,dz + \int_{|zh|\le\delta} K(z)\Big|f(x) - f(x-zh)\Big|\,dz \\ &\le \int_{|zh|>\delta} K(z)\Big|f(x) - f(x-zh)\Big|\,dz + \varepsilon\int_{|zh|\le\delta} K(z)\,dz \\ &\le \int_{|zh|>\delta} K(z)\Big|f(x) - f(x-zh)\Big|\,dz + \varepsilon. \end{aligned} \tag{2}$$

We estimate the other integral in the two cases separately. First suppose $|f(x)| \le B$ for all x. Then,

$$\begin{aligned} \int_{|zh|>\delta} K(z)\Big|f(x) - f(x-zh)\Big|\,dz &\le \int_{|zh|>\delta} K(z)\Big(f(x) + f(x-zh)\Big)\,dz \\ &\le 2B\int_{|z|>\delta/h} K(z)\,dz. \end{aligned}$$

Combine this with (2) to find that

$$\int_{-\infty}^{\infty} K(z)\Big|f(x) - f(x-zh)\Big|\,dz \le 2B\int_{|z|>\delta/h} K(z)\,dz + \varepsilon.$$

As $h \to 0$, the second integral vanishes. Because the $h \to 0$-limit of the left-hand side is independent of ε it must be zero.

Next, suppose K vanishes at infinity. Choose and fix $\eta > 0$ small. Then, $K(z) \le \eta$ whenever $|z|$ is sufficiently large. Thus, for all h small,

$$\begin{aligned} \int_{|zh|>\delta} K(z)\Big|f(x) - f(x-zh)\Big|\,dz &\le \int_{|zh|>\delta} K(z)\Big(f(x) + f(x-zh)\Big)\,dz \\ &\le \eta\int_{|z|>\delta/h}\Big(f(x) + f(x-zh)\Big)\,dz \le 2\eta. \end{aligned}$$

Therefore,

$$\lim_{h\to 0}\int_{-\infty}^{\infty} K(z)\Big|f(x) - f(x-zh)\Big|\,dz \le 2\eta + \varepsilon.$$

The left-hand side is independent of ε and η. Therefore it must be zero. $\square$

Theorem 2 really requires some form of smoothness on the part of f. However, there are versions of this theorem that require nothing more than the fact that f is a density. Here is one such version. Roughly speaking, it states that for "most" values of $x \in \mathbf{R}$, $(K_h * f)(x) \approx f(x)$ as $h \to 0$. The proof is similar to that of Theorem 2.

Theorem 3 *Suppose f and K are density functions. Then,*

$$\lim_{h\to 0} \int_{-\infty}^{\infty} |(K_h * f)(x) - f(x)|\, dx = 0.$$

There is also a "uniform" version of this. Recall that f is *uniformly continuous* if

$$\lim_{\varepsilon \to 0} \max_x |f(x+\varepsilon) - f(x)| = 0.$$

Then, the following can also be proved along the lines of Theorem 2.

Theorem 4 *Suppose f and K are density functions, and f is uniformly continuous. Then, $\lim_{h\to 0} K_h * f = f$ uniformly; i.e.,*

$$\lim_{h\to 0} \max_x |(K_h * f)(x) - f(x)| = 0.$$

3 The Kernel Density Estimator

Now suppose $X_1, \ldots, X_n$ are i.i.d. with density f. Choose and fix a bandwidth $h > 0$ (small), and define

$$\hat{f}(x) := \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right) = \frac{1}{n} \sum_{j=1}^{n} K_h(x - X_j).$$

We can easily compute the mean and variance of $\hat{f}(x)$, viz.,

$$\begin{aligned}
\mathrm{E}\,\hat{f}(x) &= \mathrm{E}\,[K_h(x - X_1)] \\
&= \int_{-\infty}^{\infty} K_h(x-y) f(y)\, dy = (K_h * f)(x); \\
\mathrm{Var}\,\hat{f}(x) &= \frac{1}{n} \mathrm{Var}\,(K_h(x - X_1)) \\
&= \frac{1}{nh^2} \int_{-\infty}^{\infty} \left| K\left(\frac{x-y}{h}\right)\right|^2 f(y)\, dy - \frac{1}{n} |(K_h * f)(x)|^2 \\
&= \frac{1}{n} \left[(K_h^2 * f)(x) - (K_h * f)^2(x)\right].
\end{aligned}$$

Now recall the *mean-squared error*:

$$\mathrm{MSE}\,\hat{f}(x) := \mathrm{E}\left[\left|\hat{f}(x) - f(x)\right|^2\right] = \mathrm{Var}\,\hat{f}(x) + \left|\mathrm{Bias}\,\hat{f}(x)\right|^2.$$

The bias is

$$\mathrm{Bias}\,\hat{f}(x) = \mathrm{E}\left[\hat{f}(x)\right] - f(x) = (K_h * f)(x) - f(x).$$

Thus, we note that for a relatively nice kernel K:

1. $\mathrm{Var}\,\hat{f}(x) \to 0$ as $n \to \infty$; whereas
2. $\mathrm{Bias}\,\hat{f}(x) \to 0$ as $h \to 0$; see Theorem 2.

The question arises: Can we let $h = h_n \to 0$ and $n \to \infty$ in such a way that $\mathrm{MSE}\,\hat{f}(x) \to 0$? We have seen that, in one form or another, all standard density estimators have a sort of "bandwidth" parameter. Optimal choice of the bandwidth is the single-most important question in density estimation, and there are no absolute answers! We will study two concrete cases next.

4 Asymptotically Optimal Bandwidth Selection

Suppose the unknown density f is smooth (three bounded and continuous derivatives, say!). Suppose also that K is symmetric [i.e., $K(a) = K(-a)$] and vanishes at infinity. Then it turns out that we can "find" the asymptotically-best value of the bandwidth $h = h_n$.

Several times in the future, we will appeal to Taylor's formula in the following form: For all h small,

$$f(x - zh) \approx f(x) - zhf'(x) + \frac{z^2h^2}{2} f''(x). \tag{3}$$

4.1 Local Estimation

Suppose we are interested in estimating f "locally." Say, we wish to know $f(x)$ for a fixed, given value of x.

We have seen already that

$$\begin{aligned}
\text{Bias } \hat{f}(x) &= (K_h * f)(x) - f(x) \\
&= \frac{1}{h} \int_{-\infty}^{\infty} K\left(\frac{x-u}{h}\right) f(u)\, du - f(x) \\
&= \int_{-\infty}^{\infty} K(z) f(x - zh)\, dz - f(x).
\end{aligned}$$

Therefore, by (3),

$$\begin{aligned}
\text{Bias } \hat{f}(x) &\approx \int_{-\infty}^{\infty} K(z) \left\{ f(x) - zhf'(x) + \frac{z^2h^2}{2} f''(x) \right\} dz - f(x) \\
&= f(x) \int_{-\infty}^{\infty} K(z)\, dz - hf'(x) \int_{-\infty}^{\infty} zK(z)\, dz + \frac{h^2}{2} f''(x) \int_{-\infty}^{\infty} z^2 K(z)\, dz - f(x) \\
&= \frac{h^2}{2} f''(x) \int_{-\infty}^{\infty} z^2 K(z)\, dz =: \frac{h^2}{2} f''(x) \sigma_K^2.
\end{aligned} \tag{4}$$

We have used the facts that: (i) $\int_{-\infty}^{\infty} K(z)\, dz = 1$ (K is a density); and (ii) $\int_{-\infty}^{\infty} zK(z)\, dz = 0$ (symmetry).

Now we turn our attention to the variance of $\hat{f}(x)$. Recall that $\text{Var } \hat{f}(x) = (K_h^2 * f)(x) - (K_h * f)^2(x)$. We begin by estimating the first term.

$$\begin{aligned}
\left(K_h^2 * f\right)(x) &= \frac{1}{h^2} \int_{-\infty}^{\infty} \left| K\left(\frac{x-u}{h}\right) \right|^2 f(u)\, du \\
&= \frac{1}{h} \int_{-\infty}^{\infty} K^2(z) f(x - zh)\, dz \\
&\approx \frac{1}{h} \int_{-\infty}^{\infty} K^2(z) \left\{ f(x) - zhf'(x) + \frac{z^2h^2}{2} f''(x) \right\} dz \\
&= \frac{1}{h} f(x) \int_{-\infty}^{\infty} K^2(z)\, dz - f'(x) \int_{-\infty}^{\infty} zK^2(z)\, dz + \frac{h}{2} f''(x) \int_{-\infty}^{\infty} z^2 K^2(z)\, dz \\
&\approx \frac{1}{h} f(x) \int_{-\infty}^{\infty} K^2(z)\, dz =: \frac{1}{h} f(x) \|K\|_2^2.
\end{aligned}$$

Because $(K_h * f)(x) \approx f(x)$ (Theorem 2), this yields the following:[2]

$$\text{Var } \hat{f}(x) \approx \frac{1}{nh} f(x) \|K\|_2^2.$$

Consequently, as $h = h_n \to 0$ and $n \to \infty$,

$$\text{MSE } \hat{f}(x) \approx \frac{1}{nh} f(x) \|K\|_2^2 + \frac{h^4}{4} \left| f''(x) \right|^2 \sigma_K^4. \tag{5}$$

Thus, we can choose $h = h_n$ as the solution to the minimization problem:

$$\min_h \left[\frac{1}{nh} f(x) \|K\|_2^2 + \frac{h^4}{4} \left| f''(x) \right|^2 \sigma_K^4 \right].$$

[2] We are writing $\|h\|_2^2 := \int_{-\infty}^{\infty} h^2(z)\, dz$ and $\sigma_h^2 := \int_{-\infty}^{\infty} z^2 h(z)\, dz$ for any reasonable function h.

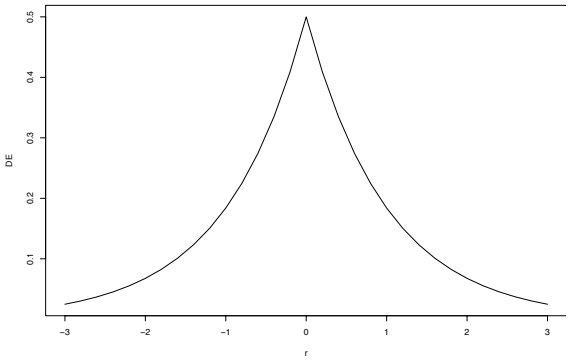

Figure 7: A plot of the double-exponential density.

Let $\psi(h)$ denote the terms in brackets. Then,

$$\psi'(h) = -\frac{1}{nh^2} f(x)\|K\|_2^2 + h^3 \left|f''(x)\right|^2 \sigma_K^4.$$

Set $\psi' \equiv 0$ to find the asymptotically-optimal value of h:

$$h_n := \frac{\alpha_f \beta_K}{n^{1/5}}, \tag{6}$$

where

$$\alpha_f := \frac{(f(x))^{1/5}}{(f''(x))^{2/5}}, \qquad \text{and} \qquad \beta_K := \frac{\|K\|_2^{2/5}}{\sigma_K^{4/5}} = \frac{\left(\int_{-\infty}^{\infty} K^2(z)\,dz\right)^{1/5}}{\left(\int_{-\infty}^{\infty} z^2 K(z)\,dz\right)^{2/5}}. \tag{7}$$

The asymptotically optimal MSE is obtained upon plugging in this h_n into (5). That is,

$$\begin{aligned} \text{MSE}_{opt}\ \hat{f}(x) &\approx \frac{1}{nh_n} f(x)\|K\|_2^2 + \frac{h_n^4}{4} |f''(x)|^2 \sigma_K^4 \\ &= \frac{1}{n^{4/5}} \left[\frac{f(x)\|K\|_2^2}{\alpha_f \beta_K} + \frac{1}{4} \alpha_f^4 \beta_K^4 |f''(x)|^2 \sigma_K^4 \right] \\ &= \frac{\|K\|_2^{8/5} \sigma_K^{4/5}}{n^{4/5}} \left[\frac{f(x)}{\alpha_f} + \frac{\alpha_f^4 |f''(x)|^2}{4} \right]. \end{aligned} \tag{8}$$

Example 5 A commonly-used kernel is the double exponential density. It is described by

$$K(x) := \frac{1}{2} e^{-|x|}.$$

See Figure 7 for a plot. By symmetry,

$$\sigma_K^2 = \int_0^\infty x^2 e^{-x}\,dx = 2, \qquad \|K\|_2^2 = \frac{1}{2}\int_0^\infty e^{-2x}\,dx = \frac{1}{4}, \qquad \beta_K = \frac{4^{-1/5}}{2^{2/5}} = \frac{1}{2^{4/5}}.$$

Therefore,

$$h_n = \frac{C}{n^{1/5}} \quad \text{where} \quad C = \frac{\alpha_f}{2^{4/5}}. \tag{9}$$

Similarly,

$$\text{MSE}_{opt}\ \hat{f}(x) \approx \frac{D}{n^{4/5}} \quad \text{where} \quad D = \frac{1}{2^{1/5}} \left[\frac{f(x)}{\alpha_f} + \frac{|f''(x)|^2 \alpha_f^4}{8} \right]. \tag{10}$$

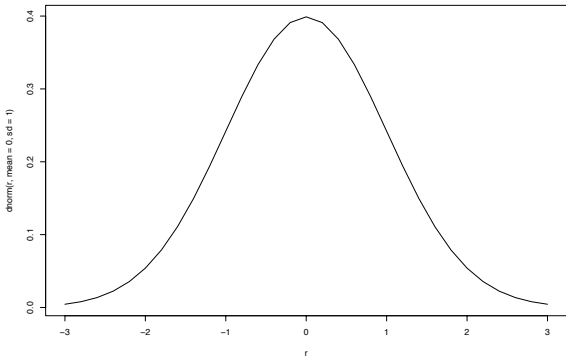

Figure 8: A plot of the $N(0,1)$ density.

Example 6 Let $\tau > 0$ be fixed. Then, the $N(0,\tau^2)$ density is another commonly-used example; i.e.,

$$K(x) = \frac{1}{\tau\sqrt{2\pi}} e^{-x^2/(2\tau^2)}.$$

See Figure 8. In this case, $\sigma_K^2 = \int_{-\infty}^{\infty} z^2 K(z)\,dz = \tau^2$, and

$$\|K\|_2^2 = \frac{1}{2\pi\tau^2}\int_{-\infty}^{\infty} e^{-x^2/\tau^2}\,dx = \frac{1}{2\pi\tau}\times\sqrt{\pi} = \frac{1}{2\tau\sqrt{\pi}}.$$

Consequently,

$$\beta_K = \frac{1}{\left(2\tau\sqrt{\pi}\right)^{1/5}}. \tag{11}$$

This yields,

$$h_n = \frac{C}{n^{1/5}}, \quad \text{where} \quad C = \frac{\alpha_f}{\left(2\tau\sqrt{\pi}\right)^{1/5}}. \tag{12}$$

Similarly,

$$\text{MSE}_{opt}\ \hat{f}(x) \approx \frac{D}{n^{4/5}} \quad \text{where} \quad D = \frac{1}{\left(2\tau\sqrt{\pi}\right)^{4/5}}\left[\frac{f(x)}{\alpha_f} + \frac{\tau^4\alpha_f^4|f''(x)|^2}{4}\right]. \tag{13}$$

4.2 Global Estimation

If we are interested in estimating f "globally," then we need a more global notion of mean-squared error. A useful and easy-to-use notion is the "mean-integrated-squared error" or "MISE." It is defined as

$$\text{MISE}\,\hat{f} := \text{E}\left[\int_{-\infty}^{\infty}\left|\hat{f}(x) - f(x)\right|^2 dx\right].$$

It is easy to see that

$$\text{MISE}\,\hat{f} = \int_{-\infty}^{\infty}\text{MSE}\left(\hat{f}(x)\right)dx.$$

Therefore, under the present smoothness assumptions,

$$\begin{aligned}\text{MISE}\,\hat{f} &\approx \frac{1}{nh}\int_{-\infty}^{\infty}K^2(z)\,dz + \frac{h^4}{4}\int_{-\infty}^{\infty}\left|f''(x)\right|^2 dx\cdot\left(\int_{-\infty}^{\infty}z^2K(z)\,dz\right)^2\\ &:= \frac{1}{nh}\|K\|_2^2 + \frac{h^4}{4}\|f''\|_2^2\sigma_K^4.\end{aligned} \tag{14}$$

See (5). Set

$$\psi(h) := \frac{1}{nh}\|K\|_2^2 + \frac{h^4}{4}\|f''\|_2^2\sigma_K^4,$$

so that

$$\psi'(h) = -\frac{1}{nh^2}\|K\|_2^2 + h^3\|f''\|_2^2\sigma_K^4.$$

Set $\psi' \equiv 0$ to find the asymptotically optimal bandwidth size for the minimum-MISE:

$$h_n := \frac{C}{n^{1/5}} \quad \text{where} \qquad C = \frac{\beta_K}{\|f''\|_2^{2/5}}. \tag{15}$$

See (7) for the notation on β_K. The asymptotically optimal MISE is obtained upon plugging in this h_n into (14). That is,

$$\begin{aligned} \text{MISE}_{opt}\,\hat{f}(x) &\approx \frac{1}{nh_n}\|K\|_2^2 + \frac{h_n^4}{4}\|f''\|_2^2\sigma_K^4 \\ &= \frac{D}{n^{4/5}} \quad \text{where} \quad D = \frac{5}{4}\|f''\|_2^{2/5}\|K\|_2^{8/5}\sigma_K^{4/5}. \end{aligned} \tag{16}$$

Example 7 (Example 5, Continued) In the special case where K is the double-exponential density,

$$h_n = \frac{C}{n^{1/5}} \quad \text{where} \quad C = \frac{1}{2^{4/5}\|f''\|_2^{2/5}}. \tag{17}$$

Also,

$$\text{MISE}_{opt}\,\hat{f}(x) \approx \frac{D}{n^{4/5}} \quad \text{where} \quad D = \frac{5}{2^{16/5}}\|f''\|_2^{2/5}. \tag{18}$$

Example 8 (Example 6, Continued) In the special case where K is the $N(0,\tau^2)$ density,

$$h_n = \frac{C}{n^{1/5}} \quad \text{where} \quad C = \frac{1}{\left(2\tau\sqrt{\pi}\right)^{1/5}\|f''\|_2^{2/5}}. \tag{19}$$

Also,

$$\text{MISE}_{opt}\,\hat{f}(x) \approx \frac{D}{n^{4/5}} \quad \text{where} \quad D = \frac{5}{2^{14/5}\pi^{2/5}}\|f''\|_2^{2/5}. \tag{20}$$

5 Problems and Some Remedies for Kernel Density Estimators

The major drawback of the preceding computations is that h_n depends on f. Typically, one picks a related value of h where the dependence on f is replaced by a similar dependency, but on a known family of densities. But there are other available methods as well. I will address two of them next.[3]

1. *The Subjective Method:* Choose various "sensible" values of h (e.g., set $h = cn^{-1/5}$ and vary c). Plot the resulting density estimators, and choose the one whose general shape matches up best with your prior belief. This can be an effective way to obtain a density estimate some times.

2. *Reference to Another Density:* To be concrete, consider h_n for the global estimate. Thus, the optimal h has the form, $h_n = \beta_K\|f''\|_2^{-2/5}n^{-1/5}$. Now replace $\|f''\|_2^{2/5}$ by $\|g''\|_2^{2/5}$ for a nice density function g. A commonly-used example is $g := N(0,\tau^2)$ density. Let $\varphi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$ be the standard-normal density. Note that $g(x) = \tau^{-1}\varphi(x/\tau)$. Therefore, $g''(x) = \tau^{-3}\varphi''(x/\tau)$, whence it follows that

$$\|g\|_2^2 = \frac{1}{\tau^6}\int_{-\infty}^{\infty}\left[\varphi''\left(\frac{x}{\tau}\right)\right]^2 dx = \frac{1}{\tau^5}\int_{-\infty}^{\infty}\left[\varphi''(y)\right]^2 dy = \frac{1}{2\pi\tau^5}\int_{-\infty}^{\infty}e^{-y^2}\left(y^2-1\right)^2 dy = \frac{3}{8\tau^5\sqrt{\pi}}.$$

   This is about $0.2115/\tau^5$. So we can choose the bandwidth $h := \beta_K\|g''\|_2^{-2/5}n^{-1/5}$; i.e.,

$$h = \frac{8^{1/5}\pi^{1/10}}{3^{1/5}}\cdot\frac{\tau\beta_K}{n^{1/5}}.$$

   To actually use this we need to know τ. But our replacement of f by g tacitly assumes that the variance of the date is τ^2; i.e., that $\tau^2 = \int_{-\infty}^{\infty}x^2 f(x)\,dx - (\int_{-\infty}^{\infty}xf(x)\,dx)^2$. So we can estimate τ^2 by traditional methods, plug, and proceed to use the resulting h. If f is truly normal, then this method works very well. Of course, you should also a normal density K as well in such cases. However, if f is "far" from normal, then $\|f''\|_2$ tends to be a lot larger than $\|g''\|_2$. Therefore, our h is much larger than the asymptotically optimal h_n. This results in *over smoothing*.

[3]We may note that by choosing K correctly, we can ensure that $\|K\|_2^2$ is small. In this way we can reduce the size of $\text{MISE}_{opt}\,\hat{f}$, for instance. But the stated problem with the bandwidth is much more serious.

6 Bias Reduction via Signed Estimators

One of the attractive features of kernel density estimators is the property that they are themselves probability densities. In particular, they have the positivity property, $\hat{f}(x) \geq 0$ for all x. If we did not need this to hold, then we can get better results. In such a case the end-result needs to be examined with extra care, but could still be useful.

So now we suppose that the kernel K has the following properties:

- [*Symmetry*] $K(x) = K(-x)$ for all x;
- $\int_{-\infty}^{\infty} K(x)\,dx = 1$;
- $\mu_2(K) = 0$, where $\mu_\ell(K) := \int_{-\infty}^{\infty} x^\ell K(x)\,dx$;
- $\mu_4(K) \neq 0$.

Then, we proceed with a four-term Taylor series expansion: If h is small then we would expect that

$$f(x-ha) \approx f(x) - ha f'(x) + \frac{h^2a^2}{2} f''(x) - \frac{h^3a^3}{6} f'''(x) + \frac{h^4a^4}{24} f^{(iv)}(x).$$

Therefore,

$$\begin{aligned}
\text{Bias}\,\hat{f}(x) &= (K_h * f)(x) - f(x) = \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{x-u}{h}\right) f(u)\,du - f(x) \\
&= \int_{-\infty}^{\infty} K(a) f(x-ah)\,da - f(x) \\
&\approx \int_{-\infty}^{\infty} K(a) \left[f(x) - ha f'(x) + \frac{h^2a^2}{2} f''(x) - \frac{h^3a^3}{6} f'''(x) + \frac{h^4a^4}{24} f^{(iv)}(x) \right] da - f(x) \\
&= \mu_4(K) \frac{h^4}{24} f^{(iv)}(x).
\end{aligned}$$

Thus, the bias is of the order h^4. This is a substantial gain from before when we insisted that K be a density function. In that case, the bias was of the order h^2 see (4).

We continue as before and compute the asymptotic variance, as well:

$$\begin{aligned}
\left(K_h^2 * f\right)(x) &= \frac{1}{h^2} \int_{-\infty}^{\infty} K^2\left(\frac{x-u}{h}\right) f(u)\,du = \frac{1}{h} \int_{-\infty}^{\infty} K^2(a) f(x-ah)\,da \\
&\approx \frac{1}{h} \int_{-\infty}^{\infty} K^2(a) \left[f(x) - ha f'(x) + \frac{h^2a^2}{2} f''(x) \right] da \\
&= \frac{1}{h} f(x) \int_{-\infty}^{\infty} K^2(a)\,da = \frac{\|K\|_2^2 f(x)}{h},
\end{aligned}$$

as before. Thus, as before,

$$\text{Var}\,\hat{f}(x) = \frac{1}{n} \left[\left(K_h^2 * f\right)(x) - (K_h * f)^2(x) \right] \approx \frac{\|K\|_2^2 f(x)}{nh}.$$

Therefore,

$$\text{MSE}\,\hat{f}(x) \approx \frac{\|K\|_2^2 f(x)}{nh} + \mu_4^2(K) \frac{h^8}{576} \left[f^{(iv)}(x) \right]^2. \tag{21}$$

Write this, as before, as $\psi(h)$, and compute

$$\psi'(h) = -\frac{\|K\|_2^2 f(x)}{nh^2} + \mu_4^2(K) \frac{h^7}{72} \left[f^{(iv)}(x) \right]^2.$$

Set $\psi'(h) \equiv 0$ to find that there exist constants C, D, and E, such that $h_n = Cn^{-1/9}$, MSE $\hat{f}(x) \approx Dn^{-8/9}$, and MISE $\hat{f} \approx En^{-8/9}$. I will leave up to you to work out the remaining details (e.g., compute C, D, and E). Instead, let us state a few examples of kernels K that satisfy the assumptions of this section.

Example 9 A classical example is

$$K(x) = \begin{cases} \frac{3}{8}(3-5x^2), & \text{if } |x| < 1, \\ 0, & \text{otherwise.} \end{cases}$$

A few lines of calculations reveal that: (i) K is symmetric; (ii) $\int_{-\infty}^{\infty} K(x)\,dx = 1$; (iii) $\int_{-\infty}^{\infty} x^2 K(x)\,dx = 0$; and (iv) $\mu_4(K) = \int_{-\infty}^{\infty} x^4 K(x)\,dx = -3/35 \neq 0$.

Example 10 We obtain another family of classical examples, due to W. R. Schucany and J. P. Sommers,[4] by first choosing a (proper probability density) kernel K, and then modifiying it as follows: Let $\nu > 1$ be fixed, and define

$$K_\nu(x) := \left(\frac{\nu^2}{\nu^2-1}\right)\left[K(x) - \frac{1}{\nu^3}K\left(\frac{x}{\nu}\right)\right].$$

Suppose K is symmetric and has four finite moments. Then, a few lines of calculations reveal that K_ν satisfies the conditions of the kernels of this section. Namely: (i) K_ν is symmetric; (ii) $\int_{-\infty}^{\infty} K_\nu(x)\,dx = 1$; (iii) $\int_{-\infty}^{\infty} x^2 K_\nu(x)\,dx = 0$; and (iv) $\mu_4(K_\nu) = \int_{-\infty}^{\infty} x^4 K_\nu(x)\,dx = -\nu^2\mu_4(K_\nu) \neq 0$. Schucany and Sommers recommend using values of ν that are > 1, but very close to one.

7 Cross-Validation

Let $\hat{f}$ denote the kernel density estimator of f based on a reasonable (density) kernel K. Define the integrated squared error as

$$\int_{-\infty}^{\infty} \left|\hat{f}(x) - f(x)\right|^2 dx = \int_{-\infty}^{\infty} \left|\hat{f}(x)\right|^2 dx - 2\int_{-\infty}^{\infty} \hat{f}(x) f(x)\,dx + \int_{-\infty}^{\infty} [f(x)]^2\,dx.$$

One way to find an optimal h, then, is to minimize this over all h, and find the (an?) h that minimizes the mentioned error. Because the last term in the display depends on f [and not on h], our problem is reduced to the following:

$$\min_h R(\hat{f}),$$

where

$$R(\hat{f}) := \int_{-\infty}^{\infty} \left|\hat{f}(x)\right|^2 dx - 2\int_{-\infty}^{\infty} \hat{f}(x) f(x)\,dx.$$

Of course, this contains f, and so need to be estimate first. The naive thing to do is to estimate $\int_{-\infty}^{\infty} \hat{f}(x) f(x)\,dx$ by $\int_{-\infty}^{\infty} |\hat{f}(x)|^2\,dx$ and then minimize over h. But then our estimate for $R(\hat{f})$ becomes $-\int_{-\infty}^{\infty} |\hat{f}(x)|^2\,dx$, whose minimum is nearly always $-\infty$, and is achieved at $h = 0$ [a bad bandwidth!].

We estimate $\int_{-\infty}^{\infty} \hat{f}(x) f(x)\,dx$ by a resampling method called "cross validation." We will return to resampling methods later on.

For all $1 \le i \le n$ define

$$\hat{f}_i(x) := \frac{1}{(n-1)h} \sum_{\substack{1\le j\le n\\ j\neq i}} K\left(\frac{x - X_j}{h}\right).$$

In words: Remove X_i from the data and then let $\hat{f}_i$ be the resulting kernel density estimate for the modified data. Note that

$$\begin{aligned}
\mathrm{E}\left[\hat{f}_i(X_i)\right] &= \frac{1}{(n-1)h} \sum_{\substack{1\le j\le n\\ j\neq i}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K\left(\frac{x-y}{h}\right) f(x) f(y)\,dx\,dy \\
&= \frac{1}{h}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K\left(\frac{x-y}{h}\right) f(x) f(y)\,dx\,dy \\
&= \int_{-\infty}^{\infty} (K_h * f)(y) f(y)\,dy.
\end{aligned}$$

This uses the fact that the random function $\hat{f}_i$ is independent of X_i. Because $\mathrm{E}[\hat{f}(x)] = (K_h * f)(x)$, we have also that

$$\mathrm{E}\left[\int_{-\infty}^{\infty} \hat{f}(x) f(x)\,dx\right] = \int_{-\infty}^{\infty} (K_h * f)(x) f(x)\,dx = \mathrm{E}\left[\hat{f}_i(X_i)\right].$$

The left-hand side does not depend on i. So we can average it to find that

$$\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \hat{f}_i(X_i)\right] = \mathrm{E}\left[\int_{-\infty}^{\infty} \hat{f}(x) f(x)\,dx\right] = \int_{-\infty}^{\infty} f^2(x)\,dx.$$

But we can expect that if n is large, then with high probability,

$$\frac{1}{n}\sum_{i=1}^{n} \hat{f}_i(X_i) \approx \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \hat{f}_i(X_i)\right] \quad \text{and} \quad \int_{-\infty}^{\infty} \hat{f}(x) f(x)\,dx \approx \int_{-\infty}^{\infty} f^2(x)\,dx.$$

[4] W. R. Schucany and J. P. Sommers (1977), Improvement of kernel type density estimators, *JASA* **72**, 420–423.

Thus, we are led to the cross-validation estimator $n^{-1}\sum_{i=1}^n \hat{f}_i(X_i)$ of $\int_{-\infty}^{\infty}\hat{f}(x)f(x)\,dx$. In order to find a good bandwidth, we solve

$$\min_h T(h),$$

where

$$T(h) := \int_{-\infty}^{\infty}\left|\hat{f}(x)\right|^2 dx - \frac{2}{n}\sum_{i=1}^n \hat{f}_i(X_i).$$

A practical method of "solving" this minimization problem is this: Compute $T(h)$ for a large group of h's of the form $h = cn^{-1/5}$, and choose the best c amongst that group. There are theoretical justifications that this method works, but they involve (very) large-sample analysis.

8 Consistency

It turns out that under some conditions on h, K, etc. the kernel density estimator is consistent. That is, there is a sense in which $\hat{f} \approx f$ for n large. We analyze three cases of this phenomenon:

1. Fix $x \in \mathbf{R}$. Then, we want to know that under some reasonable conditions, $\lim_n \hat{f}(x) = f(x)$ in probability. This is "pointwise consistency."
2. We want to know that under reasonable conditions, $\hat{f} \approx f$ in some global sense. A strong case can be made for the so-called "L^1 distance" between $\hat{f}$ and f. That is, we wish to know that under some natural conditions, $\lim_{n\to\infty}\int_{-\infty}^{\infty}|\hat{f}(x) - f(x)|\,dx = 0$ in probability. This is "consistency in L^1."
3. For some applications (e.g., mode-finding), we need to know that $\max_x|\hat{f}(x) - f(x)| \to 0$ in probability. This is the case of "uniform consistency."

8.1 Consistency at a Point

In this subsection we study that case where we are estimating $f(x)$ *locally*. That is, we fix some point $x \in \mathbf{R}$, and try to see if $\hat{f}(x) \approx f(x)$ for large values of n. For this to make sense we need to bandwidth h to depend on n, and go to zero as $n \to \infty$. We shall write h_n in place of h, but this h_n need not be the asymptotically optimal one that was referred to earlier. This notation will be adopted from here on.

The following is a stronger form of a classical consistency theorem of E. Parzen.[5]

Theorem 11 (Parzen) *Let us assume the following:*

1. *K vanishes at infinity, and $\int_{-\infty}^{\infty}K^2(x)\,dx < \infty$;*
2. *$h_n \to 0$ as $n \to \infty$; and*
3. *$nh_n \to \infty$ as $n \to \infty$.*

Then, whenever f is continuous in an open neighborhood of x we have $\hat{f}(x) \xrightarrow{\text{P}} f(x)$, as $n \to \infty$.

Proof: Throughout, we choose and fix an x around which f is continuous.

Recall from page 10 that

$$\begin{aligned}\text{E}\,\hat{f}(x) &= (K_{h_n} * f)(x),\\ \text{Var}\,\hat{f}(x) &= \frac{1}{n}\left[\left(K_{h_n}^2 * f\right)(x) - (K_{h_n} * f)^2(x)\right].\end{aligned}$$

It might help to recall the notation on convolutions. In particular, we have

$$\left(K_{h_n}^2 * f\right)(x) = \frac{1}{h_n^2}\int_{-\infty}^{\infty}K^2\left(\frac{x-y}{h_n}\right)f(y)\,dy.$$

Note that $K_{h_n}^2$ is really short-hand for $(K_{h_n})^2$. Let $G(x) := K^2(x)$ to find then that

$$\left(K_{h_n}^2 * f\right)(x) = \frac{1}{h_n}\left(G_{h_n} * f\right)(x).$$

[5] E. Parzen (1962). On estimation of a probability density function and mode, *Ann. Math. Statist.* **33**, 1065–1076.

Now, $G(x)/\int_{-\infty}^{\infty} K^2(u)\,du$ is a probability density that vanishes at infinity. Therefore, we can apply Theorem 2 to G to find that

$$\left(K_{h_n}^2 * f\right)(x) \sim \frac{f(x)}{h_n} \int_{-\infty}^{\infty} K^2(u)\,du.$$

Another application of Theorem 2 shows that $(K_{h_n} * f)(x) \to f(x)$. Therefore,

$$\operatorname{Var} \hat{f}(x) \sim \frac{1}{n}\left[\frac{f(x)}{h_n}\int_{-\infty}^{\infty} K^2(u)\,du - f(x)\right] \sim \frac{f(x)}{nh_n}\int_{-\infty}^{\infty} K^2(u)\,du. \tag{22}$$

Since $nh_n \to 0$, this proves that $\operatorname{Var} \hat{f}(x) \to 0$ as $n \to \infty$. Thanks to the Chebyshev inequality,

$$\hat{f}(x) - \mathrm{E}\,\hat{f}(x) \xrightarrow{\mathrm{P}} 0.$$

But another application of Theorem 2 shows that $\lim_{n\to\infty} \mathrm{E}\,\hat{f}(x) = f(x)$, because $h_n \to 0$. The theorem follows. □

8.2 L^1-Consistency

Here we prove a weaker formulation of a theorem of L. Devroye.[6] The following is a global counterpart of the local Theorem 11.

Theorem 12 (Devroye) *Suppose K is bounded, $h_n \to 0$, and $nh_n \to \infty$ as $n \to \infty$. Then, as $n \to \infty$,*

$$\int_{-\infty}^{\infty} \left|\hat{f}(x) - f(x)\right| dx \xrightarrow{\mathrm{P}} 0.$$

Before we prove this let us state two facts: One from real analysis, and one from probability.

Lemma 13 (The Cauchy–Schwarz Inequality) *Let h be a nonnegative function and $m > 0$ a fixed number. Then,*

$$\int_{-m}^{m} \sqrt{h(x)}\,dx \le \sqrt{2m\int_{-m}^{m} h(x)\,dx}.$$

Proof: Let U be distributed uniformly on $(-m, m)$. Then,

$$\frac{1}{2m}\int_{-m}^{m} \sqrt{h(x)}\,dx = \mathrm{E}\left[\sqrt{h(U)}\right], \quad \text{and} \quad \frac{1}{2m}\int_{-m}^{m} h(x)\,dx = \mathrm{E}\left[h(U)\right].$$

But $\mathrm{E}Z \le \sqrt{\mathrm{E}(Z^2)}$ for all nonnegative random variables Z.[7] Apply this with $Z = \sqrt{h(U)}$ to finish. □

Lemma 14 *If K is bounded above everywhere by some constant B, then*

$$\operatorname{Var} \hat{f}(x) \le \frac{B}{nh_n}\left(K_{h_n} * f\right)(x).$$

Proof: During the course of the proof of Theorem 11 we saw that

$$\operatorname{Var} \hat{f}(x) = \frac{1}{n}\left[\left(K_{h_n}^2 * f\right)(x) - \left(K_{h_n} * f\right)^2(x)\right] \le \frac{1}{n}\left(K_{h_n}^2 * f\right)(x).$$

Because $0 \le K(a) \le B$ for all a, it follows that $K_{h_n}^2(a) \le (B/h_n)K_{h_n}(a)$ for all a. Thus, the lemma follows. □

Proof of Theorem 12: If h and g are density functions, we consider

$$\|h - g\|_1 := \int_{-\infty}^{\infty} |h(x) - g(x)|\,dx.$$

This is the so-called L^1-norm, and forms a global measure of how far h and g are.

Define

$$M(x) := \mathrm{E}\,\hat{f}(x).$$

[6] L. Devroye (1983). The equivalence of weak, strong and complete convergence in density estimation in L_1 for kernel density estimates, *Ann. Statis.* **11**, 896–904.

[7] This follows immediately from the fact that $0 \le \operatorname{Var} Z = \mathrm{E}(Z^2) - |\mathrm{E}Z|^2$.

Note that $\int_{-\infty}^{\infty} M(x)\,dx = \mathrm{E}\int_{-\infty}^{\infty} \hat{f}(x)\,dx = \int_{-\infty}^{\infty}(K_h * f)(x)\,dx = 1$. Therefore, M is a probability density function. Moreover,

$$\mathrm{E}\|\hat{f} - M\|_1 = \int_{-\infty}^{\infty} \mathrm{E}|\hat{f}(x) - \mathrm{E}\hat{f}(x)|\,dx = \int_{-m}^{m} \mathrm{E}|\hat{f}(x) - \mathrm{E}\hat{f}(x)|\,dx + \int_{\{|x|>m\}} \mathrm{E}|\hat{f}(x) - \mathrm{E}\hat{f}(x)|\,dx,$$

where $m > 0$ is large but fixed. Now recall that if Z is a nonnegative random variable then $\mathrm{E}Z \le \sqrt{\mathrm{E}(Z^2)}$.Therefore,

$$\begin{aligned}\mathrm{E}\|\hat{f} - M\|_1 &\le \int_{-m}^{m} \sqrt{\mathrm{Var}\,\hat{f}(x)}\,dx + \int_{\{|x|>m\}} \mathrm{E}|\hat{f}(x) - \mathrm{E}\hat{f}(x)|\,dx \le \int_{-m}^{m} \sqrt{\mathrm{Var}\,\hat{f}(x)}\,dx + 2\int_{\{|x|>m\}} \mathrm{E}\hat{f}(x)\,dx \\ &= \int_{-m}^{m} \sqrt{\mathrm{Var}\,\hat{f}(x)}\,dx + 2\int_{\{|x|>m\}} (K_{h_n} * f)(x)\,dx.\end{aligned}$$

The last inequality follows from the facts that: (i) $f, \hat{f}$ are nonnegative; and so (ii) $|\hat{f}(x) - \mathrm{E}\hat{f}(x)| \le \hat{f}(x) + \mathrm{E}\hat{f}(x)$. Lemma 14 applies and we have the following bound:

$$\mathrm{E}\|\hat{f} - M\|_1 \le \sqrt{\frac{B}{nh_n}} \int_{-m}^{m} \sqrt{(K_{h_n} * f)(x)}\,dx + 2\int_{\{|x|>m\}} (K_{h_n} * f)(x)\,dx.$$

By the Cauchy–Schwarz inequality,

$$\int_{-m}^{m} \sqrt{(K_{h_n} * f)(x)}\,dx \le \sqrt{2m\int_{-m}^{m} (K_{h_n} * f)(x)\,dx} \le \sqrt{2m\int_{-\infty}^{\infty} (K_{h_n} * f)(x)\,dx} = \sqrt{2m}.$$

Hence,

$$\lim_{n\to\infty} \mathrm{E}\|\hat{f} - M\|_1 \le \lim_{n\to\infty} \sqrt{\frac{2mB}{nh_n}} + 2\lim_{n\to\infty} \int_{\{|x|>m\}} (K_{h_n} * f)(x)\,dx = 2\lim_{n\to\infty} \int_{\{|x|>m\}} (K_{h_n} * f)(x)\,dx.$$

But

$$\int_{\{|x|>m\}} (K_{h_n} * f)(x)\,dx = \int_{\{|x|>m\}} \frac{1}{h_n}\int_{-\infty}^{\infty} K\left(\frac{x-y}{h_n}\right) f(y)\,dy\,dx = \int_{-\infty}^{\infty} K(z)\left(\int_{\{|y+zh_n|>m\}} f(y)\,dy\right) dz.$$

Because $h_n \to 0$, a standard theorem of integration theory implies the following:[8]

$$\lim_{n\to\infty} \int_{\{|x|>m\}} (K_{h_n} * f)(x)\,dx = \int_{-\infty}^{\infty} K(z)\left(\int_{\{|y|>m\}} f(y)\,dy\right) dz = \int_{\{|y|>m\}} f(y)\,dy.$$

As a result, the following holds for all $m > 0$:

$$\lim_{n\to\infty} \mathrm{E}\|\hat{f} - M\|_1 \le 2\int_{\{|y|>m\}} f(y)\,dx.$$

Let $m \to \infty$ to deduce that $\lim_{n\to\infty} \mathrm{E}\|\hat{f} - M\|_1 = 0$. On the other hand, it is not hard to verify the "triangle inequality," $\|\hat{f} - f\|_1 \le \|\hat{f} - M\|_1 + \|M - f\|_1$. Therefore, it remains to prove that $\|M - f\|_1 \to 0$ as $n \to \infty$. But

$$\|M - f\|_1 = \|\mathrm{E}\hat{f} - f\|_1 = \|(K_{h_n} * f) - f\|_1 \to 0,$$

thanks to Theorem 3. □

8.3 Remarks on the L^1-Norm

Suppose X_n converges in distribution to X, and assume that X has a density f. Let $F_n(x) := \mathrm{P}\{X_n \le x\}$ and $F(x) = \int_0^x f(u)\,du = \mathrm{P}\{X \le x\}$. Then, convergence in distribution, in this setting, amounts to the limit theorem: $\lim_{n\to\infty} F_n(x) = F(x)$ for all $x \in \mathbf{R}$. A little observation of Polýa asserts that a slightly stronger statement is true in this case. Namely, that $\max_x |F_n(x) - F(x)| \to 0$ as $n \to \infty$. [This really needs the continuity of F.]

There is another notion of convergence "in distribution." We say that X_n converges to X in *total variation* if

$$\lim_{n\to\infty} \max_A |\mathrm{P}\{X_n \in A\} - \mathrm{P}\{X \in A\}| = 0.$$

The maximum is taken over all (measurable) sets A. Note that

$$\max_x |F_n(x) - F(x)| = \max_x \Big|\mathrm{P}\{X_n \in (-\infty, x]\} - \mathrm{P}\{X \in (-\infty, x]\}\Big| \le \max_A \Big|\mathrm{P}\{X_n \in A\} - \mathrm{P}\{X \in A\}\Big|.$$

[8] The requisite theorem is the so-called "dominated convergence theorem" of H. Lebesgue.

Therefore, convergence in total variation certainly implies convergence in distribution. The converse, however, is false. A notable counter-example in statistics is the most important classical CLT for binomial random variables.

Let $B_n \sim \text{binomial}(n, 1/2)$. Then we know that

$$X_n := \frac{B_n - (n/2)}{\sqrt{n/4}} \xrightarrow{d} X := N(0,1).$$

Let F_n and F denote the respective distribution functions of X_n and X. Because X has a nice density function it follows from Polýa's theorem that $\max_x |F_n(x) - F(x)| \to 0$ as $n \to \infty$. However, it is not the case that X_n converges to X in total variation. Indeed, $\max_A |\mathrm{P}\{X_n \in A\} - \mathrm{P}\{X \in A\}| = 1$. To see this, set A to be the collection of all real numbers of the form $(\ell - (n/2))/\sqrt{n/4}$, as ℓ ranges over all integers in $\{0, \ldots, n\}$. Evidently, $\mathrm{P}\{X_n \in A\} = 1$ and $\mathrm{P}\{X \in A\} = 0$. It follows that *convergence in total variation is a much stronger statement than convergence in distribution.* The following shows why the L^1-norm is a natural measure of distance between densities.

Theorem 15 (Scheffé) *Let X and Y be two continuous random variables with respective density functions f_X and f_Y. Then,*

$$\int_{-\infty}^{\infty} |f_X(a) - f_Y(a)|\, da = 2 \max_A \Big| \mathrm{P}\{X \in A\} - \mathrm{P}\{Y \in A\} \Big|.$$

Proof: For simplicity, define

$$d_{\mathrm{TV}}(X, Y) := \max_A |\mathrm{P}\{X \in A\} - \mathrm{P}\{Y \in A\}|.$$

For any set A,

$$\int_A (f_X(a) - f_Y(a))\, da = \int_{A^c} (f_Y(a) - f_X(a))\, da. \tag{23}$$

To prove this collect the terms involving f_X on one side and those involving f_Y on the other to find that $1 = 1$.

Now define the (measurable) set

$$A_* := \{a \in \mathbf{R} :\ f_X(a) > f_Y(a)\}.$$

Notice that

$$\Big| \mathrm{P}\{X \in A_*\} - \mathrm{P}\{Y \in A_*\} \Big| = \int_{A_*} (f_X(a) - f_Y(a))\, da = \int_{A_*} |f_X(a) - f_Y(a)|\, da.$$

But $|\mathrm{P}\{X \in A\} - \mathrm{P}\{Y \in A\}| = |\mathrm{P}\{X \in A^c\} - \mathrm{P}\{Y \in A^c\}|$. Therefore,

$$\Big| \mathrm{P}\{X \in A_*\} - \mathrm{P}\{Y \in A_*\} \Big| = \int_{A_*^c} (f_Y(a) - f_X(a))\, da = \int_{A_*^c} |f_X(a) - f_Y(a)|\, da.$$

Add the two displays to find that

$$2 d_{\mathrm{TV}}(X, Y) \geq 2 \Big| \mathrm{P}\{X \in A_*\} - \mathrm{P}\{Y \in A_*\} \Big| = \int_{-\infty}^{\infty} |f_X(a) - f_Y(a)|\, da.$$

This proves half of the theorem.

For the converse, suppose A is an arbitrary (measurable) set, and recall the definition of A_*. We have

$$\begin{aligned}
\left| \int_A f_X(a)\, da - \int_A f_Y(a)\, da \right| &= \left| \int_{A \cap A_*} (f_X(a) - f_Y(a))\, da + \int_{A \cap A_*^c} (f_X(a) - f_Y(a))\, da \right| \\
&= \left| \int_{A \cap A_*} |f_X(a) - f_Y(a)|\, da - \int_{A \cap A_*^c} |f_X(a) - f_Y(a)|\, da \right| \\
&\leq \max \left\{ \int_{A \cap A_*} |f_X(a) - f_Y(a)|\, da\ ,\ \int_{A \cap A_*^c} |f_X(a) - f_Y(a)|\, da \right\},
\end{aligned}$$

because $|z - w| \leq \max\{z, w\}$ for any two positive numbers w and z. If we now replace $A \cap A_*$ by A_* the right-most term in the display increases. The same is valid if we replace $A \cap A_*^c$ by A_*^c. Therefore,

$$|\mathrm{P}\{X \in A\} - \mathrm{P}\{Y \in A\}| \leq \max \left\{ \int_{A_*} |f_X(a) - f_Y(a)|\, da\ ,\ \int_{A_*^c} |f_X(a) - f_Y(a)|\, da \right\}.$$

Thanks to (23) the two integrals are equal. Therefore, they must equal half of their sum. However, their sum is $\int_{-\infty}^{\infty} |f_X(a) - f_Y(a)|\, da$. Maximize over A to obtain the result. □

Now let us return to kernel density estimation in the context of Theorem 12. Recall that $\hat{f}$ and f are densities, although $\hat{f}$ is random.

Define a second empirical distribution function that is based on our kernel density estimate; i.e., let

$$\widetilde{F}_n(a) := \int_{-\infty}^{a} \hat{f}(x)\,dx, \qquad \text{for all } a \in \mathbf{R}.$$

By Scheffé's theorem,

$$\max_A \left| \int_A d\widetilde{F}_n(x) - \int_A dF(x) \right| \xrightarrow{\text{P}} 0 \quad \text{as } n \to \infty.$$

[If you want to understand this rigorously, then be aware that these are Stieldtjes integrals.] However, it is not hard to see that this fails if we replace $\widetilde{F}_n$ by the usual empirical distribution function $\hat{F}_n$. This shows that the empirical distribution function based on $\hat{f}$ provides a better approximation to the usual empirical distribution function.

8.4 Uniform Consistency

Theorem 12 is a natural global-consistency theorem. But it falls shy of addressing an important application of density estimation to which we will come in the next subsection. That is, estimating the mode of a density. [This was one of the original motivations behind the theory of kernel density estimation. See E. Parzen (1962), On estimation of a probability density function and mode, *Ann. Math. Statist.* **33**, 1065–1076.] Here we address the important issue of *uniform consistency*. That is, we seek to find reasonable conditions under which $\max_x |\hat{f}(x) - f(x)|$ converges to zero in probability.

First we recall a few facts from Fourier analysis. If h is integrable then its *Fourier transform* is the function $\mathscr{F}h$ defined by

$$(\mathscr{F}h)(t) := \int_{-\infty}^{\infty} e^{itx} h(x)\,dx, \qquad \text{for all } t \in \mathbf{R}.$$

Note that whenever $h := f$ is a density function, and it is the case for us, then,

$$(\mathscr{F}f)(t) = \text{E}\left[e^{itX_1}\right], \tag{24}$$

and so $\mathscr{F}f$ is just the characteristic function of X. [See the Probability Notes.] We need the following deep fact from Fourier/harmonic analysis. In rough terms, the following tells us that after multiplying by $(2\pi)^{-1}$, the definition of $\mathscr{F}h$ can be formally inverted to yield a formula for h in terms of its Fourier transform.

Theorem 16 (Inversion Theorem) *Suppose h and $\mathscr{F}h$ are both [absolutely] integrable. Then, for all $x \in \mathbf{R}$,*

$$h(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} (\mathscr{F}h)(t)\,dt.$$

The condition that h is integrable is very natural. For us, h is a probability density, after all. However, it turns out that the absolute integrability of $\mathscr{F}h$ implies that h is *uniformly continuous*. So this can be a real restriction.

Now note that

$$\begin{aligned}(\mathscr{F}\hat{f})(t) &= \int_{-\infty}^{\infty} e^{itx} \hat{f}(x)\,dx = \frac{1}{nh_n} \sum_{j=1}^{n} \int_{-\infty}^{\infty} e^{itx} K\left(\frac{x - X_j}{h_n}\right) dx = \frac{1}{n} \sum_{j=1}^{n} e^{itX_j} \int_{-\infty}^{\infty} e^{ih_n ty} K(y)\,dy \\ &= \frac{1}{n} \sum_{j=1}^{n} e^{itX_j} (\mathscr{F}K)(h_n t).\end{aligned}$$

In particular, $\mathscr{F}\hat{f}$ is integrable as soon as $\mathscr{F}K$ is. If so, then the inversion theorem (Theorem 16) tell us that

$$\hat{f}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} (\mathscr{F}\hat{f})(t)\,dt = \frac{1}{2\pi n} \sum_{j=1}^{n} \int_{-\infty}^{\infty} e^{it(X_j - x)} (\mathscr{F}K)(h_n t)\,dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \frac{1}{n} \sum_{j=1}^{n} e^{itX_j} (\mathscr{F}K)(h_n t)\,dt.$$

Take expectations also to find that

$$\text{E}\hat{f}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \text{E}\left[e^{itX_1}\right] (\mathscr{F}K)(h_n t)\,dt.$$

Therefore,

$$\hat{f}(x) - \text{E}\hat{f}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \left(\phi_n(t) - \text{E}\phi_n(t)\right) (\mathscr{F}K)(h_n t)\,dt,$$

where ϕ_n is the "empirical characteristic function,"

$$\phi_n(t) := \frac{1}{n}\sum_{j=1}^{n} e^{itX_j}, \qquad \text{for all } t \in \mathbf{R}.$$

Because $|e^{itx}| \le 1$, the triangle inequality yields,

$$\max_x \left|\hat f(x) - \mathrm{E}\hat f(x)\right| \le \frac{1}{2\pi}\int_{-\infty}^{\infty} |\phi_n(t) - \mathrm{E}\phi_n(t)| \cdot |(\mathscr{F}K)(h_n t)|\, dt. \tag{25}$$

Take expectations and use the by-now familiar bound $\mathrm{E}|Z| \le \sqrt{\mathrm{E}(Z^2)}$ [see the footnote on page 18] to find that

$$\mathrm{E}\left(\max_x \left|\hat f(x) - \mathrm{E}\hat f(x)\right|\right) \le \frac{1}{2\pi}\int_{-\infty}^{\infty} \sqrt{\mathrm{Var}\,\phi_n(t)}\ |(\mathscr{F}K)(h_n t)|\, dt.$$

[Caution: When Z is complex-valued, by $\mathrm{Var}\,Z$ we really mean $\mathrm{E}|Z - \mathrm{E}Z|^2$.] Now, we can write $e^{itX_j} = \cos(tX_j) + i\sin(tX_j)$. Therefore (check!),

$$\mathrm{Var}\, e^{itX_j} = \mathrm{Var}\cos(tX_j) + \mathrm{Var}\sin(tX_j) \le \mathrm{E}\left[\cos^2(tX_j) + \sin^2(tX_j)\right] = 1.$$

Even it $Z_1, \ldots, Z_n$ are complex-valued, as long as they are i.i.d., $\mathrm{Var}\sum_{j=1}^n Z_j = \sum_{j=1}^n \mathrm{Var}\, Z_j$ (why?). Therefore, $\mathrm{Var}\,\phi_n(t) \le 1/n$. It follows then that

$$\begin{aligned}\mathrm{E}\left(\max_x \left|\hat f(x) - \mathrm{E}\hat f(x)\right|\right) &\le \frac{1}{2\pi\sqrt{n}}\int_{-\infty}^{\infty} |(\mathscr{F}K)(h_n t)|\, dt \\ &= \frac{1}{2\pi h_n\sqrt{n}}\int_{-\infty}^{\infty} |(\mathscr{F}K)(s)|\, ds.\end{aligned}$$

This and Chebyshev's inequality together implies that if $h_n\sqrt{n} \to \infty$ then $\max_x |\hat f(x) - \mathrm{E}\hat f(x)| \to 0$ in probability. Next we prove that if f is uniformly continuous and $h_n \to 0$, then

$$\max_x \left|\mathrm{E}\hat f(x) - f(x)\right| \to 0, \qquad \text{as } n \to \infty. \tag{26}$$

If this is the case, then we have proved the following celebrated theorem of Parzen (1962).

Theorem 17 (Parzen) *Suppose f is uniformly continuous, $\mathscr{F}K$ is integrable, $h_n \to 0$, and $h_n\sqrt{n} \to \infty$. Then,*

$$\max_x \left|\hat f(x) - f(x)\right| \xrightarrow{\mathrm{P}} 0, \qquad \text{as } n \to \infty.$$

Proof: It remains to verify (26). But this follows from Theorem 4 and the fact that $\mathrm{E}\hat f(x) = (K_{h_n} * f)(x)$. □

Remark 18 The condition that $h_n\sqrt{n} \to \infty$ can be improved (slightly more) to the following:

$$h_n\sqrt{\frac{n}{\log n}} \to \infty \quad \text{as } n \to \infty.$$

This improvement is due to M. Bertrand-Retali.[9] But this requires more advanced methods.

What does the integrability condition on $\mathscr{F}K$ mean? To start with, the inversion theorem can be used to show that if $\int_{-\infty}^{\infty} |(\mathscr{F}K)(t)|\, dt < \infty$ then K is uniformly continuous. But the integrability of $\mathscr{F}K$ is a little bit more stringent than the uniform continuity of K. This problem belongs to a course in harmonic analysis. Therefore, rather than discussing this issue further we show two useful classes of examples where this condition is verified. Both are the examples that have made several appearances in these notes thus far.

Remark 19 Suppose K is the $N(0, \tau^2)$ density, where $\tau > 0$ is fixed. Then, $\mathscr{F}K$ is the characteristic function of a $N(0, \tau^2)$ random variable; see (24). But we saw this earlier in the Probability Notes. The computation is as follows:

$$(\mathscr{F}K)(t) = e^{-\tau^2 t^2/2}, \qquad \text{for all } t \in \mathbf{R}.$$

Obviously, $\mathscr{F}K$ is integrable. In fact,

$$\int_{-\infty}^{\infty} |(\mathscr{F}K)(t)|\, dt = \int_{-\infty}^{\infty} e^{-\tau^2 t^2/2}\, dt = \sqrt{2\pi/\tau}.$$

[9] M. Bertrand-Retali (1978). Convergence uniforme d'un estimateur de la densité par la méthode de noyau, *Rev. Roumaine Math. Pures. Appl.* **23**, 361–385.

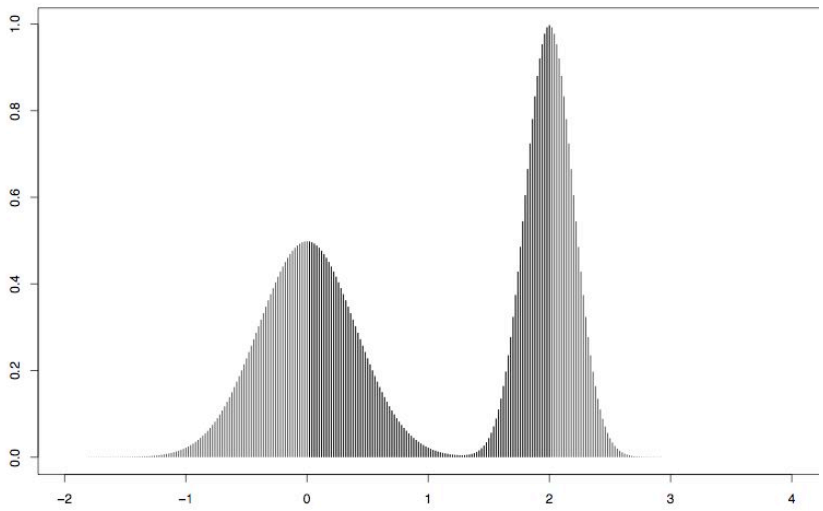

Figure 9: An Example of Modes.

Remark 20 Suppose $K(x) = \frac{1}{2}e^{-|x|}$ is the double exponential density. Then,

$$(\mathscr{F}K)(t) = \frac{1}{2}\int_{-\infty}^{\infty} e^{-|x|+itx}\,dx = \frac{1}{2}\int_0^{\infty} e^{-x+itx}\,dx + \frac{1}{2}\int_{-\infty}^{0} e^{x+itx}\,dx = \frac{1}{2}\int_0^{\infty} e^{-x+itx}\,dx + \frac{1}{2}\int_0^{\infty} e^{-x-itx}\,dx.$$

The first integral is the characteristic function of an exponential random variable with mean one. Therefore, it is given by $\int_0^{\infty} e^{-x+itx}\,dx = 1/(1-it)$. Plug $-t$ in place to t to find the second integral: $\int_0^{\infty} e^{-x-itx}\,dx = 1/(1+it)$. Add and divide by two to find that

$$(\mathscr{F}K)(t) = \frac{1}{2}\left[\frac{1}{1-it} + \frac{1}{1+it}\right] = \frac{1}{1+t^2}, \qquad \text{for all } t \in \mathbf{R}.$$

Evidently, this is integrable. In fact,

$$\int_{-\infty}^{\infty} |(\mathscr{F}K)(t)|\,dt = \int_{-\infty}^{\infty} \frac{dt}{1+t^2} = \pi.$$

9 Hunting for Modes

Let f be a density function on $\mathbf{R}$. A *mode* for f is a position of a local maximum. For example, Figure 9 depicts a density plot for the density function

$$f(x) = \frac{1}{2}\phi_1(x) + \frac{1}{2}\phi_2(x),$$

where ϕ_1 is the $N(0\,,0.4)$ density and ϕ_2 is the $N(2\,,0.2)$ density function. Because f has two local maxima, it has two modes: One is $x = 0$; and the other is $x = 2$.

In general, the question is: How can we use data to estimate the mode(s) of an unknown density function f? The answer is very simple now: If we know that $\hat{f} \approx f$ uniformly (and with very high probability), then the mode(s) of $\hat{f}$ *have* to approximate those of f with high probability. This requires an exercise in real analysis, and is omitted.