# Math 6070
# Elements of Density Estimation

Davar Khoshnevisan
University of Utah

Spring 2014

## 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 we have discussed earlier in Math 5080–5090. For example, if $X_1, \ldots, X_n \sim N(\mu\,, \sigma^2)$, then the density is

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

And 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 MLEs for $\mu$ and $\sigma$ [or you can use $S^2$ in place of $\hat{\sigma}^2$, as well]. This method is a kind of "plug-in" technique; it is a density estimation method, but a rather simple one.

Here, we are studying the more interesting case that $f$ is unknown. In this more general case, there are several different approaches to density estimation. We shall concentrate our efforts on the socalled "kernel density estimators." But for now, let us begin with a commonly-used first 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,

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

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,

$$\begin{aligned}\hat{f}(x) &= \frac{1}{18} \times \begin{cases} 3 & \text{if } 0 \le x < 1.5,\\ 4 & \text{if } 1.5 \le x < 3,\\ 4, & \text{if } 3 \le x < 4.5,\\ 1, & \text{if } 4.5 \le x < 6,\end{cases}\\ &= \begin{cases} 1/6 & \text{if } 0 \le x < 1.5,\\ 2/9 & \text{if } 1.5 \le x < 3,\\ 2/9, & \text{if } 3 \le x < 4.5,\\ 1/18, & \text{if } 4.5 \le x < 6.\end{cases}\end{aligned}$$

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

$$\hat{f}(x) = \frac{1}{18} \times \begin{cases} 7 & \text{if } 1 \le x < 2.5,\\ 4, & \text{if } 2.5 \le x < 4,\\ 1, & \text{if } 4 \le x < 5.5.\end{cases}$$

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, the command "`hist(X,breaks=6)`" produces Figure 2. [Figure 1 can be produced also with "`hist(X,breaks=3)`."]

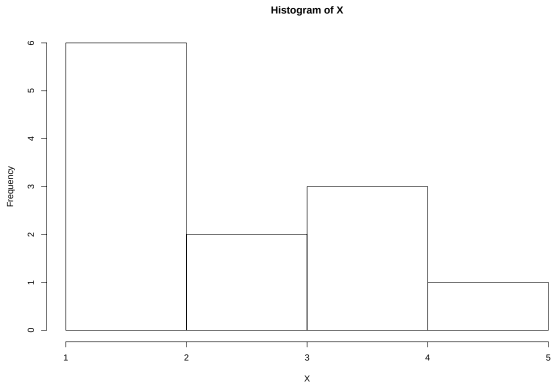

**Figure 1:** Histogram of the data of Example 1. Three breaks (automatic).

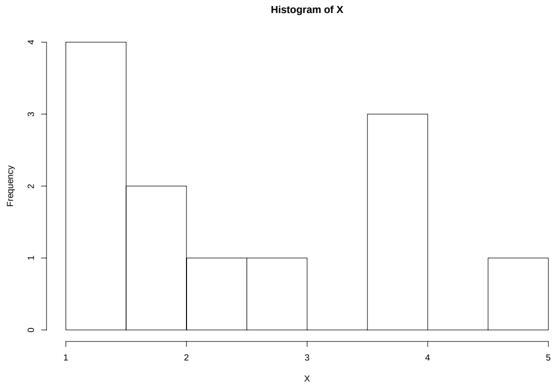

**Figure 2:** Histogram of the data of Example 1. Six breaks (manual).

## 1.2 The Kernel Density Estimator

Kernel density estimators are smooth substitutes 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 with very high probability,

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

Therefore, 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

$$W_h(x) := \frac{1}{h} W\left(\frac{x-X_j}{h}\right)$$

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. Our strategy is to replace the role of $W$ by a smoother function so that a smoother delta function is obtained.

Consider next a "kernel" $K$. For the time being, a kernel $K$ is simply a function such that $K(x) \ge 0$ and $\int_{-\infty}^{\infty} K(x)\,\mathrm{d}x = 1$. Then, define

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

The parameter $h$ is used to tune the estimator. It is alternatively called the *window width*, the *bandwidth*, and/or the *smoothing/tuning 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, we define $\rho_1(t) \le \rho_2(t) \le \cdots \le \rho_n(t)$ to be the ordered distances from $t$ to the samples $X_1, \ldots, X_n$.[1] The *nearest-neighbor density estimator* is the random function

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

This is also known as the "NN density estimator."

In order to see why this can be a sensible choice, 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 the 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 also that if $n$ is large and $r$ is small, then [in the right regime], the following holds with high probability:

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

[This is not obvious, since $\rho_k(x)$ is random; we are making non-rigorous, heuristic remarks here.] Because

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

this leads us to our NN density estimator in (1).

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

1. $\hat{f}$ is not smooth. Typically this setback can be addressed by replacing $\hat{f}$ with the following:

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

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

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

2. $\hat{f}$ is a better estimator of $f$ "locally." For instance, this is a better method for estimating $f$ when we are interested in the values/shape of $f$ near a given point $x$. The global behavior of NN estimators are sometimes bad. For instance, $\hat{f}$ is not itself a pdf:

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

since $\rho_k(x) \sim |x|$ as $|x| \to \infty$.

## 1.4 Variable Kernel Density Estimation

Let $K$ denote a nice kernel, and choose and fix a positive integer $k \geq 2$. Define $\delta_{j,k}$ to be the distance between $X_j$ and the $k$th-nearest point in $\{X_1\,,\ldots,X_n\}\setminus\{X_k\}$. Then we may consider the *variable kernel density estimator*,

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

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

## 1.5 The Orthogonal Series Method

Suppose $f$ is a nice pdf on $[0\,,1)$, and define

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

The theory of Fourier series tells us that

$$f(x) \sim \sum_{j=0}^{\infty} f_j\phi_j(x), \quad \text{where} \quad f_j := \int_0^1 f(x)\phi_j(x)\,\mathrm{d}x,$$

and $f \sim \sum_{j=1}^{\infty} f_j\phi_j$" means that "the infinite sum converges in $\mathscr{L}^2([0\,,1])$ to $f$." Stated more succintly, we have

$$\lim_{N\to\infty}\int_0^1 \left| f(x) - \sum_{j=0}^{N} f_j\phi_j(x)\right|^2 \mathrm{d}x = 0.$$

Now suppose $X$ has pdf $f$. In this case, $f_j$ is nothing but $\mathrm{E}\phi_j(X)$, and we are led to the law-of-large-numbers estimator

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

This, in turn, leads us to the orthogonal-series density estimator,

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

where $N$ is a large and fixed constant. This estimator is better used for local purposes. Globally, it is not a pdf. In fact $\hat{f}(x)$ is not in general even non negative everywhere!

## 1.6 Maximum Penalized Likelihood Estimation

Let $X_1, \ldots, X_n$ be iid with unknown common density $f$. The "likelihood" of $g$ [as a potential pdf] is

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

We can now try to find a probability density function $g$ that maximizes $\mathscr{L}(g)$.

Unfortunately, this enterprise is doomed to fail. Indeed, let $g(x) := \hat{f}(x)$ denote the histogram with origin $x_0 := 0$ and bandwidth $h > 0$. Then, it is evident that $\hat{f}$ is a pdf that satisfies $\hat{f}(X_i) \geq (nh)^{-1}$, whence

$$\max_{g \text{ a pdf}} \mathscr{L}(g) \geq \prod_{i=1}^{n} \hat{f}(X_i) \geq (nh)^{-n}.$$

The left-most quantity is independent of $h$. Therefore, we may send $h \to 0$ in the right-most quantity in order to see that the maximum likelihood estimator of $f$ is always infinity!

Although the preceding attempt failed, it is not without its merits. The reason that our first attempt failed was that we are maximizing the likelihood $\mathscr{L}(g)$ over too many pdfs $g$. Therefore, we may try to restrict the class of $g$'s over which the maximization is taken.

Maximum penalized likelihood estimation [MPLE] is one such possible approach to fixing the mentioned problem with the maximum likelihood density estimator. The idea is to maximize a *penalized log-likelihood* of the form

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

where $\lambda > 0$ is a smoothing parameter, and $F(g)$ measures the roughness of $g$ (say!).[2] The statistics $\sum_{i=1}^{n} \ln g(X_i)$ corresponds to the goodness of fit; $F(g)$ to smoothness; and $\lambda$ to how much of each we opt for.

Two major drawbacks of this method are: (i) The method depends critically on our choice of the penalization scheme $F$; and (ii) the method can be very hard to implement efficiently.

[2] An example to bear in mind is $F(g) := \int_{-\infty}^{\infty} [g''(x)]^2 \, \mathrm{d}x$.

## 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 bandwidth $h$, and then define our kernel density estimate as

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

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.

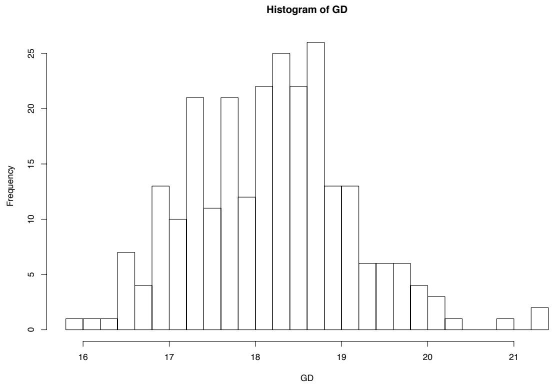

**Figure 3:** Histogram of the variable "GD".
Thirty breaks.

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., $K(x) = \frac{1}{2}\mathrm{e}^{-|x|}$. 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 $K(x) := (2\pi)^{-1/2}\exp(-x^2/2)$ is the $\mathrm{N}(0\,,1)$ 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 $\mathrm{N}(0\,,1)$ density.

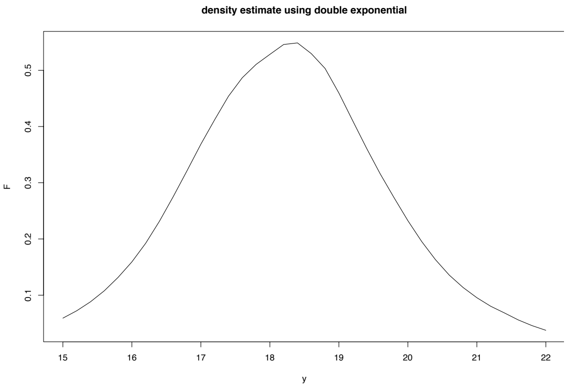

**Figure 4:** Kernel density estimate using DE ($h = 0.5$).

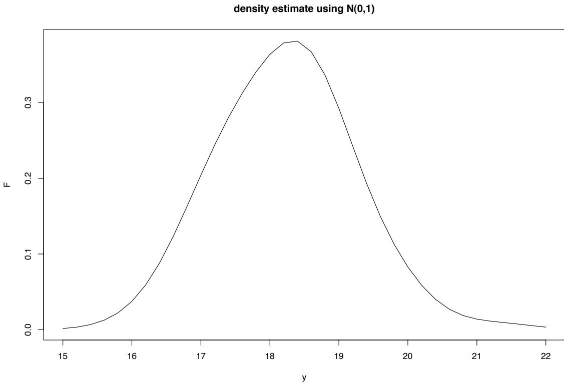

**Figure 5:** Kernel density estimate using $\mathrm{N}(0\,,1)$ ($h = 0.5$).

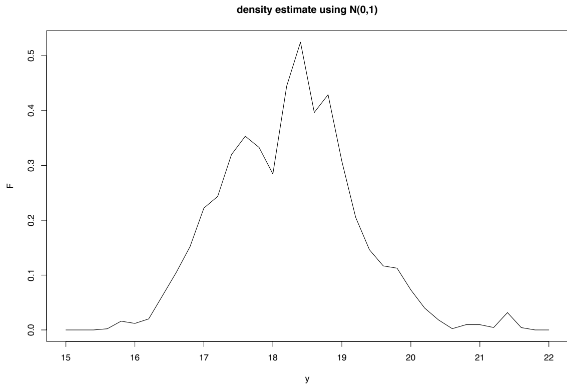

**Figure 6:** Kernel density estimate using $\mathrm{N}(0\,,1)$ $(h = 0.1)$.

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

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

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 [Math 5010] 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

$$\begin{aligned}\int_{-\infty}^{\infty} (f * g)(x)\,\mathrm{d}x &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(y)g(x-y)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} g(x-y)\,\mathrm{d}x\right) f(y)\,\mathrm{d}y \\ &= 1,\end{aligned}$$

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

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

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 $\int_{-\infty}^{\infty} K(x)\, \mathrm{d}x = 1$. 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 $\mathrm{N}(0\,,h^2)$ 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

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

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 $|f(x)| \leq B$ for all $x$; or*
2. *$K$ vanishes at infinity; that is, $\lim_{z\to\pm\infty} K(z) = 0$.*

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

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

*Proof (of Part 1 only).* Choose and fix an $x$ such that $f$ is continuous in an open neighborhood of $x$. Now let us choose and fix an arbitrary $\varepsilon > 0$. There exists $\delta > 0$, sufficiently small, such that

$$\max_{y\in(x-\delta,x+\delta)} |f(y) - f(x)| \leq \varepsilon. \tag{2}$$

Now we holds these constants $\varepsilon$ and $\delta$ fixed.

We can write

$$f(x) = \int_{-\infty}^{\infty} K(y) f(x)\,\mathrm{d}y.$$

Therefore,

$$\begin{aligned}(K_h * f)(x) - f(x) &= \int_{-\infty}^{\infty} K_h(y)\left[f(x-y) - f(y)\right]\mathrm{d}y \\ &= \frac{1}{h}\int_{-\infty}^{\infty} K(y/h)\left[f(x-y) - f(y)\right]\mathrm{d}y \\ &= \int_{-\infty}^{\infty} K(z)\left[f(x-hz) - f(x)\right]\mathrm{d}z.\end{aligned}$$

The triangle inequality for integral [Jensen's inequality] implies that

$$\begin{aligned}|(K_h * f)(x) - f(x)| &\le \int_{-\infty}^{\infty} K(z)\left|f(x-zh) - f(x)\right|\mathrm{d}z \\ &= \int_{|zh|\le\delta} K(z)\left|f(x-zh) - f(x)\right|\mathrm{d}z \\ &\qquad + \int_{|zh|>\delta} K(z)\left|f(x-zh) - f(x)\right|\mathrm{d}z.\end{aligned}$$

Because $|f(x-zh) - f(x)| \le 2B$,

$$\int_{|zh|>\delta} K(z)\left|f(x-zh) - f(x)\right|\mathrm{d}z \le 2B\int_{|z|>\delta/h} K(z)\,\mathrm{d}z.$$

And thanks to (2),

$$\int_{|zh|\le\delta} K(z)\left|f(x-zh) - f(x)\right|\mathrm{d}z \le \varepsilon\int_{-\infty}^{\infty} K(z)\,\mathrm{d}z = \varepsilon.$$

Because $K$ is a pdf, $\int_{|z|>\delta/h} K(z)\,\mathrm{d}z \to 0$ as $h \to 0$. Therefore,

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

Because $\varepsilon$ is arbitrary and the left-hand side is independent of $\varepsilon$, the left-hand side must be zero. This proves the result. □

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.

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 that satisfy the conditions of Theorem 2. Then,*

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

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

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

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

**Theorem 4.** *Suppose $f$ and $K$ are density functions as in Theorem 2, and $f$ is uniformly continuous. Then, $\lim_{h\to 0} K_h * f = f$ uniformly; i.e.,*

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

## 2.3 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

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

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

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

where

$$K_h^2(z) := \left|K_h(z)\right|^2 = \frac{1}{h^2}\left|K\left(\frac{z}{h}\right)\right|^2.$$

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

$$\text{Bias}(\hat{f}(x)) = f(x) - \text{E}[\hat{f}(x)] = f(x) - (K_h * f)(x).$$

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

1. $\text{Var}\,(\hat{f}(x)) \to 0$ as $n \to \infty$; whereas
2. $\text{Bias}(\hat{f}(x)) \to 0$ 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 $\text{MSE}(\hat{f}(x)) \to 0$ as $n \to \infty$? 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.

# 3 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: When $h$ small,

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

## 3.1 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

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

Therefore, by (3),

$$\begin{aligned} -\mathrm{Bias}(\hat{f}(x)) &\approx \int_{-\infty}^{\infty} K(z)\left\{ f(x) - zhf'(x) + \frac{z^2h^2}{2} f''(x)\right\} \mathrm{d}z - f(x) \\ &= f(x) \overbrace{\int_{-\infty}^{\infty} K(z)\,\mathrm{d}z}^{=1, \text{ since } K \text{ is a pdf}} - hf'(x) \overbrace{\int_{-\infty}^{\infty} zK(z)\,\mathrm{d}z}^{=0, \text{ by symmetry}} \\ &\qquad + \frac{h^2}{2} f''(x) \int_{-\infty}^{\infty} z^2 K(z)\,\mathrm{d}z - f(x). \end{aligned}$$

Simplify to obtain

$$\begin{aligned} -\mathrm{Bias}(\hat{f}(x)) &\approx \frac{h^2}{2} f''(x) \int_{-\infty}^{\infty} z^2 K(z)\,\mathrm{d}z \\ &:= \frac{h^2}{2} f''(x)\sigma_K^2. \end{aligned} \tag{4}$$

Now we turn our attention to the variance of $\hat{f}(x)$. Recall that $\mathrm{Var}\,(\hat{f}(x)) =$

$(K_h^2 * f)(x) - (K_h * f)^2(x)$. We begin by estimating the first term.

$$\begin{aligned}
\left(K_h^2 * f\right)(x) &= \frac{1}{h^2}\int_{-\infty}^{\infty} \left|K\left(\frac{x-u}{h}\right)\right|^2 f(u)\,\mathrm{d}u \\
&= \frac{1}{h}\int_{-\infty}^{\infty} K^2(z) f(x-zh)\,\mathrm{d}z \\
&\approx \frac{1}{h}\int_{-\infty}^{\infty} K^2(z)\left\{f(x) - zhf'(x) + \frac{z^2h^2}{2}f''(x)\right\}\mathrm{d}z \\
&= \frac{1}{h}f(x)\int_{-\infty}^{\infty} K^2(z)\,\mathrm{d}z - f'(x)\int_{-\infty}^{\infty} zK^2(z)\,\mathrm{d}z \\
&\qquad\qquad + \frac{h}{2}f''(x)\int_{-\infty}^{\infty} z^2K^2(z)\,\mathrm{d}z \\
&\approx \frac{1}{h}f(x)\int_{-\infty}^{\infty} K^2(z)\,\mathrm{d}z \qquad\qquad \text{[the other terms are bounded]} \\
&:= \frac{1}{h}f(x)\|K\|_2^2.
\end{aligned}$$

Because $(K_h * f)(x) \approx f(x)$ (Theorem 2), this yields the following:[3]

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

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

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

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

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

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 $h$:

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

where

$$\alpha_f := \frac{(f(x))^{1/5}}{(f''(x))^{2/5}}, \qquad \text{and} \qquad \beta_K := \frac{\|K\|_2^{2/5}}{\sigma_K^{4/5}} = \frac{\left(\int_{-\infty}^{\infty} K^2(z)\,\mathrm{d}z\right)^{1/5}}{\left(\int_{-\infty}^{\infty} z^2K(z)\,\mathrm{d}z\right)^{2/5}}. \tag{6}$$

[3] We are writing $\|g\|_2^2 := \int_{-\infty}^{\infty} g^2(z)\,\mathrm{d}z$ and $\sigma_g^2 := \int_{-\infty}^{\infty} z^2g(z)\,\mathrm{d}z$ for all nice functions $g$.

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

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

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

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

See Figure 7 for a plot.

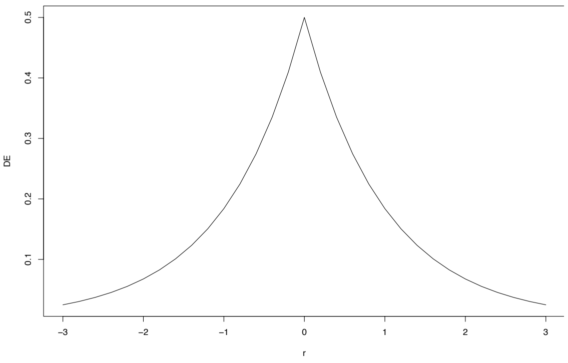

**Figure 7:** A plot of the double-exponential density.

By symmetry,

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

Therefore,

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

Similarly,

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

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

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

See Figure 8.

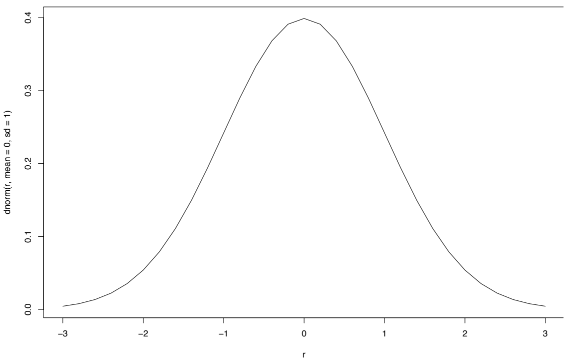

**Figure 8:** A plot of the $\mathrm{N}(0\,,1)$ density.

In this case, $\sigma_K^2 = \int_{-\infty}^{\infty} z^2 K(z)\,\mathrm{d}z = \tau^2$, and

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

Consequently,

$$\beta_K = \frac{1}{\left(2\tau\sqrt{\pi}\right)^{1/5}}. \tag{7}$$

This yields,

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

Similarly,

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

## 3.2 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

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

It is easy to see that

$$\text{MISE}\,\hat{f} = \int_{-\infty}^{\infty} \text{MSE}\,(\hat{f}(x))\,\mathrm{d}x.$$

Therefore, under the present smoothness assumptions,

$$\begin{aligned}\text{MISE}\,\hat{f} &\approx \frac{1}{nh}\int_{-\infty}^{\infty} K^2(z)\,\mathrm{d}z + \frac{h^4}{4}\int_{-\infty}^{\infty} |f''(x)|^2\,\mathrm{d}x \cdot \left(\int_{-\infty}^{\infty} z^2K(z)\,\mathrm{d}z\right)^2 \\ &:= \frac{1}{nh}\|K\|_2^2 + \frac{h^4}{4}\|f''\|_2^2\sigma_K^4. \end{aligned} \tag{8}$$

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:

$$h_n := \frac{C}{n^{1/5}} \quad \text{where} \qquad C = \frac{\beta_K}{\|f''\|_2^{2/5}}. \tag{9}$$

See (6) for the notation on $\beta_K$. The asymptotically optimal MISE is obtained upon plugging in this $h_n$ into (8). That is,

$$\begin{aligned}\text{MISE}_{opt}\,\hat{f}(x) &\approx \frac{1}{nh_n}\|K\|_2^2 + \frac{h_n^4}{4}\|f''\|_2^2\sigma_K^4 \\ &= \frac{D}{n^{4/5}} \quad \text{where} \quad D = \frac{5}{4}\|f''\|_2^{2/5}\|K\|_2^{8/5}\sigma_K^{4/5}.\end{aligned} \tag{10}$$

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

$$h_n = \frac{C}{n^{1/5}} \quad \text{where} \quad C = \frac{1}{2^{4/5}\|f''\|_2^{2/5}}. \tag{11}$$

Also,

$$\text{MISE}_{opt}\,\hat{f}(x) \approx \frac{D}{n^{4/5}} \quad \text{where} \quad D = \frac{5}{2^{16/5}}\|f''\|_2^{2/5}. \tag{12}$$

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

$$h_n = \frac{C}{n^{1/5}} \quad \text{where} \quad C = \frac{1}{(2\tau\sqrt{\pi})^{1/5}\,\|f''\|_2^{2/5}}. \tag{13}$$

Also,

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

# 4 Problems and Remedies

A 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 some of them next.[4]

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. *Making References 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 $g := \mathrm{N}(0\,,\tau^2)$ density. Let

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-x^2/2}$$

denote 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

$$\begin{aligned}\|g\|_2^2 &= \frac{1}{\tau^6}\int_{-\infty}^{\infty}\left[\varphi''\left(\frac{x}{\tau}\right)\right]^2\,\mathrm{d}x\\ &= \frac{1}{\tau^5}\int_{-\infty}^{\infty}\left[\varphi''(y)\right]^2\,\mathrm{d}y\\ &= \frac{1}{2\pi\tau^5}\int_{-\infty}^{\infty}\mathrm{e}^{-y^2}\left(y^2-1\right)^2\,\mathrm{d}y\\ &= \frac{3}{8\tau^5\sqrt{\pi}}.\end{aligned}$$

This is about $0.2115/\tau^5$. So we can choose the bandwidth

$$h := \beta_K \|g''\|_2^{-2/5} n^{-1/5};\ \text{ that is,}$$

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

In order for us to actually be able to use this, we need to know $\tau$. But our replacement of $f$ by $g$ tacitly assumes that the variance of the date is $\tau^2$; i.e., that $\tau^2 = \int_{-\infty}^{\infty} x^2\, f(x)\,\mathrm{d}x - (\int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x)^2$. So we can estimate $\tau^2$ 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

[4] 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 $\mathrm{MISE}_{opt}\hat{f}$, for instance. But the stated problem with the bandwidth is much more serious.

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 *oversmoothing.*

3. *Plug-in estimates:* Consider the case where our asymptotically-optimal bandwidth has the form $h = C(n\,, K)/\|f''\|_2^{2/5}$ where $C(n\,, K)$ is known [e.g., see Example 8 on page 21]. One can just find some sort of "plug-in estimator" of $\|f''\|_2$ and plug that in to obtain an estimated $h$. One reasonable possibility is to estimate $\|f''\|_2$ by $\|\tilde{f}_h''\|_2$, where $\tilde{f}_h$ denotes the kernel density estimator of $f$ that uses some prior estimator of $h$ in place of $h$. [For example, use one of the preceding crude methods to start the process]. In general, we will want $h$ that satisfies $h = C(n\,, K)/\|\tilde{f}_h''\|_2^{2/5}$, which can sometimes be computed numerically.

4. *Bootstrap:* We will discuss bootstrapping later on, and in a slightly different context. But there are ways to incorporate a version of the bootstrap in order to find good choices of $h$; see Jones et al [*JASA* **91***(433)*, 1996], for instance.

5. *Cross validation:* There are computationally-efficient cross-validation methods for choosing $h$ as well. See Chapter 32 of Das Gupta's comprehensive book, *Asymptotic Theory of Statistics and Probability* [Springer, 2008].

# 5 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 $x$;
- $\int_{-\infty}^{\infty} K(x)\,\mathrm{d}x = 1$;
- $\mu_2(K) = 0$, where $\mu_\ell(K) := \int_{-\infty}^{\infty} x^\ell K(x)\,\mathrm{d}x$;
- $\mu_4(K) \neq 0$.

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

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

Therefore,

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

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:

$$\begin{aligned}
\left(K_h^2 * f\right)(x) &= \frac{1}{h^2}\int_{-\infty}^{\infty} K^2\left(\frac{x-u}{h}\right) f(u)\,\mathrm{d}u\\
&= \frac{1}{h}\int_{-\infty}^{\infty} K^2(a)f(x-ah)\,\mathrm{d}a.
\end{aligned}$$

Then, we apply a Taylor expansion,

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

as before. Thus, as before,

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

Therefore,

$$\operatorname{MSE}(\hat{f}(x)) \approx \frac{\|K\|_2^2 f(x)}{nh} + \mu_4^2(K)\frac{h^8}{576}\left[f^{(iv)}(x)\right]^2. \tag{15}$$

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 $h_n = Cn^{-1/9}$, $\operatorname{MSE}(\hat{f}(x)) \approx Dn^{-8/9}$, and $\operatorname{MISE}(\hat{f}) \approx En^{-8/9}$. 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) $\int_{-\infty}^{\infty} K(x)\,\mathrm{d}x = 1$; (iii) $\int_{-\infty}^{\infty} x^2K(x)\,\mathrm{d}x = 0$; and (iv) $\mu_4(K) = \int_{-\infty}^{\infty} x^4K(x)\,\mathrm{d}x = -3/35 \neq 0$.

**Example 10.** We obtain another family of classical examples, due to W. R. Schucany and J. P. Sommers,[5] 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 the function $K_\nu$ satisfies the conditions of the kernels of this section. Namely: (i) $K_\nu$ is symmetric; (ii) $\int_{-\infty}^{\infty} K_\nu(x)\,\mathrm{d}x = 1$; (iii) $\int_{-\infty}^{\infty} x^2K_\nu(x)\,\mathrm{d}x = 0$; and (iv) $\mu_4(K_\nu) = \int_{-\infty}^{\infty} x^4K_\nu(x)\,\mathrm{d}x = -\nu^2\mu_4(K_\nu) \neq 0$. Schucany and Sommers recommend using values of $\nu$ that are $> 1$, but very close to one.

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

# 6 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. I mention three important examples 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, $\lim_{n\to\infty} \int_{-\infty}^{\infty} |\hat{f}(x) - f(x)| \, \mathrm{d}x = 0$ 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."

## 6.1 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.[6]

**Theorem 11** (Parzen, 1962)**.** *Let us assume the following:*

1. *$K$ vanishes at infinity, and $\int_{-\infty}^{\infty} K^2(x)\,\mathrm{d}x < \infty$;*
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 $\hat{f}(x) \xrightarrow{\mathrm{P}} f(x)$, as $n \to \infty$.*

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

Recall from page 15 that

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

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

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

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

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

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

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

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

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

$$\operatorname{Var}\left(\hat{f}(x)\right) \sim \frac{1}{n}\left[\frac{f(x)}{h_n}\int_{-\infty}^{\infty} K^2(u)\,\mathrm{d}u - f(x)\right] \sim \frac{f(x)}{nh_n}\int_{-\infty}^{\infty} K^2(u)\,\mathrm{d}u. \qquad (16)$$

Since $nh_n \to 0$, this proves that $\operatorname{Var}\left(\hat{f}(x)\right) \to 0$ 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. □

Next, I state [without proof] a weaker formulation of a theorem of L. Devroye.[7] 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 $n \to \infty$,*

$$\int_{-\infty}^{\infty} |\hat{f}(x) - f(x)|\,\mathrm{d}x \xrightarrow{\mathrm{P}} 0.$$

## 6.2 Uniform Consistency

Theorem 12 is a natural global-consistency theorem. But it falls short 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.

[7] 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.

First we recall a few facts from Fourier analysis. If $h$ is an *integrable* function [that is, if $\int_{-\infty}^{\infty} |h(x)|\,\mathrm{d}x < \infty$], then its *Fourier transform* is the function $\mathscr{F}h$ defined by

$$(\mathscr{F}h)(t) := \int_{-\infty}^{\infty} \mathrm{e}^{itx} h(x)\,\mathrm{d}x \qquad \text{for } -\infty < t < \infty.$$

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

$$(\mathscr{F}f)(t) = \mathrm{E}\left[\mathrm{e}^{itX_1}\right], \tag{17}$$

and so $\mathscr{F}f$ is the socalled "characteristic function of $X$." 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 13** (Inversion Theorem)**.** *If $h$ and $\mathscr{F}h$ are integrable, then*

$$h(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{-itx}(\mathscr{F}h)(t)\,\mathrm{d}t \qquad \textit{for } -\infty < x < \infty.$$

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 the Fourier transform of our kernel density estimate $\hat{f}$ is

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

after an interchange of the sum with the integral. In other words, the Fourier transform of $\hat{f}$ can be written in terms of the Fourier transform of $K$ as follows:

$$(\mathscr{F}\hat{f})(t) = \frac{1}{n}\sum_{j=1}^{n} \mathrm{e}^{itX_j}(\mathscr{F}K)(h_nt).$$

In particular, $\mathscr{F}\hat{f}$ is integrable as soon as $\mathscr{F}K$ is. If so, then the inversion theorem (Theorem 13) tell us that

$$\begin{aligned}\hat{f}(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{-itx}(\mathscr{F}\hat{f})(t)\,\mathrm{d}t\\ &= \frac{1}{2\pi n}\sum_{j=1}^{n}\int_{-\infty}^{\infty} \mathrm{e}^{it(X_j-x)}(\mathscr{F}K)(h_nt)\,\mathrm{d}t\\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{-itx}\frac{1}{n}\sum_{j=1}^{n} \mathrm{e}^{itX_j}(\mathscr{F}K)(h_nt)\,\mathrm{d}t.\end{aligned}$$

Take expectations also to find that

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

Therefore,

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

where $\phi_n$ is the "empirical characteristic function,"

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

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

$$\max_x |\hat{f}(x) - \mathrm{E}[\hat{f}(x)]| \le \frac{1}{2\pi}\int_{-\infty}^{\infty} |\phi_n(t) - \mathrm{E}[\phi_n(t)]| \cdot |(\mathscr{F}K)(h_n t)|\,\mathrm{d}t. \tag{18}$$

Take expectations and use the "Cauchy–Schwarz inequality," $\mathrm{E}(|Z|) \le \sqrt{\mathrm{E}(Z^2)}$ to find that

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

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

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

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

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

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

$$\max_x |\mathrm{E}[\hat{f}(x)] - f(x)| \to 0, \qquad \text{as } n \to \infty. \tag{19}$$

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

**Theorem 14** (Parzen)**.** *Suppose $f$ is uniformly continuous, $\mathscr{F}K$ is integrable, $h_n \to 0$, and $h_n\sqrt{n} \to \infty$. Then,*

$$\max_x |\hat{f}(x) - f(x)| \overset{\mathrm{P}}{\to} 0, \qquad \textit{as } n \to \infty.$$

**Proof:** It remains to verify (19). But this follows from Theorem 4 and the fact that $\mathrm{E}[\hat{f}(x)] = (K_{h_n} * f)(x)$. □

**Remark 15.** 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.[8] 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 $\int_{-\infty}^{\infty} |(\mathscr{F}K)(t)|\,\mathrm{d}t < \infty$ 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 16.** Suppose $K$ is the $\mathrm{N}(0\,,\tau^2)$ density, where $\tau > 0$ is fixed. Then, $\mathscr{F}K$ is the characteristic function of a $\mathrm{N}(0\,,\tau^2)$ random variable; see (17). We can compute it as easily as the MGF of a normal:

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

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

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

**Remark 17.** Suppose $K(x) = \frac{1}{2}\mathrm{e}^{-|x|}$ is the double exponential density. Then,

$$\begin{aligned}
(\mathscr{F}K)(t) &= \frac{1}{2}\int_{-\infty}^{\infty} \mathrm{e}^{-|x|+itx}\,\mathrm{d}x\\
&= \frac{1}{2}\int_{0}^{\infty} \mathrm{e}^{-x+itx}\,\mathrm{d}x + \frac{1}{2}\int_{-\infty}^{0} \mathrm{e}^{x+itx}\,\mathrm{d}x\\
&= \frac{1}{2}\int_{0}^{\infty} \mathrm{e}^{-x+itx}\,\mathrm{d}x + \frac{1}{2}\int_{0}^{\infty} \mathrm{e}^{-x-itx}\,\mathrm{d}x.
\end{aligned}$$

The first integral is the characteristic function of an exponential random variable with mean one. Therefore, it is given by $\int_0^\infty \mathrm{e}^{-x+itx}\,\mathrm{d}x = 1/(1-it)$. Plug $-t$

[8] 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.

in place to $t$ to find the second integral: $\int_0^\infty \mathrm{e}^{-x-itx}\,\mathrm{d}x = 1/(1+it)$. Add and divide by two to find that

$$(\mathscr{F}K)(t) = \frac{1}{2}\left[\frac{1}{1-it} + \frac{1}{1+it}\right] = \frac{1}{1+t^2} \qquad \text{for } -\infty < t < \infty.$$

Evidently, this is integrable. In fact,

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

# 7 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 $\phi_1$ is the $\mathrm{N}(0\,, 0.4)$ density and $\phi_2$ is the $\mathrm{N}(2\,, 0.2)$ density function.

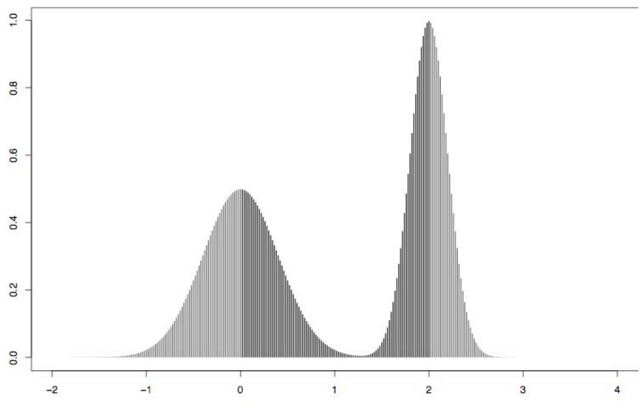

**Figure 9:** An Example of Modes.

One can understand this distribution as follows: We toss an independent fair coin; if the coin comes up heads, then we sample from $\phi_1$; if the coin comes up tails, then we sample from $\phi_2$. 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.] Therefore, we may estimate the number of modes of $f$ with the number of modes of $\hat{f}$. This process can be even carried out visually!

# 8 There are no unbiased density estimators

There is a beautiful theorem of Rosenblatt (*Ann. Math. Stat.* **27**, 1956) that says that there are no universally-unbiased density estimators of any kind. This gives us some hope about, say, kernel density estimation which we know is biased.

**Theorem 18** (Rosenblatt, 1956)**.** *Let $T$ be a function-valued function of $n$ variables such that $x \mapsto T(z_1\,, \ldots , z_n)(x)$ is a pdf for all $z_1, \ldots , z_n \in \mathbf{R}$. Given an iid sample $X_1, \ldots , X_n$ from a continuous pdf $f$, we may use the density estimator*

$$\hat{f}(x) := T(X_1, \ldots , X_n)(x).$$

*Then $\hat{f}$ is not generically unbiased. That is,* $\mathrm{E}_f[\hat{f}(x)] \neq f(x)$ *for some continuous pdf $f$ and $x \in \mathbf{R}$.*

*Sketch of proof.* Let $\Theta$ denote the parameter space of all continuous pdfs $f$. Material similar to that of 5080 and 5090 tells us that the order statistics $X_{1:n}, \ldots , X_{n:n}$ are complete and sufficient for estimating $f \in \Theta$. Let $F_n$ denote the empirical cdf; i.e.,

$$F_n(b) := \frac{1}{n}\sum_{j=1}^{n} \mathbf{I}\{X_j \leq b\}.$$

Because $F_n$ can be written directly as a function of the order statistics and $\mathrm{E}_f[F_n(b)] = \int_{-\infty}^{b} f(y)\,\mathrm{d}y$ for all $f \in \Theta$, it follows that $F_n$ is the only unbiased estimator of $F$ that is based on the order statistics.

Now suppose there exists a procedure $T$, as stated. We will derive a contradiction from this assumption.

Define

$$\tilde{f}(x) := \frac{1}{n!}\sum_{\pi} T\left(X_{\pi_1}\,, \ldots , X_{\pi_n}\right)(x),$$

where the sum is taken over all $n!$ permutations $(\pi_1\,, \ldots , \pi_n)$ of $(1\,, \ldots , n)$. Then, it is easy to see that $\tilde{f}$ is unbiased for $f$ and is a function of the order statistics alone [since it is permutation invariant]. Consequently,

$$\tilde{F}(b) := \int_{-\infty}^{b} \tilde{f}(x)\,\mathrm{d}x$$

is an unbiased estimator of $F(b)$ and depends only on the order statistics. By the first paragraph of the proof, $\tilde{F} = F_n$. But this cannot be: $\tilde{F}$ is continuous [in fact differentiable] whereas $F_n$ is pure-jump! □