Week 5

5.1 Accuracy of a kernel density estimator

Let $Y_1, Y_2, \ldots, Y_n$ denote i.i.d. random variables, each with a continuous distribution with density $p(\cdot)$. For a given kernel $K(\cdot)$, and a given value of the smoothing parameter h, let

$$\widehat{p}(y) = \frac{1}{nh}\sum_{j=1}^{n} K(\frac{y - Y_j}{h}), \quad -\infty < y < \infty.$$

We now consider the accuracy of $\widehat{p}(\cdot)$ as an estimator of $p(\cdot)$. To evaluate the accuracy of $\widehat{p}(\cdot)$, we use the same general approach used to evaluate the accuracy of the histogram density estimator.

Note that, because $Y_1, Y_2, \ldots, Y_n$ are i.i.d. random variables,

$$\mathrm{E}\left(\widehat{p}(y)\right) = \frac{1}{nh}\sum_{j=1}^{n} \mathrm{E}\left(K(\frac{y - Y_j}{h})\right) = \frac{1}{h}\mathrm{E}\left(K(\frac{y - Y_1}{h})\right) = \int_{-\infty}^{\infty} \frac{1}{h} K(\frac{y-t}{h}) p(t)dt \qquad (5.1)$$

and

$$\begin{aligned}\mathrm{Var}\left(\widehat{p}(y)\right) &= \frac{1}{(nh)^2}\sum_{j=1}^{n} \mathrm{Var}\left(K(\frac{y - Y_j}{h})\right) \\ &= \frac{1}{n}\left(\frac{1}{h^2}\int_{-\infty}^{\infty} K(\frac{y-t}{h})^2 p(t)dt - \left(\frac{1}{h}\int_{-\infty}^{\infty} K(\frac{y-t}{h})p(t)dt\right)^2\right). \qquad (5.2)\end{aligned}$$

These are exact expressions and, although it is possible, in principle, to evaluate them for specfic choices of $K(\cdot)$ and $p(\cdot)$, general results are based on approximations that are valid as $n \to \infty$ and $h \to 0$.

To derive such results, we need to approximate integrals such as

$$\frac{1}{h}\int_{-\infty}^{\infty} K(\frac{y-t}{h})p(t)dt$$

when n is large and h is small, that is, as $n \to \infty$ and $h \to 0$.

Such approximations have a few steps.

(1) Use the change-of-variable $u = (y - t)/h$ to rewrite the integral

$$\frac{1}{h}\int_{-\infty}^{\infty} K(\frac{y-t}{h})p(t)dt$$

as

$$\int_{-\infty}^{\infty} K(u)p(y-uh)du.$$

(2) Expand the function $p(y - uh)$ around $uh = 0$. Then

$$\int_{-\infty}^{\infty} K(u)p(y-uh)du$$

may be written in the form

$$\int_{-\infty}^{\infty} K(u)\left(p(y) + p'(y)(-uh) + \frac{1}{2}p''(y)(-uh)^2 + \cdots\right)du.$$

(3) Simplify this expression, using the fact that the terms $p(y), p'(y)$, and so on, do not depend on u, along with the properties of the kernel,

$$\int_{-\infty}^{\infty} K(u)du = 1, \qquad \int_{-\infty}^{\infty} uK(u)du = 0,$$

and

$$\int_{-\infty}^{\infty} u^2K(u)du = 1.$$

Then

$$\frac{1}{h}\int_{-\infty}^{\infty} K(\frac{y-t}{h})p(t)dt = \int_{-\infty}^{\infty} K(u)p(y-uh)du = p(y) + \frac{1}{2}p''(y)h^2 + O(h^4). \quad (5.3)$$

It follows that

$$\mathrm{E}\left(\widehat{p}(y)\right) = p(y) + \frac{1}{2}p''(y)h^2 + O(h^4);$$

that is, the bias of $\widehat{p}(y)$ as an estimator of $p(y)$ is of the form

$$\frac{1}{2}p''(y)h^2 + O(h^4) \quad \text{as} \quad h \to 0.$$

Thus, for the bias of the estimator to be small, we need h to be small.

Now consider the variance of the kernel estimator. Recall that

$$\mathrm{Var}\left(\widehat{p}(y)\right) = \frac{1}{n}\left(\frac{1}{h^2}\int_{-\infty}^{\infty} K(\frac{y-t}{h})^2p(t)dt - \left(\frac{1}{h}\int_{-\infty}^{\infty} K(\frac{y-t}{h})p(t)dt\right)^2\right).$$

Note that, because of the presence of the $1/n$ term in this expression, we do not need to approximate the integrals in (5.2) to the same level of accuracy used when approximating $\mathrm{E}\left(\widehat{p}(y)\right)$.

The analysis is based on the same basic steps used when approximating

$$\frac{1}{h}\int_{-\infty}^{\infty} K(\frac{y-t}{h})p(t)dt :$$

use the change-of-variable $u = (y-t)/h$, expand $p(y-uh)$ around $uh = 0$, and simplify the result, using properties of the kernel.

This leads to the result

$$\mathrm{Var}\left(\widehat{p}(y)\right) = \frac{K_2}{nh}p(y) + O(n^{-1}),$$

where

$$K_2 = \int_{-\infty}^{\infty} K(u)^2 du.$$

Thus, for the variance of the estimator to be small, we need nh to be large.

Because the MSE of an estimator is equal to its squared bias plus its variance, it follows that the MSE of $\widehat{p}(y)$ as an estimator of $p(y)$ may be expanded

$$\begin{aligned}
\mathrm{MSE}(y) &= \left(\frac{1}{2}p''(y)h^2 + O(h^4)\right)^2 + \frac{K_2}{nh}p(y) + O(n^{-1}) \\
&= \frac{1}{4}p''(y)^2h^4 + p''(y)h^2O(h^4) + O(h^6) + \frac{K_2}{nh}p(y) + O(n^{-1}) \\
&= \frac{1}{4}p''(y)^2h^4 + K_2p(y)\frac{1}{nh} + O(h^6) + O(n^{-1}).
\end{aligned}$$

As we did with the histogram density estimator, we can summarize the MSE by integrating over the range of the variable, which yields the integrated MSE (IMSE); the leading terms in the IMSE are known as the asymptotic integrated MSE (AIMSE). Thus, for the kernel density estimator, the AIMSE is given by

$$\mathrm{AIMSE} = \frac{1}{4}\int_{-\infty}^{\infty} p''(y)^2 dy \; h^4 + K_2\frac{1}{nh}\int_{-\infty}^{\infty} p(y)dy = \frac{1}{4}R_2h^4 + K_2\frac{1}{nh} \tag{5.4}$$

where

$$R_2 = \int_{-\infty}^{\infty} p''(y)^2 dy.$$

The AIMSE gives useful information regarding the accuracy of kernel density estimators.

(1) As with the histogram estimator, the AIMSE of the kernel estimator approaches 0 as $n \to \infty$ only if $h \to 0$ as $n \to \infty$ **and** $nh \to 0$ as $n \to \infty$. That is, for large n, h should be small, but not too small.

(2) The accuracy of $\widehat{p}(\cdot)$ as an estimator of $p(\cdot)$ depends on $p(\cdot)$ primarily through the quantity

$$R_2 = \int_{-\infty}^{\infty} p''(y)^2 dy.$$

Recall that the accuracy of the histogram estimator depends on $p(\cdot)$ through the quantity

$$R_1 = \int_{-\infty}^{\infty} p'(y)^2 dy.$$

(3) The effect of using different kernels is captured in the term $K_2/(nh)$. Thus, a kernel with a small value of

$$K_2 = \int_{-\infty}^{\infty} K(u)^2 du$$

will yield the most accurate kernel estimators, in a certain sense. The optimal kernel in terms of K_2 is the Epanechnikov kernel; however, the other three kernels considered here all have values of K_2 close to the optimal value.

(4) The first term in the AIMSE for the kernel estimator,

$$\frac{1}{4} \int_{-\infty}^{\infty} p''(y)^2 dy \; h^4$$

is a measure of the asymptotic squared bias of $\widehat{p}(\cdot)$. Note that for the histogram estimator, the asymptotic squared bias is of order $O(h^2)$ as $h \to 0$. For both estimators, the asymptotic variance is of order $O(1/(nh))$.

Thus, the advantage of using a kernel estimator instead of the histogram is captured primarily in the bias of the estimators: the bias of the kernel estimator tends to be smaller than the bias of the histogram estimator, at least when n is large so that h is small (recall that we require that $h \to 0$ as $n \to \infty$).

(5) The rate at which the AIMSE approaches 0 is maximized by taking $h = O(n^{-\frac{1}{5}})$ and, for such a choice, the squared bias and variance are both of order $O(1/n^{\frac{4}{5}})$. Recall that for the histogram estimator the optimal rate at which the AIMSE approaches 0 is $O(1/n^{\frac{2}{3}})$. Note that, for any constants c_1 and c_2, $c_1/n^{\frac{4}{5}} < c_2/n^{\frac{2}{3}}$ for sufficiently large n; thus, the kernel estimator tends to be more accurate than the histogram estimator for large n.

(6) The results in this section can be used to justify the statement that estimation of a density is a more difficult problem than estimating the distribution function. Consider estimation of the distribution function evaluated at y, $F(y)$, and let $\widehat{F}(y)$ denote the estimator based on the empirical distribution function $\widehat{F}(\cdot)$. We have seen that $\widehat{F}(y)$

is an unbiased estimator of $F(y)$ and it is straightforward to show that the variance is of order $O(1/n)$. Hence, the MSE of $\widehat{F}(y)$ is $O(1/n)$. This is a faster rate than that of the MSE of the kernel estimator of the density $p(y)$, which is $O(1/n^{\frac{4}{5}})$.

Example 5.1 Suppose that the density $p(\cdot)$ is given by

$$p(y) = \frac{1}{(1+y)^2}, \quad y > 0;$$

estimation of this density using a histogram estimator was considered in Example 4.1.

For this density,

$$R_2 = \int_0^\infty p''(y)^2 dy = \int_0^\infty \frac{36}{(1+y)^8} dy = \frac{36}{7}.$$

Suppose that the Gaussian kernel is used; it is straightforward to show that, for the Gaussian kernel, $K_2 = 1/(2\sqrt{\pi})$. It follows that the AIMSE of the kernel estimator is

$$\frac{9}{7}h^4 + \frac{1}{2\sqrt{\pi}}\frac{1}{nh}.$$

This can be compared to the AIMSE of the histogram estimator,

$$\frac{h^2}{15} + \frac{1}{nh}.$$

Suppose that, in each case, the optimal value of h is used. For the kernel estimator it is

$$(\frac{7}{72\sqrt{\pi}})^{\frac{1}{5}}\frac{1}{n^{\frac{1}{5}}} \doteq \frac{0.560}{n^{\frac{1}{5}}};$$

for the histogram estimator it is

$$(\frac{15}{2n})^{\frac{1}{3}} \doteq \frac{1.957}{n^{\frac{1}{3}}}.$$

Evaluating the AIMSE of the kernel estimator at its optimal value of h yields the optimal value of the AIMSE for the kernel estimator of

$$\frac{0.630}{n^{\frac{4}{5}}}.$$

Evaluating the AIMSE of the histogram estimator at its optimal value of h yields the optimal value of the AIMSE for the histogram estimator of

$$\frac{0.766}{n^{\frac{2}{3}}}.$$

Here, because $0.630 < 0.766$, based on the AIMSE, the kernel estimator is more accurate than histogram estimator for all n. □

5.2 Selection of the smoothing parameter

We have seen that kernel density estimates depend heavily on the value of the smoothing parameter h used in their construction. Hence, a useful density estimation procedure needs a reliable method of determining the "best" h to use for a given set of data.

One approach is to simply choose the value of h subjectively, by analyzing the results for several different choices and choosing the one that balances the desired level of smoothness of the estimate with the goal of capturing important features of the data. Although that approach is often useful, it is also useful to have available a more objective procedure for choosing h.

Recall that the AIMSE of the kernel estimator based on a smoothing parameter h is given by

$$\frac{1}{4}R_2h^4 + K_2\frac{1}{nh};$$

it follows that the AIMSE of the kernel estimator is minimized by choosing $h = h^*$ where

$$h^* = \left(\frac{K_2}{R_2}\right)^{\frac{1}{5}} \frac{1}{n^{\frac{1}{5}}}.$$

Note that K_2 depends only on the kernel; if a Gaussian kernel with standard deviation 1 is used, then

$$K_2 = \frac{1}{2\sqrt{\pi}}$$

so that the Gaussian-kernel expression for h^* is

$$h^*_G = \left(\frac{1}{2\sqrt{\pi}R_2}\right)^{\frac{1}{5}} \frac{1}{n^{\frac{1}{5}}}.$$

On the other hand,

$$R_2 = \int_{-\infty}^{\infty} p''(y)^2 dy$$

depends only on the unknown density $p(\cdot)$. Because of this dependence of h^* on $p(\cdot)$, h^* cannot be used directly in data analysis. However, there are two ways in which the expression for h^* may be used to guide the selection of the smoothing parameter.

First, note that R_2 depends on the unknown density $p(\cdot)$ only through the real-valued quantity R_2; that is, we do not need to know the entire density $p(\cdot)$ in order to calculate h^*. Suppose that, although we do not know $p(\cdot)$, we believe that its general features are similar in some respects to those corresponding to the those of a "reference distribution" with density $p_0(\cdot)$.

Then, in calculating h^*, we can replace R_2 by its value for the reference distribution,

$$R_{20} = \int_{-\infty}^{\infty} p_0''(y)^2 dy.$$

The most commonly-used choice for this reference distribution is the normal distribution with mean μ and standard deviation σ; taking $p_0(\cdot)$ to be the density of this distribution, it may be shown that

$$R_{20} = \frac{3}{8}\frac{1}{\sqrt{\pi}}\frac{1}{\sigma^5}.$$

Using this value in place of R_2 in the Gaussian-kernel expression for h^* leads to

$$\left(\frac{4}{3}\right)^{\frac{1}{5}} \frac{\sigma}{n^{\frac{1}{5}}} \doteq 1.059\frac{\sigma}{n^{\frac{1}{5}}}. \tag{5.5}$$

To use this choice, we need only to estimate σ using either the sample standard deviation or another estimator, such as the IQR. This approach leads to a value of the smoothing parameter that is easily calculated in practice and which should work well for smooth distributions such as the normal; the drawback is that if R_{20} is not close to the value of R_2 for the true density $p(\cdot)$, $\hat{h}_G^*$ may lead to inaccurate estimates of $p(\cdot)$.

Example 5.2 Consider computation of a kernel density estimator for the data in the variable `speed`. There are two implementations of the reference-distribution approach to choosing h available in R. Taking the argument `bw` to be `"nrd"` uses (5.5) with the estimator of σ taken to be one based on the IQR of the data; taking `bw` to be `"nrd0"` uses (5.5) with σ estimated by the minimum of the sample standard deviation and the IQR-based estimator, with the constant 1.059 modified to 0.9.

Figure 5.1 gives the kernel estimate based on the command

```
> plot(density(speed, bw="nrd"), main="", xlab="speed")
```

Note that the default value of `kernel` is `"gaussian"` so it does not need to be specified. The value of h chosen by the "nrd" method is 26.8; this can be obtained from either

```
> density(speed, bw="nrd")$bw
[1] 26.8
```

or

```
> bw.nrd(speed)
[1] 26.8
```

although for computing the density estimate, it is enough to include `bw="nrd"` in the call to the function `density`. The

Figure 5.2 gives the kernel estimate based on the command

```
> plot(density(speed, bw="nrd0"), main="", xlab="speed")
```

The value of h chosen by the "nrd0" method is 22.7 and we can see from Figures 5.1 and 5.2 that the estimates are very similar, although the one based on the "nrd0" value of h is not quite as smooth as the one based on the "nrd" value. □

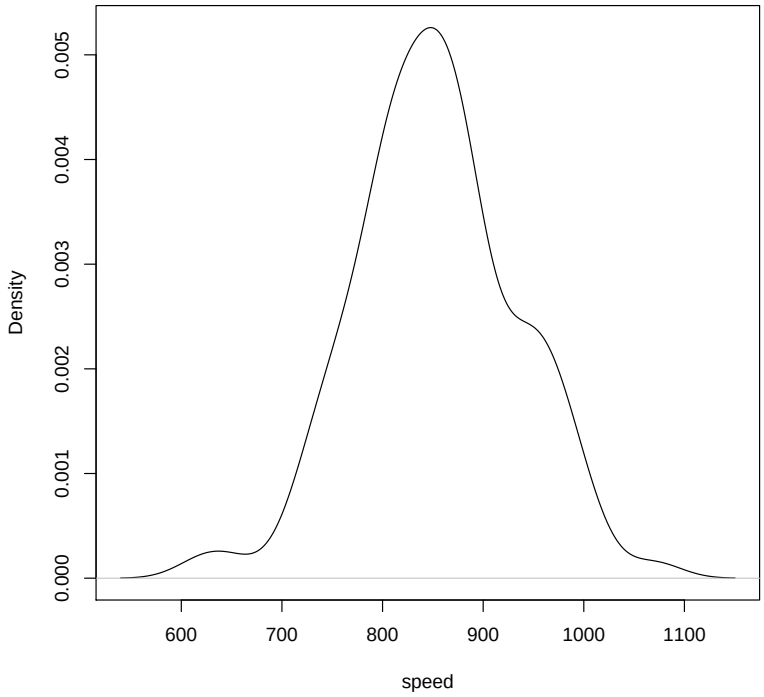

Figure 5.1: Density Estimate of the Michelson Speed of Light Data using the nrd Method

Sheather-Jones method

A second approach is to use a version of the plug-in method discussed in Week 2. The idea of the plug-in method in the present context is to use a "preliminary" estimator $\widehat{p}_1(\cdot)$ of $p(\cdot)$ and then use that estimator to estimate R_2 by

$$\hat{R}_2 = \int_{-\infty}^{\infty} \widehat{p}_1''(y)^2 dy;$$

then $\hat{R}_2$ can be used to estimate h^* by replacing R_2 by $\hat{R}_2$.

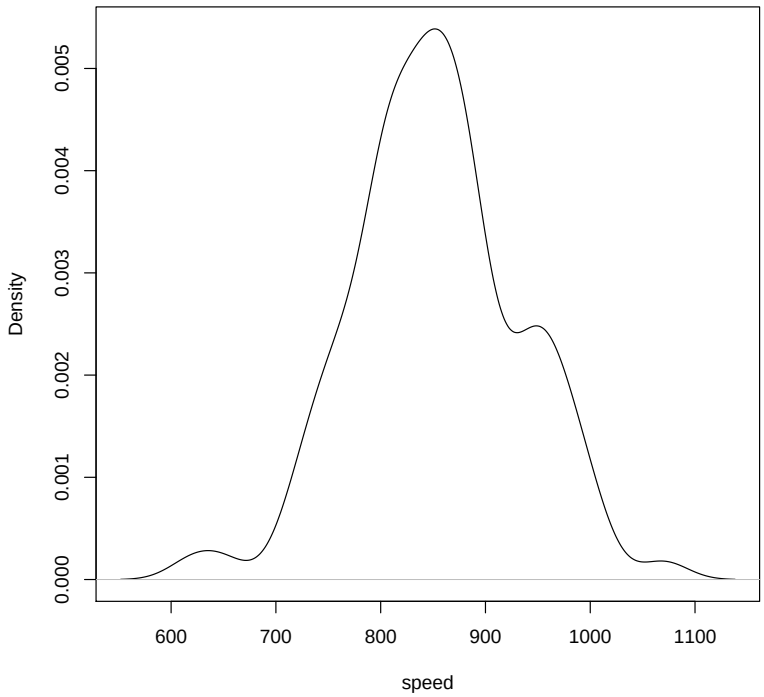

Figure 5.2: Density Estimate of the Michelson Speed of Light Data using the nrd0 Method

In choosing the preliminary estimator, it is important to keep in mind that standard methods of density estimation, such as the kernel method considered here, are based on the goal of accurately estimating the density $p(\cdot)$. However, in the present context, we are interested in estimating $p''(\cdot)$ or, more precisely, the function of $p''(\cdot)$ given by

$$\int_{-\infty}^{\infty} p''(y)^2 dy.$$

Hence, the preliminary estimator should be chosen not to be a good estimator of $p(\cdot)$ but to lead to a good estimator of $p''(\cdot)$. Although discussion of this type of estimator is beyond the scope of this course, appropriate methodology has been implemented in R, where it is called the *Sheather-Jones method*.

Example 5.3 A plot of the kernel density estimate for the data in the variable `speed` using a Gaussian kernel with the smoothing parameter chosen using the Sheather-Jones method may be obtained from the command

```
> plot(density(speed, bw="SJ"), main="", xlab="speed")
```

The plot is given in Figure 5.3. The value of h used here is 25.6, which falls between the two values given by the `nrd` and `nrd0` methods. □

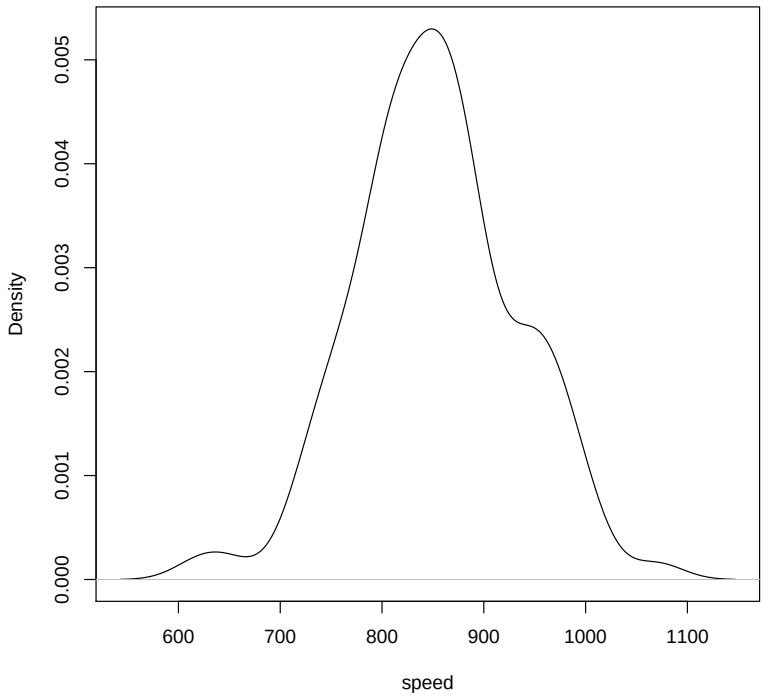

Figure 5.3: Density Estimate of the Michelson Speed of Light Data using the Sheather-Jones method

Direct approximation of the IMSE

The two methods described so for in this section are based on the approximation to the integrated mean squared error given by the AIMSE. If the AIMSE is not an accurate approximation to the integrated mean squared error then the smoothing parameter selected by these methods may not work well, even if R_2 is accurately estimated using either a reference distribution or a preliminary estimator. Hence, a third approach to choosing the smoothing parameter is based on the idea of estimating the mean integrated squared error (IMSE) directly, without relying on the AIMSE.

Let $\widehat{p}(\cdot; h)$ denote a kernel estimator of $p(\cdot)$ based on the smoothing parameter h. Then the IMSE of $\widehat{p}(\cdot; h)$ is given by

$$\mathrm{E}\left(\int_{-\infty}^{\infty}(\widehat{p}(y;h)-p(y))^2dy\right)=\mathrm{E}\left(\int_{-\infty}^{\infty}\widehat{p}(y;h)^2dy-2\int_{-\infty}^{\infty}\widehat{p}(y;h)p(y)dy+\int_{-\infty}^{\infty}p(y)^2dy\right) \tag{5.6}$$

and our goal is to choose h to make this value of the expression small. Note that the third term on the right-hand side of (5.6) does not depend on h; hence it is sufficient to choose h to minimize

$$\mathrm{E}\left(\int_{-\infty}^{\infty}\widehat{p}(y;h)^2dy-2\int_{-\infty}^{\infty}\widehat{p}(y;h)p(y)dy\right). \tag{5.7}$$

Of course, (5.7) cannot be calculated directly because it depends on the unknown density $p(\cdot)$ as well as the expectation operator $\mathrm{E}(\cdot)$ which also depends on $p(\cdot)$. Hence, we consider estimation of (5.7).

Clearly, we may estimate

$$\mathrm{E}\left(\int_{-\infty}^{\infty}\widehat{p}(y;h)^2dy\right)$$

by the unbiased estimator

$$\int_{-\infty}^{\infty}\widehat{p}(y;h)^2dy.$$

Using the fact that, for any function $g(\cdot)$,

$$\mathrm{E}\left(g(Y_1)\right)=\int_{-\infty}^{\infty}g(y)p(y)dy=\frac{1}{n}\mathrm{E}\left(\sum_{j=1}^{n}g(Y_j)\right), \tag{5.8}$$

a reasonable choice to estimate

$$\mathrm{E}\left(\int_{-\infty}^{\infty}\widehat{p}(y;h)p(y)dy\right) \tag{5.9}$$

is

$$\frac{1}{n}\sum_{j=1}^{n}\widehat{p}(Y_j;h).$$

However, unfortunately, this is a biased estimator of (5.9).

The problem arises from the fact that $\widehat{p}(y;h)$ depends $Y_1,Y_2,\ldots,Y_n$. Thus, in $\widehat{p}(Y_j;h)$, Y_j appears twice: once as the argument of the kernel estimator and once as the data used to form the kernel estimator.

To see why this is results in a biased estimator, first note that because $Y_1,Y_2,\ldots,Y_n$ are i.i.d. random variables and they are treated symmetrically in the estimator $\widehat{p}(\cdot;h)$,

$$\mathrm{E}\left(\frac{1}{n}\sum_{j=1}^{n}\widehat{p}(Y_j;h)\right)=\mathrm{E}\left(\widehat{p}(Y_1;h)\right); \tag{5.10}$$

hence, consider

$$\mathrm{E}\left(\widehat{p}(Y_1;h)\right)=\mathrm{E}\left(\frac{1}{nh}\sum_{j=1}^{n}K(\frac{Y_1-Y_j}{h})\right). \tag{5.11}$$

Recall that, for a given fixed value of y,

$$\mathrm{E}\left(\widehat{p}(y;h)\right)=\mathrm{E}\left(\frac{1}{nh}\sum_{j=1}^{n}K(\frac{y-Y_j}{h})\right)=\frac{1}{h}\mathrm{E}\left(K(\frac{y-Y_1}{h})\right), \tag{5.12}$$

and, using the argument given in Section 5.1,

$$\frac{1}{h}\mathrm{E}\left(K(\frac{y-Y_1}{h})\right)=p(y)+O(h^2), \tag{5.13}$$

so that

$$\mathrm{E}\left(\widehat{p}(y;h)\right) = p(y) + O(h^2).$$

However, the analysis is more complicated when the argument y in $\widehat{p}(y;h)$ is replaced by the random variable Y_1, because

$$K(\frac{Y_1 - Y_j}{h}) = 0 \quad \text{when} \quad j = 1.$$

It follows that

$$\sum_{j=1}^{n} K(\frac{Y_1 - Y_j}{h}) = K(0) + \sum_{j=2}^{n} K(\frac{Y_1 - Y_j}{h}).$$

The analysis of the term

$$\sum_{j=2}^{n} K(\frac{Y_1 - Y_j}{h})$$

proceeds as might be expected, leading to the result

$$\mathrm{E}\left(\frac{1}{n}\sum_{j=1}^{n}\widehat{p}(Y_j;h)\right) = \mathrm{E}\left(\widehat{p}(Y_1;h)\right) = \mathrm{E}\left(\int_{-\infty}^{\infty}\widehat{p}(y;h)p(y)dy\right) + \frac{1}{nh}K(0) + O(\frac{1}{n}).$$

It follows that

$$\frac{1}{n}\sum_{j=1}^{n}\widehat{p}(Y_j;h)$$

is a biased estimator of

$$\mathrm{E}\left(\int_{-\infty}^{\infty}\widehat{p}(y;h)p(y)dy\right),$$

with the bias due primarily to

$$\frac{1}{nh}K(0),$$

the term that appears in the expression because of the "double-use" of Y_1.

Hence,

$$\int_{-\infty}^{\infty}\widehat{p}(y;h)^2dy - \frac{2}{n}\sum_{j=1}^{n}\widehat{p}(Y_j;h)$$

is a biased estimator of

$$\mathrm{E}\left(\int_{-\infty}^{\infty}\widehat{p}(y;h)^2dy - 2\int_{-\infty}^{\infty}\widehat{p}(y;h)p(y)dy\right), \tag{5.14}$$

with bias of the form

$$-\frac{2}{nh}K(0) + O(\frac{1}{n}).$$

In many cases, a small bias of an estimator does not have an important effect on its properties and, hence, it can be safely ignored. Unfortunately, that is the not the case in the

present context. Recall that our goal is to choose h to minimize (5.14). Note that the first term in the bias depends on h and becomes large (and negative) as h approaches 0. Thus, (5.14) is typically minimized by choosing h to be near 0.

Hence, we consider the following alternative approach to estimating

$$\mathrm{E}\left(\int_{-\infty}^{\infty}\widehat{p}(y;h)p(y)dy\right).$$

Let

$$\widehat{p}_{-i}(y;h)$$

denote the kernel density estimator based on all of the data points **except** Y_i:

$$\widehat{p}_{-i}(y;h)=\frac{1}{(n-1)h}\sum_{j\neq i}K(\frac{y-Y_j}{h}).$$

We may then estimate

$$\mathrm{E}\left(\int_{-\infty}^{\infty}\widehat{p}(y;h)p(y)dy\right)$$

by

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{p}_{-i}(Y_i;h)$$

so that in $\widehat{p}_{-i}(Y_i;h)$, Y_i is used only once, as the argument of the density.

This cures the problems that arise from the double-use of Y_i in $\widehat{p}(Y_i;h)$ and it may be shown that

$$\int_{-\infty}^{\infty}\widehat{p}(y;h)^2dy-\frac{2}{n}\sum_{i=1}^{n}\widehat{p}_{-i}(Y_i;h) \tag{5.15}$$

is an unbiased estimator of

$$\mathrm{E}\left(\int_{-\infty}^{\infty}\widehat{p}(y;h)^2dy-2\int_{-\infty}^{\infty}\widehat{p}(y;h)p(y)dy\right).$$

Therefore, the value of h that minimizes (5.15) may be viewed as an estimate of the value of the smoothing parameter that minimizes the IMSE.

The expression (5.15) is known as the *cross-validation criterion* and the value of h that minimizes (5.15), which we will denote by h_{CV}, is known as the *cross-validation smoothing parameter*. Note that the minimization of (5.15) is typically done numerically.

Some authors use a slightly different definition of the cross-validation criterion, in which $\widehat{p}(y;h)^2$ is estimated by

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{p}_{-i}(y;h)^2.$$

However, the basic idea is the same and the differences between the two methods are generally minor.

Example 5.4 To use cross-validation to choose the smoothing parameter, the argument `bw="ucv"` is included in the `density` command; here "ucv" denotes "unbiased cross-validation". Hence, to construct a plot of the kernel density estimate for the `speed` data, choosing h by cross-validation, we use the command

```
> plot(density(speed, bw="ucv"), main="", xlab="speed")
```

The plot is in Figure 5.4; for the speed-of-light data, $h_{CV} = 32$. □

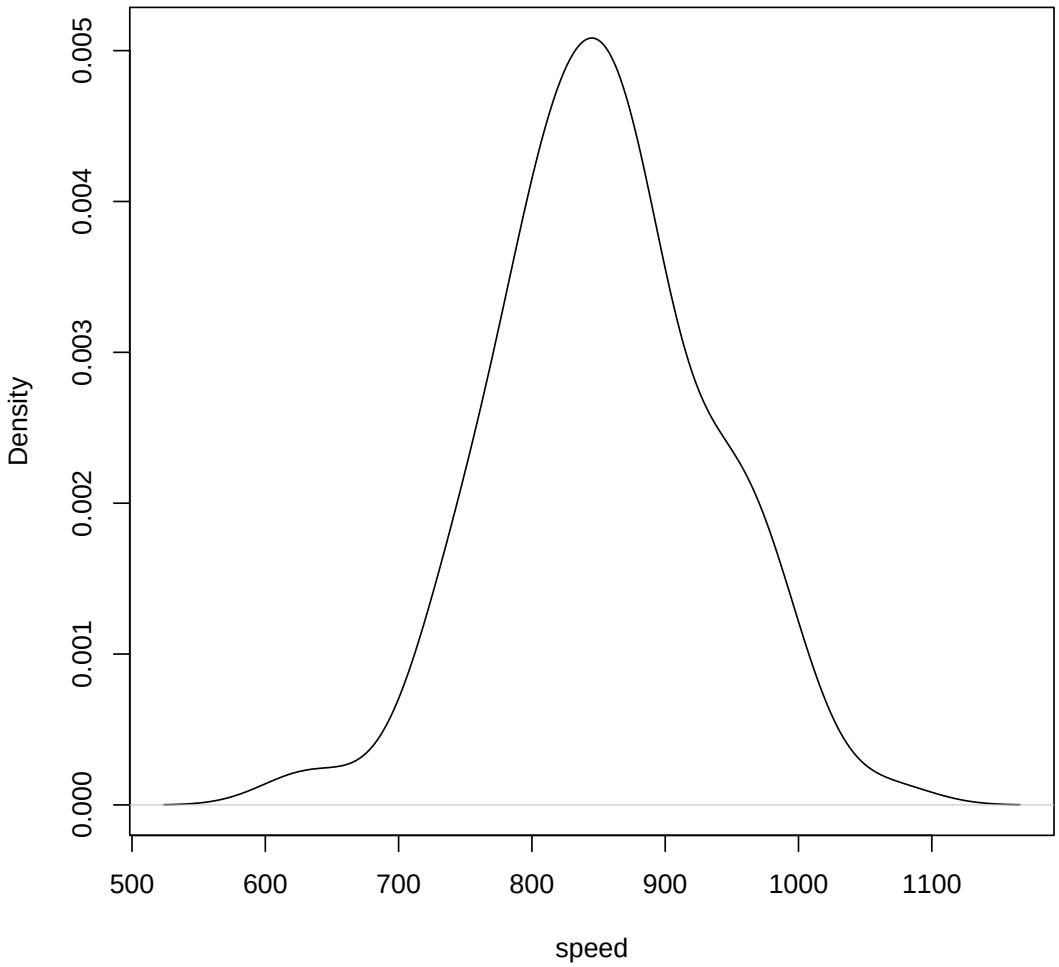

Figure 5.4: Density Estimate of the Michelson Speed of Light Data using $h = h_{CV}$

Comparison of the methods

Of the methods discussed, the Sheather-Jones and cross-validation methods are considered to be those that are generally most useful. Empirical results suggest that the Sheather-Jones method works well in many, if not most, examples. Cross-validation also often works well

although it has a tendency to undersmooth somewhat; that is, it has a tendency to choose a value of h that is too small (however that is not the case in the speed of light example).

On the other hand, cross-validation has the advantage that the basic idea it is applicable to a wide range of methods, including multivariate density estimation (considered in Week 6) and kernel estimation of a regression function (considered in the following chapter). Also, it is known to be more likely to result in poor choices for h, perhaps because it requires numerical minimization of an objective function, which can be a challenging computational problem (depending on the shape of the objective function). This situation is easily detected if one plots the estimate.

It is important to note that often the Sheather-Jones and cross-validation methods yield very similar values of h and in such cases the differences between the estimates are quite minor.

5.3 Applications of density estimates

Although often density estimation is used simply as a way to better understand the properties of the observed data, a density estimator may also be used to estimate other quantities that are functions of the unknown underlying density function of the data. In this section, we consider a few of these.

Integration of the density estimate

Let Y denote a random variable with a continuous distribution with density $p(\cdot)$. Many properties of Y, or of the density $p(\cdot)$, can be expressed in terms an integral of $p(\cdot)$ times another function. For instance,

$$\mathrm{E}(Y) = \int_{-\infty}^{\infty} y\, p(y)\, dy.$$

Thus, an integral of $\widehat{p}(\cdot)$ might give useful information regarding $\widehat{p}(\cdot)$ (for instance, the mean or standard deviation of the distribution with density $\widehat{p}(\cdot)$) or it might be used to estimate certain properties of Y. That is, given a kernel estimate $\widehat{p}(\cdot)$ of $p(\cdot)$, an estimate of

$$\int_{-\infty}^{\infty} g(y)\, p(y)\, dy$$

for some given function $g(\cdot)$ can be estimated by

$$\int_{-\infty}^{\infty} g(y)\, \widehat{p}(y)\, dy.$$

This can be viewed as an application of the plug-in method.

Although, in some cases, an analytic expression for such an integral can be determined using the expression for the kernel estimate (the following subsection discusses one example of this), in many cases, the most convenient approach is to integrate $\widehat{p}(\cdot)$ using numerical integration. In R, the simplest (though not the most efficient) way to perform numerical integration is to use the function `integrate`.

For instance, the Michelson speed of light data (stored in the variable `speed`) and suppose that we are interested in estimating the mean of the distribution. An estimate based on $\widehat{p}(\cdot)$ is given by

$$\int_{-\infty}^{\infty} y\,\widehat{p}(y)\,dy.$$

The function `integration` takes three main arguments: `f`, the function to be integrated, and `lower` and `upper`, which define the interval over which the integral is be computed. Although, technically, these can be taken to be plus or minus infinity, in practice, finite values based on the range of the data are used.

Suppose that the information regarding the kernel density estimate is stored in the variable `denout`:

```
> denout<-density(speed, bw="nrd")
```

To use `integrate`, we need the value of $\widehat{p}(y)$ for any value of y. We can obtain such values using linear interpolation, using the R function `approxfun`. `approxfun(denout)` returns a function that performs linear interpolation using the density estimate information in `denout`. Hence, for a given value of y, `approxfun(denout)(y)` returns $\widehat{p}(y)$. For instance,

```
> approxfun(denout)(800)
[1] 0.0042299
```

To integrate $y\widehat{p}(y)$, first we define function a `f` that corresponds to

$$y\widehat{p}(y).$$

Consider

```
> f<-function(y){
+ y*approxfun(denout)(y)
+ }
```

Then `f(y)` returns the value of $y\widehat{p}(y)$.

`denout` includes components `$x` and `$y` which, in the present context, contain the values of the speed of light variable (`$x`) for which the density estimate is calculated and the corresponding density estimate (`$y`) based on the kernel estimate for the speed of light data. Here

```
> summary(denout$x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    540     692     845     845     998    1150
```

so we can take 540 and 1150 as the limits for any integral. Note that these values fall outside the range of the data:

```
> summary(speed)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    620     808     850     852     892    1070
```

The integral

$$\int_{540}^{1150} y\,\widehat{p}(y)\,dy$$

can be computed using the function `integrate`:

```
> integrate(f=f, lower=540, upper=1150)
853.21 with absolute error < 0.096
```

This can be compared to the sample mean

```
> mean(speed)
[1] 852.4
```

This type of integration is most useful when it is not possible to estimate the integral using a simple average as can done to estimate $\mathrm{E}(Y)$. For instance, consider the *differential entropy* of a distribution with density $p(\cdot)$:

$$-\int_{-\infty}^{\infty} \log(p(y))p(y)\,dy.$$

This can be estimated by

$$-\int_{-\infty}^{\infty} \log(\widehat{p}(y))\widehat{p}(y)\,dy.$$

Another example occurs when comparing two density functions. One measure of the "distance" between densities $p_1(\cdot)$ and $p_2(\cdot)$ is

$$\int_{-\infty}^{\infty} (p_1(y) - p_2(y))^2\,dy.$$

Suppose that we are interested in comparing the density function $p(\cdot)$ to some "standard" density $p_0(\cdot)$, such as the density function of a normal distribution.

We can estimate the distance between $p(\cdot)$ and $p_0(\cdot)$ by

$$\int_{-\infty}^{\infty} (\widehat{p}(y) - p_0(y))^2\,dy$$

where $\widehat{p}(\cdot)$ is a kernel estimator of $p(\cdot)$. This type of problem will be considered in detail later in this section.

Smoothed estimators of a distribution function

Let $Y_1, Y_2, \ldots, Y_n$ denote i.i.d. random variables, each with distribution function $F(\cdot)$. In Week 2, we considered the empirical distribution function as an estimator of $F(\cdot)$.

Suppose that the Y_j have a continuous distribution with density $p(\cdot)$. Although the empirical distribution function is easy to compute (e.g., no smoothing parameter is needed) and easy to analyze, it has the drawback that it is the distribution function of a discrete distribution and, hence, it is not a continuous function, even though the true distribution function we are estimating is continuous.

An alternative approach is to use an estimator of the density function $p(\cdot)$ in order to construct an estimator of $F(\cdot)$. Specifically, given a kernel density estimator $\widehat{p}(\cdot)$ of $p(\cdot)$, we may estimate $F(\cdot)$ by computing the distribution function corresponding to $\widehat{p}(\cdot)$:

$$\widehat{F}_K(y) = \int_{-\infty}^{y} \widehat{p}(t)dt, \quad -\infty < y < \infty. \tag{5.16}$$

Although $\widehat{F}_K(y)$ is easily determined numerically, using numerical integration, here we consider an analytic expression for $\widehat{F}_K(\cdot)$.

Consider the case of a Gaussian kernel. Then

$$\widehat{p}(y) = \frac{1}{nh}\sum_{j=1}^{n} K(\frac{y - Y_j}{h}) = \frac{1}{nh}\sum_{j=1}^{n} \phi(\frac{y - Y_j}{h}), \quad -\infty < y < \infty,$$

where $\phi(\cdot)$ denotes the density function of the standard normal distribution. It follows that

$$\widehat{F}_K(y) = \int_{-\infty}^{y} \frac{1}{nh}\sum_{j=1}^{n} \phi(\frac{t - Y_j}{h})dt = \frac{1}{n}\sum_{j=1}^{n} \int_{-\infty}^{y} \frac{1}{h}\phi(\frac{t - Y_j}{h})dt.$$

Using the change-of-variable $u = (t - Y_j)/h$,

$$\int_{-\infty}^{y} \frac{1}{h}\phi(\frac{t - Y_j}{h})dt = \int_{-\infty}^{\frac{y-Y_j}{h}} \phi(u)du = \Phi(\frac{y - Y_j}{h}),$$

where $\Phi(\cdot)$ is the distribution function of the standard normal distribution; it follows that

$$\widehat{F}_K(y) = \frac{1}{n}\sum_{j=1}^{n} \Phi(\frac{y - Y_j}{h}), \quad -\infty < y < \infty. \tag{5.17}$$

Note that

$$\mathrm{E}\left(\widehat{F}_K(y)\right) = \mathrm{E}\left(\Phi(\frac{y - Y_1}{h}) = \int_{-\infty}^{\infty} \Phi(\frac{y - t}{h})p(t)dt\right) \tag{5.18}$$

so that, in constrast to $\widehat{F}(\cdot)$, $\widehat{F}_K(\cdot)$ is, in general, a biased estimator of $F(\cdot)$.

Although the bias of $\widehat{F}_K(y)$ is larger than that of $\widehat{F}(y)$, it may be shown that its variance is smaller and the mean squared error of $\widehat{F}_K(y)$ is smaller than that of $\widehat{F}(y)$.

Example 5.5 Consider the data on the returns on Apple stock, stored in the variable `apple`; see Example 2.10. A plot of the kernel density estimate based on a Gaussian kernel and $h = h_{CV}$ is given in Figure 5.5. Here $h_{CV} = 0.0246$; $h_{SJ} = 0.0238$, yielding similar results.

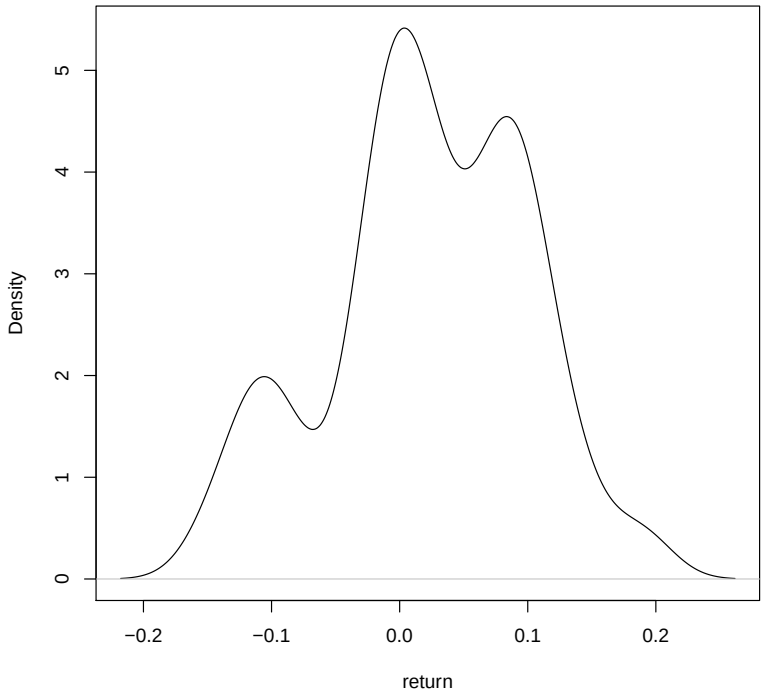

Figure 5.5: Density Estimate of the Apple Return Data using $h = h_{CV}$

To compute the smoothed estimate of the distribution function, $\widehat{F}_K(\cdot)$, for these data, we need to calculate the expression in (5.17) for many values of y.

Note that the minimum and maximum values in `apple` are -0.144 and 0.188, respectively. Hence, construct a vector `yv` by

```
> yv<-(-.20) + (.44)*(0:500)/500
```

Thus, `yv` contains 501 equally-spaced values from -0.20 to 0.24, roughly the minimum return value minus $2h_{CV}$ to the maximum return value plus $2h_{CV}$.

We can use the `outer` function to compute a matrix of values corresponding to $(y_k - Y_j)/h_{CV}$ where y_k is an element of the vector `yv` and Y_j is an observed return value, that is, an element of `apple`:

```
> mat<-outer(yv, apple, "-")/0.0246
```

Thus, `mat` is a 501×36 matrix.

We may now calculate $\widehat{F}_K(y)$ for the values of y in `yv` by using the matrix `mat` as the argument to the function `pnorm`, which computes the standard normal distribution function of an argument, and then averaging across rows using the `apply` function, which applies a function to margins of a matrix or, more generally, an array:

```
> dist_fun<-apply(pnorm(mat), MARGIN=1, FUN=mean)
```

Here `pnorm(mat)` returns a matrix of values of the form $\Phi((y_k - Y_j)/h_{CV})$ and `apply` takes the sample mean of the values in each row (specified by the second argument `1`, which denotes the first margin of the matrix, rows). It follows that `dist_fun` contains 501 values of the kernel distribution function estimate $\widehat{F}_K(\cdot)$ given by (5.16); a plot of $\widehat{F}_K(\cdot)$ is given in Figure 5.6. □

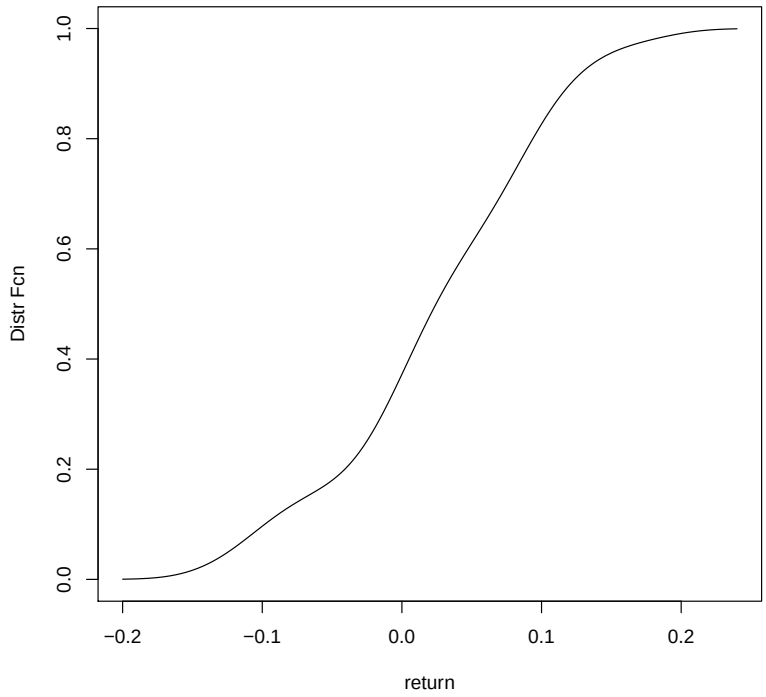

Figure 5.6: Kernel Estimate of the Distribution Function of Apple Return Data

Goodness-of-fit tests

Suppose that $Y_1, Y_2, \ldots, Y_n$ are i.i.d. random variables each continuously distributed with density $p(\cdot)$; a density estimate based on $Y_1, Y_2, \ldots, Y_n$ gives useful information regarding

the properties of $p(\cdot)$. Let $p_0(\cdot)$ denote a density thought to be a reasonable choice for $p(\cdot)$; for instance, we might think that the Y_j are normally distributed so that $p_0(\cdot)$ is a normal density. We can compare a kernel density estimate $\widehat{p}(\cdot)$ to $p_0(\cdot)$ to determine if $p_0(\cdot)$ is a plausible choice for $p(\cdot)$.

Example 5.6 Consider the speed-of-light data stored in the variable `speed` and suppose we are interested in determining if the speed-of-light measurements are approximately normally distributed.

Recall that a plot of the kernel density estimate for the `speed` data, using $h = h_{CV}$ is given in Figure 5.4. Based on this plot, the speed-of-light measurements appear to be approximately normally distributed. □

The purpose of the goodness-of-fit test is to formally assess how close an estimate $\widehat{p}(\cdot)$ is to $p_0(\cdot)$.

To measure the "distance" between the kernel estimate $\widehat{p}(\cdot)$ and a hypothesized density for the data, $p_0(\cdot)$, we may use the statistic

$$\int_{-\infty}^{\infty} \left(\widehat{p}(u) - p_0(u)\right)^2 du. \tag{5.19}$$

The statistic (5.19) may be viewed as an estimator of

$$\int_{-\infty}^{\infty} \left(p(u) - p_0(u)\right)^2 du,$$

which is 0 whenever $p(u) = p_0(u)$ for all u, that is, whenever $p(\cdot)$ and $p_0(\cdot)$ describe the same probability distribution. The statistic (5.19) may be calculated using numerical integration, as discussed previously in this section.

A formal hypothesis test based on the test statistic (5.19) may be conducted by computing a p-value based on the asymptotic distribution of the statistic under the hypothesis that $p(\cdot) = p_0(\cdot)$. Such a test can be conducted using the function `fan.test` in the package "GoFKernel".

Example 5.7 Consider the comparison of the distribution of the speed-of-light data with the normal distribution, as discussed in Example 5.6. The following commands can be used to compute the p-value of the test that the data are normally distributed, with mean 852.4 and standard deviation 79.01 (the sample mean and standard deviation of the data in `speed`).

```
> library(GoFKernel)
> fan.test(speed, fun.den=dnorm,
+   par=list(mean=852.4, sd=79.01), bw=bw.SJ(speed),
```

```
+    lower=500, upper=1200)

        Fan's test

data:  speed
Ig = -1.39, p-value = 0.92
```

□

The function `fan.test` has a number of arguments:

- `fun.den` specifies a function that computes the density p_0, that is, the density of the data under the null hypothesis. R contains a number of such functions, such as `dnorm`, the density of the normal distribution, and `dgamma`, the density of the gamma distribution. The command `?Distributions` provides a list of the density functions available in the base package of R. Alternatively, you can write your own function.

- `par` specifies the parameter values for the function given in `fun.den`; the parameter values are given as a "list", with the components taking the same names as the parameter values used in the density specified by `fun.den`.

- `bw` specifies the smoothing parameter to be used in the kernel density estimate. The function `bw.SJ` chooses h using the Sheather-Jones method; cross-validation (`bw.ucv`) could also be used.

- `lower` and `upper` specify the range of values over which the integration is to be done when calculating the statistic (5.19); the default values are $-\infty$ and ∞, but these are generally poor choices. Choosing too long of an interval over which to calculate the test statistic often includes values of the argument y for which $p(y)$ is not of much interest. This can lead to a large value for (5.19) and, hence, rejection of the null hypothesis even when $\widehat{p}(y)$ is close to $p_0(y)$ for relevant values of y. Hence, `lower` and `upper` should be chosen to include the density values of interest; an interval slightly longer than that given by the range of the data often works well.

The result of the function `fan.test` is the value of the standardized test statistic for testing $p(\cdot) = p_0(\cdot)$, such that a larger value indicates less agreement between $\widehat{p}(\cdot)$ and $p_0(\cdot)$. Under the null hypothesis, the distribution of this statistic is approximately standard normal and the stated p-value is simply $1 - \Phi(T_0)$ where T_0 is the observed value of the statistic (5.19). Thus, in the example considered here it is

```
> 1-pnorm(-1.42)
[1] 0.9222
```

It follows that the null hypothesis is not rejected so that the estimated density of the data is consistent with a normal distribution; given the plot in Figure 5.7, such a result is not surprising.

Note that the results of the test may be sensitive to the choice of smoothing parameter. Consider the speed of light example; if, for example, cross-validation is used to select the smoothing parameter, the p-value changes to 0.988:

```
> fan.test(speed, fun.den=dnorm, par=list(mean=852.4, sd=79.01),
> bw=bw.ucv(speed), lower=500, upper=1200)
        Fan's test
data:  speed
Ig = -2.254, p-value = 0.988
```

Of course, in this case, the conclusion of the test is unchanged.

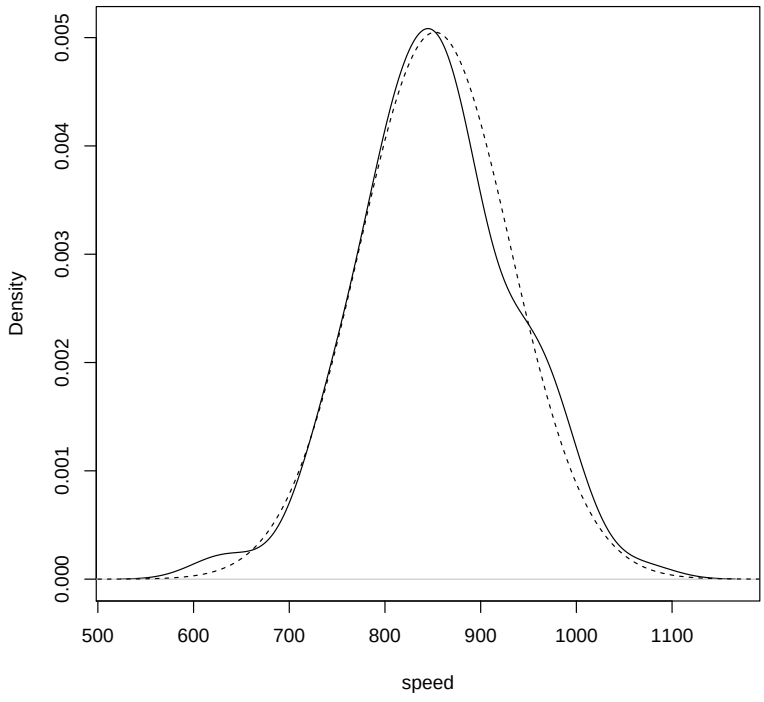

Figure 5.7: Kernel Estimate of the Density of the Speed of Light Data with a Normal Density as a Reference

Comparing two distributions

The method discussed above considers the comparison of the distribution of the observed data and a hypothesized distribution. The same general approach may be used to compare the distributions of two sets of data.

Suppose that we observe two samples of data, $Y_1, Y_2, \ldots, Y_n$, which are i.i.d. random variables each distributed according to a density $p_Y(\cdot)$ and $X_1, X_2, \ldots, X_m$, which are i.i.d. random variables each distributed according to a density $p_X(\cdot)$. Furthermore, we assume that $(Y_1, Y_2, \ldots, Y_n)$ and $(X_1, X_2, \ldots, X_m)$ are independent. Here we consider a test of the null hypothesis $H_0 : p_Y(\cdot) = p_X(\cdot)$, that is, the hypothesis that for each j, k, Y_j and X_k have the same distribution.

Let $\widehat{p}_Y(\cdot; h_Y)$ and $\widehat{p}_X(\cdot; h_X)$ denote kernel estimators of $p_Y(\cdot)$ and $p_X(\cdot)$, respectively, where h_Y and h_X denote the respective smoothing parameters. We assume that both kernel estimators are based on a Gaussian kernel.

Consider the test statistic

$$\int_{-\infty}^{\infty} \left(\widehat{p}_Y(u; h_Y) - \widehat{p}_X(u; h_X)\right)^2 du, \tag{5.20}$$

which may be viewed as a measure of the "distance" between the function $\widehat{p}_Y(\cdot; h_Y)$ and the function $\widehat{p}_X(\cdot; h_X)$. The statistic (5.20) may be viewed as an estimator of

$$\int_{-\infty}^{\infty} \left(p_Y(u) - p_X(u)\right)^2 du,$$

which is 0 if and only if $p_Y(\cdot)$ and $p_X(\cdot)$ describe the same probability distribution.

Although h_Y and h_X could be chosen individually for each density, based on the samples $Y_1, Y_2, \ldots, Y_n$ and $X_1, X_2, \ldots, X_m$, respectively, it is generally desirable to use the same smoothing parameter for both estimators. One reason for this is the properties of $\widehat{p}_Y(\cdot) - \widehat{p}_X(\cdot)$ under the null hypothesis that $p_X(\cdot) = p_Y(\cdot)$.

Using the fact that

$$\mathrm{E}\left(\widehat{p}(y)\right) = \frac{1}{nh} \sum_{j=1}^{n} \mathrm{E}\left(K(\frac{y - Y_j}{h})\right) = \int_{-\infty}^{\infty} \frac{1}{h} K(\frac{y-t}{h}) p(t) dt = \int_{-\infty}^{\infty} K(u) p(y - uh) du,$$

it follows that

$$\mathrm{E}\left(\widehat{p}_Y(y; h_Y)\right) = \int_{-\infty}^{\infty} K(u) p_Y(y - uh_Y) du$$

and

$$\mathrm{E}\left(\widehat{p}_X(x; h_X)\right) = \int_{-\infty}^{\infty} K(u) p_X(x - uh_X) du.$$

If $p_X(\cdot) = p_Y(\cdot) \equiv p(\cdot)$, then these expressions become

$$\mathrm{E}\left(\widehat{p}_Y(y; h_Y)\right) = \int_{-\infty}^{\infty} K(u)p(y - uh_Y)du$$

and

$$\mathrm{E}\left(\widehat{p}_X(x; h_X)\right) = \int_{-\infty}^{\infty} K(u)p(x - uh_X)du,$$

respectively. Choosing $h_Y = h_X$ ensures that, under the null hypothesis,

$$\mathrm{E}\left(\widehat{p}_Y(u; h_Y)\right) = \mathrm{E}\left(\widehat{p}_X(u; h_X)\right)$$

for each u; that is, $\widehat{p}_Y(u; h_Y) - \widehat{p}_X(u; h_X)$ has expected value 0 under the null hypothesis.

One method of choosing such a common value of the smoothing parameter is to first choose h_Y and h_X using one of the methods discussed previously, and then combine these into a single value by computing their geometric mean, leading to h^* given by

$$\frac{1}{h^*} = \frac{1}{2}\left(\frac{1}{h_Y} + \frac{1}{h_X}\right).$$

The test statistic

$$\int_{-\infty}^{\infty} \left(\widehat{p}_Y(u; h_Y) - \widehat{p}_X(u; h_X)\right)^2 du$$

may be calculated using numerical integration. To calculate a p-value for the test, we use the following reasoning. Suppose that $p_Y(\cdot) = p_X(\cdot)$. Then $Y_1, Y_2, \ldots, Y_n, X_1, \ldots, X_m$ are i.i.d. random variables and any two groups of these random variables, of size n and m respectively, yields a test statistic with the same distribution as the one based on the partition $Y_1, Y_2, \ldots, Y_n$ and $X_1, X_2, \ldots, X_m$.

That is, permuting $(Y_1, Y_2, \ldots, Y_n, X_1, X_2, \ldots, X_m)$ and taking the first n values of the permutation as the first group and the remaining values as the second group, leads to a test statistic with the same distribution as the one based on the groups $Y_1, Y_2, \ldots, Y_n$ and $X_1, X_2, \ldots, X_m$.

For instance, suppose that $n = 3$ and $m = 2$. Then, under the null hypothesis, the statistic (5.19) based on the groups Y_1, Y_2, Y_3 and X_1, X_2 has the same distribution as the statistic based on the permutation

$$(X_2, Y_2, Y_3, X_1, Y_1),$$

which leads to the groups X_2, Y_2, Y_3 and X_1, Y_1. More generally, any of the $5! = 32$ permutations of $(Y_1, Y_2, Y_3, X_1, X_2)$ leads to a test statistic with the same distribution as the one based on the observed groups Y_1, Y_2, Y_3 and X_1, X_2.

Then, the null distribution of the test statistic may be taken to be the discrete distribution obtained by computing the test statistic for each such permutation of the data. We reject the null hypothesis of equal distributions if the observed value of the test statistic is large relative to those values of the test statistic based on different permutations of the data. Hence, such a test is often called a *permutation test*.

In practice, we do not need to use all possible permutations of the data (which, for large n and m, would require excessive computation). Instead, the test is based on a random sample of all such permutations, leading to a procedure similar to those based on the bootstrap method.

Example 5.8 The R variable `tri` contains measurements on the plasma triglycerides (in mg per dl) for 371 male patients being evaluated for chest pain. The variable `grp` is an indicator variable taking the value 0 if the patient showed no evidence of heart disease and taking the value 1 if the patient exhibited narrowing of the arteries; 51 patients were in the no-disease group and 320 patients were in the disease group. These data are available in the dataset "blood".

Figure 5.8 contains kernel density estimates for the data from each group; in both cases, the smoothing parameter h was chosen by the Sheather-Jones method. It is worth noting that the values of h used were fairly different for the two groups, with $h_{SJ} = 19.5$ for the no-heart-disease group and $h_{SJ} = 14.7$ for the heart-disease group.

```
> bw.SJ(tri[grp==0])
[1] 19.5
> bw.SJ(tri[grp==1])
[1] 14.7
```

The sample means and standard deviations for the groups follow:

```
> mean(tri[grp==0])
[1] 140.4
> sd(tri[grp==0])
[1] 74.3
> mean(tri[grp==1])
[1] 179.4
> sd(tri[grp==1])
[1] 101.8
```

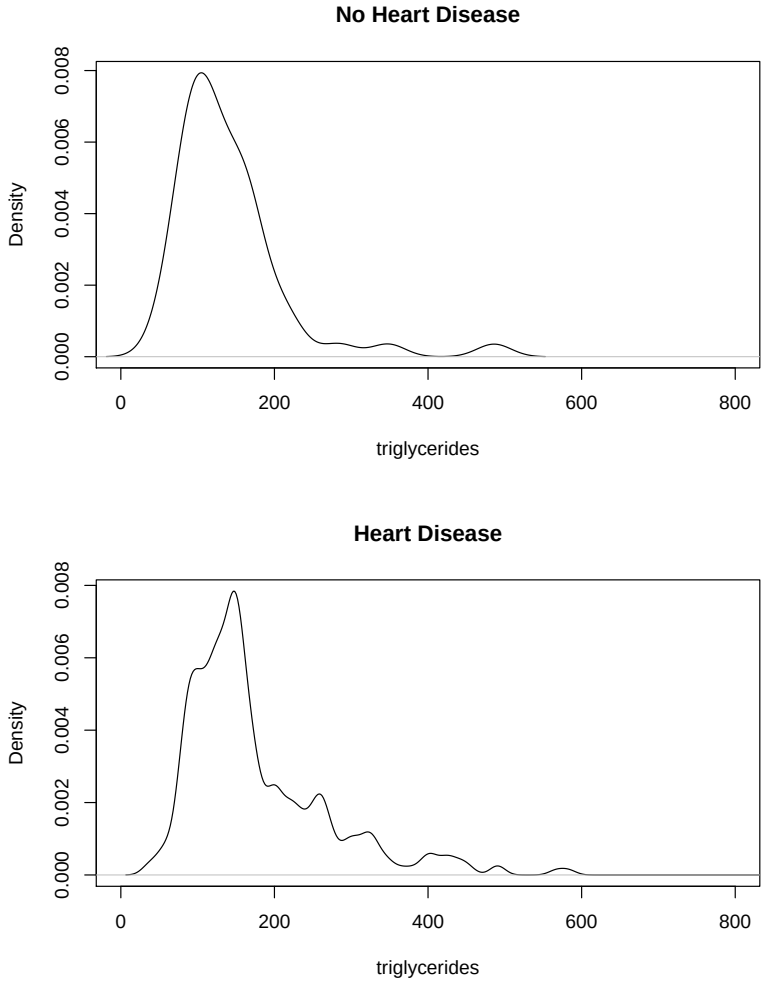

Figure 5.8: Kernel Estimate of the Density of the Plasma Triglyceride Levels for Patients With and Without Heart Disease

To test the hypothesis that the density functions of plasma triglycerides are the same in the two groups, we can use the command `sm.density.compare` in the package "sm". The method implemented by this command uses the same smoothing parameter for each group, which can be chosen by taking the geometric mean of h_{SJ} for the two groups:

$$h^* = 2(\frac{1}{19.5} + \frac{1}{14.7})^{-1} = 16.8.$$

The form of the command is

```
> sm.density.compare(tri, group=grp, model="equal", bw=16.8, nboot=10000)
Test of equal densities:  p-value =  0.011
```

Here `model="equal"` specifies that a test of equal densities is to be conducted; the argument `bw` specifies the value of the smoothing parameter to be used and `nboot` specifies the number of permutations to use when calculating the p-value.

Using the standard 0.05 criterion for significance, the p-value of 0.011 observed here indicates that we would reject the hypothesis the density functions for the two groups are equal; thus, there appears to be a difference in the distribution of triglyceride levels for patients with and without heart disease. □

In some cases, we may be interested in testing the hypothesis that two density functions have the same "shape" but with different location and scale parameters. In such cases, we may first standardize the data using measures of the location and scale of the data, such as the sample mean and sample standard deviation, and then apply the test to the standardized data. For instance, to test that $p_Y(\cdot)$ and $p_X(\cdot)$ have the same shape, we may first standardize $Y_1, Y_2, \ldots, Y_n, X_1, X_2, \ldots, X_m$ to

$$\frac{Y_1 - \bar{Y}}{S_Y}, \frac{Y_2 - \bar{Y}}{S_Y}, \ldots, \frac{Y_n - \bar{Y}}{S_Y}, \frac{X_1 - \bar{X}}{S_X}, \frac{X_2 - \bar{X}}{S_X}, \ldots, \frac{X_m - \bar{X}}{S_X},$$

where $\bar{Y}, \bar{X}$ denote the respective sample means an S_Y, S_X denote the respective sample standard deviations, and then base the test on these standardized data.

Example 5.9 Consider the data on plasma triglyceride levels analyzed in Example 5.8. Let `tri_st` denote the standardized triglyceride values, formed by subtracting the group sample mean from each observation and then dividing by the group sample standard deviation:

```
> tri_st<-c((tri[grp==0]-mean(tri[grp==0]))/sd(tri[grp==0]),
+  (tri[grp==1]-mean(tri[grp==1]))/sd(tri[grp==1]))
> grp_st<-c(rep(0, 51), rep(1, 320))
```

The values of h_{SJ} for the data in `tri_st` for the two groups are given by

```
> bw.SJ(tri_st[grp==0])
[1] 0.264
> bw.SJ(tri_st[grp==1])
[1] 0.145
```

so that their geometric mean is 0.187.

Thus, a test that the densities of the standardized triglyceride levels are equal can be carried out using the command

```
> sm.density.compare(tri_st, group=grp_st, model="equal", bw=0.187, nboot=10000)

Test of equal densities:  p-value =  0.580
```

Thus, there is no evidence to reject the hypothesis that there is no difference in the shapes of the distributions of triglyceride levels for patients with and without heart disease. □

5.4 Exercises

5.1. Consider the failure data for sample of ball bearings stored in the dataset "failure". Each value is the failure time, in cycles of use, of a ball bearing. Plot the kernel density estimate based on the smoothing parameter chosen by the Sheather-Jones method. Add a curve to the plot, representing the density function of a gamma distribution, which can be calculated using the R function `dgamma`.

Recall that the gamma distribution has two parameters, corresponding to the arguments `shape` and `scale` to the function `dgamma`. The mean of the distribution is (shape) × (scale) and the variance of the distribution is (shape) × (scale)^2 ; see the help file for `rgamma` for further details.

Use these relationships, along with the sample mean and sample variance of the failure data, to choose the values of the arguments (that is, we are using the method of moments estimators of the parameters).

Based on this plot, does it appear that the failure data follow a gamma distribution?

5.2. Consider the scores on the Peabody Picture Vocabulary Test contained in the dataset "Peabody".

Plot the kernel density estimates for these data based on the nrd0, Sheather-Jones, and cross-validation methods, together with the estimate based on the value of h you chose in Exercise 4.6.

In your opinion, do any of the three methods produce a density estimate that is preferable to the one you chose subjectively?

5.3. The dataset "geyser" contains data on the time between eruptions of the Old Faithful geyser at Yellowstone National Park, over the period from August 1 to August 15, 1985.

Estimate the density of the time between eruptions using a kernel estimate. Consider four methods of choosing the smoothing parameter: `nrd`, `nrd0`, `SJ`, and `ucv`. Give the value of the smoothing parameter found by each method and plot the density estimate that you consider to be the best choice (from among the four possibilities).

Comment on the general features of the density.

5.4. Consider the data in the dataset "Peabody".

(a) Let $\widehat{p}(\cdot)$ denote the kernel density estimate with the smoothing parameter taken to be $h = 2$. Use numerical integration to find $\widehat{\mu}$ and $\widehat{\sigma}$, where

$$\widehat{\mu} = \int_{-\infty}^{\infty} y\,\widehat{p}(y)\,dy$$

and

$$\widehat{\sigma}^2 = \int_{-\infty}^{\infty} (y - \widehat{\mu})^2 \widehat{p}(y) dy.$$

When calculating the integrals numerically, use the method described in Section 5.3 to find the limits of integration.

(b) Repeat part (a) for $h = 3, 4$, and 5.

(c) Summarize how the values of $\widehat{\mu}$ and $\widehat{\sigma}$ change as h increases.

5.5. For the failure data analyzed in Exercise 5.1, use `fan.test` to test the hypothesis that the data follow a gamma distribution, choosing the parameters of the distribution using the same approach used in Exercise 5.1. Choose the smoothing parameter using the Sheather-Jones method and compare the densities over the interval $(15, 185)$.

Based on the results, is the assumption of a gamma distribution for the failure data seem reasonable?

5.6. For the data in the Excel file "peabody", use `fan.test` to test the hypothesis that the data are normally distributed, using the sample mean and sample standard deviation for the parameter values. Compare the densities over a range that is slightly larger than the range of the data and use the Sheather-Jones method to choose the smoothing parameter.

What do you conclude about the assumption that the test scores are normally distributed?

5.7. The dataset "rats" contains the lifetime (in days) of 195 rats, each which received either an unrestricted diet (diet = 1) or an restricted diet (diet = 0). The goal of this exercise is to determine if there is evidence that the distribution of lifetimes is different for the two diets.

Following the procedure described in Example 5.8, compute the p-value for the test of the hypothesis that the density function of lifetime is the same for the two groups defined by the type of diet. Choose the smoothing parameters using the Sheather-Jones method. Based on this result, what do you conclude regarding the hypothesis?