Lec 7: Density estimation

Weiping Zhang

2018.10.15

o_p and O_p

- If $X_n \to 0$ in probability, then we write $X_n = o_p(1)$ The expression $O_p(1)$ denotes a sequence that is bounded in probability, say, write $X_n = Op(1)$: for all $\epsilon > 0$, there exists some $M > 0$ such that

$$P(|X_n| \geq M) < \epsilon$$

- More generally, for a given sequence of random variables R_n:

$$X_n = o_p(R_n) \textit{ means } X_n = Y_n R_n \textit{ and } Y_n \to 0 \textit{ in probability};$$
$$X_n = O_p(R_n) \textit{ means } X_n = Y_n R_n \textit{ and } Y_n = O_p(1)$$

- This expresses that the sequence X_n converges in probability to zero or is bounded in probability "at the rate R_n".
- Obviously, $X_n = o_p(R_n)$ implies that $X_n = O_p(R_n)$.

Results on o_p and O_p

- For some sequence a_n, if $a_n X_n \to 0$ in probability, then we write $X_n = o_p(a_n^{-1})$; if $a_n X_n = O_p(1)$, then we write $X_n = O_p(a_n^{-1})$.
- There are many rules of calculus with o and O symbols, which we will apply without comment. For instance,

$$
\begin{aligned}
&o_p(1) + o_p(1) = o_p(1), o_p(1) + O_p(1) = O_p(1),\\
&O_p(1)o_p(1) = o_p(1), (1 + o_p(1))^{-1} = O_p(1),\\
&O_p(R_n) = R_n O_p(1), o_p(R_n) = R_n o_p(1), o_p(O_p(1)) = o_p(1).
\end{aligned}
$$

- Particularly, if $X_n \rightsquigarrow F$, then $X_n = O_p(1)$, $X_n + o_p(1) \rightsquigarrow F$, $X_n \cdot o_p(1) = o_p(1)$.

Density estimation

- Let $X_1, \ldots, X_n$ be a sample from a distribution F with density f. The goal of nonparametric density estimation is to estimate f with as few assumptions about f as possible.
- Density estimation used for: regression, classification, clustering and unsupervised prediction. For example, if $\hat{f}(x, y)$ is an estimate of $f(x, y)$ then we get the following estimate of the regression function:

$$\hat{m}(x) = \hat{E}[Y|x] = \int y\hat{f}(y|x)dy$$

where $\hat{f}(y|x) = \hat{f}(x, y)/\hat{f}_X(x)$.

Bart Simpson density

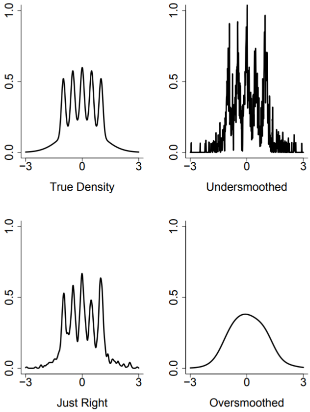
Consider the density

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

where $\phi(x; \mu, \sigma)$ denotes a Normal density with mean μ and standard deviation σ. Such density is called "the claw" or "Bart Simpson" density.

- Based on 1,000 draws from f, we computed a kernel density estimator, which depends on a tuning parameter called the bandwidth.

Density estimation



Top left: true density. The other plots are kernel estimators based on $n = 1,000$ draws. Bottom left: bandwidth $h = 0.05$ chosen by leave-one-out cross-validation. Top right: bandwidth $h/10$. Bottom right: bandwidth $10h$.

Error for Density Estimates

Our first step is to get clear on what we mean by a "good" density estimate. There are three leading ideas:

- $\int [\hat{f}(x) - f(x)]^2 dx$ should be small: the squared deviation from the true density should be small, averaging evenly over all space.
- $\int |\hat{f}(x) - f(x)| dx$ should be small: minimize the average absolute, rather than squared, deviation.
- $\int f(x) log \frac{f(x)}{\hat{f}_n(x)} dx$ should be small: the average log-likelihood ratio should be kept low.

- Option (1) is reminiscent of the MSE criterion we've used in regression.
- Option (2) looks at what's called the L1 or **total variation** distance between the true and the estimated density. It has the nice property that $\frac{1}{2}\int |f(x) - \hat{f}_n(x)|dx$ is exactly the maximum error in our estimate of the probability of *any set*. Unfortunately it's a bit tricky to work with, so we'll skip it here.
- Finally, minimizing the log-likelihood ratio is intimately connected to maximizing the likelihood. This is not a good loss function to use for nonparametric density estimation. The reason is that the Kullback-Leibler loss is completely dominated by the tails of the densities.

we will give more attention to minimizing (1), because it's mathematically tractable.

- Given the sample $X_1, \ldots, X_n$, our goal is to estimate f nonparametrically. Finding the best estimator $\hat{f}_n$ in some sense is equivalent to finding the optimal smoothing parameter h.
- Notice that the **Risk/Integrated Mean Square Error (IMSE,MISE)**:

$$\begin{aligned} R(\hat{f}_n, f) &= \int E(\hat{f}_n(x) - f(x))^2 \\ &= \int E(\hat{f}_n(x) - E\hat{f}_n(x) + E\hat{f}_n(x) - f(x))^2 \\ &= \int Var(\hat{f}_n(x))dx + \int (E\hat{f}_n(x) - f(x))^2 dx \end{aligned}$$

- One can find an optimal estimator that minimizes the risk function:

$$\hat{f}_n^*(x) = \arg\min R(\hat{f}_n, f)$$

Cross validation

- Use leave-one-out cross validation to estimate the risk function.
- One can express the loss function as a function of the smoothing parameter h

$$ISE(h) = \int (\hat{f}_n(x) - f(x))^2 dx$$
$$= \underbrace{\int (\hat{f}_n(x))^2 dx - 2\int \hat{f}_n(x) f(x) dx}_{J(h)} + \int f^2(x) dx$$

- (Least-Square) Cross-validation estimator (LSCV) of the risk function $J(h)$ (up to constant)

$$cv(h) = \int \hat{f}_n^2(x) dx - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{n,-i}(X_i)$$

where $\hat{f}_{n,-i}$ is the density estimator obtained after removing ith observation.

- *Biased Cross-Validation (BCV)* The difference between the LSCV and the biased cross-validation method is the fact that here, minimization is based on the AMISE (discussed later).
- *(Pseudo)-Likelihood Cross-Validation (LCV)* The LCV-selector was maybe the first commonly used automatic bandwidth selector because it is based on a basic statistic concept, the maximum-likelihood optimization. The criterion to maximize is

$$LCV(h) = \frac{1}{n} \prod_{i=1}^{n} \hat{f}_{n,-i}(X_i)$$

Histogram

- The oldest density estimator is the histogram.
- Without loss of generality, we assume that the support of f is [0,1]. Divide the support into m equally sized bins

$$B_1 = \left[0, \frac{1}{m}\right), B_2 = \left[\frac{1}{m}, \frac{2}{m}\right), \cdots, B_m = \left[\frac{m-1}{m}, 1\right]$$

- Let $h = \frac{1}{m}$, $p_j = \int_{B_j} f(x)dx$ and $Y_j = \sum_{j=1}^{n} I(X_i \in B_j)$
- The histogram estimator is defined by

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

where $\hat{p}_j = \frac{Y_j}{n}$.

Theorem

Suppose that f' is absolutely continuous and $\|f'(x)\|^2 < \infty$, then

$$R(\hat{f}_n, f) = \frac{h^2}{12}\|f'\|^2 + \frac{1}{nh} + o(h^2) + O(\frac{1}{n})$$

The optimal bandwidth is

$$h_{opt} = \frac{1}{n^{1/3}}\left(\frac{6}{\|f'\|^2}\right)^{1/3} = kn^{-1/3}$$

with the optimal bandwidth,

$$R(\hat{f}_n, f) \approx \frac{C}{n^{2/3}}$$

where $\|g\|^2 = \int (g(x))^2 dx$, $C = (\frac{3}{4})^{2/3}\|f'\|^{2/3}$.

Proof. For any $x, u \in B_j$,

$$f(u) = f(x) + (u - x)f'(x) + \frac{(u - x)^2}{2} f''(\tilde{x})$$

for some $\tilde{x}$ between x and u. Hence,

$$\begin{aligned} p_j &= \int_{B_j} f(u)du \\ &= \int_{B_j} \Big(f(x) + (u - x)f'(x) + \frac{(u - x)^2}{2} f''(\tilde{x})\Big)du \\ &= f(x)h + hf'(x)\Big(h(j - \frac{1}{2}) - x\Big) + O(h^3). \end{aligned}$$

Therefore, the bias of $\hat{f}_n(x)$ is

$$\begin{aligned}
b(x) &= E(\hat{f}_n(x) - f(x)) = \frac{p_j}{h} - f(x) \\
&= \frac{1}{h}\Big(f(x)h + hf'(x)\Big(h(j-1/2) - x\Big) + O(h^3)\Big) - f(x) \\
&= f'(x)\Big(h(j-1/2) - x\Big) + O(h^2)
\end{aligned}$$

By the mean value theorem, for some $\tilde{x} \in B_j$,

$$\begin{aligned}
\int_{B_j} b^2(x)dx &= \int_{B_j} (f'(x))^2\Big(h(j-1/2) - x\Big)^2 dx + O(h^4) \\
&= (f'(\tilde{x}))^2 \int_{B_j} \Big(h(j-1/2) - x\Big)^2 dx + O(h^4) \\
&= (f'(\tilde{x}))^2 \frac{h^3}{12} + O(h^4).
\end{aligned}$$

Hence,

$$
\begin{aligned}
\int_0^1 b^2(x)dx &= \sum_{j=1}^{m} \int_{B_j} b^2(x)dx \\
&= \sum_{j=1}^{m} (f'(\tilde{x}))^2 \frac{h^3}{12} + O(h^3) \\
&= \frac{h^2}{12} \|f'(x)\|^2 + o(h^2)
\end{aligned}
$$

For the variance of $\hat{f}_n$:

$$
v(x) = Var(\hat{f}_n(x)) = \frac{1}{h^2} Var(\hat{p}_j) = \frac{p_j(1-p_j)}{nh^2}
$$

By the mean value theorem, for some $x_j \in B_j$,

$$
p_j = \int_{B_j} f(x)dx = hf(x_j).
$$

Therefore,

$$
\begin{aligned}
\int_0^1 v(x)dx &= \sum_{j=1}^m \int_{B_j} v(x)dx = \sum_{j=1}^m \int_{B_j} \frac{p_j(1-p_j)}{nh^2}dx \\
&= \frac{1}{nh^2}\sum_{j=1}^m \int_{B_j} p_j dx - \frac{1}{nh^2}\sum_{j=1}^m \int_{B_j} p_j^2 dx \\
&= \frac{1}{nh} - \frac{1}{nh}\sum_{j=1}^m p_j^2 = \frac{1}{nh} - \frac{1}{nh}\sum_{j=1}^m h^2 f^2(x_j) \\
&= \frac{1}{nh} - \frac{1}{n}(\|f\|^2 + o(1)) = \frac{1}{nh} + O(\frac{1}{n}).
\end{aligned}
$$

This completes the proof.

Now, note that if we minimize the asymptotic integrated squared error,

$$AMISE(h) = \frac{h^2}{12}\|f'\|^2 + \frac{1}{nh}$$

we obtain the optimal bandwith $h_{opt} = cn^{-1/3}$.

- if $X \sim N(\mu, \sigma^2)$, then we have Scott's $c \approx 3.5\sigma$
- Freedman and Diaconis proposed a robust estimator of σ by using the interquartile range IQR, then $h^* = 2IQRn^{-1/3}$.

- the **R** *hist* command uses $h = 1/(log_2(n) + 1)$ which R calls Sturges rule and is sometimes also called Doane's Rule.
- Since the number of bars in a histogram is $k = O(h^{-1})$, we have $k = O(log_2(n) + 1)$ bars while for optimal method we have $k = O(c^{-1}n^{1/3})$.
- So the number of bars increases much faster for optimal choice. For $n < 500$ it doesn't matter much but for n larger than 500 it does matter.
- R allows the user to specify one of these alternative rules by specifying breaks = "Scott" for the rule $k = 3.5\hat{\sigma}n^{-1/3}$ or breaks = "FD" for the rule $k = 2IQRn^{-1/3}$.

Theorem

The cross-validation estimator of risk for the histogram is

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

- It turns out that if we pick h by cross-validation, then we attain this optimal rate in the large-sample limit.
- By contrast, if we knew the correct parametric form and just had to estimate the parameters, we'd typically get an error decay of $O(n^{-1})$.
- This is substantially faster than histograms, so it would be nice if we could make up some of the gap, without having to rely on parametric assumptions.

Naive density estimator

- Since

$$f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h} = \lim_{h \to 0} \frac{1}{2h} P(x-h < X \leq x+h)$$

- One could imagine estimating f by picking a small value of h and taking

$$\begin{aligned} \hat{f}_h(x) &= \frac{1}{2h}[\hat{F}_n(x+h) - \hat{F}_n(x-h)] \\ &= \frac{1}{2hn} \sum_{i=1}^{n} I(x - h < X_i \leq x + h) \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K(\frac{X_i - x}{h}) \end{aligned}$$

where $K(x) = \frac{1}{2} I(-1 < x \leq 1)$.

- This is the *naive density estimate*.

Theorem

If $h = h_n \to 0$ and $nh_n \to \infty$, as $n \to \infty$, then,for any x,

$$\hat{f}_h \to f(x) \ in \ P$$

- $\hat{f}_h$ *is a probability density function.*
- *The fact that* $n(\hat{F}_n(x+h) - \hat{F}_n(x-h)) \sim B(n, F(x+h) - F(x-h))$ *leads to*

$$E\hat{f}_h = \frac{F(x+h) - F(x-h)}{2h}$$

$$Var(\hat{f}_h) = \frac{(F(x+h) - F(x-h))(1 - F(x+h) + F(x-h))}{4nh^2}$$

- It amounts to estimating $f(x)$ by a superposition (sum) of boxcar functions centered at the observations, each with width $2h$ and area $1/n$.
- This sum is also blocky and discontinuous, but it avoids one of the arbitrary choices in constructing a histogram: the choice of locations of the bins.
- As $h \to 0$, the naive estimate converges weakly to the sum of point masses at the data; for $h > 0$, the naive estimator smooths the data.
- The tuning parameter h is analogous to the bin width in a histogram. Larger values of h give smoother density estimates. Whether "smoother" means "better" depends on the true density f; generally, there is a tradeoff between bias and variance: increasing the smoothness increases the bias but decreases the variance.

Kernel estimation

- Obviously, whenever $K(x)$ is itself a probability density function, then $\hat{f}_K$ is a probability density function.
- Using a smoother kernel function K, such as a Gaussian density, leads to a smoother estimate $\hat{f}_K$.
- Estimates that are linear combinations of such kernel functions centered at the data are called **kernel density estimates**. We denote the kernel density estimate with bandwidth (smoothing parameter) h by

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{X_i - x}{h}).$$

Is $\hat{f}_h(x)$ a legitimate density function? It needs to satisfy:
(1) nonnegative
(2) integrate to one
Easy to do: Require the Kernel function, $K(\cdot)$ to satisfy:

- $K(u) \geq 0$ for all u
- $\int K(u)du = 1$

Additionally, the kernel K is also assumed to satisfy

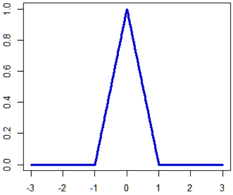
$$K(u) = K(-u), \int uK(u)du = 0$$

$$0 < \kappa_{21} = \int u^2 K(u)du < \infty, \kappa_{02} = \|K\|^2 = \int K^2(u)du < \infty$$

where $\kappa_{ij} = \int u^i K^j(u)du$.

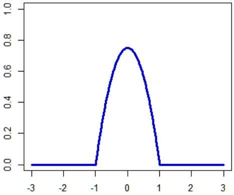
Popular kernels

Triangle



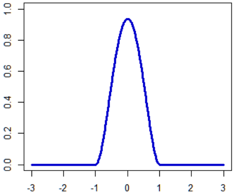
$$K(u) = (1 - |u|)I(|u| \leq 1)$$

Epanechnikov



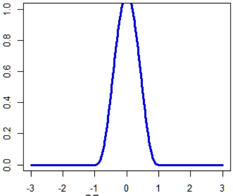
$$K(u) = \frac{3}{4}(1 - u^2)I(|u| \leq 1)$$

Quartic (biweight)



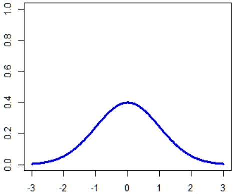
$$K(u) = \frac{15}{16}(1 - u^2)^2 I(|u| \leq 1)$$

Triweight



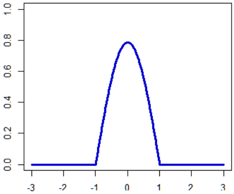
$$K(u) = \frac{35}{32}(1 - u^2)^3 I(|u| \leq 1)$$

Gaussian



$$K(u) = (2\pi)^{-1/2} exp(-\frac{1}{2}u^2)$$

Cosin



$$K(u) = \frac{\pi}{4} cos(\frac{\pi u}{2})I(|u| \leq 1)$$

Properties of KDE

To see the performance of the estimator, consider the bias and the mean square error of $\hat{f}_h(x)$ for fixed x.

Theorem

Let f be twice continuously differentiable in a neighborhood of x. Let the kernel K satisfy the above assumptions. If $\lim_{n\to\infty} h = 0$, then,

$$E(\hat{f}_h(x) - f(x)) = \frac{1}{2} h_n^2 f''(x)\kappa_{21} + o(h^2)$$

If in addition, $\lim_{n\to\infty} nh = \infty$, then

$$Var(\hat{f}_h(x)) = \frac{1}{nh} f(x)\kappa_{02} + o(\frac{1}{nh})$$

Properties of KDE

Thus,

$$
\begin{aligned}
MSE(\hat{f}_h(x)) &= E(\hat{f}_h(x) - f(x))^2 \\
&= \underbrace{\frac{1}{4} h_n^4 (f''(x))^2 \kappa_{21}^2 + \frac{1}{nh} f(x) \kappa_{02}}_{AMSE} + o(h^4 + \frac{1}{nh})
\end{aligned}
$$

Proof. By Taylor expansion of $f(x+uh)$ at x:

$$f(x+uh) = f(x) + f'(x)uh + \frac{1}{2}f''(x)(uh)^2 + o((uh)^2)$$

Therefore,

$$\begin{aligned} E\hat{f}_h(x) &= E\Big[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\frac{X_i - x}{h})\Big] = \frac{1}{h}EK(\frac{X_1 - x}{h}) \\ &= \int K(u)f(x+uh)du \\ &= \int K(u)[f(x) + f'(x)uh + \frac{1}{2}f''(x)u^2h^2 + u^2o(h^2)]du \\ &= f(x) + \frac{1}{2}f''(x)\kappa_{21}h^2 + o(h^2) \end{aligned}$$

and

$$
\begin{aligned}
Var(\hat{f}_h(x)) &= Var\Big[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\frac{X_i - x}{h})\Big] = \frac{1}{n}Var\Big(\frac{1}{h}K(\frac{X_1 - x}{h})\Big) \\
&= \frac{1}{n}E\Big[\frac{1}{h}K(\frac{X_1 - x}{h})\Big]^2 - \frac{1}{n}\Big[E\frac{1}{h}K(\frac{X_1 - x}{h})\Big]^2 \\
&= \frac{1}{nh}\int K^2(u)f(x+uh)du - \frac{1}{n}\Big(\int K(u)f(x+uh)du\Big)^2 \\
&= \frac{1}{nh}\int K^2(u)f(x+uh)du - \frac{1}{n}\Big(f(x) + \frac{1}{2}f''(x)\kappa_{21}h^2 + o(h^2)\Big)^2 \\
&= \frac{f(x)}{nh}\kappa_{02} + o(\frac{1}{nh})
\end{aligned}
$$

This completes the proof.

Optimal bandwidth

Observe that as h increases, the bias becomes large while the variance decreases. In order to find the optimal value of h, we minimize the AMSE. This leads to:

$$h_{opt1}(x) = \Big(\frac{f(x)\kappa_{02}}{(f''(x))^2\kappa_{21}^2}\Big)^{1/5} n^{-1/5}$$

It follows that the corresponding AMSE and variance are both of the order $n^{-4/5}$.

Global behavior $\hat{f}_h$

Observe that

$$MISE(\hat{f}_h) = \int MSE(\hat{f}_h(x))dx = \int E(\hat{f}_h(x) - f(x))^2 dx$$

It can be shown

$$MISE(\hat{f}_h) = \underbrace{\frac{\kappa_{02}}{nh} + \frac{1}{4}\|f''\|^2 \kappa_{21}^2 h^4}_{AMISE} + o(h^4 + \frac{1}{nh})$$

Thus $MISE(\hat{f}_h) \to 0$ and further

$$ISE_h(\hat{f}_h) = \int (\hat{f}_h(x) - f(x))^2 dx \to 0.$$

Global optimal bandwidth

Minimizing the AMISE leads to the following optimal bandwidth,

$$h_{opt2} = \left(\frac{\kappa_{02}}{\|f''\|^2 \kappa_{21}^2}\right)^{1/5} n^{-1/5}.$$

The resulting MISE is of the order $n^{-4/5}$.

- Both locally and globaly, the optimal bandwidth is of the order $n^{-1/5}$, and the convergence rate is $n^{-4/5}$.
- Bandwidth plays a more important role than the kernel. The choice of kernel does not effect the order of bandwidth or the rate of mean square convergence. Any kernel from a large class satisfying the assumptions can be used.

Practical bandwidth choices

The theoretically optimal bandwidth, h_{opt2}, depends on the unknown density f through $\|f''\|^2$. The actual choice of h is a critical issue. There are different approaches to choose h in practice. Write $h_{opt2} = n^{-1/5}\frac{C(K)}{\|f''\|^{2/5}}$, where $C(K)$ is the constant depending only on K.

- **Rule of thumb** Choose an auxiliary parametric family, say normal distributions, to choose h, not to estimate f.
  - We plug in the density of $N(0, \sigma^2)$ into the formula of h_{opt2}, then

    $$h_{opt} \approx 1.06\hat{\sigma}n^{-1/5}$$

    where $\hat{\sigma}$ is the sample standard deviation.
  - It is recommended to estimate σ with $min(\hat{\sigma}, R/1.35)$, where $\hat{\sigma}$ is the sample standard deviation and R is the sample interquantile range, that is $R = \hat{F}_n^{-1}(0.75) - \hat{F}_n^{-1}(0.25)$ $(\Phi^{-1}(0.75) - \Phi^{-1}(0.25) = 1.35)$.

$$h_{opt} = 1.06min\{\hat{\sigma}, \frac{R}{1.35}\}n^{-1/5}$$

Practical bandwidth choices

- **Cross-validation** Cross-validation score function:

$$cv(h) = \int \hat{f}_h^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{h,-i}(X_i)$$

Since the first term

$$\begin{aligned}
\int \hat{f}_h^2(x)dx &= \int \frac{1}{nh}\sum_{i=1}^{n} K(\frac{X_i - x}{h})\frac{1}{nh}\sum_{j=1}^{n} K(\frac{X_j - x}{h})dx \\
&= \frac{1}{n^2h^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int K(\frac{X_i - x}{h})K(\frac{X_j - x}{h})dx \\
&= \frac{1}{n^2h}\sum_{i=1}^{n}\sum_{j=1}^{n}\int K(u)K(u - \frac{X_i - X_j}{h})dx \\
&= \frac{1}{n^2h}\sum_{i=1}^{n}\sum_{j=1}^{n} K * K(\frac{X_i - X_j}{h})
\end{aligned}$$

where $K * K(v) = \int K(u)K(v-u)du$ is the convolution of kernel K.

For the second term,

$$\frac{2}{n}\sum_{i=1}^{n}\hat{f}_{h,-i}(X_i) = \frac{2}{n(n-1)h}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}K(\frac{X_i - X_j}{h})$$

Therefore, the optimal h is $\hat{h} = \arg\min_h cv(h)$.

Theorem (Stone's Theorem)

Suppose f is bounded. Let $\hat{f}_n$ denote the kernel estimator with bandwidth h and let h_* denote the bandwidth chosen by cross-validation. Then

$$\frac{ISE_{h_*}(\hat{f}_{h_*})}{\inf_h ISE_h(\hat{f}_h)} \to 1, a.s.$$

- **Biased cross-validation**. This was proposed by Scott and George (1987), which has as its immediate target the AMISE. They proposed to estimate $R(f'') = \|f''\|^2$ by

$$\hat{R}(f'') = \|\hat{f}_h''\|^2 - \frac{\|K''\|^2}{nh^5}$$

The biased cross-validation for bandwidth choice is

$$BCV(h) = \frac{\|K\|^2}{nh} + \frac{\kappa_{21}^2}{4n(n-1)h} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} K'' * K''(\frac{X_j - X_i}{h})$$

- There is another version of BCV by Jones and Kappenman (1991).
- Other variants include Maximum likelihood cross-validation, Complete cross-validation, Modified cross-validation, Trimmed cross-validation. (See R package **kedd**)

Choosing the Kernel

- To discuss the choice of the kernel we will consider equivalent kernels, i.e. kernel functions that lead to exactly the same kernel density estimator.
- Consider a kernel function $K(\cdot)$ and the following modification:

$$K_\delta(\cdot) = \frac{1}{\delta} K(\frac{\cdot}{\delta})$$

- If $h = \delta \tilde{h}$, then the following two KDEs are equivalent:

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(X_i - x) = \frac{1}{n} \sum_{i=1}^{n} K_{\tilde{h}\delta}(X_i - x) = \tilde{f}_{\tilde{h}\delta}(x)$$

This means, all rescaled versions K_δ of a kernel function K are equivalent if the bandwidth is adjusted accordingly.

- Different values of δ correspond to different members of an equivalence class of kernels.
- Recall the AMISE criterion, i.e.

$$AMISE = \frac{\|K\|^2}{nh} + \frac{h^4}{4}\|f''\|^2\mu_2^2(K),$$

  where $\mu_2(K) = \kappa_{21}$.
- We rewrite this formula for some equivalence class of kernel functions K_δ:

$$AMISE(K_\delta) = \frac{\|K_\delta\|^2}{nh} + \frac{h^4}{4}\|f''\|^2\mu_2^2(K_\delta)$$

- In each of the two components of this sum there is a term involving K_δ. The idea for separating the problems of choosing h and K is to find δ such that

$$\|K_\delta\|^2 = \mu_2^2(K_\delta)$$

This is fulfilled if

$$\delta_0 = \Big(\frac{\|K\|^2}{\mu_2^2(K)}\Big)^{1/5} = \Big(\frac{\kappa_{02}}{\kappa_{21}^2}\Big)^{1/5}$$

The value δ_0 is called the *canonical bandwidth* corresponding to the kernel function K.

- Let $T(K) = \kappa_{02}/\delta_0$, then

$$AMISE(h) = \Big[\frac{1}{nh} + \frac{h^4}{n}\|f''\|^2\Big]T(K)$$

- This has an interesting implication: Even though T(K) is not the same for different kernels, it does not matter for the asymptotic behavior of AMISE (since it is just a multiplicative constant).
- Hence, AMISE will be asymptotically equal for different equivalence classes if we use K_{δ_0} to represent each class, and call it *canonical kernel* of an equivalent class.

Kernel			δ_0
Uniform	$\left(\frac{9}{2}\right)^{1/5}$	$\approx$	1.3510
Epanechnikov	$15^{1/5}$	$\approx$	1.7188
Quartic	$35^{1/5}$	$\approx$	2.0362
Triweight	$\left(\frac{9450}{143}\right)^{1/5}$	$\approx$	2.3122
Gaussian	$\left(\frac{1}{4\pi}\right)^{1/10}$	$\approx$	0.7764

- Suppose now that we have estimated an unknown density f using some kernel K^A and bandwidth h_A, what bandwidth h_B should we use in the estimation with kernel K^B when we want to get approximately the same degree of smoothness as we had in the case of K^A and h_A?
- The answer is given by the following formula:

$$h_B = h_A \frac{\delta_0^B}{\delta_0^A}$$

- A question of immediate interest is to find the kernel that minimizes $T(K)$, Epanechnikov (1969, the person, not the kernel) has shown that under all nonnegative kernels with compact support, the kernel of the form

$$K(u) = \frac{3}{4} \frac{1}{15^{1/5}} \Big(1 - (\frac{u}{15^{1/5}})^2\Big) I(|u| \leq 15^{1/5})$$

minimizes the function $T(K)$.

- Compare the values of $T(K)$ of other kernels with the value of $T(K)$ for the Epanechnikov kernel:

Kernel	$T(K)$	$T(K)/T(K_{Epa})$
Uniform	0.3701	1.0602
Triangle	0.3531	1.0114
Epanechnikov	0.3491	1.0000
Quartic	0.3507	1.0049
Triweight	0.3699	1.0595
Gaussian	0.3633	1.0408
Cosine	0.3494	1.0004

- After all, we can conclude that for practical purposes the choice of the kernel function is almost irrelevant for the efficiency of the estimate.

- Suppose that f'' exists and $h = cn^{-1/5}$. Then

$$n^{2/5}(\hat{f}_h - f(x)) \rightsquigarrow N\Big(\underbrace{\frac{c^2}{2} f''(x)\kappa_{21}}_{b_x}, \underbrace{\frac{1}{c} f(x)\kappa_{02}}_{v_x^2}\Big)$$

- We then get

$$\begin{aligned} 1-\alpha &\approx P(b_x - z_{1-\alpha/2}v_x \le n^{2/5}(\hat{f}_h(x) - f(x)) \le b_x + z_{1-\alpha/2}v_x) \\ &= P(\hat{f}_h(x) - n^{-2/5}(b_x + z_{1-\alpha/2}v_x) \\ &\le f(x) \le \hat{f}_h(x) + n^{-2/5}(b_x - z_{1-\alpha/2}v_x)) \end{aligned}$$

- Using $h = cn^{-1/5}$ we get the asymptotic confidence interval for $f(x)$:

$$\Big[\hat{f}_h(x) - \frac{h^2}{2} f''(x)\kappa_{21} - z_{1-\alpha/2}\sqrt{\frac{f(x)\kappa_{02}}{nh}}),$$
$$\hat{f}_h(x) - \frac{h^2}{2} f''(x)\kappa_{21} + z_{1-\alpha/2}\sqrt{\frac{f(x)\kappa_{02}}{nh}})\Big]$$

- Unfortunately, the interval boundaries still depend on $f(x)$ and $f''(x)$. If h is small relative to $n^{-1/5}$ we can neglect the second term of each boundary. Replacing $f(x)$ with $\hat{f}_h(x)$ gives an approximate confidence interval that is applicable in practice.

$$\Big[\hat{f}_h(x) - z_{1-\alpha/2}\sqrt{\frac{\hat{f}_h(x)\kappa_{02}}{nh}}),$$
$$\hat{f}_h(x) + z_{1-\alpha/2}\sqrt{\frac{\hat{f}f_n(x)\kappa_{02}}{nh}})\Big]$$

- Confidence bands for f have only been derived under some rather restrictive assumptions.

- Suppose that f is a density on $[0,1]$ and given that certain regularity conditions are satisfied, then for $h = n^{-\delta}, \delta \in (1/5, 1/2)$, and for all $x \in [0,1]$ the following formula has been derived by Bickel & Rosenblatt (1973)

$$\lim_{n\to\infty} P\Big(\hat{f}_h(x) - \Big\{\frac{\hat{f}_h(x)\kappa_{02}}{nh}\Big\}^{1/2}\Big\{\frac{z}{(2\delta logn)^{1/2}} + d_n\Big\}^{1/2} \le f(x)$$

$$\le \hat{f}_h(x) + \Big\{\frac{\hat{f}_h(x)\kappa_{02}}{nh}\Big\}^{1/2}\Big\{\frac{z}{(2\delta logn)^{1/2}} + d_n\Big\}^{1/2}\Big)$$

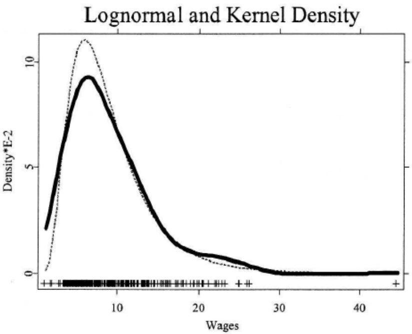
$$= exp\{-2exp(-z)\},$$

where $d_n = (2\delta logn)^{1/2} + (2\delta logn)^{-1/2} log\Big(\frac{1}{2\pi}\frac{\|K\ '\|}{\|K\|}\Big)$.

- A confidence band for a given significance level α can be found by searching the value of z that satisfies $exp\{-2exp(-z)\} = 1 - \alpha$. If $\alpha = 0.05$, then $z \approx 3.663$.

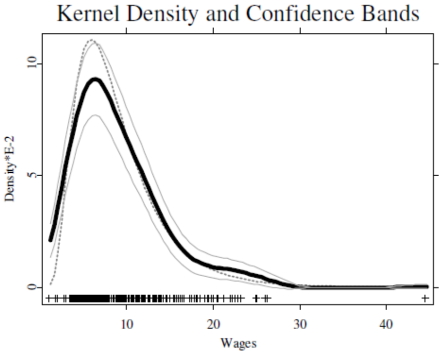
What can we use the confidence intervals for $\hat{f}_h(x)$ in practice?

In the following example we check if a parametric estimate can describe the data



Sample of 534 randomly selected U.S workers average hourly earnings from May 1985. (Nonpar.: thick solid line, Lognormal: thin line)

Compute the 95% confidence bands around the nonpar. estimate



Lognormal density exceeds the upper limit of the confidence band in the mode- reject the lognormal distribution as "true", even though the lognormal distribution fit the shape quite well.
Checking whether the parametric density do not exceed the conf. bands is a very conservative test-not the best way to check it.