KERNEL DENSITY ESTIMATION (See chapter 1 of Tsybakov 2008)

Let $X_1, \dots, X_n$ be a r.s. with cdf F.

If the cdf F has a continuous derivative F', then the corresponding pdf is $f = F'$. In this case

$$f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h} \quad \text{for all } x \in \mathbb{R}.$$

This suggests the estimator of $f(x)$ given by

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

for some small value of h, where $\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(X_i \le x)$.

We may rewrite $\hat{f}_n(x)$ as

$$\begin{aligned}
\hat{f}_n(x) &= \frac{1}{nh} \sum_{i=1}^n \frac{1}{2}\left[\mathbb{1}(X_i \le x+h) - \mathbb{1}(X_i \le x-h)\right] \\
&= \frac{1}{nh} \sum_{i=1}^n \frac{1}{2} \mathbb{1}(x-h < X_i \le x+h) \\
&= \frac{1}{nh} \sum_{i=1}^n \frac{1}{2} \mathbb{1}\left(-1 < \frac{X_i - x}{h} \le 1\right) \\
&= \frac{1}{nh} \sum_{i=1}^n K_0\left(\frac{X_i - x}{h}\right),
\end{aligned}$$

where $K_0(u) = \frac{1}{2} \mathbb{1}(-1 < u \le 1)$.

This is called the Rosenblatt estimator of f.

Exercise: Check whether $\hat{f}_n$ is a legitimate pdf.

Solution: We have $\hat{f}_n(x) \geq 0 \;\forall\, x \in \mathbb{R}$, since $K_0(u) \geq 0 \;\forall\, u \in \mathbb{R}$.

We have

$$\int_{\mathbb{R}} \hat{f}_n(x)\, dx = \int_{\mathbb{R}} \frac{1}{nh} \sum_{i=1}^{n} K_0\left(\frac{X_i - x}{h}\right) dx$$

$u = \frac{X_i - x}{h}$ $\quad dx = -h\,du$

$x = X_i - hu$

$$= -\frac{1}{n} \sum_{i=1}^{n} \int_{\infty}^{-\infty} K_0(u)\, du$$

$$= \frac{1}{n} \sum_{i=1}^{n} \underbrace{\int_{-\infty}^{\infty} \frac{1}{2} \mathbb{1}(-1 < u \leq 1)\, du}_{=1}$$

$$= 1.$$

We can generalize the Rosenblatt estimator by choosing other K:

$$K(u) = (1 - |u|)\, \mathbb{1}(|u| \leq 1) \quad [\text{Triangular}]$$

$$K(u) = \frac{3}{4}(1 - u^2)\, \mathbb{1}(|u| \leq 1) \quad [\text{Epanechnikov}]$$

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \quad [\text{Gaussian}]$$

Then we have the so-called Kernel density estimator (KDE)

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \quad \text{for } x \in \mathbb{R}.$$

The function K is called the Kernel and h the bandwidth.

For $\hat{f}_n$ to be a legitimate density, we require

(K0) $K(u) \geq 0 \quad \forall \; u \in \mathbb{R}$.

(K1*) $\int_{\mathbb{R}} K(u)\, du = 1$.

<u>Exercise</u>: Generate data and plot KDEs along with true density. Try different kernels and bandwidths.

<u>Mean squared error of KDE at a point x_0 (upper bound):</u>

<u>Variance</u>: Let $\sigma_n^2(x_0) = \text{Var}\, \hat{f}_n(x_0)$ for $x_0 \in \mathbb{R}$.

Make the assumptions

(F1) $\exists \; f_{max} > 0$ s.t. $f(x) \leq f_{max} \quad \forall \; x \in \mathbb{R}$.

(K2) $\int_{\mathbb{R}} K^2(u)\, du \leq \kappa^2 < \infty$.

<u>Result</u>: Under (F1) and (K2) we have

$$\sigma_n^2(x_0) \leq \frac{1}{nh} f_{max} \cdot \kappa^2$$

for each $x_0 \in \mathbb{R}$.

Proof: For a point $x_0 \in \mathbb{R}$ we have

$$\begin{aligned}
\mathrm{Var}\left[K\left(\frac{X_1 - x_0}{h}\right)\right] &= \mathbb{E}\left[K^2\left(\frac{X_1 - x_0}{h}\right)\right] - \left[\mathbb{E}\, K\left(\frac{X_1 - x_0}{h}\right)\right]^2 \\
&\leq \mathbb{E}\left[K^2\left(\frac{X_1 - x_0}{h}\right)\right] \\
&= \int_{\mathbb{R}} K^2\left(\frac{x - x_0}{h}\right) f(x)\, dx \\
&\leq f_{max} \int_{\mathbb{R}} K^2\left(\frac{x - x_0}{h}\right) dx \\
&= f_{max} \cdot h \cdot \int_{\mathbb{R}} K^2(u)\, du \\
&\leq h \cdot f_{max} \cdot \kappa^2 .
\end{aligned}$$

Now we have

$$\begin{aligned}
\mathrm{Var}\, \hat{f}_n(x_0) &= \frac{1}{n^2 h^2} \sum_{i=1}^{n} \mathrm{Var}\left[K\left(\frac{X_i - x_0}{h}\right)\right] \\
&\leq \frac{1}{n^2 h^2} \sum_{i=1}^{n} h \cdot f_{max} \cdot \kappa^2 \\
&= \frac{1}{nh} \cdot f_{max} \cdot \kappa^2 .
\end{aligned}$$

$\square$

Bias: Let $b_n(x_0) = \mathbb{E}\, \hat{f}_n(x_0) - f(x_0)$ for some $x_0 \in \mathbb{R}$.

Trickier than the variance, must consider wiggliness of f.

Defn: Let T be an interval in $\mathbb{R}$ and $L > 0$. The Lipschitz class of functions Lipschitz(L) on T is the set of functions $f: T \to \mathbb{R}$ satisfying

$$|f(x) - f(x')| \leq L\,|x - x'| \quad \forall \; x, x' \in T.$$

Let the set of pdfs that belong to Lipschitz (L) on $\mathbb{R}$ be denoted by

$$\mathcal{P}_{\mathcal{L}}(L) = \left\{ f : f \geq 0, \int_{\mathbb{R}} f(x)\,dx = 1, \; f \in \text{Lipschitz}(L) \text{ on } \mathbb{R} \right\}$$

We require another assumption on K, which we will later expand on:

$$(K3^*) \quad \int_{\mathbb{R}} |u|\,|K(u)|\,du \leq K_1 < \infty$$

Result: Under $(K1^*)$ and $(K3^*)$, and if $f \in \mathcal{P}_{\mathcal{L}}(L)$, then

$$|b(x_0)| = h \cdot L \cdot K_1$$

for each $x_0 \in \mathbb{R}$.

Proof: We have

$$\mathbb{E}\hat{f}_n(x_0) - f(x_0) = \frac{1}{nh} \sum_{i=1}^{n} \left[\mathbb{E}\, K\left(\frac{X_i - x_0}{h}\right) \right] - f(x_0),$$

$$= \frac{1}{h} \mathbb{E}\, K\left(\frac{X_1 - x_0}{h}\right) - f(x_0)$$

$$= \frac{1}{h} \int_{\mathbb{R}} K\left(\frac{x - x_0}{h}\right) f(x)\,dx - f(x_0)$$

$$u = \frac{x - x_0}{h}, \quad x = x_0 + uh, \quad dx = h\,du$$

$$= \int_{\mathbb{R}} K(u)\, f(x_0 + uh)\, du - f(x_0)$$

by $(K1^*)$

$$= \int_{\mathbb{R}} K(u) \left[f(x_0 + uh) - f(x_0) \right] du .$$

So we have

$$|b_n(x_0)| \le \int_{\mathbb{R}} K(u) \left| f(x_0 + uh) - f(x_0) \right| du$$

by $(F2^*)$

$$\le \int_{\mathbb{R}} |K(u)|\, L\, |uh|\, du$$

$$= L \cdot h \int_{\mathbb{R}} |u| \, |K(u)|\, du$$

by $(K3^*)$

$$\le L \cdot h \cdot K_1 . \qquad \square$$

Putting the bias and variance together we get the following result:

Result: Under $(K1^*)$, $(K2)$, $(K3^*)$, $(F1)$, and if $f \in \mathcal{P}_{\mathcal{L}}(L)$, then

$$MSE\ \hat{f}_n(x_0) \le h^2 \cdot L^2 \cdot K_1^2 + \frac{1}{nh} \cdot f_{max} \cdot K^2 .$$

for each point $x_0 \in \mathbb{R}$.

Optimal h: To minimize $MSE\ \hat{f}_n(x_0)$ in the setting of the previous result, we would choose

$$h_{opt} = n^{-1/3}\, C_1 \qquad \left(f \in \mathcal{P}_{\mathcal{L}}(L) \right)$$

where C_1 depends on unknown constants.

Why??

$$\frac{\partial}{\partial h}\left(\text{bound on MSE } \hat{f}_n(x_0)\right) = 2h \cdot L^2 \cdot K_1^2 - \frac{1}{nh^2} \cdot f_{max} \cdot K^2 = 0$$

$$\Longleftrightarrow \quad 2h^3 \cdot L^2 \cdot K_1^2 = \frac{f_{max} K^2}{n}$$

$$\Longleftrightarrow \quad h = n^{-1/3} \underbrace{\left(\frac{f_{max} \cdot K^2}{L^2 \cdot K_1^2 \cdot 2}\right)^{1/3}}_{C_1}$$

Optimal order of MSE $\hat{f}_n(x_0)$: Plugging in $h_{opt} = n^{-1/3} C_1$ gives

$$\text{MSE } \hat{f}_n(x) \leq n^{-2/3} C_1^* . \quad \left(f \in \mathcal{P}_{\mathcal{L}}(L)\right)$$

We now introduce a more general class of functions than Lipschitz.

Defn: Let T be an interval in $\mathbb{R}$, β a positive integer, and $L > 0$. The Hölder class $\mathcal{H}(\beta, L)$ on T is the set of $\ell = \beta - 1$ times differentiable functions $f: T \to \mathbb{R}$ of which the derivative $f^{(\ell)}$ satisfies

$$\left| f^{(\ell)}(x) - f^{(\ell)}(x') \right| \leq L \left| x - x' \right| \qquad \forall \; x, x' \in T .$$

* Can define Hölder classes for non-integer β, but we ignore these.

For $\beta = 2$, the condition is $\left| f^{(1)}(x) - f^{(1)}(x') \right| \leq L \left| x - x' \right|$.

Exercise: Determine if $f \in \mathcal{H}(\beta, L)$ for some β, L or if $f \in \text{Lipschitz}(L)$.

(a) $f(x) = x^2$ (b) $f(x) = 1/x$ (c) $f(x) = \log x$

(d) $f(x) = e^x$ (e) $f(x) = e^{-x} \mathbb{1}(x > 0)$

Let the set of all pdfs belonging to $\mathcal{H}(\beta, L)$ on $\mathbb{R}$ be represented by

$$\mathcal{P}_{\mathcal{H}}(\beta, L) = \left\{ f : f \geq 1, \int_{\mathbb{R}} f(x)\,dx, \quad f \in \mathcal{H}(\beta, L) \text{ on } \mathbb{R} \right\}.$$

To consider estimating $f \in \mathcal{P}_{\mathcal{H}}(\beta, L)$, we must expand $(K1^*)$. We will need a new definition:

<u>Defn:</u> Let $\ell \geq 1$ be an integer. We call $K : \mathbb{R} \to \mathbb{R}$ a kernel of order ℓ if the functions $u \mapsto u^j K(u)$, $j = 0, 1, \dots, \ell$ are integrable and satisfy

$$\int_{\mathbb{R}} K(u)\,du = 1, \qquad \int_{\mathbb{R}} u^j K(u)\,du = 0, \quad j = 1, \dots, \ell.$$

In the following we will use the assumptions

(K1) K is a kernel of order ℓ

(K3) $\displaystyle\int_{\mathbb{R}} |u|^\beta |K(u)|\,du \leq \kappa_\beta < \infty$

with ℓ and β taken from the relevant Hölder class.

We now revisit the bias:

<u>Result:</u> Under (K1) and (K3) and if $f \in \mathcal{P}_{\mathcal{H}}(\beta, L)$ on $\mathbb{R}$, we have

$$|b_n(x_0)| \leq h^\beta \cdot \frac{L \cdot \kappa_\beta}{\ell!}$$

for each $x_0 \in \mathbb{R}$.

<u>Proof:</u> Following our previous work, we can write

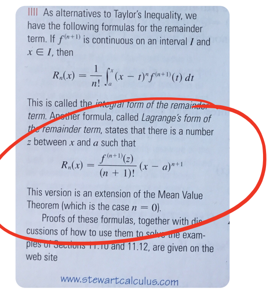
$$\mathbb{E}\hat{f}_n(x_0) - f(x_0) = \frac{1}{nh} \sum_{i=1}^{n} \left[\mathbb{E}\, K\left(\frac{X_i - x_0}{h}\right) \right] - f(x_0),$$

$$\vdots$$

$$= \int_{\mathbb{R}} K(u) \left[f(x_0 + uh) - f(x_0) \right] du.$$

Now we write $f(x_0 + uh)$ via Taylor expansion as

$$f(x_0 + uh) = f(x_0) + \sum_{j=1}^{\ell-1} \frac{f^{(j)}(x_0)(uh)^j}{j!} + \frac{f^{(\ell)}(x_0 + \tau u h)\,(uh)^\ell}{\ell!}$$


|||| As alternatives to Taylor's Inequality, we have the following formulas for the remainder term. If $f^{(n+1)}$ is continuous on an interval I and $x \in I$, then

$$R_n(x) = \frac{1}{n!}\int_a^x (x-t)^n f^{(n+1)}(t)\,dt$$

This is called the *integral form of the remainder term.* Another formula, called *Lagrange's form of the remainder term,* states that there is a number z between x and a such that

$$R_n(x) = \frac{f^{(n+1)}(z)}{(n+1)!}(x-a)^{n+1}$$

This version is an extension of the Mean Value Theorem (which is the case $n = 0$).

Proofs of these formulas, together with discussions of how to use them to solve the examples of Sections 11.10 and 11.12, are given on the web site

www.stewartcalculus.com

for some $0 \leq \tau \leq 1$. Note that

$$\int_{\mathbb{R}} K(u) \sum_{j=1}^{\ell-1} f^{(j)}(x_0)(uh)^j \, du = 0,$$

since K, by (K1) is a kernel of order ℓ. This gives

$$b_n(x_0) = \int_{\mathbb{R}} K(u) \frac{f^{(\ell)}(x_0 + \tau u h)\,(uh)^\ell}{\ell!} \, du$$

by (K1)

$$= \frac{h^\ell}{\ell!} \int_{\mathbb{R}} u^\ell K(u) \left[f^{(\ell)}(x_0 + \tau u h) - f^{(\ell)}(x_0) \right] du .$$

So we have

$$|b_n(x_0)| \leq \frac{h^\ell}{\ell!} \int_{\mathbb{R}} |u|^\ell K(u) \left| f^{(\ell)}(x_0 + \tau u h) - f^{(\ell)}(x_0) \right| du$$

$$\leq \frac{h^\ell}{\ell!} \int_{\mathbb{R}} |u|^\ell |K(u)| \, L \, |\tau u h| \, du$$

$(\beta = \ell + 1)$

$$\leq \frac{h^\beta}{\ell!} \cdot L \int_{\mathbb{R}} |u|^\beta |K(u)| \, du$$

$$\leq \frac{h^\beta}{\ell!} \cdot L \cdot K_\beta . \qquad \square$$

The variance depends only on f_{max} and K^2 from (F1), (K2).

Result: Under (K1), (K2), (K3), (F1), and if $f \in \mathcal{P}_{H}(\beta, L)$, then

$$\text{MSE}\ \hat{f}_n(x_0) \le h^{2\beta}\left(\frac{L \cdot K_\beta}{\ell!}\right)^2 + \frac{1}{nh} f_{max} \cdot K^2$$

for each point $x_0 \in \mathbb{R}$.

Optimal h: To minimize $\text{MSE}\ \hat{f}_n(x_0)$ in the setting of the previous result, we would choose

$$h_{opt} = n^{-\frac{1}{2\beta+1}} C_2, \qquad \left(f \in \mathcal{P}_{H}(\beta, L)\right)$$

where C_2 depends on unknown constants.

Why??

$$\frac{\partial}{\partial h}\left(\text{Bound on MSE}\ \hat{f}_n(x_0)\right) = 2\beta h^{2\beta-1}\left(\frac{L \cdot K_\beta}{\ell!}\right)^2 - \frac{1}{nh^2} f_{max} \cdot K^2 = 0$$

$$\Leftrightarrow \quad h^{2\beta+1} = \frac{1}{n} \frac{f_{max} \cdot K^2}{2\beta\,(L \cdot K_\beta / \ell!)^2}$$

$$\Leftrightarrow \quad h = n^{-\frac{1}{2\beta+1}} \underbrace{\left[\frac{f_{max} \cdot K^2}{2\beta\,(L \cdot K_\beta / \ell!)^2}\right]^{\frac{1}{2\beta+1}}}_{C_2}.$$

Optimal order of $\text{MSE}\ \hat{f}_n(x_0)$: Plugging h_{opt} into the MSE bound gives

$$\text{MSE}\ \hat{f}_n(x_0) \le n^{-\frac{2\beta}{2\beta+1}} C_2^*. \qquad \left(f \in \mathcal{P}_{H}(\beta, L)\right)$$

Exercise: Suppose f is $2\times$ differentiable with a bounded second derivative. What is h_{opt} and the resulting bound on MSE $\hat{f}_n(x_0)$?

Solution: We have $f \in \mathcal{P}_H(2, L)$ for some L, so

$$h_{opt} \asymp n^{-\frac{1}{5}} \qquad \text{and} \qquad \text{MSE}\, \hat{f}_n(x_0) \asymp n^{-\frac{4}{5}}$$

We now present a summarizing theorem for this section

Theorem 1: Under (K1), (K2), and (K3), if $h = \alpha n^{-\frac{1}{2\beta+1}}$ for some $\alpha \geqslant 0$, then for all $n \geqslant 1$, we have

$$\sup_{x_0 \in \mathbb{R}} \ \sup_{f \in \mathcal{P}_H(\beta, L)} \mathbb{E}_f\left[\left(\hat{f}_n(x_0) - f(x_0)\right)^2\right] \leq C\, n^{-\frac{2\beta}{2\beta+1}}$$

where $C > 0$ depends only on β, L, α, and the kernel $K(\cdot)$.

This bounds MSE $\hat{f}_n(x_0)$ uniformly over the class of densities $\mathcal{P}_H(\beta, L)$.

Remark: Where did assumption (F1) go? We can show that $\exists\, f_{max}$ s.t.

$$\sup_{x_0 \in \mathbb{R}} \ \sup_{f \in \mathcal{P}_H(\beta, L)} f(x_0) \leq f_{max} < \infty, \qquad \text{[see Tsyb. pg 9]}$$

so (F1) is redundant if $f \in \mathcal{P}_H(\beta, L)$.

Mean integrated squared error of KDE (upper bound):

Instead of looking at a single point x_0, look at entire $\mathbb{R}$.

Defn: The mean integrated squared error (MISE) of $\hat{f}_n$ is defined as

$$\text{MISE } \hat{f}_n = \mathbb{E} \int_{\mathbb{R}} [\hat{f}_n(x) - f(x)]^2 \, dx \, .$$

We have the following decomposition:

$$\begin{aligned}
\text{MISE } \hat{f}_n &= \mathbb{E} \int_{\mathbb{R}} [\hat{f}_n(x) - f(x)]^2 \, dx \\
&\overset{\text{(Fubini-Tonelli)}}{=} \int_{\mathbb{R}} \mathbb{E} [\hat{f}_n(x) - f(x)]^2 \, dx \\
&= \int_{\mathbb{R}} \text{MSE } \hat{f}_n(x) \, dx \\
&= \underbrace{\int_{\mathbb{R}} b^2(x) \, dx}_{\text{bias term}} + \underbrace{\int_{\mathbb{R}} \sigma^2(x) \, dx}_{\text{variance term}}
\end{aligned}$$

Result: Under (K2) we have

$$\int_{\mathbb{R}} \sigma^2(x) \, dx \le \frac{1}{nh} K^2 .$$

Proof:

$$\sigma^2(x) = \text{Var } \hat{f}_n(x)$$

$$
\begin{aligned}
&= \operatorname{Var}\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\right) \\
&= \frac{1}{nh^2}\operatorname{Var}\left(K\left(\frac{X_1 - x}{h}\right)\right) \\
&= \frac{1}{nh^2}\left[\mathbb{E}\, K^2\left(\frac{X_1 - x}{h}\right) - \left(\mathbb{E}\, K\left(\frac{X_1 - x}{h}\right)\right)^2\right] \\
&\le \frac{1}{nh^2}\,\mathbb{E}\, K^2\left(\frac{X_1 - x}{h}\right) \\
&= \frac{1}{nh^2}\int_{\mathbb{R}} K^2\left(\frac{z - x}{h}\right) f(z)\, dz \\
&= \frac{1}{nh}\int_{\mathbb{R}} K^2(u)\, f(x + uh)\, du
\end{aligned}
$$

Now we have

$$
\begin{aligned}
\int_{\mathbb{R}} \sigma^2(x)\, dx &\le \frac{1}{nh}\int_{\mathbb{R}}\int_{\mathbb{R}} K^2(u)\, f(x + uh)\, du\, dx \\
&= \frac{1}{nh}\int_{\mathbb{R}} K^2(u)\underbrace{\left[\int_{\mathbb{R}} f(x + uh)\, dx\right]}_{=1} du \\
&= \frac{1}{nh} K^2. \qquad \square
\end{aligned}
$$

The bias, as before, is much trickier, as we must consider the smoothness of the function.

We find that we need another function class to describe smoothness under the L_2-norm:

<u>**Defn:**</u> For $\beta > 0$ an integer and $L > 0$ the Nikol'ski class of functions $\mathcal{N}(\beta, L)$ is the set of functions $f : \mathbb{R} \to \mathbb{R}$ of which the derivatives $f^{(\ell)}$ of order $\ell = \beta - 1$ exist and satisfy

$$\left(\int_{\mathbb{R}} \left[f^{(\ell)}(x+t) - f^{(\ell)}(x) \right]^2 dx \right)^{1/2} \le L |t| \qquad \forall\, t \in \mathbb{R}.$$

Let the set of densities $f \in \mathcal{N}(\beta, L)$ be represented by

$$\mathcal{P}_{\mathcal{N}}(\beta, L) = \left\{ f \in \mathcal{N}(\beta, L) : \quad f \ge 0, \quad \int_{\mathbb{R}} f(x)\, dx = 1 \right\}$$

<u>**Result:**</u> Under (K1) and (K3), if $f \in \mathcal{P}_{\mathcal{N}}(\beta, L)$, then

$$\int_{\mathbb{R}} b^2(x)\, dx \le h^{2\beta} \left(\frac{L \cdot K_\beta}{\ell!} \right)^2.$$

In order to prove the result, we need this inequality:

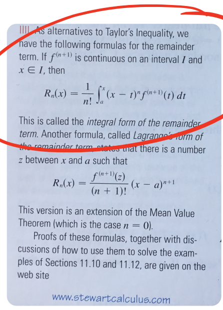
<u>**Generalized Minkowski inequality:**</u> For any Borel function on $\mathbb{R} \times \mathbb{R}$, we have

$$\left[\int_{\mathbb{R}} \left(\int_{\mathbb{R}} g(u,x)\, du \right)^2 dx \right]^{1/2} \le \int_{\mathbb{R}} \left(\int_{\mathbb{R}} g^2(u,x)\, dx \right)^{1/2} du.$$

Proof of bound on the bias term:

First consider the Taylor expansion

$$f(x+uh) = f(x) + \sum_{j=1}^{\ell-1} \frac{f^{(j)}(x)\,(uh)^j}{j!} + \underbrace{\frac{1}{(\ell-1)!}\int_x^{x+uh} (x+uh-t)^{\ell-1} f^{(\ell)}(t)\,dt}_{= \frac{(uh)^\ell}{(\ell-1)!}\int_0^1 (1-\tau)^{\ell-1} f^{(\ell)}(x+\tau uh)\,d\tau}$$


As alternatives to Taylor's Inequality, we have the following formulas for the remainder term. If $f^{(n+1)}$ is continuous on an interval I and $x \in I$, then

$$R_n(x) = \frac{1}{n!}\int_a^x (x-t)^n f^{(n+1)}(t)\,dt$$

This is called the *integral form of the remainder term*. Another formula, called *Lagrange's form of the remainder term*, states that there is a number z between x and a such that

$$R_n(x) = \frac{f^{(n+1)}(z)}{(n+1)!}(x-a)^{n+1}$$

This version is an extension of the Mean Value Theorem (which is the case $n = 0$).

Proofs of these formulas, together with discussions of how to use them to solve the examples of Sections 11.10 and 11.12, are given on the web site

www.stewartcalculus.com

Now we write

$$b(x) = \mathbb{E}\,\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) - f(x)$$

$$= \frac{1}{h}\,\mathbb{E}\,K\left(\frac{X_1 - x}{h}\right) - f(x)$$

$$= \frac{1}{h}\int_{\mathbb{R}} K\left(\frac{z-x}{h}\right) f(z) - f(x)$$

$$= \int_{\mathbb{R}} K(u)\, f(x+uh)\,du - f(x)$$

$$= \int_{\mathbb{R}} K(u)\left[f(x+uh) - f(x)\right] du$$

$$= \int_{\mathbb{R}} K(u)\left[\sum_{j=1}^{\ell-1} \frac{f^{(j)}(x)\,(uh)^j}{j!} + \frac{1}{(\ell-1)!}\int_0^1 (uh)^{\ell-1} (1-\tau)^{\ell-1} f^{(\ell)}(x+\tau uh)\,d\tau\right] du$$

K a kernel of order ℓ

$$= \int_{\mathbb{R}} K(u)\,\frac{(uh)^\ell}{(\ell-1)!}\int_0^1 (1-\tau)^{\ell-1} f^{(\ell)}(x+\tau uh)\,d\tau\,du.$$

$$= \int_{\mathbb{R}} K(u) \frac{(uh)^{\ell}}{(\ell-1)!} \int_0^1 (1-\tau)^{\ell-1} \left[f^{(\ell)}(x+\tau uh) - f^{(\ell)}(x) \right] d\tau \, du$$

Can subtract because kernel of order ℓ.

Now we get a bound on $\int b^2(x)\,dx$ by using Minkowski's 2x and the smoothness class:

$$\int_{\mathbb{R}} b^2(x)\,dx \le \int_{\mathbb{R}} \Bigg[\int_{\mathbb{R}} \underbrace{K(u) \frac{(uh)^{\ell}}{(\ell-1)!} \int_0^1 (1-\tau)^{\ell-1} \left| f(x+\tau uh) - f(x) \right| d\tau \, du}_{g(u,x)\,du} \Bigg]^2 dx$$

$$\le \left[\int_{\mathbb{R}} \left(\int_{\mathbb{R}} \underbrace{\left[K(u) \frac{(uh)^{\ell}}{(\ell-1)!} \int_0^1 (1-\tau)^{\ell-1} \left| f(x+\tau uh) - f(x) \right| d\tau \right]^2 dx}_{g^2(u,x)\,dx} \right)^{1/2} du \right]^2$$

$$= \left[\int_{\mathbb{R}} \left(|K(u)| \frac{|uh|^{\ell}}{(\ell-1)!} \int_{\mathbb{R}} \left[\int_0^1 \underbrace{(1-\tau)^{\ell-1} \left| f(x+\tau uh) - f(x) \right| d\tau}_{g(\tau,x)\,d\tau} \right]^2 dx \right)^{1/2} du \right]^2$$

$$\le \left[\int_{\mathbb{R}} \left(|K(u)| \frac{|uh|^{\ell}}{(\ell-1)!} \int_0^1 \left[\int_{\mathbb{R}} \underbrace{(1-\tau)^{2(\ell-1)} \left| f(x+\tau uh) - f(x) \right|^2 dx}_{g^2(\tau,x)\,dx} \right]^{1/2} d\tau \, du \right]^2$$

$$\le \left[\int_{\mathbb{R}} \left(|K(u)| \frac{|uh|^{\ell}}{(\ell-1)!} \int_0^1 |1-\tau|^{(\ell-1)} \left[\int_{\mathbb{R}} \left| f(x+\tau uh) - f(x) \right|^2 dx \right]^{1/2} d\tau \, du \right]^2$$

$$\leq \left[\int_{\mathbb{R}} \left(|K(u)| \frac{|uh|^{\ell}}{(\ell-1)!} \underbrace{\int_0^1 (1-\tau)^{(\ell-1)} d\tau}_{\frac{1}{\ell}} \; L \, \underbrace{|\tau u h|}_{\leq 1} \; du \right] \right]^2$$

$(\beta = \ell + 1)$

$$\leq h^{2\beta} \frac{L^2}{\ell!} \left[\int_{\mathbb{R}} |u|^{\beta} |K(u)| \, du \right]^2 .$$

$\square$

Result: Under (K1), (K2), and (K3), if $f \in \mathcal{P}_{\mathcal{N}}(\beta, L)$, we have

$$\text{MISE } \hat{f}_n \leq h^{2\beta} \left(\frac{L \cdot K_\beta}{\ell!} \right)^2 + \frac{1}{nh} K^2 .$$

We can also state a uniform result:

Theorem 2: Under (K1), (K2), and (K3), if $h = \alpha n^{-\frac{1}{2\beta+1}}$ for some $\alpha > 0$, then for all $n \geq 1$,

$$\sup_{f \in \mathcal{P}_{\mathcal{N}}} \mathbb{E}_f \int_{\mathbb{R}} [\hat{f}_n(x) - f(x)]^2 dx \leq C n^{-\frac{2\beta}{2\beta+1}} ,$$

where $C > 0$ depends only on β, L, α and the kernel $K(\cdot)$.

Data-based bandwidth selection:

A "plug-in" method: the Sheather–Jones method uses the following result:

Theorem 3: If K is a kernel of order 1 s.t.

$$K^2 = \int_{\mathbb{R}} K^2(u)\,du < \infty, \quad \int_{\mathbb{R}} u^2 |K(u)|\,du < \infty, \quad \sigma_K^2 = \int_{\mathbb{R}} u^2 K(u)\,du < \infty$$

and f is differentiable on $\mathbb{R}$ with f' a.c. on $\mathbb{R}$ and with

$$\|f''\|_2^2 = \int_{\mathbb{R}} [f''(x)]^2\,dx < \infty,$$

then

$$\text{MISE}\,\hat{f}_n = \left[h^4 \left(\frac{\|f''\|_2 \cdot \sigma_K^2}{2} \right)^2 + \frac{1}{nh} K^2 \right] \Big(1 + \underbrace{o(1)}_{\to 0 \text{ as } h \to 0} \Big)$$

as $n \to \infty$, $h \to 0$.

See pg. 192 of Tsybakov for a proof.

The proof of the above result is more tedious than the proof of the MISE result under the Nikol'ski class.

The above result is also not a uniform result over any class of functions, but rather assumes a single, fixed density f.

Exercise: Get an expression for the optimal bandwidth using Thm 3.

Solution: We have

$$\frac{\partial}{\partial h}\Big(\text{Dominating term of MISE} \Big) = h^3 \cdot \|f''\|_2^2 \cdot \sigma_K^4 - \frac{1}{nh} K^2 = 0$$

So the MISE-minimizing choice of h is

$$h_{opt} = n^{-\frac{1}{5}} \left(\frac{K^2}{\|f''\|_2^2 \cdot \sigma_K^4} \right)^{1/5}.$$

The SJ method "plugs in" an estimator of $\|f''\|_2^2$. Hence "plug-in" method. See option bw="SJ" of the density() function in R.

A Crossvalidation Method:

The idea is to estimate from the data the MISE of $\hat{f}_n$ at given bandwidth values and then to choose h which minimizes it.

Consider

$$\begin{aligned} MISE_h \hat{f}_n &= \mathbb{E}_f \int_{\mathbb{R}} [\hat{f}_n(x) - f(x)]^2 \, dx \\ &= \mathbb{E}_f \int_{\mathbb{R}} \hat{f}_n^2(x) \, dx \quad - 2\, \mathbb{E}_f \int_{\mathbb{R}} \hat{f}_n(x) f(x) \, dx \; - \; \int_{\mathbb{R}} f^2(x) \, dx \end{aligned}$$

Let

$$A_n = \mathbb{E}_f \int_{\mathbb{R}} \hat{f}_n^2(x) \, dx .$$

$$B_n = \mathbb{E}_f \int_{\mathbb{R}} \hat{f}_n(x) f(x) \, dx$$

Note that $\int_{\mathbb{R}} \hat{f}_n^2(x) \, dx$ is, trivially, an unbiased estimator of A_n.

Now, an unbiased estimator of B_n can be constructed as

$$\hat{B}_n = \frac{1}{n} \sum_{i=1}^{n} \hat{f}_{n,-i}(X_i),$$

where

"Leave-one-out" crossvalidation.

$$\hat{f}_{n,-i}(x) = \frac{1}{(n-1)h} \sum_{j \neq i} K\left(\frac{X_j - x}{h}\right)$$

is the KDE of f computed after removing obs i from the data.

We see that

$$\mathbb{E}_f \hat{B} = \mathbb{E}_f \frac{1}{n} \sum_{i=1}^{n} \hat{f}_{n,-i}(X_i)$$

$$= \mathbb{E}_f f_{n,-1}(X_1)$$

$$= \mathbb{E}_f \left[\frac{1}{(n-1)h} \sum_{j \neq 1} K\left(\frac{X_j - X_1}{h}\right) \right]$$

$$= \frac{1}{h} \mathbb{E}_f K\left(\frac{X_2 - X_1}{h}\right)$$

$$= \frac{1}{h} \int_{\mathbb{R}} \int_{\mathbb{R}} K\left(\frac{z-x}{h}\right) f(z) f(x)\, dz\, dx$$

We also have

$$B_n = \mathbb{E}_f \int_{\mathbb{R}} \hat{f}_n(x) f(x)\, dx$$

$$= \mathbb{E}_f \int_{\mathbb{R}} \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) f(x)\, dx$$

$$= \frac{1}{h} \int_{\mathbb{R}} \mathbb{E}_f K\left(\frac{X_i - x}{h}\right) f(x)\, dx$$

$$= \frac{1}{h} \int_{\mathbb{R}} \int_{\mathbb{R}} K\left(\frac{z-x}{h}\right) f(x)\, dx\ f(z)\, dz,$$

so that $\mathbb{E}\hat{B}_n = B_n$.

Since $\underset{h>0}{\operatorname{argmin}}\ MISE_h\ \hat{f}_n = \underset{h>0}{\operatorname{argmin}}\ A_n - 2B_n$, choose

$$h_{CV} = \underset{h>0}{\operatorname{argmin}}\ CV(h),$$

where

$$CV(h) = \int_{\mathbb{R}} \hat{f}_n^2(x)\,dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{n,-i}(X_i).$$

Some assumptions implicit in our work are given in the following proposition:

__Result:__ Assume that $K:\mathbb{R}\to\mathbb{R}$ and f satisfy

$$\int_{\mathbb{R}} f^2(x)\,dx < \infty \quad \text{and} \quad \int_{\mathbb{R}}\int_{\mathbb{R}} \left|K\left(\frac{z-x}{h}\right)\right| f(z)\,dz\, f(x)\,dx$$

for all $h>0$. Then

$$\mathbb{E}\,CV(h) = MISE_h\ \hat{f}_n - \int_{\mathbb{R}} f^2(x)\,dx.$$