MS = DENSITY ESTIMATION

I - PRELIMINARIES

Let $X_1, \dots, X_n$ iid $\sim P_X$, absolutely continuous RVs, with density $f_X : \mathbb{R}^d \to [0, +\infty)$:

$$\forall B \in \mathcal{B}(\mathbb{R}^d), \quad P_X(B) = \int_B f_X(x)\, dx .$$

The goal of density estimation is to produce a function $\hat{f}_n : \mathbb{R}^d \to [0, +\infty)$ based on $X_1, \dots, X_n$ that is a "good" approximation to the unknown f_X

- Can be measured by either one of the following =

(i) MSE at a particular point $x_0 \in \mathbb{R}^d$:

$$MSE(\hat{f}_n ; x_0) := \mathbb{E}\left\{ \left(\hat{f}_n(x_0) - f_X(x_0) \right)^2 \right\}$$

(ii) Mean Integrated Squared Error (MISE)

$$MISE(\hat{f}_n) = \int_{\mathbb{R}^d} MSE(\hat{f}_n ; x)\, dx$$

(iii) KL divergence $KL(\hat{f}_n \| f_X)$

(iv) ... /...

x Estimation Strategy = Let $K : \mathbb{R}^d \to \mathbb{R}$ be a function (called a KERNEL) satisfying the following properties :

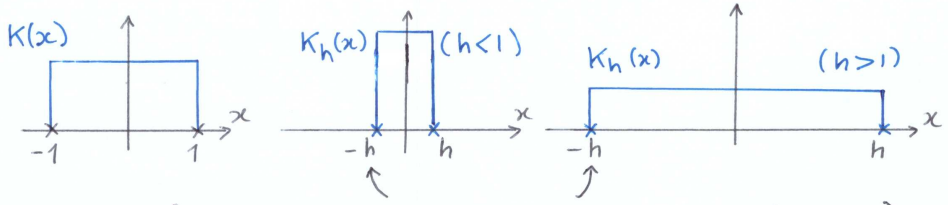
(i) $K \geq 0$

(ii) $K(x) = 0$ if $\|x\| > 1$ ← Can be weakened

(iii) $\int_{\mathbb{R}^d} K(x)\, dx = 1$

For $h > 0$ (called the BANDWITH), put

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



(The bandwidth controls the spread of the kernel.)

Define the convolution $K_h * f$ between K_h and any function $f : \mathbb{R}^d \to \mathbb{R}$ by

$$K_h * f(x) = \int_{\mathbb{R}^d} K_h(x - u)\, f(u)\, du$$

It is a well known result that if $f \in \mathcal{L}^p$ ($\int_{\mathbb{R}^d} |f(x)|^p dx < \infty$) ($1 \leq p < +\infty$), and if K satisfies conditions (i), (ii) & (iii) mentioned above, then $K_h * f \to f$ in $\mathcal{L}^p$, as $h \to 0$; that is

$$\int_{\mathbb{R}^d} |K_h * f(x) - f(x)|^p dx \to 0 , \text{ as } h \to 0 .$$

In other words, for h small enough, $K_h * f \approx f$.

Key point in non parametric density estimation. Indeed, if $f = f_X$ represent the density of $X_1, \dots, X_n$, then

$$K_h * f_X(x) = \int K_h(x-u) f_X(u)\, du = \mathbb{E}\left\{ K_h(x - X) \right\} ; \quad X \sim f_X .$$

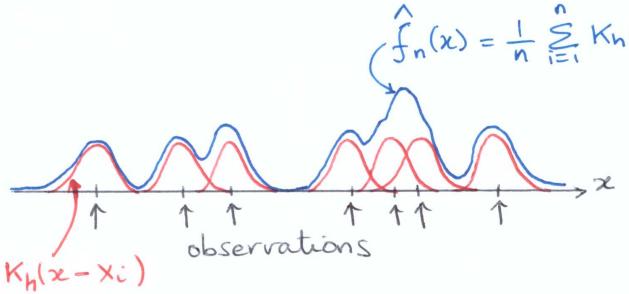
Given $X_1, \dots, X_n$, the last written expectation may be approximated by the empirical mean

$$f_X(x) \approx K_h * f(x) \approx \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) \xrightarrow[n\to+\infty]{\text{SLLN}} \mathbb{E}\{K(x-X)\}.$$

h small enough — n large enough

"Place a kernel function around each observation and sum the result."

Note that conditions (i) and (iii) ensure that our estimate $\frac{1}{n}\sum_{i=1}^{n} K(x - X_i)$ is a density (≥ 0 and integrates to 1). Condition (ii) can be weakened so that the support of K is infinite (e.g. take $K(x) = \mathcal{N}(0, h^2)$).



× Remark: Influence of h.

"h small" → the averaging is performed at each x on few observations → rough estimate $\hat{f}_n$.

"h large" → each observation has an increased impact on the averaging, even for values of x far from it → smooth estimate $\hat{f}_n$.

⇒ Requires a tradeoff if one wants to avoid over/under smoothing.

II. KERNEL DENSITY ESTIMATION

II.1. Pointwise kernel estimation in dimension 1.

In this section, we fix $x_0 \in \mathbb{R}$, and we want to estimate $f_X(x_0)$. We denote

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

where K is a kernel function with properties precised below.

× Bias-variance decomposition of $\mathrm{MSE}(\hat{f}_n; x_0)$

$$\begin{aligned}
\mathrm{MSE}(\hat{f}_n; x_0) &= \mathbb{E}\left\{\left(\hat{f}_n(x_0) - f_X(x_0)\right)^2\right\} \\
&= \mathbb{E}\left\{\left(\hat{f}_n(x_0) - \mathbb{E}\,\hat{f}_n(x_0) + \mathbb{E}\,\hat{f}_n(x_0) - f_X(x_0)\right)^2\right\} \\
&= \mathbb{E}\left\{\left(\hat{f}_n(x_0) - \mathbb{E}\,\hat{f}_n(x_0)\right)^2\right\} \\
&\quad + \left(\mathbb{E}\,\hat{f}_n(x_0) - f_X(x_0)\right)^2 \\
&\quad + 2\left(\mathbb{E}\,\hat{f}_n(x_0) - f_X(x_0)\right) \times \underbrace{\mathbb{E}\left\{\hat{f}_n(x_0) - \mathbb{E}\,\hat{f}_n(x_0)\right\}}_{=0}.
\end{aligned}$$

$:= \sigma_n^2(x_0)$ = variance of $\hat{f}_n$ evaluated at x_0

$:= b_n^2(x_0)$ = squared bias of $\hat{f}_n$ at x_0.

$$\mathrm{MSE}(\hat{f}_n; x_0) = b_n^2(x_0) + \sigma_n^2(x_0)$$

We analyse these two terms separately.

(i) Analysis of the variance term $\sigma_n^2(x_0)$.

Theorem. Suppose that there exists $C>0$ such that
$\forall x \in \mathbb{R}$, $f_X(x) \leq C < +\infty$.
Suppose in addition that $K: \mathbb{R} \to \mathbb{R}$ is chosen such that
$\int K^2(u)\,du < +\infty$.
Then, $\forall x_0 \in \mathbb{R}$, $\forall h>0$, $\forall n \geq 1$,

$$\sigma_n^2(x_0) \leq \frac{C}{nh} \int_{\mathbb{R}} K^2(u)\,du.$$

Consequence: If we allow h to vary with n, then we should have

(i) $\lim_{n\to+\infty} h = 0$ (so that $f_X \approx K_h * f_X$ holds)

(ii) $\lim_{n\to+\infty} nh = +\infty$ (so that $\sigma_n^2(x_0) \to 0$)

The bandwidth should tend to 0 at a rate smaller than n^{-1}.

proof $\sigma_n^2(x_0) = \mathbb{E}\left\{\left(\frac{1}{nh}\sum_{i=1}^n K\left(\frac{x_0-X_i}{h}\right) - \mathbb{E}(\dots)\right)^2\right\}$

$$= \frac{1}{n^2h^2}\,\mathrm{Var}\left\{\sum_{i=1}^n K\left(\frac{x_0-X_i}{h}\right)\right\}$$

$$= \frac{1}{nh^2}\,\mathrm{Var}\left\{K\left(\frac{x_0-X_i}{h}\right)\right\}$$

$$\leq \frac{1}{nh^2}\,\mathbb{E}\left\{K^2\left(\frac{x_0-X_i}{h}\right)\right\}$$

$$= \frac{1}{nh^2}\int K^2\left(\frac{x_0-u}{h}\right) f_X(u)\,du$$

$$\leq \frac{C}{nh^2}\int K^2\left(\frac{x_0-u}{h}\right)du = \frac{C}{nh}\int K^2(v)\,dv. \quad \blacksquare$$

(ii) Analysis of the bias term $b_n(x_0)$.

The bias term requires a bit more work. We introduce the following class of functions:

HÖLDER CLASS $\mathbb{H}(\ell, \alpha, L)$

Let . $\ell \geq 0$ be an integer
. $\alpha \in [0,1]$, $\ell + \alpha > 0$
. $L > 0$ be a constant

The class $\mathbb{H}(\ell, \alpha, L)$ is the class of functions $h: \mathbb{R} \to \mathbb{R}$ such that

a) h is ℓ-times continuously differentiable on $\mathbb{R}$

b) We have $|h^{(\ell)}(x) - h^{(\ell)}(y)| \leq L|x-y|^\alpha \quad \forall x,y$

($h^{(\ell)}$: ℓ-th derivative of h. (if $\ell=0$, then $h^{(0)} = h$))

We denote $\beta := \ell + \alpha > 0$, and refer to β as the REGULARITY PARAMETER of $h \in \mathbb{H}(\ell, \alpha, L)$.

As β increases, functions in $\mathbb{H}(\ell, \alpha, L)$ are smoother and smoother.

We use next the following well known result:

Lemma (TAYLOR-LAGRANGE)

Let $a<b$ and $f: [a,b] \to \mathbb{R}$ be ℓ times continuously differentiable ($\ell \geq 1$ integer) on $[a,b]$.
Then there exists $\theta \in [a,b]$ s.t.

$$f(b) = \sum_{k=0}^{\ell-1} \frac{(b-a)^k}{k!} f^{(k)}(a) + \frac{(b-a)^\ell}{\ell!} f^{(\ell)}(\theta).$$

Theorem

Let $\ell \geq 0$ be an integer, $\alpha \in [0,1]$ s.t. $\beta = \ell + \alpha > 0$, and $L > 0$.

Suppose $f_X \in \mathbb{H}(\ell, \alpha, L)$.

Suppose in addition that $K: \mathbb{R} \to \mathbb{R}$ is chosen such that

(i) $\forall k \in \{0, 1, \dots, \ell, \ell+\alpha\} \quad \int |u^k K(u)| \, du < \infty$

(ii) $\int K(u) \, du = 1$

(iii) If $\ell \geq 1$, then for $k = 1, \dots, \ell$, $\int u^k K(u) \, du = 0$.

Then, $\forall x_0 \in \mathbb{R}$, $\forall h > 0$, $\forall n \geq 1$,

$$|b_n(x_0)| \leq L \frac{h^{\ell+\alpha}}{\ell!} \int |u^{\ell+\alpha} K(u)| \, du .$$

proof = We have

$$\begin{aligned} b_n(x_0) &= \mathbb{E}\, \hat{f}_n(x_0) - f_X(x_0) \\ &= \frac{1}{nh} \sum_{i=1}^{n} \mathbb{E}\left\{ K\left(\frac{x_0 - X_i}{h} \right) \right\} - f_X(x_0) \\ &= \frac{1}{h} \mathbb{E}\left\{ K\left(\frac{x_0 - X}{h} \right) \right\} - f_X(x_0) \\ &= \frac{1}{h} \int K\left(\frac{x_0 - u}{h} \right) f_X(u) \, du - \int f_X(x_0) K(u) \, du \end{aligned}$$

K is symmetrical →

$$\begin{aligned} &= \int K(u) f_X(x_0 + uh) \, du - \int f_X(x_0) K(u) \, du \\ &= \int K(u) \left[f_X(x_0 + uh) - f_X(x_0) \right] du \end{aligned}$$

↘ If $\ell = 0$, $|f_X(x_0 + uh) - f_X(x_0)| \leq L h^\alpha |u|^\alpha$ since $f \in \mathbb{H}(\ell, \alpha, L)$. Thus

$$|b_n(x_0)| \leq \int |K(u)| \, |f_X(x_0 + uh) - f_X(x_0)| \, du \leq L h^\alpha \int |u^\alpha K(u)| \, du .$$

↘ If $\ell \geq 1$, we have

$$f_X(x_0 + uh) - f_X(x_0) = \sum_{k=1}^{\ell-1} \frac{(uh)^k}{k!} f_X^{(k)}(x_0) + \frac{(uh)^\ell}{\ell!} f_X^{(\ell)}(x_0 + \theta uh),$$

for some $\theta \in [0,1]$ (Taylor-Lagrange lemma).

Thus

$$b_n(x_0) = \sum_{k=1}^{\ell-1} \frac{h^k}{k!} f_X^{(k)}(x_0) \int u^k K(u) \, du + \frac{h^\ell}{\ell!} \int K(u) u^\ell f_X^{(\ell)}(x_0 + \theta u h) \, du$$

($\int u^k K(u) \, du = 0$)

$$= \frac{h^\ell}{\ell!} \int K(u) u^\ell f_X^{(\ell)}(x_0 + \theta u h) \, du$$

↓ $\int K(u) u^\ell f_X^{(\ell)}(x_0) = 0$ from (iii)

$$= \frac{h^\ell}{\ell!} \int K(u) u^\ell \left[f_X^{(\ell)}(x_0 + \theta u h) - f_X^{(\ell)}(x_0) \right] du$$

$$|b_n(x_0)| \leq \frac{|h|^\ell}{\ell!} \int |K(u) u^\ell| \underbrace{\left| f_X^{(\ell)}(x_0 + \theta u h) - f_X^{(\ell)}(x_0) \right|}_{\leq L |\theta u h|^\alpha} du$$

$$\leq \frac{L h^{\ell+\alpha}}{\ell!} \int |u^{\ell+\alpha} K(u)| \, du \qquad (|\theta| < 1)$$

■

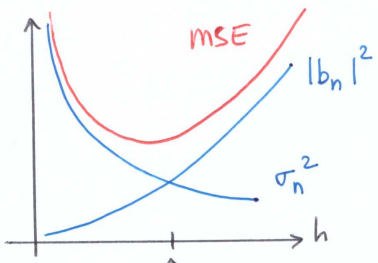
Remark: The upper bounds on the bias & variance have opposite behaviours in h: the variance decreases as h grows ($\sigma_n^2(x_0) = O(h^{-1})$), while the bias increases with h ($b_n(x_0) = O(h^\beta)$).

- small h $\leftrightarrow$ large variance $\leftrightarrow$ undersmoothing
- large h $\leftrightarrow$ large bias $\leftrightarrow$ oversmoothing.

In fact, we see that

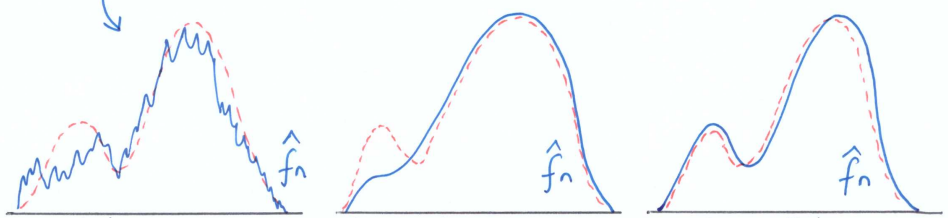
$$\mathrm{MSE}(\hat{f}_n; x_0) \leq \frac{c}{nh}\int K^2(u)\,du \quad (*)$$
$$+ L^2 \frac{h^{2\beta}}{(\ell!)^2}\left(\int |u^{\ell+\alpha} K(u)|\,du\right)^2$$

(under the conditions of the theorems on page 5 & 7).



"optimal" bandwidth h.

Assume that the true f_X is bimodal (in red)



small h
wiggly estimate, not representative of f_X

large h
smooth estimate
$\Rightarrow$ First bump is missed

"optimal" h
$\Rightarrow$ correctly identifies the two bumps.

x Optimal choice of the bandwidth: minimize (as a function of h) the right-hand-side of $(*)$. We easily find that h must be of order $O(n^{-\frac{1}{2\beta+1}})$, and that the resulting MSE is of order $O(n^{-\frac{2\beta}{2\beta+1}})$, $n \to +\infty$.

↑ of order larger than n^{-1} (the "parametric rate")

The bound gets better (closer to the parametric rate) as β gets larger. (the more regular f_X, the better).
For example, for a twice continuously differentiable density f_X, we obtain the rate $O(n^{-4/5})$ with a bandwidth $h = O(n^{-1/5})$.

II.2. Average estimation in dimension 1.

Recall that the Mean Integrated Square Error (MISE) is defined as

$$\mathrm{MISE}(\hat{f}_n) = \int_{\mathbb{R}} \mathrm{MSE}(\hat{f}_n; x)\,dx$$
$$= \int_{\mathbb{R}} \sigma_n^2(x)\,dx + \int_{\mathbb{R}} b_n^2(x)\,dx.$$

↑ We study these two terms separately.

x Analysis of the integrated variance term.

Theorem: Suppose $K: \mathbb{R} \to \mathbb{R}$ is chosen such that $\int K^2(u)\,du < +\infty$. Then $\forall h > 0$, $\forall n \geq 1$,

$$\int_{\mathbb{R}} \sigma_n^2(x)\,dx \leq \frac{1}{nh}\int K^2(u)\,du.$$

↑ Boundedness of f_X is not required here (compare with the theorem on page 5).

proof: We have $\sigma_n^2(x) \leq \frac{1}{nh^2}\int K^2\left(\frac{x_0-u}{h}\right) f_X(u)\,du$ (p. 5)

Thus,

$$\int \sigma_n^2(x)\,dx \leq \frac{1}{nh^2}\int\left(\underbrace{\int K^2\left(\frac{x-u}{h}\right)dx}_{=h\int K^2(v)\,dv}\right) f_X(u)\,du$$

$$= \frac{1}{nh}\int\left(\int K^2(v)\,dv\right) f_X(u)\,du.$$

$$= \frac{1}{nh}\int K^2(v)\,dv. \quad \blacksquare$$

x Analysis of the integrated bias term.

To study the term $\int b_n^2(x)\,dx$, we introduce the Nikol'ski class of functions:

NIKOL'SKI CLASS $\mathbb{N}(\ell,\alpha,L)$

Let $\ell \geq 0$ be an integer
- $\alpha \in [0,1]$ such that $\beta := \ell+\alpha > 0$
- $L > 0$ a constant.

The Nikol'ski class $\mathbb{N}(\ell,\alpha,L)$ is defined as the set of functions $h:\mathbb{R}\to\mathbb{R}$ such that

a) h is ℓ-times continuously differentiable on $\mathbb{R}$

b) $\forall x \in \mathbb{R}$ $\left(\int\left[h^{(\ell)}(u+x)-h^{(\ell)}(u)\right]^2 du\right)^{1/2} \leq L|x|^\alpha$

↑ Since the MISE is defined in terms of the $\mathcal{L}^2$ norm, Nikol'ski class assumes naturally that the density f_X is smooth with respect to this norm.

We need two technical lemmas:

Lemma (TAYLOR).
Let $f:[a,b]\to\mathbb{R}$ be ℓ-times continuously differentiable on $[a,b]$ ($\ell \geq 1$ integer). Then

$$f(b) = \sum_{k=0}^{\ell-1}\frac{(b-a)^k}{k!}f^{(k)}(a) + \frac{(b-a)^\ell}{(\ell-1)!}\int_0^1 (1-u)^{\ell-1} f^{(\ell)}(a+u(b-a))\,du$$

Lemma (Generalized MINKOWSKI)
For any measurable function $f:\mathbb{R}\times\mathbb{R}\to\mathbb{R}$, we have

$$\int\left(\int g(u,x)\,du\right)^2 dx \leq \left(\int\left(\int |g(u,x)|^2\,dx\right)^{1/2} du\right)^2$$

↑ For a proof of this result, see Lemma A.1 p. 191 in A. B. Tsybakov (2009). Introduction to non-parametric estimation. Springer.

Theorem.
Let $\ell \geq 0$ be an integer, $\alpha \in [0,1]$ s.t. $\beta := \ell+\alpha > 0$, and $L > 0$.
Suppose that $f_X \in \mathbb{N}(\ell,\alpha,L)$.
Suppose in addition that $K:\mathbb{R}\to\mathbb{R}$ is chosen such that

(i) $\forall k \in \{0,1,\dots,\ell,\ell+\alpha\}$, $\int |u^k K(u)|\,du < +\infty$

(ii) $\int K(u)\,du = 1$

(iii) If $\ell \geq 1$, for $k=1,\dots,\ell$, $\int u^k K(u)\,du = 0$.

Then $\forall h>0$, $\forall n \geq 1$,

$$\int b_n^2(x)\,dx \leq h^{2\beta}\left(\frac{L}{\ell!}\int |u^\beta K(u)|\,du\right)^2$$

proof = We have that (p.7)

$$b_n(x) = \int_{\mathbb{R}} K(u)\underbrace{\left[f_x(x+uh) - f_x(x)\right]}_{\text{= (Taylor)}} du$$

$$\sum_{k=1}^{\ell-1} \frac{(uh)^k}{k!} f_x^{(k)}(x) + \frac{(uh)^\ell}{(\ell-1)!}\int_0^1 (1-v)^{\ell-1} f^{(\ell)}(x+vuh)\, dv.$$

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

↑ since $\int u^k K(u) = 0$ for $k = 1, \dots, \ell-1$

Next, using that $\int u^\ell K(u)\, du = 0$, we have that

$$\int_{\mathbb{R}} K(u)\frac{(uh)^\ell}{(\ell-1)!}\int_0^1 (1-v)^{\ell-1} f^{(\ell)}(x)\, dv\, du = 0,$$

so that

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

↓ squaring & integrating with respect to x:

$$\int b_n^2(x)\, dx = \int\left(\int_{\mathbb{R}} K(u)\frac{(uh)^\ell}{(\ell-1)!}\int_0^1 (1-v)^{\ell-1}\left[f^{(\ell)}(x+vuh) - f^{(\ell)}(x)\right] dv\, du\right)^2 dx$$

$=: g(u,x)$

Making use of the generalized Minkowski inequality,

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

$$= \left(\int_{\mathbb{R}}\left(\int_{\mathbb{R}} \frac{K^2(u)}{[(\ell-1)!]^2}(uh)^{2\ell} \times \left(\int_0^1 (1-v)\left[f^{(\ell)}(x+vuh) - f^{(\ell)}(x)\right] dv\right)^2 dx\right)^{1/2} du\right)^2$$

$$= \left(\int_{\mathbb{R}} \frac{|K(u)|}{(\ell-1)!}|uh|^\ell \left(\int_{\mathbb{R}}\left(\int_{\mathbb{R}} h(v,x)\, dv\right)^2 dx\right)^{1/2} du\right)^2$$

where
$h(v,x) := (1-v)^{\ell-1}\left[f^{(\ell)}(x+vuh) - f^{(\ell)}(x)\right]\mathbb{1}(v \in [0,1])$

↳ Applying a second time the generalized Minkowski inequality to $\int_{\mathbb{R}}\left(\int h(v,x)\, dv\right)^2 dx$ yields:

$$\int_{\mathbb{R}}\left(\int h(v,x)\, dv\right)^2 dx \le \left(\int_{\mathbb{R}}\left(\int h^2(v,x)\, dx\right)^{1/2} dv\right)^2$$

$$= \left(\int_{\mathbb{R}}\left(\int_{\mathbb{R}} (1-v)^{2(\ell-1)}\left[f^{(\ell)}(x+vuh) - f^{(\ell)}(x)\right]^2 \mathbb{1}(\dots)\, dx\right)^{1/2} dv\right)^2$$

$$= \left(\int_0^1 (1-v)^{\ell-1}\underbrace{\left(\int_{\mathbb{R}}\left[f^{(\ell)}(x+vuh) - f^{(\ell)}(x)\right]^2 dx\right)^{1/2}}_{\le L|vuh|^\alpha \le L|uh|^\alpha \text{ since } f \in \mathbb{N}(\ell,\alpha,L)} dv\right)^2$$

We get

$$\int_{\mathbb{R}} \left(\int h(v,x)\, dv \right)^2 dx \leq \left(\int_0^1 (1-v)^{\ell-1} L\, |uh|^{\alpha}\, dv \right)^2$$

$$= L^2 h^{2\alpha} |u|^{2\alpha} \left(\int_0^1 (1-v)^{\ell-1} dv \right)^2$$

$$= L^2 |hu|^{2\alpha} \frac{1}{\ell^2} .$$

Plugging this expression back into $\int b_n^2(x)\, dx$ gives:

$$\int b_n^2(x)\, dx \leq \left(\int_{\mathbb{R}} \frac{|K(u)|}{(\ell-1)!} |uh|^{\ell} L\, |hu|^{\alpha} \frac{1}{\ell}\, du \right)^2$$

$$= \left(\int_{\mathbb{R}} |u^{\beta} K(u)|\, h^{\ell+\alpha} \frac{L}{\ell!}\, du \right)^2$$

$$= h^{2\beta} \left(\frac{L}{\ell!} \int |u^{\beta} K(u)|\, du \right)^2 .$$

Putting terms together, we obtain

$$\mathrm{MISE}(\hat{f}_n) \leq \frac{1}{nh} \int K^2(u)\, du + h^{2\beta} \left(\frac{L}{\ell!} \int |u^{\beta} K(u)|\, du \right)^2$$

↑ For a density f_X & kernel K satisfying the conditions of the theorems p. 10 and 12.

– Compare with expression on page 9 –

⇒ Optimal rate of convergence $O\left(n^{-\frac{2\beta}{2\beta+1}}\right)$ is obtained for $h = O\left(n^{-\frac{1}{2\beta+1}}\right)$, just as for the pointwise case. Faster convergence rates are achieved as the regularity parameter β increases.

• Remark = The MSE/MISE can be controlled under other assumptions than the ones introduced previously. For example, integrability conditions of the density can replace Lipschitz-like conditions of the Nikol'ski class of functions. To fix ideas, suppose that

(i) f is twice continuously differentiable, with
$R(f'') := \int (f''(x))^2 dx < +\infty$. ($\equiv$ "$\beta = 2$")

In addition, we assume that the kernel K satisfies

(ii) $\int K(x)\, dx = 1$, $\int x\, K(x)\, dx = 0$, and has a finite fourth moment.

Also, assume that the bandwith $h = h_n$ is such that

(iii) $h \to 0$ and $nh \to +\infty$ as $n \to +\infty$.

For $\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$, we have that

• $\mathbb{E}\{\hat{f}_n(x)\} = \mathbb{E}\{K_h(x - X)\}$

$$= \int K(u) f(x - uh)\, du$$

$$= f(x) + \frac{1}{2} h^2 f''(x) \mu_2(K) + o(h^2),$$

where $\mu_2(K) := \int u^2 K(u)\, du$, and

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

• $\mathrm{Var}\{\hat{f}_n(x)\} = n^{-1} \mathrm{Var}\{K_h(x - X)\}$

$$= \frac{1}{nh} \int K^2(u) f(x - hu)\, du - \frac{1}{n} \left(\mathbb{E}\{\hat{f}_n(x)\} \right)^2$$

$$\mathrm{Var}\left\{\hat{f}_n(x)\right\} = \frac{1}{nh}\int K^2(u)\,(f(x)+o(1))\,du - n^{-1}(f(x)+o(1))$$

$$= \frac{1}{nh}R(K)\,f(x) + o\left(\frac{1}{nh}\right)$$

$R(K) = \int K^2(u)\,du$.

- Putting terms together,

$$\mathrm{MSE}(\hat{f}_n; x) = \frac{R(K)}{nh}f(x) + \frac{h^4}{4}\mu_2^2(K)(f''(x))^2 + o\left(h^4 + \frac{1}{nh}\right).$$

$$\mathrm{MISE}(\hat{f}_n) = \frac{R(K)}{nh} + \frac{1}{4}h^4\mu_2^2(K)R(f'') + o\left(h^4+\frac{1}{nh}\right)$$

Compare with the upper bound page 15

This term is commonly referred to as the Asymptotic MISE ; or simply AMISE $(\hat{f}_n)$.

Based on $\mathrm{AMISE}(\hat{f}_n) = \frac{R(K)}{nh} + \frac{1}{4}h^4\mu_2^2(K)R(f'')$,

we calculate the "optimal" bandwidth minimizing this term, and we get $h = \left(\frac{R(K)}{\mu_2^2(K)R(f'')\,n}\right)^{1/5}$, for which

we deduce that the smallest possible AMISE is given by

$\frac{5}{4}\left(\mu_2^2(K)\,R^4(K)\,R(f'')\right)^{1/5} n^{-4/5}$ (compare with the rate p. 15)

x Example: If $f_X \sim \mathcal{N}(0,\sigma^2)$ and $K = \mathcal{N}(0,1)$, we obtain the optimal bandwidth $h \approx 1.06\,\hat{\sigma}\,n^{-1/5}$ ("h SILVERMAN"), for some estimator $\hat{\sigma}$ of σ.

II.3 Choosing / constructing the kernel K

The results obtained (p. 9 & 15) rely on the selection of a kernel function $K: \mathbb{R} \to \mathbb{R}$ satisfying:

$$[K] = \begin{cases} \text{(i) } \forall k = 0, 1, \dots, \ell, \ell+\alpha, & \int |x^k K(x)|\,dx < +\infty \\ \text{(ii) } & \int K(x)\,dx = 1 \\ \text{(iii) If } \ell \geq 1, \text{ for } k = 1, \dots, \ell, & \int x^k K(x)\,dx = 0 \end{cases}$$

Any probability density function having infinitely many moments satisfy properties (i) and (ii)

Ex: . Gaussian kernel $K(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$

. Parabolic kernel $K(x) = \frac{3}{4}(1-x^2)\mathbb{1}(x \in [-1,1])$

The question we address here concerns the construction of a function K satisfying in addition property (iii). This requires a bit more work, and involves the use of ORTHOGONAL POLYNOMIALS.

x Let $w: \mathbb{R} \to [0, +\infty)$ ("weight function") be a function satisfying

$$[W] = \forall \text{integer } k \geq 0, \quad \int_{\mathbb{R}} |x|^k w(x)\,dx < +\infty. \quad \& \; w(0) = 1.$$

Let $\mathcal{L}^2_w := \left\{ f: \mathbb{R} \to \mathbb{R} \;\middle|\; \int f^2(x)\,w(x)\,dx < +\infty \right\}$.

$\forall f, g \in \mathcal{L}^2_w$, define $\langle f, g \rangle_w := \int_{\mathbb{R}} f(x)\,g(x)\,w(x)\,dx$ (inner product),

which is finite since $|\langle f, g\rangle_w| \leq \|f\|_w \cdot \|g\|_w < +\infty$,

for $\|f\|_w^2 := \langle f, f\rangle_w = \int_{\mathbb{R}} f^2(x)\,w(x)\,dx < +\infty$. (norm)

Examples (i) $w(x) = \mathbb{1}(x \in [-1,1])$

(ii) $w(x) = e^{-x^2}$

(iii) $w(x) = e^{-x}\,\mathbb{1}(x \geq 0)$

(iv) $w(x) = \frac{1}{\sqrt{1-x^2}}\,\mathbb{1}(x \in [-1,1])$

Fact = For any weight function w satisfying condition [W], there exists a unique sequence $\{P_n\}_{n\geq 0}$ of polynomials st:

↘ P_n = unitary polynomial of degree n

($P_n(x) = x^n + \sum_{j=0}^{n-1} a_j^n x^j$, $n \geq 1$ and $P_0(x) = 1$)

↘ P_n are orthogonal: $\forall n \neq m \quad \langle P_n, P_m \rangle_w = 0$.

Indeed, supposing we have constructed $P_0, \dots, P_n$, let's construct P_{n+1}.

We can look for an expression of the form

$$P_{n+1}(x) = x^{n+1} + \sum_{j=0}^{n} \alpha_j^{n+1} P_j(x), \quad \text{for some } \alpha_j^{n+1} \in \mathbb{R},\ j = 0, \dots, n.$$

(P_{n+1} is unitary; $P_0, \dots, P_n$ form a basis of the set of polynomials of degree $\leq n$.)

P_{n+1} is such that $\langle P_{n+1}, P_j \rangle_w = 0$ for $j = 0, \dots, n$.

Thus

$$\langle P_{n+1}, P_j \rangle_w = \langle x^{n+1}, P_j \rangle_w + \sum_{k=0}^{n} \alpha_k^{n+1} \langle P_k, P_j \rangle_w = 0$$

↓ orthogonality of the $P_0, \dots, P_n$

$$\langle x^{n+1}, P_j \rangle_w + \alpha_j^{n+1} \underbrace{\langle P_j, P_j \rangle_w}_{>0 \text{ (since } w(0)=1)} = 0$$

Thus $\alpha_j^{n+1} = -\frac{\langle x^{n+1}, P_j \rangle_w}{\langle P_j, P_j \rangle_w}$, $j = 0, \dots, n$.

$\Rightarrow P_{n+1}$ exists and is uniquely defined given $P_0, \dots, P_n$. ∎

Examples = (i) $w(x) = \mathbb{1}(x \in [-1,1])$ leads to LEGENDRE POLYNOMIALS

$$P_n(x) = \frac{1}{2^n n!} \left[(x^2-1)^n \right]^{(n)}$$

((n): n-th derivative)

(ii) $w(x) = e^{-x^2}$ leads to HERMITE POLYNOMIALS

$$P_n(x) = (-1)^n e^{x^2} \left[e^{-x^2} \right]^{(n)}$$

(iii) $w(x) = e^{-x}\mathbb{1}(x \geq 0)$ leads to LAGUERRE POLYN.

$$P_n(x) = \frac{e^x}{n!} \left[x^n e^{-x} \right]^{(n)}$$

(iv) $w(x) = \frac{1}{\sqrt{1-x^2}}\mathbb{1}(x \in [-1,1])$ leads to TCHEBYSHEV POLYNOMIALS

The closed form expressions are not necessarily the most convenient ones. Alternatively, there exists a recursive formula that can be used to compute more efficiently these polynomials.

Fact = For any weight function w satisfying condition [W], the unitary polynomial sequence $\{P_n\}_{n\geq 0}$ satisfies the recursion:

$$\forall n \geq 2 \qquad P_n(x) = (x - \lambda_n) P_{n-1}(x) - \mu_n P_{n-2}(x)$$

where

$$\lambda_n = \frac{\langle x P_{n-1}, P_{n-1} \rangle_w}{\langle P_{n-1}, P_{n-1} \rangle_w}, \quad \mu_n = \frac{\|P_{n-1}\|_w^2}{\|P_{n-2}\|_w^2},$$

and $P_0(x) = 1$

$$P_1(x) = x - \frac{\int u\, w(u)\, du}{\int w(u)\, du}$$

Check that the result holds for $n = 2$, and then for $n \geq 3$, making use of the fact that for any polynomial Q of degree at most n, $\langle Q, P_j \rangle_w = 0 \quad \forall j \geq n+1$.

Theorem: Let $\{P_n\}_{n\geq 0}$ be the sequence of orthogonal polynomials associated to a weight function w satisfying [W].

Put $\varphi_n(x) := \dfrac{P_n(x)}{\|P_n\|_w}$, $n \geq 0$.

Then, the kernel $K(x) := \sum_{j=0}^{\ell} \varphi_j(0)\,\varphi_j(x)\,w(x)$ satisfies condition [K].

proof = x Property (i). We need to check that for $k = 0, \dots, \ell$, $\ell + \alpha$, holds $\int |x^k K(x)|\,dx < +\infty$.

Since w satisfies [W], $\forall k \geq 0$ $\int |x|^k w(x)\,dx < +\infty$

↓ Linearity

$\forall$ polynomial P, $\int P(x)\,w(x)\,dx < +\infty$

We can write

$$\int |x^k K(x)|\,dx = \int \underbrace{\left| x^k \sum_{j=0}^{\ell} \varphi_j(0)\varphi_j(x) \right|}_{\text{polynomial function}} w(x)\,dx < +\infty$$

x Property (ii)

$$\int K(x)\,dx = \sum_{j=0}^{\ell} \varphi_j(0) \int \varphi_j(x)\,w(x)\,dx \qquad (\ell \geq 1)$$

$$= \underbrace{\varphi_0(0) \int \varphi_0(x)\,w(x)\,dx}_{=} + \sum_{j=1}^{\ell} \varphi_j(0) \underbrace{\int 1 \cdot \frac{P_j(x)}{\|P_j\|_w}\,w(x)\,dx}_{\frac{1}{\|P_j\|_w}\langle P_0, P_j\rangle_w = 0}$$

$$= \frac{1}{\|P_0\|_w} \int \frac{1}{\|P_0\|_w}\,w(x)\,dx = \frac{1}{\|P_0\|_w^2} \int w(x)\,dx = 1$$

x Property (iii)

For $k = 1, \dots, \ell$, write $x^k = \sum_{j=0}^{k} \alpha_j^k P_j(x)$ for some α_j^k.

We have

$$\int x^k K(x)\,dx = \int \left(\sum_{j=0}^{k} \alpha_j^k P_j(x) \right) \left(\sum_{n=0}^{\ell} \varphi_n(0)\,\varphi_n(x)\,w(x) \right) dx$$

$$= \sum_{j=0}^{k} \sum_{n=0}^{\ell} \alpha_j^k \varphi_n(0) \int P_j(x)\,\varphi_n(x)\,w(x)\,dx$$

$$= \sum_{j=0}^{k} \sum_{n=0}^{\ell} \alpha_j^k \frac{\varphi_n(0)}{\|P_n\|_w} \underbrace{\int P_j(x)\,P_n(x)\,w(x)\,dx}_{= \langle P_j, P_n\rangle_w = 0 \text{ if } j \neq n}$$

$$= \sum_{j=0}^{k} \alpha_j^k \frac{\varphi_j(0)}{\|P_j\|_w} \|P_j\|_w^2$$

$$= \sum_{j=0}^{k} \alpha_j^k \underbrace{\varphi_j(0)\,\|P_j\|_w}_{= P_j(0)}$$

$$= \sum_{j=0}^{k} \alpha_j^k P_j(0) = 0^k$$

↑ definition of x^k. ■

• Example = For $w(x) = \mathbb{1}(x \in [-1,1])$ (Legendre polynomials), we have

$$\varphi_0(x) = 1/\sqrt{2}$$

$$\varphi_1(0) = 0$$

$$\varphi_2(x) = \sqrt{\frac{5}{2}}\left(\frac{3x^2-1}{2}\right)$$

$\Rightarrow K(u) = \left(\frac{9}{8} - \frac{15}{8}u^2\right)\mathbb{1}(u \in [-1,1])$ satisfies [K] with $\ell = 2$

II.4. Higher dimensions ($d \geq 2$).

In this section, we mention how the results established in dimension 1 generalize to higher dimensions ($d \geq 2$).

→ $X_1, \cdots, X_n \in \mathbb{R}^d$ have density $f_X : \mathbb{R}^d \to [0, +\infty)$.

→ The kernel density estimator is

$$\hat{f}_n(x) := \frac{1}{n h^d} \sum_{i=1}^{n} \mathbb{K}\left(\frac{x - X_i}{h} \right), \quad h > 0,$$

where $\mathbb{K} : \mathbb{R}^d \to \mathbb{R}$.

(for $x = (x_1, \cdots, x_d) \in \mathbb{R}^d$, take for example $\mathbb{K}(x) := K(x_1) \times \cdots \times K(x_d)$, for univariate kernel K.)

In this context, the bias-variance decomposition still holds, and

- $\int_{\mathbb{R}^d} \sigma_n^2(x)\, dx \leq \frac{1}{n h^d} \int_{\mathbb{R}^d} \mathbb{K}^2(u)\, du$, $\forall n \geq 1$, $\forall h > 0$.

  ↖ Same as in dimension 1. Requires $\int_{\mathbb{R}^d} \mathbb{K}^2(u)\, du < +\infty$. The proof is omitted.

- $\int_{\mathbb{R}^d} b_n^2(x)\, dx \leq C h^{2\beta}$

  ↖ To establish this result, we need to introduce a generalized version of Nikol'ski class, in terms of the smoothness of the partial derivatives of the density f_X. In particular, f_X should have continuous partial derivatives up to order $\lfloor \beta \rfloor$ (integer part of β). In addition, the kernel $\mathbb{K}$ should be such that $\int_{\mathbb{R}^d} |u^s \mathbb{K}(u)|\, du < +\infty$, $|s| = s_1 + \cdots + s_d \leq \lfloor \beta \rfloor$, $u^s = u_1^{s_1} \times \cdots \times u_d^{s_d}$,

and $\forall 1 \leq |s| \leq \ell$, $\int_{\mathbb{R}^d} u^s \mathbb{K}(u)\, du = 0$.

The main message here is that combining the bounds on the integrated square bias and variance, the choice of a bandwidth $h = O\left(n^{-\frac{1}{2\beta+d}}\right)$ leads to a mean integrated square error smaller than $C\, n^{-\frac{2\beta}{2\beta+d}}$, for some $C > 0$.

↑ The impact of dimension on convergence rates: for fixed n and β, the upper bounds becomes larger as d increases.

Remark: This result is optimal in some sense: there exists some f_X for which the required conditions holds (belongs to the multivariate Nikol'ski class, etc), and for which $\mathrm{MISE}(\hat{f}) \geq C\, n^{-\frac{2\beta}{2\beta+d}}$, $C > 0$, for any estimator $\hat{f}$ of f_X.

"Density estimation is more challenging in higher dimensions".

II.5. Bandwidth selection.

So far, the choice of the bandwidth h is made based on regularity assumptions on the unknown density f_X. For example, we derived on page 17 the optimal bandwith for a gaussian density, minimising the asymptotic MISE; it is given by $h \approx 1.06\, \hat{\sigma}\, n^{-1/5}$, where $\hat{\sigma}$ is an estimator of the population variance.

We discuss in this section two techniques for bandwidth selection, which are data driven: Leave One Out Cross Validation (LOOCV), and Biased Cross Validation (BCV).

× Leave One Out Cross Validation (LOOCV)

Recall that

$$\mathrm{MISE}(\hat{f}_n) = \mathbb{E}\left\{ \int (\hat{f}_n(x) - f_x(x))^2 dx \right\} \quad \text{(page 10)}$$

$$= \mathbb{E} \int \hat{f}_n^2(x)\,dx - 2\,\mathbb{E} \int \hat{f}_n(x) f_x(x)\,dx + \boxed{\int f_x^2(x)\,dx}$$

$\boxed{\int f_x^2(x)\,dx}$: independent of h

⇒ Minimizing the MISE with respect to h is equivalent to minimizing

$$\mathbb{E}\left\{ \int \hat{f}_n^2(x)\,dx - 2 \int \hat{f}_n(x) f_x(x)\,dx \right\}$$

↑ unknown, since it depends on $f_x(x)$.

A strategy consists in observing that if one is able to construct a quantity $\mathrm{LOOCV}(h)$ such that

$$\mathbb{E}\{\mathrm{LOOCV}(h)\} = \mathbb{E}\left\{ \int \hat{f}_n^2(x)\,dx - 2 \int \hat{f}_n(x) f_x(x)\,dx \right\},$$

then a reasonable choice for h would be

$$h_{\mathrm{LOOCV}} \in \underset{h>0}{\operatorname{argmin}}\ \mathrm{LOOCV}(h)$$

⇒ $\mathbb{E}\left\{ \int \hat{f}_n^2(x)\,dx \right\} - 2\,\mathbb{E}\left\{ \int \hat{f}_n(x) f_x(x)\,dx \right\}$

Take simply $\int \hat{f}_n^2(x)\,dx$

$= \mathbb{E}\left\{ \hat{f}_n(x) \mid X_1, \dots, X_n \right\}$

To construct an unbiased estimator of this quantity, we cannot use directly the sample sample twice: $\frac{1}{n}\sum_{i=1}^n \hat{f}_n(X_i)$

The reason is that $\hat{f}_n$ and X_i are not independent

⇒ Consider instead:

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

↑ "Leave observation i out"

and take

$$\left\{ \frac{1}{n} \sum_{i=1}^n \hat{f}_n^{-i}(X_i) \right\}$$ as an unbiased estimator of $\int \hat{f}_n(x) f_x(x)\,dx$

$\hat{f}_n^{-i}(X_i)$: independent.

- **Summary:** $h_{\mathrm{LOOCV}} \in \underset{h>0}{\operatorname{argmin}}\ \mathrm{LOOCV}(h)$,

where

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

with

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

LOOCV bandwith selector

↑ Can have more than one local minimum

Big minus = computationally demanding.

x Biased Cross Validation (BCV)

BCV is based on the expression of the AMISE (p. 17) :

$$\text{AMISE}(\hat{f}_n) = \frac{R(K)}{nh} + \frac{1}{4} h^4 \mu_2^2(K) \underbrace{R(f'')}$$

A natural estimator of this term would be $R(\hat{f}_n'')$, where

$$\hat{f}_n''(x) = \frac{1}{nh^3} \sum_{i=1}^{n} K''\left(\frac{x - X_i}{h}\right).$$

However, this estimator is a biased estimator of $R(f'')$.

Under some regularity conditions on the density f_X, and under some conditions on the kernel K, Scott & Terrell (1987) (Lemma 3.2 p. 9) established that

$$\mathbb{E}\{R(\hat{f}_n'')\} = R(f'') + \frac{R(K'')}{nh^5} + O(h^2).$$

$\Rightarrow$ Take $R(\hat{f}_n'') - \frac{R(K'')}{nh^5}$ as an estimator of $R(f'')$, plug this expression into $\text{AMISE}(\hat{f}_n)$, and select the value of $h>0$ minimizing it.

- Summary : $h_{BCV} \in \underset{h>0}{\text{argmin}}\ BCV(h)$,

where

$$BCV(h) := \frac{R(K)}{nh} + \frac{1}{4} h^4 \mu_2^2(K) \hat{R}_n(f''),$$

with

$$\hat{R}_n(f'') := R(\hat{f}_n'') - \frac{R(K'')}{nh^5}.$$

- Proof of (*)

We have $\hat{f}_n''(x) = \frac{1}{nh^3} \sum_{i=1}^{n} K''\left(\frac{x - X_i}{h}\right)$.

Thus

$$R(\hat{f}_n'') = \int \left[\hat{f}_n''(x)\right]^2 dx$$

$$= \frac{1}{n^2 h^6} \sum_{i=1}^{n} \int \left[K''\left(\frac{x-X_i}{h}\right)\right]^2 dx + \frac{1}{n^2 h^6} \sum_{j \neq i} \int K''\left(\frac{x-X_i}{h}\right) \times K''\left(\frac{x-X_j}{h}\right) dx$$

$$\mathbb{E}\{R(\hat{f}_n'')\} = \frac{1}{nh^6} \mathbb{E}\left\{\int \left[K''\left(\frac{x-X}{h}\right)\right]^2 dx\right\} + \frac{n-1}{nh^6} \mathbb{E}\left\{\int K''\left(\frac{x-X_1}{h}\right) K''\left(\frac{x-X_2}{h}\right) dx\right\}$$

$$= \frac{R(K'')}{nh^5} + \frac{n-1}{nh^4} \int \left(\int K''(u) f(x-hu)\, du\right)^2 dx$$

Expand $f(x-hu) = f(x) - hu f'(x) + \frac{(hu)^2}{2} f''(x) - \frac{1}{6}(hu)^3 f'''(x) + O(h^4)$

Under the assumption that K is such that K is compactly supported on $[-1,1]$, $K(\pm 1) = 0$, $K'(\pm 1) = 0$, K symmetric, we see that

$$\int u K''(u)\, du = \int u^3 K''(u)\, du = 0 \quad ; \quad \int u^2 K''(u)\, du = 2,$$

and so on, so that

$$\int K''(u) f(x - hu)\, du = h^2 f''(x) + O(h^4), \text{ and we get}$$

$$\mathbb{E}\{R(\hat{f}_n'')\} = \frac{R(K'')}{nh^5} + R(f'') + O(h^2). \quad \blacksquare$$

x Comparison of h_{LOOCV} and h_{BCV}

Bandwith h_{LOOCV} and h_{BCV} have asymptotic normal properties. Since both are based on minimizing the MISE, we compare their performance with the "optimal" one :

$$h_{MISE} := \underset{h>0}{\operatorname{argmin}} \; MISE(\hat{f}_n),$$

and present asymptotic distributions for $\frac{h_{LOOCV}}{h_{MISE}}$ & $\frac{h_{BCV}}{h_{MISE}}$.

Under certain regularity conditions, we have

- $n^{1/10}\left(\frac{h_{LOOCV}}{h_{MISE}} - 1\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{LOOCV})$

(Hall & Marron (1987) ; Scott & Terrel (1987))

- $n^{1/10}\left(\frac{h_{BCV}}{h_{MISE}} - 1\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{BCV})$

(Scott & Terrel (1987))

The ratio of the asymptotic variances for the gaussian kernel is $\sigma^2_{LOOCV} / \sigma^2_{BCV} \approx 15.7$.

$\Rightarrow$ Expect h_{LOOCV} to be more variable than h_{BCV}.
" h_{BCV} is more stable than h_{LOOCV} ".

However, h_{BCV} is known to be a biased estimator of h_{MISE}.
The usual bias-variance tradeoff.

Remark: The relative rate of convergence $n^{-1/10}$ of the bandwith selectors to h_{MISE} is far from optimal, as Hall & Marron (1991) showed that it can go as fast as $n^{-1/2}$.

REFERENCES

- A.B. Tsybakov (2009). Introduction to non-parametric estimation. Springer.
- M.P. Wand & M.C. Jones (1995). Kernel Smoothing. Chapman & Hall.
- P. Hall & J.S. Marron (1987). Extent to which least-squares Cross Validation Minimizes Integrated Square Error in Non-Parametric Density Estimation. Proba. Theory Rel. Fields 74, p. 567-581
- P. Hall & J.S. Marron (1991). Lower Bounds for Bandwidth Selection in Density Estimation. Proba Theory Rel. Fields 90, p. 149-173.
- D.W. Scott & G.R. Terrel (1987). Biased and Unbiased Cross-Validation in Density Estimation. JASA 82, p. 1131-1146.