STAT 824 sp 2023 Lec 02 slides

Kernel density estimation

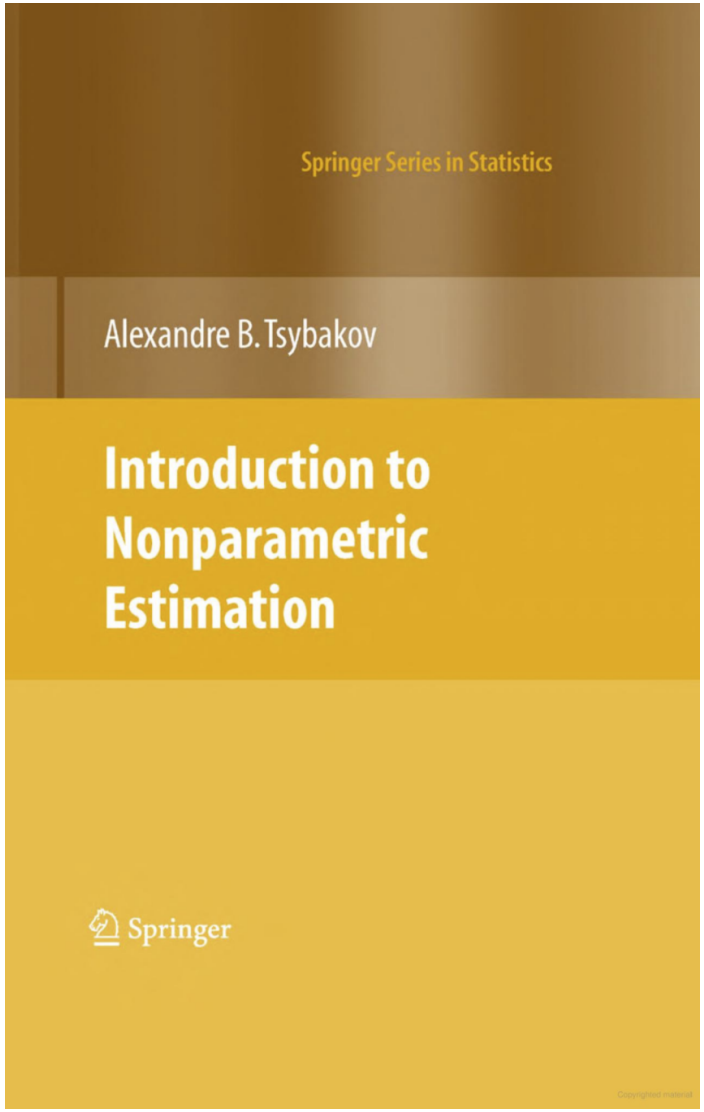
Karl B. Gregory

University of South Carolina

These slides are an instructional aid; their sole purpose is to display, during the lecture, definitions, plots, results, etc. which take too much time to write by hand on the blackboard. They are not intended to explain or expound on any material.

Table of Contents

Much of this lecture comes from Chapter 1 of the book [2]:



Let $X_1, \ldots, X_n$ be a rs with cdf F.

If F has continuous derivative F', then the pdf is $f = F'$. So

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

Rosenblatt estimator

The *Rosenblatt* estimator of f is given by

$$\hat{f}_n(x) = \frac{\hat{F}_n(x + h) - \hat{F}_n(x - h)}{2h} \quad \text{for all } x \in \mathbb{R}$$

for some small $h > 0$.

Exercise: Put this estimator into the form

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

Exercise: Check whether the Rosenblatt estimator gives a legitimate pdf.

Kernel density estimator (KDE)

We generalize the Rosenblatt estimator with the *KDE*, given by

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

where $K(\cdot)$ is the *kernel* function and h is the *bandwidth*.

Some examples of kernel functions:

$$K(u) = (1 - |u|) \cdot \mathbf{1}(|u| \leq 1)$$
$$K(u) = 3/4 \cdot (1 - u^2) \cdot \mathbf{1}(|u| \leq 1)$$
$$K(u) = (2\pi)^{-1/2} e^{-u^2/2}$$

Discuss: What assumptions are needed about $K(\cdot)$ for $\hat{f}_n$ to be a pdf?

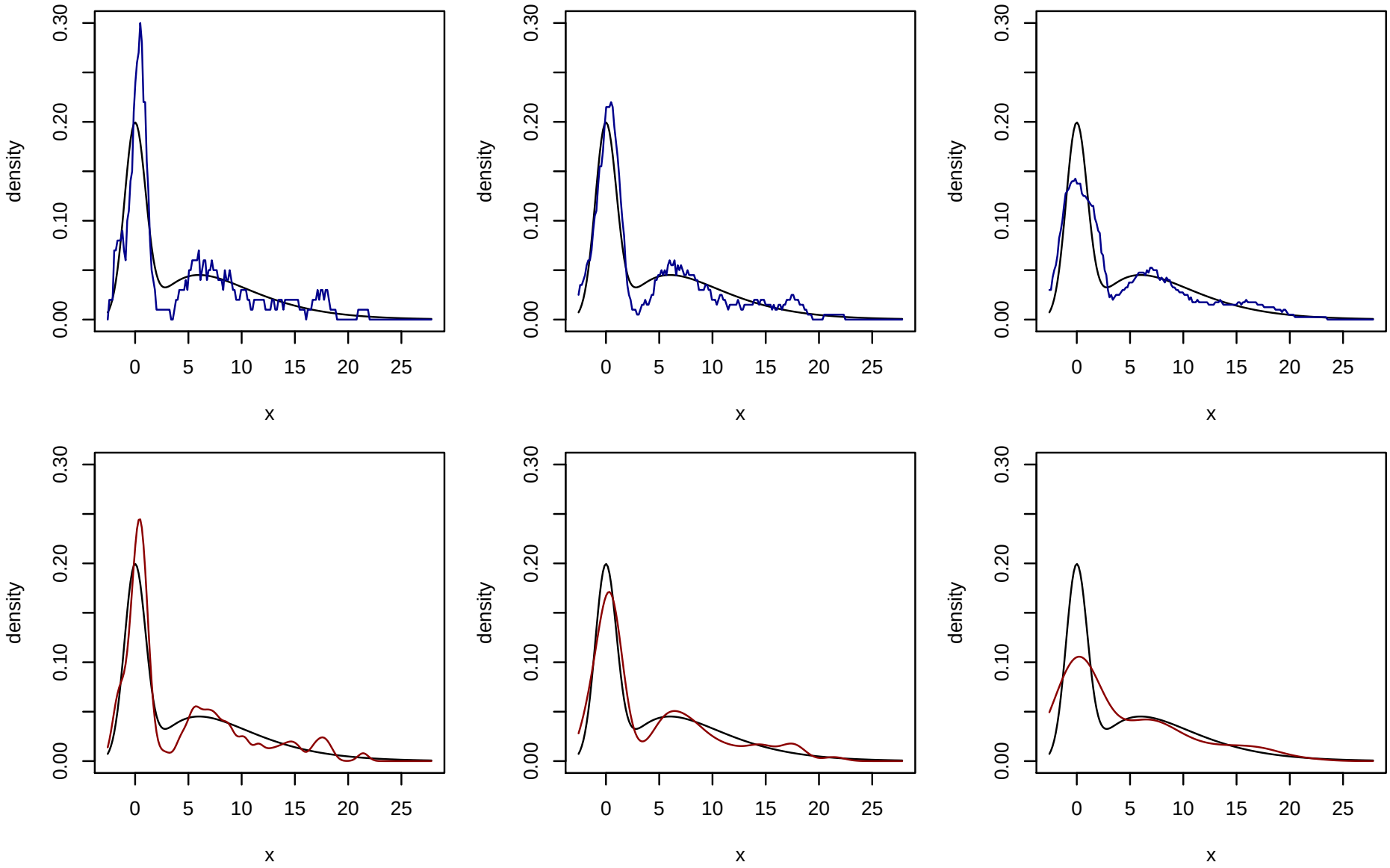
Assumptions under which the KDE $\hat{f}_n$ is a legitimate pdf

(K0) $K(u) \geq 0$ for all $u \in \mathbb{R}$.

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

Exercise: Verify that the kernels on the previous slide satisfy these assumptions.

Exercise: Generate some data and plot the KDE for different K and h.



Consider MSE $\hat{f}_n(x_0) = \underbrace{b_n^2(x_0)}_{\text{bias}} + \underbrace{\sigma_n^2(x_0)}_{\text{variance}}$ at a point $x_0 \in \mathbb{R}$.

Assumptions that allow us to bound the variance

(F1) There exists $f_{\max} > 0$ such that $f(x) \leq f_{\max} < 0$ for all $x \in \mathbb{R}$.

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

Bound for the variance of $\hat{f}_n(x_0)$

Under (F1) and (K2) we have

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

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

Exercise: Prove the above.

Bounding the bias $b_n(x_0)$ is more complicated; must consider smoothness of f.

Lipschitz class of functions

For an interval $T \subset \mathbb{R}$ and $L > 0$, the *Lipschitz class* of functions Lipschitz(L) on T is the set of functions $f : T \to \mathbb{R}$ satisfying

$$|f(x) - f(x')| \leq L|x - x'| \text{ for all } x, x' \in T.$$

Exercise: Check whether $f \in$ Lipschitz(L) on $\mathbb{R}$ for some L:

1. $f(x) = |x|$
2. $f(x) = \sin(x)$
3. $f(x) = x^2$
4. $f(x) = e^x$

Let $\mathcal{P}_{\mathcal{L}}(L)$ denote the set of densities in Lipschitz(L) on $\mathbb{R}$, that is, let

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

We also need another assumption on the kernel to bound the bias:

$$(\text{K3}^*) \quad \int_{\mathbb{R}} |u||K(u)|du \leq \kappa_1 < \infty.$$

Bound for the bias of $\hat{f}_n(x_0)$

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

$$|b_n(x_0)| \leq h \cdot L \cdot \kappa_1$$

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

Exercise: Prove the above.

Bound for the MSE of $\hat{f}_n(x_0)$

Under (K1*), (K2), (K3*), and (F1), and if $f \in \mathcal{P}_{\mathcal{L}}(L)$, we have

$$\text{MSE}\, \hat{f}_n(x_0) \leq h^2 \cdot L^2 \cdot \kappa_1^2 + \frac{1}{nh} \cdot f_{\max} \cdot \kappa^2$$

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

Follows from bounds on the bias and variance.

Exercise:

- Find the optimal bandwidth h_{opt} (that minimizes the MSE bound)
- Give the MSE bound under h_{opt}

We now consider a more general class of smooth functions.

Hölder class of functions

For an interval $T \subset \mathbb{R}$, $\beta > 0$ an integer, and $L > 0$, the *Hölder class* of functions $\mathcal{H}(\beta, L)$ on T is the set functions $f : T \to \mathbb{R}$ with $\ell = \beta - 1$ derivatives such that $f^{(\ell)}$ satisfies

$$|f^{(\ell)}(x) - f^{(\ell)}(x')| \leq L|x - x'| \text{ for all } x, x' \in T.$$

Can have Hölder classes with non-integer β, but we ignore these.

Exercise: Check whether $f \in \mathcal{H}(\beta, L)$ for some β, L for

1. $f(x) = |x|$
2. $f(x) = e^x$
3. $f(x) = x^2 \mathbf{1}(0 \leq x < 1/3) + [2x - 2x^2 - 1/3]\mathbf{1}(1/3 \leq x < 2/3)$
   $+(1 - x)^2 \mathbf{1}(2/3 \leq x < 1)$

Let $\mathcal{P}_{\mathcal{H}}(\beta, L)$ denote the set of densities in $\mathcal{H}(\beta, L)$ on $\mathbb{R}$, that is, let

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

To accommodate the Hölder class of functions, we need the following definition:

Kernel of order ℓ

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

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

Exercise: Check whether these are kernels of some order $\ell \geq 1$:

$$K(u) = (1/2) \cdot \mathbf{1}(|u| \leq 1)$$
$$K(u) = (1 - |u|) \cdot \mathbf{1}(|u| \leq 1)$$
$$K(u) = 3/4 \cdot (1 - u^2) \cdot \mathbf{1}(|u| \leq 1)$$
$$K(u) = (2\pi)^{-1/2} e^{-u^2/2}$$

We now analyze the bias when $f \in \mathcal{P}_{\mathcal{H}}(\beta, L)$. But first:

Updated assumptions for bounding the bias—when $f \in \mathcal{P}_{\mathcal{H}}(\beta, L)$

(K1) K is a kernel of order ℓ

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

Bound for the bias of $\hat{f}_n(x_0)$

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

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

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

Exercise: Prove the above.

Bound for the MSE of $\hat{f}_n(x_0)$

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

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

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

Follows from bounds on the bias and variance.

Exercise: Show that the optimal bound in the above result is of the form

$$\text{MSE}\, \hat{f}_n(x_0) \leq C \cdot n^{-\frac{2\beta}{2\beta+1}} \quad \text{for all } x_0 \in \mathbb{R}.$$

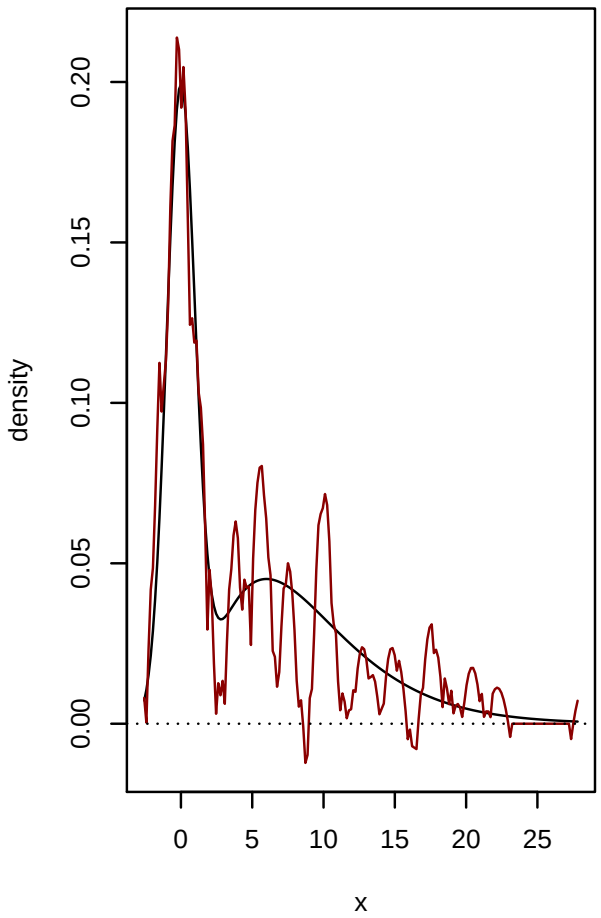
Kernels of order $\ell \geq 2$ must take negative values on a set of positive Leb. measure!

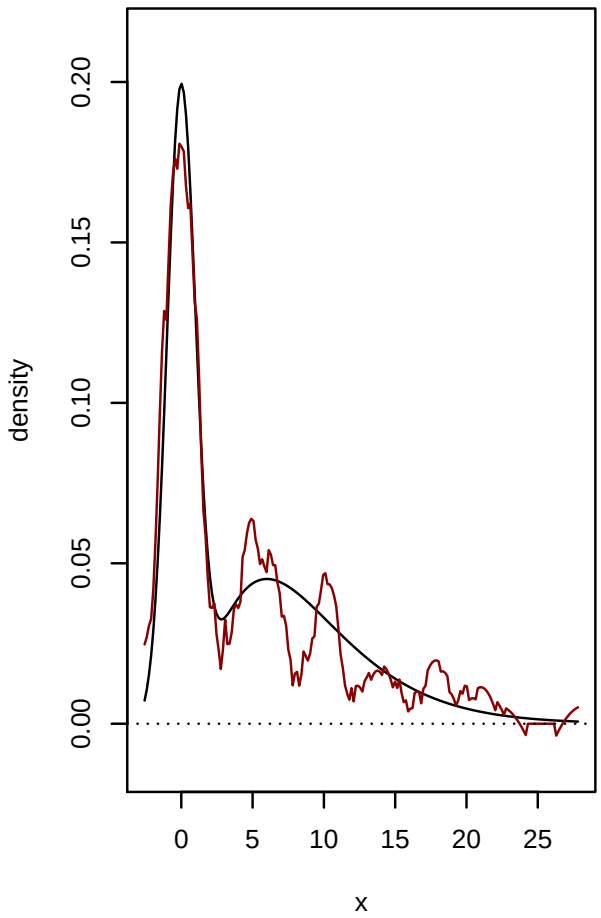
Why??

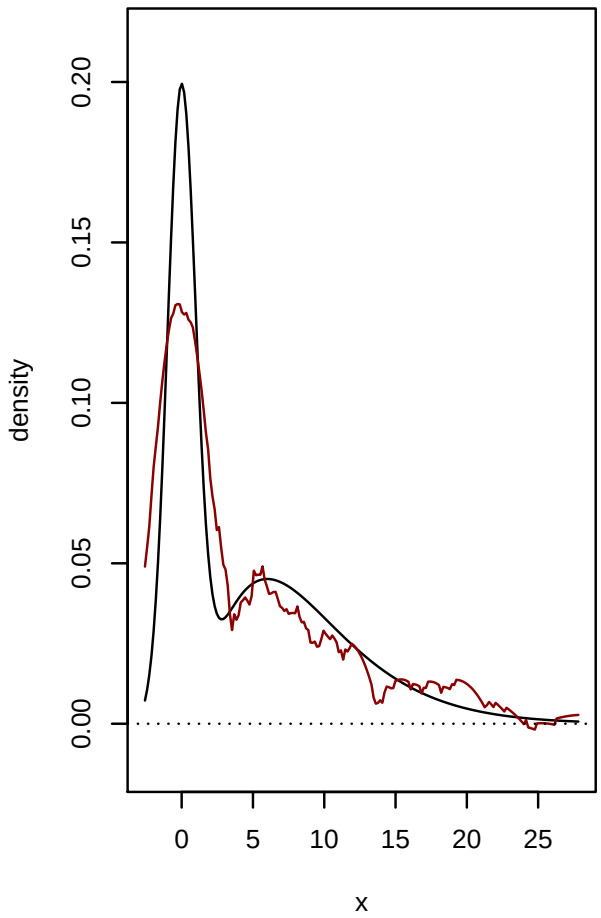
Exercise: Consider the kernel

$$K(u) = \left(\frac{9}{8} - \frac{15}{8}u^2\right) \mathbf{1}(|u| \leq 1).$$

1. Show that this is a kernel of order 2.
2. Generate some data and use this as the kernel in a KDE. Plot results.
3. Discuss using $\hat{f}_n^+(x) = \max\{0, \hat{f}_n(x)\}$.


density
0.00
0.05
0.10
0.15
0.20
0
5
10
15
20
25
x


density
0.00
0.05
0.10
0.15
0.20
0
5
10
15
20
25
x


density
0.00
0.05
0.10
0.15
0.20
0
5
10
15
20
25
x

Theorem 1: Uniform bound for MSE $\hat{f}_n(x)$ over $x \in \mathbb{R}$, $f \in \mathcal{P}_{\mathcal{H}}(\beta, L)$

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

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

where $C > 0$ is a constant depending only on β, L, α, and the kernel K.

Where is (F1)? Can show $\exists\ f_{\max} < \infty$ s.t. $\sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta,L)} \sup_{x \in \mathbb{R}} f(x) \leq f_{\max}$.

The $n^{-\frac{2\beta}{2\beta+1}}$ is the *rate of convergence* of $\hat{f}_n(x)$.

Discuss:

1. Can there be an estimator with a better rate of convergence?
2. How does the smoothness of the function class affect the rate?

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

$$\text{MISE}\, \hat{f}_n = \mathbb{E} \int_{\mathbb{R}} [\hat{f}_n(x) - f(x)]^2 dx = \underbrace{\int_{\mathbb{R}} b^2(x) dx}_{\text{bias term}} + \underbrace{\int_{\mathbb{R}} \sigma^2(x) dx}_{\text{variance term}}.$$

Bound on MISE variance term

Under (K2) we have

$$\int_{\mathbb{R}} \sigma^2(x) dx \leq \frac{1}{nh} \kappa^2.$$

Exercise: Prove the above.

Tsybakov presents a bound on the bias term under the following function class:

Nikol'ski class of functions

For $\beta > 0$ an integer, $L > 0$, 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 $\ell = \lfloor \beta \rfloor$ exist and satisfy

$$\left(\int_{\mathbb{R}} [f^{(\ell)}(x+t) - f^{(\ell)}(x)]^2 dx \right)^{1/2} \leq L|t| \quad \text{for all } t \in \mathbb{R}.$$

Similar to $\mathcal{H}(\beta, L)$, but enables bounding an L_2-norm appearing in the bias term.

Let $\mathcal{P}_{\mathcal{N}}(\beta, L)$ denote the set of densities in $\mathcal{N}(\beta, L)$, that is, let

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

Bound on MISE bias term

Under (K1) and (K3), if $f \in \mathcal{P}_{\mathcal{N}}(\beta, L)$, then

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

For the proof see pg. 14 of [2] or the handwritten notes.

Putting together the bounds on the bias and variance terms gives:

Bound on MISE

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

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

We conclude this section with a uniform result over $\mathcal{P}_{\mathcal{N}}(\beta, L)$.

Theorem 2: Uniform bound for MISE over $f \in \mathcal{P}_{\mathcal{N}}(\beta, L)$

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

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

where $C > 0$ is a constant depending only on β, L, α, and the kernel K.

The Sheather-Jones bandwidth selection method is based on this result [1]:

Theorem 3: "Usual" asymptotic expansion of MISE $\hat{f}_n$

If K is a kernel of order 1 with

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

and if f differentiable on $\mathbb{R}$, f' a.c. on $\mathbb{R}$, $\|f''\|_2^2 = \int_{\mathbb{R}}[f''(x)]^2dx < \infty$, then

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

Proof on pg. 192 of Tsybakov [2].

SJ propose to plug in an estimate of $\|f''\|_2^2$. In R: `density(x, bw = "SJ")`

Exercise: Find optimal bandwidth in terms of σ_K, κ^2, and $\|f''\|_2^2$.

Let the *leave-one-out crossvalidation estimators* of f be given by

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

and define the function

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

Then a leave-one-out crossvalidation choice of h is

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

Exercise: Derive the above by expanding $\mathrm{MISE}_h\, \hat{f}_n$.

CV estimate of MISE

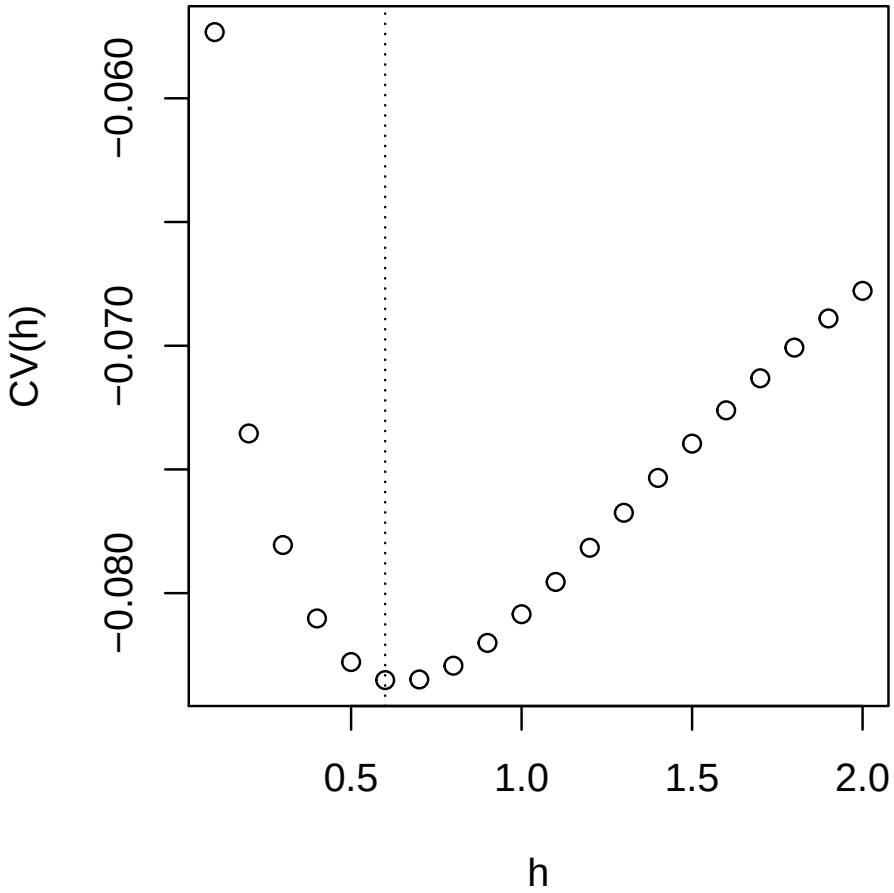
Assume that $K : \mathbb{R} \to \mathbb{R}$ and f satisfy

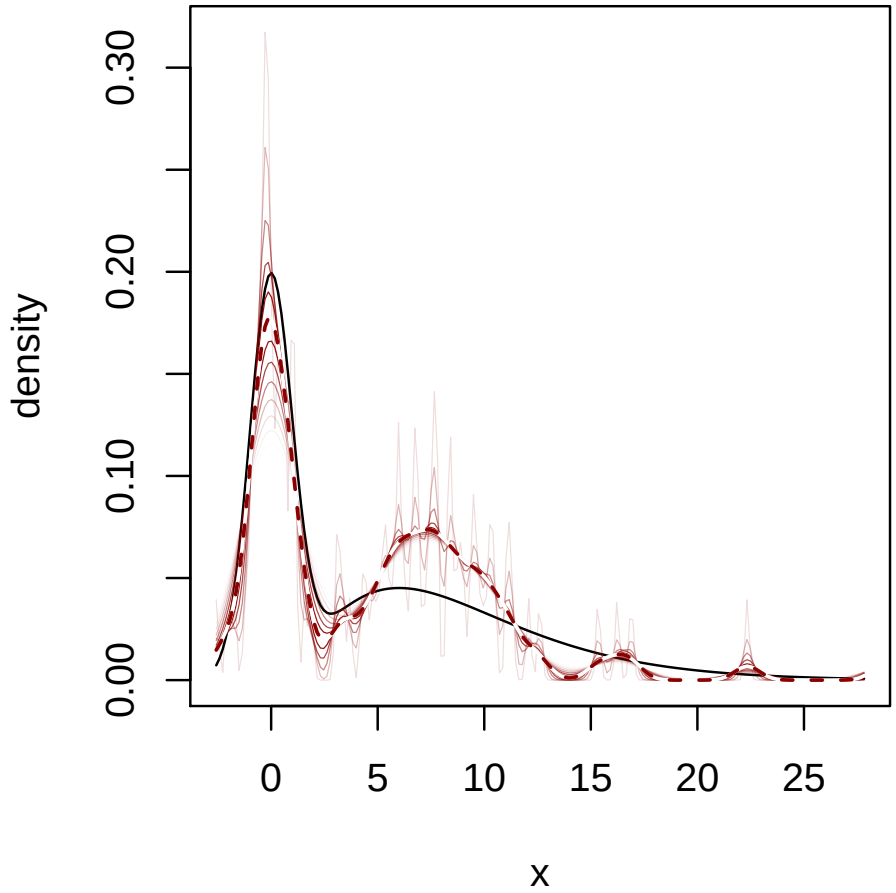
$$\int_{\mathbb{R}} f^2(x)dx < \infty \quad \text{and} \quad \int_{\mathbb{R}} \int_{\mathbb{R}} \left| K\left(\frac{z-x}{h}\right) \right| f(z)dz \; f(x)dx < \infty$$

for all $h > 0$. Then

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

Exercise: Prove the above result.


CV(h)
−0.060
−0.070
−0.080
0.5
1.0
1.5
2.0
h


density
0.30
0.20
0.10
0.00
0
5
10
15
20
25
x


Simon J Sheather and Michael C Jones.
A reliable data-based bandwidth selection method for kernel density estimation.
Journal of the Royal Statistical Society: Series B (Methodological), 53(3):683–690, 1991.


Alexandre B Tsybakov.
Introduction to nonparametric estimation.
Springer Science & Business Media, 2008.