Lecture 2

Recall the definition

Definition 1.1. Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$ with a given density $F' = f$. A **kernel density estimator** for f is defined via

$$\widehat{f}(x; h) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \; x \in \mathbb{R}, \; h > 0.$$

Thereby $K : \mathbb{R} \to \mathbb{R}$, such that $\int_{-\infty}^{\infty} K(x)dx = 1$ is known as **kernel** and $h > 0$ is called **bandwidth**. A **jth moment** of a kernel K is defined as $\mu_j = \int_{-\infty}^{\infty} x^j K(x)dx$.

Some classical kernels:

1. $K(x) = 0.5\, \mathbb{I}(|x| \leq 1)$ (the rectangular kernel)
2. $K(x) = (1 - |x|)\, \mathbb{I}(|x| \leq 1)$ (the triangular kernel)
3. $K(x) = 0.75(1 - x^2)\, \mathbb{I}(|x| \leq 1)$ (the Epanechnikov kernel)
4. $K(x) = (2\pi)^{-1/2} \exp(-x^2/2)$ (the Gaussian kernel)
5. $K(x) = 0.5 \exp(-|x|/\sqrt{2}) \sin(|x|/\sqrt{2} + \pi/4)$ (the Silverman kernel)

To study the point-wise bias and variance of kernel density estimators, the following definitions are needed.

Definition 1.2. Let $\beta > 0$ and $L > 0$. The **Hölder class** $\Sigma(\beta, L)$ on T is defined as set of $\ell = \lfloor \beta \rfloor$ times differentiable functions $f : T \to \mathbb{R}$ whose ℓ-th derivative satisfies

$$\left| f^{(\ell)}(x) - f^{(\ell)}(y) \right| \leq L|x - y|^{\beta - \ell}, \;\; \forall x, y \in T$$

Definition 1.3. Let $\ell \geq 1$ be an integer. We say that $K : \mathbb{R} \to \mathbb{R}$ is a **kernel of order** ℓ, if the functions $x \mapsto x^j K(x)$, $j = 0, \ldots, \ell$ are integrable and satisfy

$$\int K(x)dx = 1, \quad \mu_j = \int x^j K(x)dx = 0, \;\; j = 1, \ldots, \ell.$$

Note that there is another definition of a kernel of order ℓ that is often used in the literature: a kernel K is said to be of order $\ell + 1$ (with an integer $\ell \geq 1$), if Definition 1.3 holds and $\int x^{\ell+1} K(x)dx \neq 0$.

Let us also define a class of densities that belong to a Hölder class.

$$\mathcal{F}(\beta, L) = \{f : \; f \geq 0, \int f(x)dx = 1 \text{ and } f \in \Sigma(\beta, L) \text{ on } \mathbb{R}\}.$$

Theorem 1.1. *Let* $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$, *where c.d.f.* F *has a Lebesgue density* f *and* $\widehat{f}(x;h) = (nh)^{-1}\sum_{i=1}^n K\{(X_i - x)/h\}$ *be a kernel density estimator.*

(i) Suppose that the density f *satisfies* $f(x) \le f_{max} < \infty$ *for all* $x \in \mathbb{R}$. *Let* $K : \mathbb{R} \to \mathbb{R}$ *be a function such that* $\int \{K(x)\}^2 dx < \infty$. *Then, for any* $x \in \mathbb{R}$, $h > 0$ *and* $n \ge 1$ *we have*

$$var\left\{\widehat{f}(x;h)\right\} \le \frac{C_1}{nh},$$

where $C_1 = f_{max}\int\{K(x)\}^2 dx$.

(ii) Assume that $f \in \mathcal{F}(\beta, L)$ *and that* K *is a kernel of order* $\ell = \lfloor \beta \rfloor$ *satisfying* $\int |x|^\beta K(x)dx < \infty$. *Then for all* $x \in \mathbb{R}$, $h > 0$ *and* $n \ge 1$ *we have*

$$\left|bias\left\{\widehat{f}(x;h)\right\}\right| = \left|E\{\widehat{f}(x;h)\} - f(x)\right| \le C_2 h^\beta,$$

where

$$C_2 = \frac{L}{\ell!}\int |x|^\beta |K(x)| dx.$$

Proof

(i) Let

$$Z_i(x) = K\left(\frac{X_i - x}{h}\right) - \mathrm{E}\left\{K\left(\frac{X_i - x}{h}\right)\right\}.$$

$Z_1(x), \ldots, Z_n(x)$ are i.i.d. random variables with zero mean and variance

$$\begin{aligned}
\mathrm{E}\{Z_i(x)\}^2 &= \mathrm{E}\left[\left\{K\left(\frac{X_i - x}{h}\right)\right\}^2\right] - \left[\mathrm{E}\left\{K\left(\frac{X_i - x}{h}\right)\right\}\right]^2 \\
&\le \mathrm{E}\left[\left\{K\left(\frac{X_i - x}{h}\right)\right\}^2\right] = \int \left\{K\left(\frac{u - x}{h}\right)\right\}^2 f(u)du \\
&\le f_{max} h \int \{K(x)\}^2 dx.
\end{aligned}$$

With this

$$\mathrm{var}\left\{\widehat{f}(x;h)\right\} = \mathrm{E}\left[\left\{\frac{1}{nh}\sum_{i=1}^n Z_i(x)\right\}^2\right] = \frac{1}{nh^2}\mathrm{E}\left[\{Z_1(x)\}^2\right] \le \frac{C_1}{nh}.$$

(ii)

$$\mathrm{bias}\left\{\widehat{f}(x;h)\right\} = \frac{1}{h}\int K\left(\frac{u - x}{h}\right) f(u)du - f(x) = \int K(u)\left\{f(x + uh) - f(x)\right\} du.$$

Now,

$$f(x + uh) = f(x) + f'(x)uh + \ldots + \frac{(uh)^\ell}{\ell!} f^{(\ell)}(x + \tau uh)$$

for some $\tau \in [0,1]$. Since K has order $\ell = \lfloor \beta \rfloor$, we obtain

$$\begin{aligned}
\text{bias}\left\{\widehat{f}(x;h)\right\} &= \int K(u)\left\{f^{'}(x)uh + \ldots + \frac{(uh)^\ell}{\ell!}f^{(\ell)}(x+\tau uh)\right\}du \\
&= \int K(u)\frac{(uh)^\ell}{\ell!}f^{(\ell)}(x+\tau uh)du \\
&= \int K(u)\frac{(uh)^\ell}{\ell!}\left\{f^{(\ell)}(x+\tau uh) - f^{(\ell)}(x)\right\}du.
\end{aligned}$$

With this,

$$\begin{aligned}
\left|\text{bias}\left\{\widehat{f}(x;h)\right\}\right| &\leq \int |K(u)|\frac{|uh|^\ell}{\ell!}\left|f^{(\ell)}(x+\tau uh) - f^{(\ell)}(x)\right|du \\
&\leq L\int |K(u)|\frac{|uh|^\ell}{\ell!}|\tau uh|^{\beta-\ell}du \leq C_2 h^\beta.
\end{aligned}$$

□

Hence, we observe the same bias-variance trade-off as in histogram estimation: the upper bound on the variance decreases with growing h, whereas the bound on the bias increases. The choice of a small h that corresponds to a large variance is called *undersmoothing*, while a large h leads to a large bias and *oversmoothing*.

If f and K satisfy assumptions of Theorem 1.1, then

$$\text{MSE}\left\{\widehat{f}(x;h)\right\} \leq \frac{C_1}{nh} + C_2^2 h^{2\beta}.$$

The right-hand side is minimized by

$$h_{MSE} = \left(\frac{C_1}{2\beta C_2^2}\right)^{\frac{1}{2\beta+1}} n^{-\frac{1}{2\beta+1}}$$

and

$$\text{MSE}\left\{\widehat{f}(x;h_{MSE})\right\} = \mathcal{O}\left(n^{-\frac{2\beta}{2\beta+1}}\right),$$

which holds uniformly for each x.

Let now discuss how kernels of order ℓ can be constructed.

Let $\{\phi_m\}_{m=0}^\infty$ be the orthonormal basis of Legendre polynomials in $L_2[-1,1]$ defined by the formulas

$$\phi_0(x) = \frac{1}{\sqrt{2}}, \quad \phi_m(x) = \sqrt{\frac{2m+1}{2}}\frac{1}{2^m m!}\frac{d^m\{(x^2-1)^m\}}{dx^m}, \quad m = 1, 2, \ldots,$$

for $x \in [-1, 1]$. In particular,

$$\int_{-1}^{1} \phi_m(x)\phi_k(x)dx = \delta_{mk} = \begin{cases} 1, & \text{if } m = k \\ 0, & \text{if } m \neq k \end{cases}.$$

Then, a function $K : \mathbb{R} \to \mathbb{R}$ defined via

$$K(x) = \sum_{m=0}^{\ell} \phi_m(0)\phi_m(x)\mathbb{I}(|x| \leq 1)$$

is a kernel of order ℓ. Indeed,

$$\begin{aligned} \int x^j K(x)dx &= \sum_{q=0}^{j}\sum_{m=0}^{\ell}\int_{-1}^{1} b_{qj}\phi_q(x)\phi_m(0)\phi_m(x)dx \\ &= \sum_{q=0}^{j} b_{qj}\phi_q(0) = \begin{cases} 1, & \text{if } j = 0 \\ 0, & \text{if } j = 1, \dots, \ell \end{cases}. \end{aligned}$$

Here we used that

$$x^j = \sum_{q=0}^{j} b_{qj}\phi_q(x), \quad \text{for all } x \in [-1, 1] \text{ and some } b_{qj} \in \mathbb{R},$$

since ϕ_q is a polynomial of degree q.

For example,

$$\phi_0(x) = \frac{1}{\sqrt{2}}, \quad \phi_1(x) = \sqrt{\frac{3}{2}}\, x, \quad \phi_2(x) = \sqrt{\frac{5}{2}}\, \frac{(3x^2 - 1)}{2}.$$

Then,

$$K(x) = \left(\frac{9}{8} - \frac{15}{8}x^2\right)\mathbb{I}(|x| \leq 1)$$

is a kernel of order 2.

Note that kernels of order ℓ constructed as above, are symmetric, that is $K(x) = K(-x)$, for all $x \in \mathbb{R}$. Indeed, $\phi_m(0) = 0$ for all odd m, while ϕ_m are symmetric functions for even m. By symmetry, for even ℓ, kernel K is of order $\ell + 1$. In particular, the above kernel of order 2 is also a kernel of order 3.