Econometrics 2: Non-parametric methods

Eric Gautier*

2024 Spring

I have a question!

*Notes by FU Zixuan, last compiled on February 21, 2024.

Contents

1 Preliminaries

1.1 Probability basics

Definition 1.1 (distribution law)**.** The distribution law of a random variable X is $\mathbb{P}_X$ is the probability on $\left(\mathbb{R}^d, \mathcal{B}\left(\mathbb{R}^d\right)\right)$ such that $\mathbb{P}_X[B] = \mathbb{P}[X \in B]$ for all $B \in \mathcal{B}\left(\mathbb{R}^d\right)$

Definition 1.2 (density)**.** X has density f_X if $\mathbb{P}_X[B] = \int_B \underbrace{f_X(x)dx}_{d\mathbb{P}_X(x)}$ for all $B \in \mathcal{B}\left(\mathbb{R}^d\right)$

Let Y be a random variable and X be a random vector in $\mathbb{R}^d$ defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We want to define and manipulate $\mathbb{E}[Y|X]$.

There are 2 particular cases

1. $\left(Y, X^\top\right)^\top$ has discrete support. Let $x \in \operatorname{spt}(X)$, then $\mathbb{E}[Y|X=x] = \sum y_j \mathbb{P}(Y = y_j \mid X = X)$. This is well defined only when $\mathbb{P}(X=x) > 0$ in which case $\mathbb{P}(Y = y_j \mid X = x) = \frac{\mathbb{P}(Y=y_j, X=x)}{\mathbb{P}(X=x)}$ is well defined.

2. $\left(Y, X^\top\right)^\top$ and X have a density then $\mathbb{E}[Y|X=x] = \int y f_{Y|X=x}(y)dy$ where $f_{Y|X=x}(y) = \frac{f_{Y,X}(y,x)}{f_X(x)}$.

Proposition 1.1 (conditional expectation)**.** The random variabel $Z \coloneqq \mathbb{E}[Y|X]$ is the unique random variable such that

1. $Z \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, that is Z is $\sigma(X)$-measurable.

2. $\mathbb{E}[Z \mathbb{1}_B] = \mathbb{E}[Y \mathbb{1}_B]$ for all $B \in \sigma(X)$.

unique means that if Z' is another random variable satisfying the same properties, then $Z = Z'$ a.s.

Remark. The random variable Z is $\sigma(X)$-measurable iff $Z = \phi(X)$ for some function $\phi : \left(\mathbb{R}^d, \mathcal{B}\left(\mathbb{R}^d\right)\right) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The corresponding function ϕ for $Z = \underbrace{E[Y|X]}_{\text{conditional expectation}}$ is denoted by $\underbrace{\mathbb{E}[Y|X=x]}_{\text{conditional expectation function}}$.

Remark. The proposition 2 is equivalent to

$$
\begin{aligned}
&\mathbb{E}[(Y-Z)\mathbb{1}_B] = 0, \quad \forall B \in \sigma(X) \\
\Leftrightarrow &\mathbb{E}[(Y-Z)\psi(X)] = 0, \quad \forall \psi \text{ bounded and meansurable.}
\end{aligned}
$$

1.2 Completeness condition

We want to understand the completeness condition when $X = Z - \eta$, where $Z \perp\!\!\!\perp \eta$ and both have densities. Recall the definition of **completeness**.

Definition 1.3 (Completeness)**.** Completeness is defined as such that

$$\forall z \in \mathbb{R}, \int_{\mathbb{R}} \varphi(x) f_\eta(z-x) dx = 0 \text{ implies that for all } x,\ \varphi(x) = 0,$$

where φ is continuous and $\int_{\mathbb{R}} |\varphi(x)|\, dx < \infty$.

Now, We make a detour to introduce some notations in function space.

Definition 1.4. Let f be a function defined on $\mathbb{R}^d$ with values in $\mathbb{R}$ or $\mathbb{C}$ and $p \leq 1$. Then $L^p(\mathbb{R}^d)$ is defined as the space of measurable function from $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ such that $\int_{\mathbb{R}^d} |f(x)|^p\, dx < \infty$. If f takes value from $\mathbb{C}$, $|\cdot|$ is the modulus.

Definition 1.5 ($L^1(\mathbb{R})$ space)**.** A function is in $L^1(\mathbb{R})$ if $\int_{\mathbb{R}} |f(x)|\, dx < \infty$.

Definition 1.6 (Fourier transform)**.** If $f \in L^1(\mathbb{R})$, the Fourier transform of f is defined for all $w \in \mathbb{R}$ by

$$\mathcal{F}[f](w) = \int_{\mathbb{R}} e^{iwx} f(x)\, dx.$$

Remark. Let $t \in \mathbb{R}$, $e^{it} = \cos(t) + i\sin(t)$ and $|e^{it}|^2 = 1$

Definition 1.7 (Convolution)**.** If f and g belong to $L^1(\mathbb{R}^d)$, the convolution of f and g is $f * g(z) = \int f(x) g(z-x) dx$.

Proposition 1.2. If f and g belong to $L^1(\mathbb{R}^d)$, then $f * g \in L^1(\mathbb{R}^d)$. Its Fourier transformation is $F[f*g](w) = F[f](w) = F[f](w)F[g](w)$ for all $w \in \mathbb{R}^d$

Remark. check this proposition as an exercise.

Proposition 1.3. If $f \in L^1(\mathbb{R}^d)$, then $F[f]$ is continuous and $\lim_{\|w\|_2 \to \infty} F[f](w) = 0$.

We introduce two properties that are useful for later cause.

Property 1.1. If $f, \mathcal{F}[f] \in L^2(\mathbb{R}) \cap L^1(\mathbb{R})$, then

1. (The Placherel equality)$\frac{1}{2\pi} \|\mathcal{F}[f]\|_2^2 = \|f\|_2^2$ (Plancherel's theorem)
2. (The Fourier inverse formula) For all $x \in \mathbb{R}$, $f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-iwx} \mathcal{F}[f](w)\, dw$, the inversion of the Fourier transform.

Question 1. Let $Z \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, then $\mathbb{E}[|z|] \leq \sqrt{\mathbb{E}[z^2]}\sqrt{\mathbb{E}[1^2]}$. Therefore, $L^2(\Omega, \mathcal{F}, \mathbb{P}) \subset L^1(\Omega, \mathcal{F}, \mathbb{P})$.

Example 1.1. Let $k(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, then $K \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. Then for all $w \in \mathbb{R}$,

$$F[K](w) = e^{-\frac{w^2}{2}}.$$

Example 1.2. Let $K(x) = \frac{1}{\sqrt{2}} \mathbb{1}_{\{|x| \leq 1\}}$, then $K \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. Then for all $w \in \mathbb{R}$,

$$\begin{aligned}
F[K](w) &= 1/2 \int_{-1}^{1} \cos(wx)dx + 1/2 \int_{-1}^{1} \sin(wx)dx \\
&= \frac{1}{2w} [\sin(wx)] \Big|_{-1}^{1} \\
&= \frac{\sin(wx)}{w}
\end{aligned}$$

Here $F[K] \notin L^1(\mathbb{R})$ but $F\left[K\right] \in L^2(\mathbb{R})$. Note also that $F[K](w) = 0$ if and only if $w = \pm k\pi$ for $k \in \mathrm{N}$.

Completeness Let us check whether the functions given in Example 1.1 and 1.2 satisfy the completeness condition 1.3 for $X = Z - \eta$.

1. Since $F[f_\eta](w) > 0$ for all w. Thus, $F[\varphi](w) = 0 \Leftrightarrow \varphi(x) = 0$ for all x.
2. Similarly, $F[\varphi](w) = 0$ for all $w \in \mathbb{R} \setminus S$. Because φ is continuous, it is 0 everywhere.

2 Density function and kernel estimation

2.1 Density function

We want to estimate the density f_X of $X \in \mathbb{R}$ and will work among classes of densities. For example,

1. **continuous densities**
2. densities such that for all $x,\ x' \in \mathbb{R},\ |f_X(x) - f_X(x')| \leq M\,|x - x'|$ for some $M > 0$
3. densities which are **monotonically increasing** on $[0, 1]$

2.2 Density function estimation

If X has a density f_X , then $f_X(x) = F'_X(x)\ a.e.$ because

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt = \mathbb{E}\left[\mathbb{1}_{\{X \leq x\}}\right].$$

A natural estimator of the CDF is the **empirical CDF**, defined as

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \le x\}}.$$

where n is the sample size. Therefore, an estimator of f_X is the derivative of the empirical CDF, which is the **empirical density function** defined as

$$\hat{f}_n(x) = \frac{\hat{F}_X\left(x + h/2\right) - \hat{F}_X\left(x - h/2\right)}{h} = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

where $K(x) = \mathbb{1}_{\{X \le \frac{1}{2}\}}$

Definition 2.1 (kernel function)**.** A kernel is a function $K : \mathbb{R} \to \mathbb{R}$ such that $K \in L^1(\mathbb{R})$ and $\int K(x)dx = 1$.

Definition 2.2 (kernel density estimator with kernel K and bandwidth h)**.** The kernel density estimator of f_X is defined as

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

2.3 Kernels estimators

Some kernels We list out some common kernels.

1. $K\left(x\right) = \frac{1}{2}\mathbb{1}_{|x| \le \frac{1}{2}}$ the rectangular
2. $K\left(x\right) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, the Gaussian kernel
3. $K\left(x\right) = \frac{\sin(x)}{\pi x}$, the sinc kernel
4. $K\left(x\right) = \frac{3}{4}\left(1 - x^2\right)\mathbb{1}_{|x| \le 1}$, the Epanechnikov kernel

Remark. Note that the Gaussian kernel is both in $L^1\left(\mathbb{R}, \mathcal{B}\left(\mathbb{R}\right), dx\right)$ and in $L^2\left(\mathbb{R}, \mathcal{B}\left(\mathbb{R}\right), dx\right)$. The sinc kernel is only in $L^2\left(\mathbb{R}, \mathcal{B}\left(\mathbb{R}\right), dx\right)$ but not in $L^1\left(\mathbb{R}, \mathcal{B}\left(\mathbb{R}\right), dx\right)$, as the absolute value fails to be integrable. However, we have

$$1 = \lim_{R\to\infty}\int_{-R}^{R} \frac{\sin\left(x\right)}{\pi x}dx.$$

2.4 Performance analysis

Definition 2.3. We introduce the quadratic **risk**

$$\mathrm{MSE}\left(x\right)=\mathbb{E}\left[\left(\hat{f}_X\left(x\right)-f_X\left(x\right)\right)^2\right],$$

where

$$\ell\left(x,y\right)=\left(x-y\right)^2$$

is the **loss** function.

Other risks include

$$\mathbb{E}\left[\sup_{x\in\mathbb{R}}\left|\hat{f}_X\left(x\right)-f_X\left(x\right)\right|\right]=\mathbb{E}\left[\left\|\hat{f}_X-f_X\right\|_\infty\right]$$

Note that $\hat{f}_X$ is a function of x and the observations $X=(X_1,\dots,X_n)$.

Definition 2.4. We define the **bias** of $\hat{f}_X\left(x\right)$ by

$$\mathrm{Bias}\left(\hat{f}_X\right)=b\left(x\right)=\mathbb{E}\left[\hat{f}_X\left(x\right)-f_X\left(x\right)\right]$$

and we denote the **variance** of $\hat{f}_X\left(x\right)$ by $\sigma^2\left(x\right)$.

Proposition 2.1. We have

$$\mathrm{MSE}\left(x\right)=b\left(x\right)^2+\sigma^2\left(x\right).$$

Proof. We have

$$\begin{aligned}\mathrm{MSE}\left(x\right)&=\mathbb{E}\left[\left(\hat{f}_X\left(x\right)-\mathbb{E}\left[\hat{f}_X\left(x\right)\right]+\mathbb{E}\left[\hat{f}_X\left(x\right)\right]-f_X\left(x\right)\right)^2\right]\\&=\mathbb{E}\left[\left(\hat{f}_X\left(x\right)-\mathbb{E}\left[\hat{f}_X\left(x\right)\right]\right)^2\right]+2\mathbb{E}\left[\left(\hat{f}_X\left(x\right)-\mathbb{E}\left[\hat{f}_X\left(x\right)\right]\right)\underbrace{\left(\mathbb{E}\left[\hat{f}_X\left(x\right)\right]-f_X\left(x\right)\right)}_{\text{not random}}\right]\\&\quad+\mathbb{E}\left[\left(\mathbb{E}\left[\hat{f}_X\left(x\right)\right]-f_X\left(x\right)\right)^2\right]\\&=\underbrace{\mathbb{E}\left[\left(\hat{f}_X\left(x\right)-\mathbb{E}\left[\hat{f}_X\left(x\right)\right]\right)^2\right]}_{=\sigma^2(x)}+2\left(\mathbb{E}\left[\hat{f}_X\left(x\right)\right]-f_X\left(x\right)\right)\underbrace{\mathbb{E}\left[\hat{f}_X\left(x\right)-\mathbb{E}\left[\hat{f}_X\left(x\right)\right]\right]}_{=0}\\&\quad+\underbrace{\left(\mathbb{E}\left[\hat{f}_X\left(x\right)\right]-f_X\left(x\right)\right)^2}_{=b(x)^2}.\end{aligned}$$

□

Proposition 2.2 (upper bound of $\sigma^2(x)$)**.** Assume that there exists $f_{\max} \in \mathbb{R}$ such that $\forall x \in \mathbb{R}$, $f_X(x) \le f_{\max}$ and $\int_{\mathbb{R}} K^2(u)\,du < \infty$. Then we have, for $C = f_{\max} \int_{\mathbb{R}} K^2(u)\,du$,

$$\forall x \in \mathbb{R} \forall n \ge 1 \forall h > 0, \sigma^2(x) \le \frac{C}{nh}.$$

Proof. First observe that, by identical distribution of $X_1, \ldots, X_n$,

$$\mathbb{E}\left[\hat{f}_X(x)\right] = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h}\mathbb{E}\left[K\left(\frac{X_i - x}{h}\right)\right] = \frac{1}{h}\mathbb{E}\left[K\left(\frac{X_1 - x}{h}\right)\right].$$

Now, using independence in the second line and identical distribution in the third line,

$$\begin{aligned}
\sigma^2(x) &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{h}K\left(\frac{X_i - x}{h}\right)\right) - \mathbb{E}\left[\hat{f}_X(x)\right]\right)^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\left(\frac{1}{h}K\left(\frac{X_i - x}{h}\right) - \mathbb{E}\left[\hat{f}_X(x)\right]\right)^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[\left(\frac{1}{h}K\left(\frac{X_1 - x}{h}\right) - \mathbb{E}\left[\hat{f}_X(x)\right]\right)^2\right]
\end{aligned}$$

Inserting equation* 2.4,

$$\begin{aligned}
\sigma^2(x) &= \frac{1}{n}\mathbb{E}\left[\left(\frac{1}{h}K\left(\frac{X_1 - x}{h}\right) - \mathbb{E}\left[\frac{1}{h}K\left(\frac{X_1 - x}{h}\right)\right]\right)^2\right] \\
&= \frac{1}{n}\mathrm{Var}\left[\frac{1}{h}K\left(\frac{X_1 - x}{h}\right)\right] \\
&= \frac{1}{n}\left(\mathbb{E}\left[\frac{1}{h^2}K^2\left(\frac{X_1 - x}{h}\right)\right] - \mathbb{E}\left[\frac{1}{h}K\left(\frac{X_1 - x}{h}\right)\right]^2\right) \\
&\le \frac{1}{n}\mathbb{E}\left[\frac{1}{h^2}K^2\left(\frac{X_1 - x}{h}\right)\right] \\
&= \frac{1}{nh}\mathbb{E}\left[\frac{1}{h}K^2\left(\frac{X_1 - x}{h}\right)\right] \\
&= \frac{1}{nh}\int_{\mathbb{R}} \frac{1}{h}K^2\left(\frac{y - x}{h}\right) f_X(y)\,dy \\
&= \frac{1}{nh}\int_{\mathbb{R}} K^2(u)\underbrace{f_X(x + hu)}_{\le f_{\max}}\,du \\
&\le \frac{1}{nh}\underbrace{f_{\max}\int_{\mathbb{R}} K^2(u)\,du}_{=C},
\end{aligned}$$

where we used the change of variables $y = x + hu$. $\square$

Definition 2.5 (β for a density function)**.** Let $\beta > 0$, $L > 0$ and set $\ell = \lfloor \beta \rfloor$, by which we mean the greatest integer **strictly** less than β. The Hölder class $\Sigma(\beta, L)$ is the class of functions $f : \mathbb{R} \to \mathbb{R}$ such that $f^{(\ell)}$ exists and for all $x, x' \in \mathbb{R}$ we have

$$\left|f^{(\ell)}(x) - f^{(\ell)}(x')\right| \leq L \left|x - x'\right|^{\beta - \ell}.$$

Definition 2.6. We define

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

Example 2.1. $\beta = 1$ gives the usual Hölder continuity condition: for all $x, x' \in \mathbb{R}$

$$\left|f(x) - f(x')\right| \leq L \left|x - x'\right|^{\beta}.$$

Remark. This Hölder condition implies continuity of f.

Definition 2.7 (β for a kernel)**.** $K : \mathbb{R} \to \mathbb{R}$ is a kernel **of order** $\ell \in \mathbb{N}_0$ if

- $u \mapsto u^j K(u)$ is integrable for any $j \in \{0, \dots, \ell\}$,
- $\int_{\mathbb{R}} K(u)\, du = 1$,
- and $\int_{\mathbb{R}} u^j K(u)\, du = 0$ for $j \in \{1, \dots, \ell\}$.

Proposition 2.3 (upper bound of $|b(x)|$)**.** Let $f_X \in \mathcal{P}(\beta, L)$ with $\beta, L > 0$ and K of order $\ell \geq \lfloor \beta \rfloor$ such that

$$\int_{\mathbb{R}} |u|^{\beta} |K(u)|\, du < \infty.$$

Then, for all $x \in \mathbb{R}$, $n \geq 1$ and $h > 0$, we have

$$|b(x)| \leq C_1 h^{\beta},$$

where

$$C_1 = \frac{L}{\ell!} \int_{\mathbb{R}} |u|^{\beta} |K(u)|\, du.$$

Proof. Reusing equation 2.4 and using $1 = \int_{\mathbb{R}} K(u)\, du$,

$$\begin{aligned}
b(x) &= \mathbb{E}\left[\hat{f}_X(x)\right] - f_X(x) \\
&= \frac{1}{h}\mathbb{E}\left[K\left(\frac{X_1 - x}{h}\right)\right] - f_X(x) \\
&= \frac{1}{h}\int_{\mathbb{R}} K\left(\frac{y - x}{h}\right) f_X(y)\, dy - f_X(x) \\
&= \frac{1}{h}\int_{\mathbb{R}} K\left(\frac{y - x}{h}\right) f_X(y)\, dy - \int_{\mathbb{R}} K(u) f_X(x)\, du.
\end{aligned}$$

With the change of variables $y = hu + x$, we obtain

$$\begin{aligned} b(x) &= \int_{\mathbb{R}} K(u) f_X(hu + x)\, du - \int_{\mathbb{R}} K(u) f_X(x)\, du \\ &= \int_{\mathbb{R}} K(u) \left(f_X(hu + x) - f_X(x)\right) du. \end{aligned}$$

By a Taylor expansion, for some $\tau \in [0,1]$, we obtain

$$f_X(hu + x) - f_X(x) = uhf'_X(x) + \cdots + \frac{(uh)^{\ell-1}}{(\ell-1)!} f_X^{(\ell-1)}(x) + \frac{(uh)^{\ell}}{\ell!} f_X^{(\ell)}(x + \tau uh).$$

Thus, recalling that $\int_{\mathbb{R}} u^j K(u)\, du = 0$ for $j \in \{1, \ldots, \ell\}$ (we use it in the second and the third step),

$$\begin{aligned} b(x) &= \int_{\mathbb{R}} K(u) \left(uhf'_X(x) + \cdots + \frac{(uh)^{\ell-1}}{(\ell-1)!} f_X^{(\ell-1)}(x) + \frac{(uh)^{\ell}}{\ell!} f_X^{(\ell)}(x + \tau uh) \right) du \\ &= \int_{\mathbb{R}} K(u) \frac{(uh)^{\ell}}{\ell!} f_X^{(\ell)}(x + \tau uh)\, du \\ &= \int_{\mathbb{R}} K(u) \frac{(uh)^{\ell}}{\ell!} \left(f_X^{(\ell)}(x + \tau uh) - f_X^{(\ell)}(x) \right) du. \end{aligned}$$

Taking absolute values, using the Hölder property $f_X \in \mathcal{P}(\beta, L)$, and recalling finally $0 \leq \tau \leq 1$,

$$\begin{aligned} |b(x)| &= \left| \int_{\mathbb{R}} K(u) \frac{(uh)^{\ell}}{\ell!} \left(f_X^{(\ell)}(x + \tau uh) - f_X^{(\ell)}(x) \right) du \right| \\ &\leq \int_{\mathbb{R}} |K(u)| \frac{|uh|^{\ell}}{\ell!} \left| f_X^{(\ell)}(x + \tau uh) - f_X^{(\ell)}(x) \right| du \\ &\leq \int_{\mathbb{R}} |K(u)| \frac{|uh|^{\ell}}{\ell!} L |\tau uh|^{\beta-\ell}\, du \\ &= \int_{\mathbb{R}} |K(u)| \frac{L |uh|^{\beta}}{\ell!} |\tau|^{\beta-\ell}\, du \\ &\leq \int_{\mathbb{R}} |K(u)| \frac{L |uh|^{\beta}}{\ell!} du \\ &= \frac{Lh^{\beta}}{\ell!} \int_{\mathbb{R}} |K(u)| \, |u|^{\beta}\, du. \end{aligned}$$

This shows the claim. □

Remark. Note that the expectation

$$\mathbb{E}\left[\hat{f}_X(x)\right] = \frac{1}{h} \int_{\mathbb{R}} K\left(\frac{y - x}{h}\right) f_X(y)\, dy$$

is the **convolution** $\frac{1}{h}K\left(\frac{-(\cdot)}{h}\right) * f_X$.

In general, the convolution of two integrable functions $f, g : \mathbb{R} \to \mathbb{R}$ is defined as

$$(f * g)(x) = \int_{\mathbb{R}} f(x-y)\, g(y)\, dy.$$

One interpretation of the convolution is the following: if f_X, f_Y are the densities of independent random variables X, Y, then the density of $X+Y$ is $f_X * f_Y$.

Indeed, let φ be bounded and continuous. Then, using independence and writing $u = x+y$, and using Fubini-Tonelli,

$$\begin{aligned}
\mathbb{E}\left[\varphi(X+Y)\right] &= \int_{\mathbb{R}\times\mathbb{R}} \varphi(x+y)\, f_{X,Y}(x,y)\, dxdy \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} \varphi(x+y)\, f_X(x)\, f_Y(y)\, dxdy \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} \varphi(u)\, f_X(y-u)\, f_Y(y)\, dudy \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} \varphi(u)\, f_X(y-u)\, f_Y(y)\, dydu \\
&= \int_{\mathbb{R}} \varphi(u) \int_{\mathbb{R}} f_X(y-u)\, f_Y(y)\, dydu \\
&= \int_{\mathbb{R}} \varphi(u)\, f_X * f_Y(u)\, du.
\end{aligned}$$

This characterises the density uniquely.

Another way to see this is to consider the characteristic function, which is the Fourier transform of the random variable, using independence:

$$\mathbb{E}\left[e^{it(X+Y)}\right] = \mathbb{E}\left[e^{itX}e^{itY}\right] = \mathbb{E}\left[e^{itX}\right]\mathbb{E}\left[e^{itY}\right].$$

The latter is the product of the characteristic functions of X and Y. The very same expression as on the right-hand side is yielded taking the characteristic function of a random variable with density $f_X * f_Y$, and the characteristic function characterises the distribution uniquely.

Result Combining proposition 2.2 and 2.3, we see

$$\mathrm{MSE}(x) \le C_1^2 h^{2\beta} + \frac{C}{nh}.$$

Minimizing the right-hand side in h yields $h_{\mathrm{opt}} = \left(\frac{C}{2\beta C_1^2 n}\right)^{\frac{1}{2\beta+1}} \sim n^{-\frac{1}{2\beta+1}}$.

Plugging this back into the right-hand side, we obtain

$$\mathrm{MSE}(x) = O\left(n^{-\frac{2\beta}{2\beta+1}}\right).$$

3 Sobolev class and symmetric kernel

3.1 Review of Fourier transform

Definition 3.1. The characteristic function of a random variable X is

$$\varphi_X(w) = \mathbb{E}\left[e^{iwX}\right] = \int_{\mathbb{R}} e^{iwx} f_X(x)\, dx.$$

Remark. It is possible as well to define the Fourier transform of $f \in L^2(\mathbb{R})$. Therefore, we take a sequence $f_m \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ such that $\|f - f_m\|_2^2 \to 0$ as $m \to \infty$ and define the Fourier transform of f as the L^2-limit of $\mathcal{F}[f_m]$. More precisely, we may take $f_m(x) = f(x)\mathbb{1}_{|x|\leq m}$. It is in L^2 as the product of an $L^2(\mathbb{R})$ function and a bounded function, and it is in $L^1(\mathbb{R})$ as a result of the Cauchy-Schwarz inequality:

$$\int_{\mathbb{R}} f(x)\mathbb{1}_{|x|\leq m} dx \leq \sqrt{\int_{\mathbb{R}} f(x)^2\, dx}\sqrt{\int_{-m}^{m} 1 dx} = \sqrt{2m}\sqrt{\underbrace{\int_{\mathbb{R}} f(x)^2\, dx}_{<\infty}}.$$

Moreover,

$$\|f_m - f\|_2^2 = \int_{-\infty}^{m} |f(x)|^2\, dx + \int_{m}^{\infty} |f(x)|^2\, dx \to 0 \tag{1}$$

as $m \to \infty$. By equation 1, for all m, m', $\|f_m - f_{m'}\|_2^2 \to 0$ as $m, m' \to \infty$, i.e. (f_m) is a Cauchy sequence. By Plancherel's theorem 1.1,

$$\|\mathcal{F}[f_m] - \mathcal{F}[f_{m'}]\|_2^2 = \|\mathcal{F}[f_m - f_{m'}]\|_2^2 = 2\pi\|f_m - f_{m'}\|_2^2.$$

Thus, $\mathcal{F}[f_m]$ is a Cauchy sequence in $L^2(\mathbb{R})$, so that it admits a limit in $L^2(\mathbb{R})$, since $L^2(\mathbb{R})$ is a complete normed space. We can then define the Fourier transform of f to be this limit.

3.2 Sobolev class

Building on this, we make the following definition.

Definition 3.2. Let $\beta > 0$, $L > 0$, the Sobolev class $\mathcal{P}_S(\beta, L)$ is defined as

$$\mathcal{P}_S(\beta, L) = \left\{ f : f \text{ is a density on } \mathbb{R} \text{ and } \int_{\mathbb{R}} |w|^{2\beta} |\mathcal{F}[f](w)|^2\, dw \leq 2\pi L^2 \right\}.$$

3.3 Symmetric kernel

Theorem 3.1 (Symmetric kernel)**.** *Let $f_X \in L^2(\mathbb{R}), K \in L^2(\mathbb{R})$ be a symmetric kernel such that*

$$\sup_{w\in\mathbb{R}\setminus\{0\}} \frac{|1 - \mathcal{F}[K](w)|}{|w|^{\beta'}} \leq A < \infty$$

for some $\beta', A > 0$. Then

$$\sup_{f_X \in \mathcal{P}_S(\beta,L)} \mathbb{E}\left[\left\|\hat{f}_X - f_X\right\|_2^2\right] \leq C n^{-\frac{2\tilde{\beta}}{2\tilde{\beta}+1}},$$

where $\tilde{\beta} = \min\{\beta, \beta'\}$, if $h = \alpha n^{-\frac{1}{2\tilde{\beta}+1}}$ for some $\alpha > 0$ and C is a constant which only depends on L, α, A, K.

Example 3.1. 1. Gaussian kernel: $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$, $\mathcal{F}[K](u) = e^{-\frac{u^2}{2}}$. We have

$$\frac{\left|1 - e^{-w^2/2}\right|}{|w|^{\beta'}} \leq \begin{cases} |w|^{-\beta'}, & |w| \geq 1 \\ \frac{w^2/2}{|w|^{\beta'}}, & |w| \leq 1 \end{cases}$$

so 3.1 holds if $\beta' \leq 2$, else the $\sup$ is ∞.

2. The sinc kernel: $K(u) = \frac{\sin(u)}{\pi u}, \mathcal{F}[K](w) = \mathbb{1}_{|w|\leq 1}$. We have

$$\frac{|1 - \mathcal{F}[K](w)|}{|w|^{\beta'}} \leq \begin{cases} |w|^{-\beta'}, & |u| > 1 \\ 0, & |u| \leq 1, \end{cases}$$

so 3.1 holds for all β'. Such a kernel is called an **infinite power kernel** or **superkernel**.

3. Trapeze kernel: Let

$$\mathcal{F}[K](w) = \begin{cases} 0, & |w| > a \\ 1, & |w| \leq b \\ \text{linear}, & \text{otherwise}, \end{cases}$$

a trapeze. Then 3.1 holds for all β'. Let us write K_2 for the trapeze (in Fourier space) and K_1 for the sinc Kernel (see 3.1). Then

$$K_2 = \frac{1}{2\pi}\mathcal{F}[\mathcal{F}[K_1] * F[K_1]] = \frac{1}{2\pi}\mathcal{F}[\mathcal{F}[K_1]]\,\mathcal{F}[\mathcal{F}[K_1]] = 2\pi K_1^2(u) = 2\pi\left(\frac{\sin u}{\pi u}\right),$$

which is in $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$.

Optimal rate of convergence It can be shown that the *sinc* kernel has the optimal rate of convergence.

A Corollary of the Theorem 3.1 that we have seen for cross-validation is

Corollary 3.2. *Let K be the sinc kernel, then*

$$\sup_{f_X \in \mathcal{P}_S(\beta,L)} \mathbb{E}\left[\left\|\hat{f}_X^{\mathrm{CV}} - f_X\right\|_2^2\right] \leq C n^{-\frac{2\beta}{2\beta+1}}$$

for all $\beta > \frac{1}{2}, L > 0$, where C only depends on β and L.

Some people have shown:

Proposition 3.1.

$$\inf_{\hat{f}} \sup_{f_X \in \mathcal{P}_S(\beta,L)} \mathbb{E}\left[\left\|\hat{f}_X - f_X\right\|_2^2\right] \geq C_* n^{-\frac{2\beta}{2\beta+1}}$$

for some absolute constant C_*.

This means that $n^{-\frac{2\beta}{2\beta+1}}$ is the "minimax" optimal rate of convergence and the cross-validated estimator is minimax adaptive (i.e. we can construct it with the data only).

Kernel comparison We end this section by the following table.

name	kernel	$\mathcal{F}[K]$	$\frac{\lvert 1-\mathcal{F}[K](w)\rvert}{\lvert w\rvert^\beta}$
Gaussian			
Epanechnikov			
Sinc			
Trapeze			

Table 1: Summary

4 MISE and Cross validation

4.1 MISE

To define the MISE, we would like

$$\mathbb{E}\left[\left\|\hat{f}_X - f_X\right\|_2^2\right] < \infty.$$

We assume $f_X \in L^2(\mathbb{R})$. We would like as well $\hat{f}_X \in L^2(\mathbb{R})$. This is true if $K \in L^2(\mathbb{R})$.

Indeed,

$$\begin{aligned}\left\|\hat{f}_X\right\|_2^2 &\leq \frac{2^{n-1}}{(nh)^2}\sum_{i=1}^{n}\int_{\mathbb{R}} K\left(\frac{X_i - x}{h}\right)^2 dx \\ &\leq \frac{2^{n-1}}{nh}\int_{\mathbb{R}} K^2(u)\,du < \infty.\end{aligned}$$

The idea behind this inequality is $(a+b)^2 \leq 2(a^2+b^2)$, and then by induction, $\left(\sum_{i=1}^n a_i\right)^2 \leq 2^{n-1}\sum_{i=1}^n a_i^2$.

4.2 Cross validation

Let us write

$$\begin{aligned}\mathrm{MISE}(h) &= \mathbb{E}\left[\int_{\mathbb{R}} \left(\hat{f}_X^h(x) - f_X(x)\right)^2 dx\right] \\ &= \underbrace{\mathbb{E}\left[\int_{\mathbb{R}} \left(\hat{f}_X^h(x)\right)^2 dx - 2\int_{\mathbb{R}} \hat{f}_X^h(x) f_X(x)\,dx\right]}_{=I(h)} + \int_{\mathbb{R}} f_X^2(x)\,dx\end{aligned}$$

introduced

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

where $\hat{f}_{X,-i} = \frac{1}{(n-1)h}\sum_{j=1,j\neq i}^n K\left(\frac{X_j - x}{h}\right)$. The cross-validated bandwidth is

$$\hat{h}_{\mathrm{CV}} = \arg\min_{h>0} \widehat{\mathrm{CV}}(h)$$

We claim

$$\frac{1}{2}\mathbb{E}\left[\hat{A}\right] = \mathbb{E}\left[\int_{\mathbb{R}} \hat{f}_X(x) f_X(x)\,dx\right]$$

We have

$$\begin{aligned}\mathbb{E}\left[\hat{f}_{X,-1}(X_1)\right] &= \mathbb{E}_{\mathbb{P}_{X_2}\otimes\cdot\otimes\mathbb{P}_{X_n}}\left[\int_{\mathbb{R}} \hat{f}_{X,-i}(x) f_X(x)\,dx\right] \\ &= \mathbb{E}\left[\frac{1}{(n-1)h}\sum_{j=2}^n \int_{\mathbb{R}} K\left(\frac{X_j - x}{h}\right) f_X(x)\,dx\right] \\ &= \frac{1}{h}\int_{\mathbb{R}}\int_{\mathbb{R}} K\left(\frac{z-x}{h}\right) f_X(z) f_X(x)\,dz dx\end{aligned}$$

As an exercise, show that this yields the claim.

Theorem 4.1 (Oracle inequality)**.** *Let $f_{\max}$ be such that for all x, $f_X(x) \le f_{\max} < \infty$. Assume the kernel K is such that $\int_{\mathbb{R}} K^2(u)\,du < \infty$. $\mathcal{F}[K] \ge 0$ and $\mathrm{supp}(\mathcal{F}[K]) \subseteq [-1,1]$. Then $\hat{f}_X^* = \hat{f}_X^{h_{\mathrm{CV}}}$ is such that for all $0 < \delta < 1$, for all $n \ge 1$,*

$$\mathbb{E}\left[\int_{\mathbb{R}} \left(\hat{f}_X^*(x) - f_X(x)\right)^2 dx\right] \le \left(1 + \frac{C}{n^\delta}\right)\min_{h>\frac{1}{n}} \mathbb{E}\left[\int \left(\hat{f}_X^h(x) - f_X(x)\right)^2 dx\right] + \frac{C(\log n)^{\frac{\delta}{2}}}{n^{1-\delta}}$$

Remark. The cross-validation bandwidth from theorem 4.1 is random. The kind of inequality in the the theorem is called **oracle inequality**, as it is not possible to obtain the values on each side. They involve the unknown $f_X(x)$. The estimation of errors in cross-validation kernel estimation is hard, but in practice it often works well.

4.3 Extension

Remark. The condition 3.1 is satisfied for an integer β if K is a kernel of order $\beta - 1$ and $\int |u|^{\beta} |K(u)| < \infty$.

Remark. We can work with a smaller class of *super smooth* density functions.

1. $\mathcal{P}_{\alpha,r} = \left\{ f \in L^2(\mathbb{R}) \quad \text{such that} \quad \int \exp(\alpha |w|^2) |\phi(w)|^2 dw \leq L^2 \right\}$ where $\phi = \mathcal{F}[f]$ is the Fourier transform of f. We can show that a MISE optimal kernel density estimatro could have a risk less than $C\frac{(\log n)^{1/r}}{n}$.

2. $\mathcal{P}_{\alpha,r} = \{ f \in L^2(\mathbb{R}) \quad \text{such that} \quad \text{supp}(\phi) \subset [-a, a] \}$. In this case, the upper bound is $\frac{a\pi}{n}$.

5 Other types of non-parametric estimators

5.1 Orthogonal series estimators

[1] Let $f_X \in L^2\left([0,1]^d\right)$, where $L^2\left([0,1]^d\right)$ can be proven to be a *separable Hilbert space* when endowed with the inner product

$$\langle f, g \rangle = \int_{\mathbb{R}} f(x)g(x)dx.$$

We write

$$\|f\|_2^2 = \langle f, f \rangle .$$

Some properties are comparable to $\mathbb{R}^d$ with $\langle x, y \rangle = x^T y$. As a separable space, $L^2\left([0,1]^d\right)$ has a countable basis $(e_j)_{j=1}^{\infty}$, which is a sequence of functions in $L^2\left([0,1]^d\right)$ such that for all

$$\langle e_j, e_k \rangle = \delta_{jk} = \begin{cases} 1, & j = k, \\ 0, & \text{else}, \end{cases}$$

and for all $f \in L^2\left([0,1]^d\right)$,

$$f = \lim_{k \to \infty} \sum_{j=1}^{k} \langle f, e_j \rangle e_j.$$

Think of $\mathbb{R}^d$, where $(e_j)_{j=1}^{d}$ is a basis for $e_j = (0, \ldots, 0, 1, 0, \ldots, 0)$ the j-th unit vector. Then $\langle e_j, e_k \rangle = e_j^T e_k = \delta_{jk}$, and for $x \in \mathbb{R}^d$,

$$x = \sum_{j=1}^{d} x_j e_j = \sum_{j=1}^{d} x^T e_j e_j = \sum_{j=1}^{d} \langle x, e_j \rangle e_j.$$

[1] Generalizations are called sieves (in Econometrics) or dictionaries in machine-learning.

Given $(e_j)_{j=1}^{\infty}$ a basis, for all $f \in L^2\left([0,1]^d\right)$,

$$\|f\|_2^2 = \sum_{j=1}^{\infty} \langle f, e_j \rangle^2 .$$

This is a version of the Pythagorean theorem. In $\mathbb{R}^d$,

$$\|x\|_2^2 = \sum_{j=1}^{d} x_j^2 = \sum_{j=1}^{d} \langle x, e_j \rangle^2 .$$

Back to our goal to estimate $f_X = \lim_{k\to\infty} \sum_{j=1}^{\infty} \langle f, e_j \rangle e_j$. For some $T \in \mathbb{N}$, consider $f_X^T \stackrel{\text{def}}{=} \sum_{j=1}^{T} \langle f, e_j \rangle e_j$. The idea is to estimate this cut-off sum instead of the limit expression for f_X. We have

$$c_j \stackrel{\text{def}}{=} \langle f_X, e_j \rangle = \int_{[0,1]^d} f_X(x) e_j(x) dx = \mathbb{E}\left[e_j(X)\right],$$

so that an unbiased estimator is

$$\hat{c}_j = \frac{1}{n} \sum_{i=1}^{n} e_j\left(X_i\right).$$

Thus, a candidate estimator for f_X is

$$\hat{f}_X^T = \sum_{j=1}^{T} \hat{c}_j e_j,$$

where

$$\mathbb{E}\left[\hat{f}_X^T\right] = \sum_{j=1}^{T} c_j e_j = f_X^T.$$

It is possible to write

$$\hat{f}_X^T = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\sum_{j=1}^{T} e_j(X_i) e_j(x),}_{q_T(X_i, x)}$$

where $q_T\left(X_i, x\right)$ plays the role of a kernel and T plays the same role as $\frac{1}{h}$. On $L^2\left([0,1]^d\right)$ we can use bases for which $e_j = f_{j_1} \cdots f_{j_d}$ where $(f_k)_{k=1}^{\infty}$ is a basis of $L^2\left([0,1]\right)$ and $(j_1, ..., j_d)$ plays the role of the index j[2]. For example, $f_k\left(x\right) = \sqrt{2}\sin\left(\pi k x\right)$ is a basis of $L^2\left([0,1]\right)$. This gives

$$e_{j_1,\ldots,j_d}\left(x\right) = 2^{\frac{d}{2}} \prod_{k=1}^{d} \sin\left(\pi j_k x_k\right).$$

[2]Note that there exists a bijection $\mathbb{N}^d \to \mathbb{N}$.

One can check that this defines an orthogonal system **(Exercise)**.
We define

$$W(\beta, L) = \Bigg\{ f : [0,1]^d \to \mathbb{R} \text{ with coefficients } c_{j_1,\ldots,j_d} \text{ w.r.t. } (f_{j_1}, \ldots, f_{j_d})$$
$$\text{such that } \sum_{j_1=1}^{\infty} \sum_{j_2=1}^{\infty} \cdots \sum_{j_d=1}^{\infty} c^2_{j_1,\ldots,j_d} \left(j_1^2 + \ldots + j_d^2\right)^{\beta} \leq L^2 \Bigg\}$$

Remark. The $\|j\|^{2\beta}$ is present due to the fact that we take derivatives of our basis functions defined above till the order of β.

In $L^2\left(\mathbb{R}^d\right)$, an analogous condition (with Fourier transform instead of Fourier series) would be:

$$\int_{\mathbb{R}^d} |\mathcal{F}[f](w_1, \ldots, w_d)|^2 \left(|w_1|^2 + \ldots + |w_d|^2\right)^{\beta} dw \leq L^2.$$

Note the usual bias-variance decomposition of the mean-squared error,

$$\mathbb{E}\left[\left\|\hat{f}_X^T - f_X\right\|_2^2\right] = \underbrace{\left\|f_X^T - f_X\right\|_2^2}_{b^2 = \text{Bias}^2} + \underbrace{\mathbb{E}\left[\left\|\hat{f}_X^T - f_X^T\right\|_2^2\right]}_{\sigma^2}.$$

Then

$$\begin{aligned}
b^2 &= \sum_{j_1=T+1}^{\infty} \cdots \sum_{j_d=T+1}^{\infty} c^2_{j_1,\ldots,j_d} \\
&\leq \sum_{j_1=T+1}^{\infty} \cdots \sum_{j_d=T+1}^{\infty} c^2_{j_1,\ldots,j_d} \left(\left(\frac{j_1}{T+1}\right)^2 + \cdots + \left(\frac{j_1}{T+1}\right)^2 \right)^{\beta} \\
&= \left(\frac{1}{T+1}\right)^{2\beta} \sum_{j_1=T+1}^{\infty} \cdots \sum_{j_d=T+1}^{\infty} c^2_{j_1,\ldots,j_d} \left(j_1^2 + \cdots + j_d^2\right)^{\beta} \\
&\leq \left(\frac{1}{T+1}\right)^{2\beta} L^2.
\end{aligned}$$

Note that $\|f_{j_1}\cdots f_{j_d}\|_2^2 = \|f_{j_1}\|_2^2\cdots\|f_{j_d}\|_2^2$, which are all $=1$. Then,

$$\begin{aligned}
\sigma^2 &= \mathbb{E}\left[\left\|\hat{f}_X^T - f_X^T\right\|_2^2\right] \\
&= \mathbb{E}\left[\sum_{j_1=1}^{T}\cdots\sum_{j_d=1}^{T}\left(\hat{c}_{j_1,\dots,j_d} - c_{j_1,\dots,j_d}\right)^2\|f_{j_1}\cdots f_{j_d}\|_2^2\right] \\
&= \mathbb{E}\left[\sum_{j_1=1}^{T}\cdots\sum_{j_d=1}^{T}\left(\hat{c}_{j_1,\dots,j_d} - c_{j_1,\dots,j_d}\right)^2\right] \\
&= \sum_{j_1=1}^{T}\cdots\sum_{j_d=1}^{T}\mathrm{Var}\left(\hat{c}_{j_1,\dots,j_d}\right) \\
&\leq \sum_{j_1=1}^{T}\cdots\sum_{j_d=1}^{T}\frac{2^d}{n} \\
&\leq \frac{\left(2\left(T+1\right)\right)^d}{n}
\end{aligned}$$

since

$$\begin{aligned}
\mathrm{Var}\left(\hat{c}_{j_1,\dots,j_d}\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(f_{j_1}\cdots f_{j_d}\left(X_i\right)\right) \\
&\leq \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[f_{j_1}^2\cdots f_{j_d}^2\left(X_i\right)\right] \\
&\leq \frac{2^d}{n},
\end{aligned}$$

where we used $f_{j_k}^2 \leq 2$ for our particular basis $f_{j_k}\left(x\right) = \sqrt{2}\sin\left(\pi j_k x\right)$. Remember $\sigma^2 \leq \frac{1}{nh^d}$ for kernel estimators. This gives an upper bound of the order of $n^{-\frac{2\beta}{2\beta+d}}$ if $T = \left\lfloor n^{\frac{1}{2\beta+d}}\right\rfloor$.

Remark. (Nonexaminable content)

- We can work with families of functions which may not be basis functions. We talk about series, dictionaries (machine learning).
- In the previous upper bound, the choice of T is infeasible because it depends on β which is unknown.
- It it classical to estimate many coefficients c_j, for T much larger than before (*e.g.* $\sqrt{n}$) and work with the estimators

$$\hat{f}_X^T\left(x\right) = \sum_{j_1=1}^{T}\cdots\sum_{j_d=1}^{T}\hat{\tau}(c_{j_1,\dots,j_d})e_{j_1}\cdots e_{j_d}\left(x\right).$$

where $\tau \propto \frac{\sqrt{\log n}}{n}$. For example

– $\tau_\rho(x) = \mathbb{1}_{\{|x|\geq\rho\}}$, where ρ is a thresholding function. This is the **hard** thresholding function.
– $\tau_\rho(x) = x\max\left(1 - \frac{\rho}{|x|}, 0\right)$. This is the **soft** thresholding function.

6 Regression Function Estimation

6.1 Introduction: average effect of X on Y

The model for a nonparametric model is

$$Y = f(X) + \varepsilon$$

, where $\mathbb{E}[\varepsilon|X] = 0$ and $\mathbb{E}[|\varepsilon|] < \infty$. The goal is to estimate f. We say it has a random design if X is random, and a fixed design if X is fixed. We will focus on the random design case.
First, we define the average effect of X on Y as $\mathbb{E}[f(X)]$ if the expectation is defined.
If $f_{y|x}$, the conditional density of Y given X exists, it is given by $f_{Y|X}(y \mid x) = \frac{f(x,y)}{f_X(x)}$ if $f_X(x) > 0$.
Also the conditional expectation function $\mathbb{E}[Y|X=x]$ is given by

$$\mathbb{E}[Y|X=x] = \int y f_{Y|X}(y \mid x)\, dy = \frac{\int y f(x,y)\, dy}{f_X(x)} = \frac{\int y f(x,y)\, dy}{\int f(x,y)\, dy}.$$

A natural idea would be to use

$$\hat{f}_X(x) = \frac{1}{nh^2}\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) K\left(\frac{Y_i - y}{h}\right),$$

where K is a kernel.
As an exercise, we can check that

$$\int y \hat{f}_{Y,X}(y,x)\, dy = \frac{1}{nh}\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) Y_i$$

and

$$\int \hat{f}_{Y,X}(y,x)\, dy = \frac{1}{nh}\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right).$$

Nadaraya-Watson estimator This leads to the

$$\hat{f}_X(x) = \frac{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}.$$

In practice, dealing with the denominator can be tricky. We propose two ideas to deal with this issue.

1. We can work with **nonnegative** kernels because

$$\sum Y_i \underbrace{\frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}}_{\in[0,1]}.$$

2. We can use a trimming factor ρ and write

$$\hat{f}_X(x) = \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i}{\max\left(\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right), \rho\right)}.$$

Suppose now supp $(X) = [a, b]$ and $\exists\, m > 0$ s.t. $f_X(x) \geq m$. Suppose I am interested in $f(b)$ and I use the rectangular kernel 1. Then $\hat{f}(b)$ where $\hat{f}$ is the N.W. estimator is biased. But a

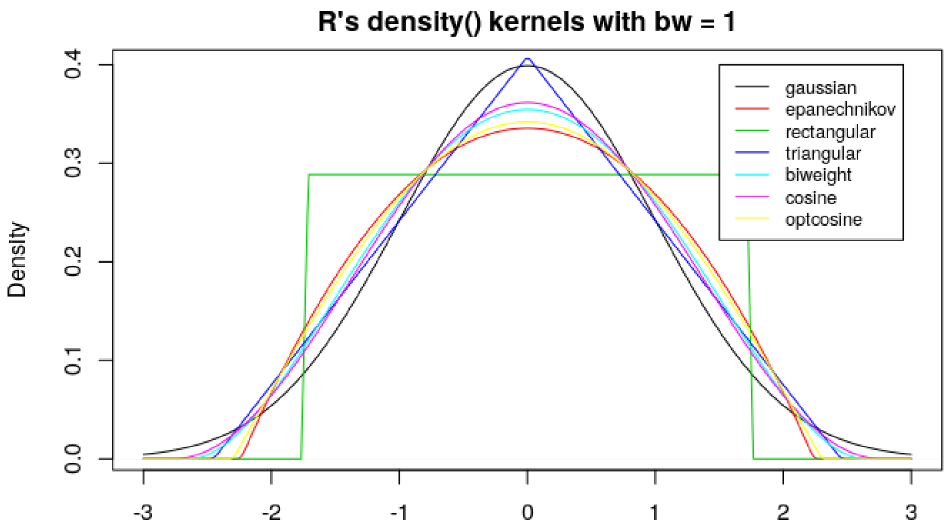

Figure 1: Kernels

local polynomial estimator of order ≥ 1 is consistent and unbiased.

Remark. In TD, we will see that we can get a fast rate of convergence with nonnegative kernels (unlike in density estimation).

6.2 Local Polynomial Estimation

7 Treatment Effects

7.1 Setup

We have a dataset $[Y_i, D_i, X_i, Z_i, W_i]_{i=1}^{n}$ following i.i.d. from a joint distribution.

- D is a binary treatment variable, $D \in [0, 1]$.
- Y is the outcome variable. Here Y is a random variable $Y \in \mathbb{R}$.
- X, Z, W are covariates/additional random variables.

The model equation (for each individual i) is $y_i = y\left(0\right)\left(1-D\right)+y\left(1\right)D$. The potential outcome is $y_i\left(1\right), y_i\left(0\right)$, which are not observed. We can only observe $y_i = y\left(D_i\right)$.

7.2 Parameters of Interest

- Average treatment effect (ATE): $\tau = \mathbb{E}\left[Y\left(1\right) - Y\left(0\right)\right]$
- Conditional average treatment effect (CATE): $\tau(x) = \mathbb{E}\left[Y\left(1\right) - Y\left(0\right)|X = x\right]$. It can be useful if we care about the effect of the treatment on a specific subgroup of the population.
- Average treatment effect on the treated (ATT): $\tau_{\mathsf{ATT}} = \mathbb{E}\left[Y\left(1\right) - Y\left(0\right)|D = 1\right]$
- Average treatment effect on the untreated (ATU): $\tau_{\mathsf{ATU}} = \mathbb{E}\left[Y\left(1\right) - Y\left(0\right)|D = 0\right]$
- Conditional average treatment effect on the treated (CATT): $\tau_{\mathsf{ATT}}(x) = \mathbb{E}\left[Y\left(1\right) - Y\left(0\right)|D = 1, X = x\right]$

7.3 Identification

We need to impose some assumptions in order to identify the parameters.

Assumption 7.3.1. $\mathbb{P}\left(D = 1\right) \in (0, 1)$

Assumption 7.3.2. *The covariates X, Z, W are such that if $X = X(0) + D\left(X(1) - X(0)\right)$, then $X(1) = X(0)$.*

Assumption 7.3.3. *The potential outcome $Y\left(1\right), Y\left(0\right)$ are independent of D given*

Assumption 7.3.4. *The potential outcome $Y\left(1\right), Y\left(0\right)$ are independent of D given X, Z, W*

We introduce a new notation for the purpose of another assumption.

Definition 7.1 (Propensity score)**.** The propensity score is defined as the conditional probability of receiving the treatment given the covariates, that is

$$\pi(x) = \mathbb{P}(D = 1 \mid X = x)$$

Remark. Later we will build estimators using propensity score, called **inverse propensity score weighting (IPSW) estimator**.

Assumption 7.3.5 (Common support)**.** *The propensity score $\pi(x)$ continuous and bounded between 0 and 1 for all $x \in$ supp(X)*

Assumption 7.3.6 (Mean independence)**.** *The potential outcome $Y(1), Y(0)$ are independent of D given X, Z, W, that is*

$$\mathbb{E}[Y(d) \mid D, X] = \mathbb{E}[Y(d) \mid X]$$

Some proposition

Proposition 7.1. For any $Y \in \mathbb{R}$, we have that

1. the expectation of $\mathbb{1}_{\{Y(1)<y\}}$ conditional on $D = 1$ equals to the expectation of $\mathbb{1}_{\{Y<y\}}$ conditional on $D = 1$, that is

$$\mathbb{E}\left[\mathbb{1}_{\{Y(1)<y\}} \mid D = 1\right] = \mathbb{E}\left[\mathbb{1}_{\{Y<y\}} \mid D = 1\right]$$

2. By mean independence, $\tau(x) = \mathbb{E}[Y(1) - Y(0) \,|X = x] = \mathbb{E}[Y(1) \,|X = x, D = 1] - \mathbb{E}[Y(0) \,|X = x, D = 0]$
3. $\mathbb{E}[Y(1) \,|X = x, D = 1] =$

Proof. Because

$$\mathbb{1}\{Y < y\} = \mathbb{1}\{Y(0) \leq y\}(1 - D) + \mathbb{1}\{Y(1) < Y\}D$$

We have that

□

Proposition 7.2. Under the assumptions 7.3.5 and 7.3.6, it can be shown that the conditional ATE(x) equals to ATT(x) and ATU(X), that is

$$\tau(x) = \tau_{\mathsf{ATT}}(x) = \tau_{\mathsf{ATU}}(x)$$

Proof. □

7.4 Regression discontinuity design

7.4.1 Sharp RDD

Preliminaries Previously, we consider D as a binary variable and impose certain conditions on whether $D_i = 1$ or $D_i = 0$ (for each individual). Now we specify how D is determined by a continuous variable X, that is

$$D_i = \mathbb{1}\{X_i \geq c\}$$

where c is a known threshold. The idea is that the treatment is assigned based on the value of X. We can think of X as a score, and D is assigned to those who score above a certain threshold.

For example, in the context of education, X can be the score of a student in a standardized test, and D is whether the student is admitted to a college. The threshold c is the cutoff score for admission.

Conditions Recall the definition of *unconfoundedness*:

Definition 7.2 (unconfoundedness)**.** The treatment D is unconfounded with the potential outcome Y given X if

$$Y(1), Y(0) \perp\!\!\!\perp D \mid X$$

It is easy to see that since D is determined by X, the potential outcome $Y(1), Y(0)$ are independent of D given X. (Given X, D is already determined, thus a constant.) Therefore, the unconfoundedness assumption is satisfied.

Since X is a continuous variable, we make an assumption on the **average potential outcome** $\mathbb{E}[Y(j) \mid X = x]$ for $j = 0, 1$. We assume that the average potential outcome is continuous at the threshold c. For example, for those who have a test score slightly above and below the threshold, the average potential earning $Y(1), Y(0)$ is similar.

Remark. We assume that $\mathbb{E}[Y(j) \mid X = x]$ is continuous at the threshold c but not $\mathbb{E}[Y \mid X = x]$ at c.

Let us now check conditional ATE at the point c. Previously, we define the conditional ATE as

$$\begin{aligned}
\tau(x) &= \mathbb{E}[Y(1) - Y(0) \mid X = x] \\
&= \mathbb{E}[Y(1) \mid X = x] - \mathbb{E}[Y(0) \mid X = x] \\
&= \underbrace{\mathbb{E}[Y \mid D = 1, X = x] - \mathbb{E}[Y \mid D = 0, X = x]}_{\text{by unconfoundedness thus mean independence}}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\tau(c) &= \lim_{x \to c^+} \mathbb{E}[Y \mid D = 1, X = x] - \lim_{x \to c^-} \mathbb{E}[Y \mid D = 0, X = x] \\
&= \lim_{x \to c^+} \mathbb{E}[Y \mid X = x] - \lim_{x \to c^-} \mathbb{E}[Y \mid X = x]
\end{aligned} \tag{2}$$

Estimation To estimate $\mathbb{E}[Y \mid X = x]$, we can use local polynomial of order 0 (Nadaraya-Watson estimator) or more:

$$\left(\hat{\alpha}_1, \hat{\beta}_1\right) = \arg\min_{\alpha,\beta} \sum_{i=1, X_i \geq c}^{n} K\left(\frac{X_i - x}{h}\right) \left(Y_i - \alpha - \beta\left(X_i - c\right)\right)^2$$

$$\left(\hat{\alpha}_0, \hat{\beta}_0\right) = \arg\min_{\alpha,\beta} \sum_{i=1, X_i < c}^{n} K\left(\frac{X_i - x}{h}\right) \left(Y_i - \alpha - \beta\left(X_i - c\right)\right)^2$$

Then we estimate $\tau(c)$ by $\hat{\tau}(c) = \hat{\alpha}_1 - \hat{\alpha}_0$.

7.4.2 Fuzzy RDD

Preliminaries First, let's recall the definition of conditional mean independence:

Definition 7.3 (Conditional mean independence)**.** We say U and V are conditionally mean independent given X if

$$\mathbb{E}\left[\phi(U)\psi(V) \mid X\right] = \mathbb{E}\left[\phi(U) \mid X\right]\mathbb{E}\left[\psi(V) \mid X\right]$$

Naturally local conditional mean independence in a neighborhood $\mathcal{N}$ is defined as

$$\mathbb{E}\left[\phi(U)\psi(V) \mid X = x\right] = \mathbb{E}\left[\phi(U) \mid X = x\right]\mathbb{E}\left[\psi(V) \mid X = x\right]$$

for almost every $x \in \mathcal{N}$.

Condition Now we are ready to move from *sharp RDD* to *fuzzy RDD*. In the fuzzy RDD, the treatment D is not exactly determined by X but instead satisfies the following condition to create a discontinuity at c: The propensity score function $\pi(x)$ is continuous on $(c - \epsilon, c)$ and $(c, c + \epsilon)$ for some $\epsilon > 0$ and $\lim_{x \to c^+} \pi(x) \neq \lim_{x \to c^-} \pi(x)$. We also loosen the mean independence condition to local conditional mean independence in a neighborhood $\mathcal{N}$ of c.

Remark. The fuzzy RDD is more general than the sharp RDD.

Since we depart from sharp RDD, the conditional ATE at c is now defined as the following:

Proposition 7.3.

$$\tau(c) = \frac{\lim_{x \to c^+} \mathbb{E}\left[Y \mid D = 1, X = x\right] - \lim_{x \to c^-} \mathbb{E}\left[Y \mid D = 0, X = x\right]}{\lim_{x \to c^+} \pi(x) - \lim_{x \to c^-} \pi(x)}$$

Proof. □

Estimation We can make use of local polynomial estimator to estimate the denominator and the numerator separately for $tau(c)$.

7.5 Instrumental variable

In this section, we are in the case of selection on *unobservables*. We have binary treatment D, covariates X, and a binary instrument Z, such that treatment is assigned based on Z (and maybe X). But the assigned treatment may not be taken by the individual (imperfect compliance). We have the following model:

$$\begin{aligned} Y = Y(0,0) &+ Z\left(Y(1,0) - Y(0,0)\right) + D\left(Y(0,1) - Y(0,0)\right) \\ &+ DZ\left(Y(1,1) - Y(0,1) - Y(1,0) + Y(0,0)\right) \end{aligned} \tag{3}$$

and

$$D = D(0) + Z\left(D(1) - D(0)\right) \tag{4}$$

Preliminaries We define the following:

- One-sided compliance: $P\left(D(0) = 0\right) = 0$ which means there is no always taker or defiers.
- two-sided compliance: $P\left(D(0) = 0\right) \in (0, 1)$ and $P\left(D(1) = 1\right) \in (0, 1)$.

Condition We need to impose that the assignment Z is independent of the potential outcome $Y(z,d)$ and the treatment D (given X). From now on we omitted the conditioning on X for simplicity. This is the *exclusion restriction* assumption. It is similar to $\mathbb{E}[\epsilon \mid X, Z] = 0$ assumption in standard linear IV model.

Definition 7.4 (Local average treatment effect (LATE))**.** The local average treatment effect is defined as

$$\tau_{\mathsf{LATE}} = \mathbb{E}[Y(Z,1) - Y(Z,0) \mid D(1) - D(0) = 1]$$

which is the average treatment effect for the compliers.

7.6 One-sided compliance

Instead of discussing ATE, we define a new set of parameters called *Intention to treat* (ITT).

Definition 7.5 (Intention to treat (ITT))**.** The intention to treat on treatment is defined as

$$\mathsf{ITT}_D = \mathbb{E}[D(1) - D(0) \mid Z = 1]$$

The intention to treat on outcome is defined as

$$\mathsf{ITT}_Y = \mathbb{E}[Y(1, D(1)) \mid Z = 1] - \mathbb{E}[Y(0, D(0)) \mid Z = 0]$$

The intention to treat on outcome for compilers is defined as

$$\mathsf{ITT}_{Y|D(1)-D(0)=1} = \mathbb{E}[Y(1,1) - Y(0,1) \mid D(1) - D(0) = 1]$$

Condition Under the one-sided compliance, we need to make the following assumption:

- If $D(1) = 0$ (never taker), then $Y(1,0) = Y(0,0)$.
- If $D(1) = 1$ (compiler), then $Y(0,d) = Y(1,d)$ for $d = 0,1$.

Because every individual is either a never taker or a compiler, we have the following:

$$Y(z,d) = Y(d)$$

This is sometimes called the an *exclusion restriction* assumption.

Then it can be shown that

$$\mathsf{ITT}_Y = \mathsf{ITT}_{Y,CO}\mathsf{ITT}_D$$

Recall that LATE is the average treatment effect for the compilers. And under the condition mentioned above,

$$\tau_{\mathsf{LATE}} = \mathbb{E}[Y(Z,1) - Y(Z,0) \mid D(1) - D(0) = 1]$$
$$\mathsf{ITT}_{Y|D(1)-D(0)=1} = \mathbb{E}[Y(1,1) - Y(0,0) \mid D(1) - D(0) = 1]$$

That is,

$$\tau_{\mathsf{LATE}} = \mathsf{ITT}_{Y|D(1)-D(0)=1}$$

Therefore,

$$\tau_{\mathsf{LATE}} = \frac{\mathsf{ITT}_Y}{\mathsf{ITT}_D} = \frac{\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]}{\mathbb{E}[D \mid Z = 1] - \mathbb{E}[D \mid Z = 0]}$$

7.6.1 Two-sided compliance

Condition This is a natural assumption.

- For never takers, $Y(0,0) = Y(1,0)$.
- For always takers, $Y(0,1) = Y(1,1)$.
- For compliers (and defiers), $Y(1,d) = Y(0,d)$.

We also need to impose the *monotonicity* assumption. It states that the instrument Z has a monotonic effect on the treatment D. That is, $D(1) \geq D(0)$ a.s. or $D(1) \leq D(0)$ a.s.

Under these conditions, we also have

$$\tau_{\mathsf{LATE}} = \frac{\mathbb{E}[Y \mid Z=1] - \mathbb{E}[Y \mid Z=0]}{\mathbb{E}[D \mid Z=1] - \mathbb{E}[D \mid Z=0]}$$

It can be shown that if there is no defiers, then $\mathbb{E}[D \mid Z=1] - \mathbb{E}[D \mid Z=0] = \mathbb{P}(D(1) - D(0) = 1)$.

7.7 Estimation methods: (Augmented) Inverse probability Weighting (AIPW)