

Econometrics 2: Non-parametric methods

Eric Gautier*

2024 Spring



I have a question!

*Notes by FU Zixuan, last compiled on February 22, 2024.

Contents

1	Preliminaries	3
1.1	Probability basics	3
1.2	Identification	4
1.3	Completeness condition	6
2	Density function and kernel estimation	7
2.1	Density function	7
2.2	Density function estimation	7
2.3	Kernels estimators	8
2.4	Performance analysis	9
3	Sobolev class and symmetric kernel	14
3.1	Review of Fourier transform	14
3.2	Sobolev class	14
3.3	Symmetric kernel	14
4	MISE and Cross validation	16
4.1	MISE	16
4.2	Cross validation	17
4.3	Extension	18
5	Other types of non-parametric estimators	18
5.1	Orthogonal series estimators	18
6	Regression Function Estimation	22
6.1	Introduction: average effect of X on Y	22
6.2	Nadaraya-Watson estimator	22
6.3	Local Polynomial Estimation	23
7	Treatment Effects	24
7.1	Setup	24
7.2	Parameters of Interest	24
7.3	Identification	24
7.4	Regression discontinuity design	26
7.4.1	Sharp RDD	26
7.4.2	Fuzzy RDD	27
7.5	Instrumental variable	28
7.5.1	One-sided compliance	28
7.5.2	Two-sided compliance	29
7.6	Estimation methods: (Augmented) Inverse probability Weighting (AIPW)	30

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 $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ such that $\mathbb{P}_X[B] = \mathbb{P}[X \in B]$ for all $B \in \mathcal{B}(\mathbb{R}^d)$

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 $B \in \mathcal{B}(\mathbb{R}^d)$

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. $(Y, X^\top)^\top$ has discrete support. Let $x \in \text{spt}(X)$, then $\mathbb{E}[Y|X = x] = \sum y_j \mathbb{P}(Y = y_j | X = x)$. This is well defined only when $\mathbb{P}(X = x) > 0$ in which case $\mathbb{P}(Y = y_j | X = x) = \frac{\mathbb{P}(Y = y_j, X = x)}{\mathbb{P}(X = x)}$ is well defined.
2. $(Y, X^\top)^\top$ 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 variable $Z := \mathbb{E}[Y|X]$ 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 $\phi : (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The corresponding function ϕ for $Z = \underbrace{\mathbb{E}[Y|X]}_{\text{conditional expectation}}$ is de-

noted by $\underbrace{\mathbb{E}[Y|X = x]}_{\text{conditional expectation function}}$.

Remark. The proposition 2 is equivalent to

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

Exercise. Let X and β be random vectors in \mathbb{R}^d such that X and β ARE independent, that is $\mathbb{P}_{X,\beta} = \mathbb{P}_X \times \mathbb{P}_\beta$. Let $g : \mathbb{R}^d \rightarrow \mathbb{R}$ be a bounded and measurable function. Define $Y = g(X^\top \beta)$. It can be shown that $\mathbb{E}[Y|X = x] = \mathbb{E}[g(x^\top \beta)]$ for all $x \in \text{supp}(X)$.

Proof. Let $B \in \sigma(X)$. Then

$$\mathbb{E}[Y \mathbb{1}\{X \in B\}] = \mathbb{E}[g(X^\top \beta) \mathbb{1}\{X \in B\}] = \int \int g(X^\top b) \mathbb{1}\{x \in B\} d\mathbb{P}_{\beta, X}(b, x).$$

Since X and β are independent, we have

$$\mathbb{P}_{\beta, X}(b, x) = \mathbb{P}_X(x) \mathbb{P}_\beta(b).$$

Therefore,

$$\begin{aligned} \mathbb{E}[Y \mathbb{1}\{X \in B\}] &= \int \int g(x^\top b) \mathbb{1}\{x \in B\} d\mathbb{P}_\beta(b) d\mathbb{P}_X(x) \\ &= \int \mathbb{E}[g(x^\top \beta)] \mathbb{1}\{x \in B\} d\mathbb{P}_X(x) \\ &= \mathbb{E}[\mathbb{E}[g(x^\top \beta)] \mathbb{1}\{X \in B\}] \\ &= \mathbb{E}[\phi(X) \mathbb{1}\{X \in B\}] \end{aligned}$$

where $\mathbb{E}[g(x^\top \beta)] \equiv \phi(x)$ takes expectation over β and is a function of x . By the uniqueness of conditional expectation, we have $\mathbb{E}[Y|X] = \mathbb{E}[\phi(X)]$. \square

Exercise. Let \mathcal{H} be a σ -algebra such that all elements of \mathcal{H} belong to $\sigma(X)$. We can show that

$$\mathbb{E}[\mathbb{E}[Y|X] | \mathcal{H}] = \mathbb{E}[Y | \mathcal{H}]$$

We can think of this in terms of projections. The projection of Y onto \mathcal{H} is $\mathbb{E}[Y | \mathcal{H}]$, and the projection of $\mathbb{E}[Y|X]$ onto \mathcal{H} is $\mathbb{E}[\mathbb{E}[Y|X] | \mathcal{H}]$. The equality says that the projection of the projection of Y onto \mathcal{H} is the same as the projection of Y onto \mathcal{H} .

1.2 Identification

We are given data consisting of draws from a distribution law $\mathbb{P}_{Y, X}$ where Y, X are observable vectors. An economic model consists of

1. An equation $v(Y, \gamma, X, \varepsilon; \zeta) = 0$ where v is a system of functions, γ is a vector of variables that is determined within the model but unobservable, ε is a vector of variables that is determined within the model and unobservable. ζ is a vector of functions and distributions.
2. Restrictions: $\zeta \in \mathcal{R}$ where \mathcal{R} is a set of restrictions.

For any $\zeta \in \mathcal{R}$, $\mathbb{P}_{Y, X; \zeta}$ is the distribution law of the observables generated by ζ . We denote the true structural parameter by ζ^* . **We often care about $\psi^* = \Psi(\zeta^*)$ where Ψ is a mapping from \mathcal{R} to \mathcal{P} and \mathcal{P} is the parameter space.** We define the identified set as

$$\Gamma_{Y, X}(\psi, \mathcal{R}) = \{\mathbb{P}_{Y, X; \zeta} : \zeta \in \mathcal{R} \text{ s.t. } \Psi(\zeta) = \psi\}.$$

It is the set of all distributions of the observables that are consistent with the model and the restrictions, that is, generated by ζ contained within the restriction.

Definition 1.3 (Identification). We say that ψ^* is identified if for any $\psi^* \in \mathcal{P}$ if $\Gamma_{Y,X}(\psi^*, \mathcal{R}) \cap \Gamma_{Y,X}(\psi, \mathcal{R}) \neq \emptyset$, then $\psi^* = \psi$.

Exercise. We specify a linear model $Y = f(X) + \varepsilon$ where f is continuous near $x_0 \in \text{supp}(X)$, and $\mathbb{E}[|\varepsilon| + |f(X)|] < \infty$ and $\mathbb{E}[\varepsilon|X] = 0$. We can show that $\psi^* = \mathbb{E}[f(X)]$ is identified because under these conditions $f(X) = \mathbb{E}[f(X)]$, the conditional expectation. The system of equations is $v(Y, \gamma, X, \varepsilon; \zeta) = Y - f(X) - \varepsilon = 0$. The restriction is $\zeta = (f, \mathbb{P}_{X,\varepsilon}) \in \mathcal{R}$.

Proof. Assume that there are two ζ that satisfy the restrictions and generate the same distribution of the observables.

$$(f, \mathbb{P}_{X,\varepsilon}), (f^*, \mathbb{P}_{X,\varepsilon}^*) \xrightarrow[\text{generate}]{} \mathbb{P}_{Y,X}$$

Then we have the following

- $\int \mathbb{P}_{Y,X(y,x)}(y, \cdot) dy = \int \mathbb{P}_{\varepsilon,X}(e, \cdot) de = \int \mathbb{P}_{\varepsilon,X}^*(e, \cdot) = \mathbb{P}_X(x)$
- $\mathbb{E}[Y|X] = f(X) = f^*(X)$
- Because f is identified, $\mathbb{P}_{\varepsilon,X} = \mathbb{P}_{Y-f(X),X} = \mathbb{P}_{Y-f^*(X),X} = \mathbb{P}_{\varepsilon,X}^*$

□

Now that we have introduced the basic nonparametric model, we introduce nonparametric model with instrumental variables, where $\mathbb{E}[\varepsilon|Z] = 0$. The identification requires an additional restriction – *Completeness*.

Definition 1.4 (Completeness). For any ϕ such that $\mathbb{E}[|\phi(X)|] < \infty$, $\mathbb{E}[\phi(X) | Z] = 0$ implies that $\phi(x) = 0$ on the support of X .

Discrete case When (X, Z) is discrete finite,

$$\text{supp}(X) = \{x_1, \dots, x_n\}, \text{supp}(Z) = \{z_1, \dots, z_m\}$$

We can write the completeness condition as

$$\mathbb{E}[\phi(X) | Z = z_j] = \sum \phi(x_i) \mathbb{P}_{X|Z}(X = x_i | Z = z_j) = 0 \quad \forall j$$

This is a system of m equations in n unknowns. We can show that the completeness condition is satisfied if and only if $m \geq n$. In the following section, we discuss a specific continuous case ($\text{supp}(X, Z) \in \mathbb{R}^2$) where $X = Z - \eta$ and Z is independent of ε . Both have densities and $\eta \sim \mathcal{U}([-1, 1])$.

Parametric VS Non-Parametric Models If $\zeta = (\gamma, p_\theta)$ where $\gamma \in \mathbb{R}^{d_\gamma}, \theta \in \mathbb{R}^{d_\theta}$, then the model is parametric. If $\zeta = (\gamma, p)$ where $\gamma \in \mathbb{R}^{d_\gamma}, p$ is a distribution, then the model is non-parametric. When we care about $\psi^* = \Psi(\zeta^*)$, we are in the framework of semi-parametric models.

Example 1.1. The model is $Y = \alpha + \beta X + \varepsilon$ where $\zeta = (\alpha, \beta, p_{\varepsilon,X})$ and $\psi^* = (\alpha^*, \beta^*)$. We are in semi-parametric models because α, β are finite-dimensional and $p_{\varepsilon,X}$ is infinite-dimensional.

1.3 Completeness condition

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

Definition 1.5 (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 1.6. 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 1.7 ($L^1(\mathbb{R})$ space). A function is in $L^1(\mathbb{R})$ if $\int_{\mathbb{R}} |f(x)| dx < \infty$.

Definition 1.8 (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 1.9 (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 $\mathcal{F}[f * g](w) = \mathcal{F}[f](w) \mathcal{F}[g](w)$ 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 $\mathcal{F}[f]$ is continuous and $\lim_{\|w\|_2 \rightarrow \infty} \mathcal{F}[f](w) = 0$.

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 1.2. 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 1.3. 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 $F[K] \in L^2(\mathbb{R})$. Note also that $F[K](w) = 0$ if and only if $w = \pm k\pi$ for $k \in \mathbb{N}$.

Completeness Let us check whether the functions given in Example 1.2 and 1.3 satisfy the completeness condition 1.5 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

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

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

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

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

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

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

Definition 2.1 (kernel function). A kernel is a function $K : \mathbb{R} \rightarrow \mathbb{R}$ 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. $K(x) = \frac{1}{2} \mathbb{1}_{|x| \leq \frac{1}{2}}$ the rectangular
2. $K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, the Gaussian kernel
3. $K(x) = \frac{\sin(x)}{\pi x}$, the sinc kernel
4. $K(x) = \frac{3}{4} (1 - x^2) \mathbb{1}_{|x| \leq 1}$, the Epanechnikov kernel

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

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

2.4 Performance analysis

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

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

where

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

is the **loss** function.

Other risks include

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

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 $\hat{f}_X(x)$ by

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

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

Proposition 2.1. We have

$$\text{MSE}(x) = b(x)^2 + \sigma^2(x).$$

Proof. We have

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

□

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

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

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

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

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

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

Inserting equation* 2.4,

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

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

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 $f : \mathbb{R} \rightarrow \mathbb{R}$ such that $f^{(\ell)}$ exists and for all $x, x' \in \mathbb{R}$ we have

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

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 $x, x' \in \mathbb{R}$

$$|f(x) - f(x')| \leq L |x - x'|^\beta.$$

Remark. This Hölder condition implies continuity of f .

Definition 2.7 (β for a kernel). $K : \mathbb{R} \rightarrow \mathbb{R}$ 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

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

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

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

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

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

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

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

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 $f, g : \mathbb{R} \rightarrow \mathbb{R}$ 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,

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

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:

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

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

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

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

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

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

3 Sobolev class and symmetric kernel

3.1 Review of Fourier transform

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

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

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 $\|f - f_m\|_2^2 \rightarrow 0$ as $m \rightarrow \infty$ 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,

$$\|f_m - f\|_2^2 = \int_{-\infty}^m |f(x)|^2 dx + \int_m^{\infty} |f(x)|^2 dx \rightarrow 0 \quad (1)$$

as $m \rightarrow \infty$. By equation 1, for all m, m' , $\|f_m - f_{m'}\|_2^2 \rightarrow 0$ as $m, m' \rightarrow \infty$, 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

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

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

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

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

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

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	$\mathcal{F}[K]$	$\frac{ 1-\mathcal{F}[K](w) }{ w ^\beta}$
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, $(\sum_{i=1}^n a_i)^2 \leq 2^{n-1} \sum_{i=1}^n a_i^2$.

4.2 Cross validation

Let us write

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

introduced

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

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

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

We claim

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

We have

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

As an exercise, show that this yields the claim.

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

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

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. $\mathcal{P}_{\alpha,r} = \{f \in L^2(\mathbb{R}) \text{ such that } \int \exp(\alpha |w|^2) |\phi(w)|^2 dw \leq L^2\}$ where $\phi = \mathcal{F}[f]$ is the Fourier transform of f . We can show that a MISE optimal kernel density estimator could have a risk less than $C \frac{(\log n)^{1/r}}{n}$.
2. $\mathcal{P}_{\alpha,r} = \{f \in L^2(\mathbb{R}) \text{ such that } \text{supp}(\phi) \subset [-a, a]\}$. 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 $f_X \in L^2([0, 1]^d)$, where $L^2([0, 1]^d)$ 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, $L^2([0, 1]^d)$ has a countable basis $(e_j)_{j=1}^\infty$, which is a sequence of functions in $L^2([0, 1]^d)$ 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 $f \in L^2([0, 1]^d)$,

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

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

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

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

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

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

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

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

so that an unbiased estimator is

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

Thus, a candidate estimator for f_X is

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

where

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

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 $q_T(X_i, x)$ plays the role of a kernel and T plays the same role as $\frac{1}{h}$. On $L^2([0, 1]^d)$ 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 $L^2([0, 1])$ and (j_1, \dots, j_d) plays the role of the index j^2 . For example, $f_k(x) = \sqrt{2} \sin(\pi k x)$ is a basis of $L^2([0, 1])$. This gives

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

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

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

We define

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

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

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

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

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

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,

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

since

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

where we used $f_{j_k}^2 \leq 2$ for our particular basis $f_{j_k}(x) = \sqrt{2} \sin(\pi j_k x)$. 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 is classical to estimate many coefficients c_j , for T much larger than before (e.g. \sqrt{n}) and work with the estimators

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

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 $f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$ if $f_X(x) > 0$.

Also the conditional expectation function $\mathbb{E}[Y|X=x]$ is given by

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

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

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

6.2 Nadaraya-Watson estimator

This leads to the following estimator called **Nadaraya-Watson estimator**

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

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 $\text{supp}(X) = [a, b]$ 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

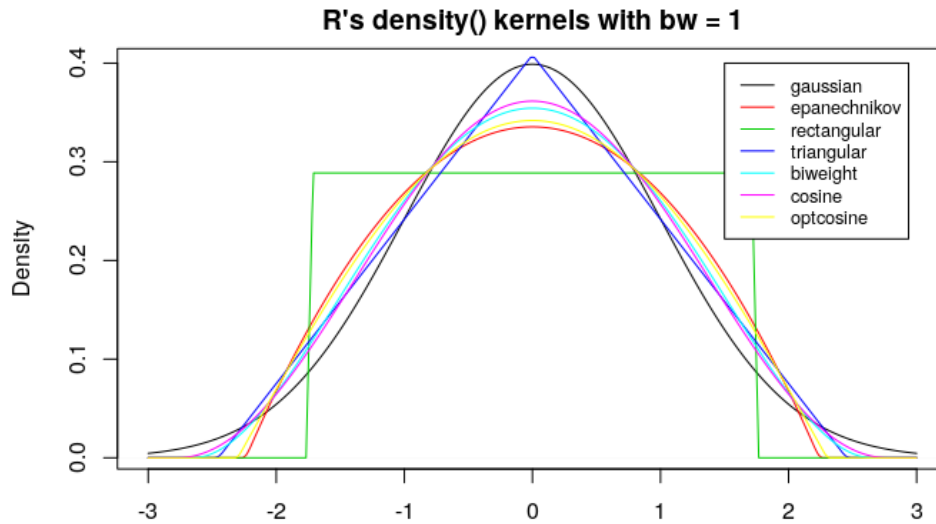


Figure 1: Kernels

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).

6.3 Local Polynomial Estimation

tutorial 3

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 $y_i = y(0)(1 - D) + y(1)D$. The potential outcome is $y_i(1), y_i(0)$, which are not observed. We can only observe $y_i = y(D_i)$.

7.2 Parameters of Interest

- Average treatment effect (ATE): $\tau = \mathbb{E}[Y(1) - Y(0)]$
- Conditional average treatment effect (CATE): $\tau(x) = \mathbb{E}[Y(1) - Y(0) | X = x]$. 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): $\tau_{\text{ATT}} = \mathbb{E}[Y(1) - Y(0) | D = 1]$
- Average treatment effect on the untreated (ATU): $\tau_{\text{ATU}} = \mathbb{E}[Y(1) - Y(0) | D = 0]$
- Conditional average treatment effect on the treated (CATT): $\tau_{\text{ATT}}(x) = \mathbb{E}[Y(1) - Y(0) | D = 1, X = x]$

7.3 Identification

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

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

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

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

Assumption 7.3.4. *The potential outcome $Y(1), Y(0)$ 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

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

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 $x \in \text{supp}(X)$*

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

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

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

3. $\mathbb{E}[Y(1) \mid X = x, D = 1] =$

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

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

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,

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

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

$$\begin{aligned} (\hat{\alpha}_1, \hat{\beta}_1) &= \arg \min_{\alpha, \beta} \sum_{i=1, X_i \geq c}^n K\left(\frac{X_i - x}{h}\right) (Y_i - \alpha - \beta(X_i - c))^2 \\ (\hat{\alpha}_0, \hat{\beta}_0) &= \arg \min_{\alpha, \beta} \sum_{i=1, X_i < c}^n K\left(\frac{X_i - x}{h}\right) (Y_i - \alpha - \beta(X_i - c))^2 \end{aligned}$$

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

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

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

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

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 $\lim_{x \rightarrow c^+} \pi(x) \neq \lim_{x \rightarrow c^-} \pi(x)$. 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.

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

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:

$$Y = Y(0, 0) + Z(Y(1, 0) - Y(0, 0)) + D(Y(0, 1) - Y(0, 0)) + DZ(Y(1, 1) - Y(0, 1) - Y(1, 0) + Y(0, 0)) \quad (3)$$

and

$$D = D(0) + Z(D(1) - D(0)) \quad (4)$$

Preliminaries We define the following:

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

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 $\mathbb{E}[\epsilon | X, Z] = 0$ assumption in standard linear IV model.

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

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

which is the average treatment effect for the compliers.

7.5.1 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

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

The intention to treat on outcome is defined as

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

The intention to treat on outcome for compliers is defined as

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

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 $D(1) = 1$ (complier), then $Y(0, d) = Y(1, d)$ for $d = 0, 1$.

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

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

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

Then it can be shown that

$$ITT_Y = ITT_{Y,CO} ITT_D$$

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

$$\begin{aligned}\tau_{LATE} &= \mathbb{E}[Y(Z, 1) - Y(Z, 0) \mid D(1) - D(0) = 1] \\ ITT_{Y|D(1)-D(0)=1} &= \mathbb{E}[Y(1, 1) - Y(0, 0) \mid D(1) - D(0) = 1]\end{aligned}$$

That is,

$$\tau_{LATE} = ITT_{Y|D(1)-D(0)=1}$$

Therefore,

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

7.5.2 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

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

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

Estimation We can estimate τ_{LATE} by

$$\frac{\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]}{\mathbb{E}[D \mid Z = 1] - \mathbb{E}[D \mid Z = 0]} = \frac{Cov(Y, Z)}{Cov(D, Z)}$$

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

tutorial 4