Lec 10: Nonparametric Model (ML)

Eric Hsienchen Chu*

Spring, 2024

(⊛) Suggested readings: Hansen (2022), Ch19.

1 Nonparametric & ML

Overview. Unless an economic model restricts the form of $m(x)$ to a parametric function, $m(x)$ can take any nonlinear shape and is therefore **nonparametric**.

$$Y = m(X) + \varepsilon,\ \mathbb{E}[\varepsilon|X] = 0 \tag{1.1}$$

Here, the parameter of interest $m(X) = \mathbb{E}[Y|X]$ is *infinite* dimensional. In particular, we may want to discuss kernel density estimators of $m(x)$.

Question. How do we estimate $\mathbb{E}[Y|X = x] = m(x)$, where X has continuum supp?
Answer. There are several ways:

① $\hat{m}(x) = \frac{1}{|\mathcal{N}(x)|} \sum_{i \in \mathcal{N}(x)} Y_i$, where $\mathcal{N}(x) \equiv \{i = 1, \cdots, n : x_i \text{ "close" to } x\}$

$\Longrightarrow$ k-nearest neighbors (KNN), Regression trees, $\cdots$

② $m(x) = \mathbb{E}[Y|X = x] = \int y \cdot f_{Y|X}(y|x)dy = \int y \frac{f_{YX}(y,x)}{f_X(x)} dy$

$\Longrightarrow$ It suffices to estimate the **density** f_{YX} & f_X!

⊛ **Machine Learning** ("Modern Nonparametrics")

- Bias–variance Trade–off (★)
- Curse of dimensionality
- Tuning Parameter Selection

*Department of Economics, University of Wisconsin-Madison. hchu38@wisc.edu. This is lecture notes from the second half of ECON710: Economic Statistics and Econometrics II. Instructor: Prof. Harold Chiang. Materials and sources: Harold's handwritten notes.

2 Kernel Density Estimation

Motivation. If x is **discrete** (finite support), then $\mathbb{P}(X = x)$ can be calculated by:

$$\hat{f}_X(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i = x\} \tag{2.1}$$

However, this does not work well if X takes many values & does not work *at all* if X has atomless distribution ($\mathbb{P}(X = x) = 0$, atomless). What are our options?

Definition 2.1 (Histogram)**.** A histrogram has:

$$\hat{f}_X(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}\{x - \frac{h}{2} \leq X_i \leq x + \frac{h}{2}\}, \ (\bigstar) \tag{2.2}$$

where $(h; h > 0)$ is "bandwith" (tuning parameter).

Remark (Empirical CDF & Histogram)**.** Recall that Empirical CDF is defined by:

$$\hat{F}_X(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq x\}, \tag{2.3}$$

which estimates $F_X(x) = \mathbb{P}(X_i \leq x)$. By definition of limits, we have:

$$f_X(x) = F'_X(x) \equiv \lim_{h \to 0} \frac{F_X(x + \frac{h}{2}) - F_X(x - \frac{h}{2})}{h} \tag{2.4}$$

Then, for a small enough h, we know that:

$$\begin{aligned} \hat{f}_X(x) &= \frac{\hat{F}_X(x + \frac{h}{2}) - \hat{F}_X(x - \frac{h}{2})}{h} \longleftarrow \text{for some small } h \qquad (2.5) \\ &= \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}\{x - \frac{h}{2} \leq X_i \leq x + \frac{h}{2}\} \longleftarrow \text{by } (\bigstar) \qquad (2.6) \end{aligned}$$

which is exactly the historgram at some fixed x!

Definition 2.2 (Kernal Density Estimation; KDE)**.** If we set $\mathcal{K}(u) = \mathbb{1}\{\frac{-1}{2} \leq u \leq \frac{1}{2}\}$, then the histogram at a fixed x is given by:

$$\hat{f}_X(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathcal{K}\left(\frac{X_i - x}{h}\right) \tag{2.7}$$

The function $\mathcal{K}$ is called rectangular/**uniform kernel**.

Remark. We can also use the other PDF's kernel as well.

Example 2.1 (Triangular Kernel). $\mathcal{K}(u) = \begin{cases} 1 - |u|, & \text{if } -1 \leq u \leq 1 \\ 0, & \text{else} \end{cases}$

Example 2.2 (Epanechnikov Kernel). $\mathcal{K}(u) = \begin{cases} \frac{3}{4}(1 - u^2), & \text{if } -1 \leq u \leq 1 \\ 0, & \text{else} \end{cases}$

Example 2.3 (Gaussian Kernel). $\mathcal{K}(u) = \phi(u) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}u^2}, \ u \in \mathbb{R}$

Fact 2.1 (Kernels). Consider the kernels of the Examples above:

	Uniform (1)	Triangular (2)	Epanechnikov (3)	Gaussian (4)
$\int \mathcal{K}(u)du$ (Prob.)	1	1	1	1
Smoothness	Discrete	$\mathcal{C}$	$\mathcal{C}$	$\mathcal{C}^\infty$
$\int \mathcal{K}(u)^2 du$ (★)	1	2/3	3/5	$1/\sqrt{2\pi}$
$\int u\mathcal{K}(u)du$ (Mean)	0	0	0	0
$\int u^2\mathcal{K}(u)du$ (★★)	1/12	1/6	1/5	1

Note: $\int \mathcal{K}(u)^2 du$ is useful for $\text{Var}(\hat{f}_X(x))$. $\int u^2\mathcal{K}(u)du$ is useful for $\text{Bias}(\hat{f}_X(x))$.

The key is that we want (★) & (★★) to be *finite* ($< \infty$). With an $\mathcal{K}$ chosen, the density estimator is then:

$$\hat{f}_X(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathcal{K}\left(\frac{X_i - x}{h}\right) \tag{2.8}$$

3 Bias–Variance

Motivation. As hinted before, we will discuss the Bias–variance trade–off (Spoiler at Fact 3.3). But we need to establish some terms first.

Definition 3.1. Fix an $x \in int\left(\text{supp}(x)\right)$, then:

- $\text{Bias}\left(\hat{f}_X(x)\right) = \mathbb{E}[\hat{f}_X(x)] - f_X(x)$ $\longleftarrow$ dist of my (exp'd) density estimator to the true density
- $\text{Var}\left(\hat{f}_X(x)\right) = \mathbb{E}\left[\left(\hat{f}_X(x) - \mathbb{E}[\hat{f}_X(x)]\right)^2\right]$
- $\text{MSE}\left(\hat{f}_X(x)\right) = \mathbb{E}\left[\left(\hat{f}_X(x) - f_X(x)\right)^2\right] = \left[\text{Bias}\left(\hat{f}_X(x)\right)\right]^2 + \text{Var}\left(\hat{f}_X(x)\right)$ $(\spadesuit)$

Example 3.1 (MSE)**.** Let's actually show $\text{MSE}\left(\hat{f}_X(x)\right) = \left[\text{Bias}\left(\hat{f}_X(x)\right)\right]^2 + \text{Var}\left(\hat{f}_X(x)\right)$ by the "add & subtract" trick **[Spring 2023 Final Q2]**:

$$\begin{aligned}
\text{MSE}\left(\hat{f}_X(x)\right) &= \mathbb{E}\left[\left(\hat{f}_X(x) - f_X(x)\right)^2\right] && (3.1)\\
&= \mathbb{E}\left[\left(\hat{f}_X(x) - \mathbb{E}[\hat{f}_X(x)] + \mathbb{E}[\hat{f}_X(x)] - f_X(x)\right)^2\right] && (3.2)\\
&= (\spadesuit) + 2\mathbb{E}\left[\left(\hat{f}_X(x) - \mathbb{E}[\hat{f}_X(x)]\right)\left(\mathbb{E}[\hat{f}_X(x)] - f_X(x)\right)\right] && (3.3)\\
&= (\spadesuit) + 2\underbrace{\mathbb{E}\left[\hat{f}_X(x) - \mathbb{E}[\hat{f}_X(x)]\right]}_{=\mathbb{E}[\hat{f}_X(x)] - \mathbb{E}[\hat{f}_X(x)] = 0}\mathbb{E}\left[\mathbb{E}[\hat{f}_X(x)] - f_X(x)\right] && (3.4)\\
&= (\spadesuit) = \left[\text{Bias}\left(\hat{f}_X(x)\right)\right]^2 + \text{Var}\left(\hat{f}_X(x)\right) \ \square && (3.5)
\end{aligned}$$

Lemma 3.1 (Bias KDE)**.** Suppose $(X_i)_{i=1}^n \overset{iid}{\sim} X \sim f_X$. If:

① $\left\|f'''\right\|_\infty < \infty$, and

② $\int u^3 \mathcal{K}(u) du < \infty$

Then, as $h \to 0$ (i.e., choosing small h), the bias of density estimator is:

$$\text{Bias}\left(\hat{f}_X(x)\right) = \frac{\mathbf{h^2}}{2} f_X''(x) \int u^2 \mathcal{K}(u) du + \text{o}(h^2) \tag{3.6}$$

Remark. Equation (3.6) means that $\text{Bias}\left(\hat{f}_X(x)\right) \sim \mathbf{h^2} = \text{O}(h^2)$.

Proof. By definition, we have $\text{Bias}\left(\hat{f}_X(x)\right) = \underbrace{\mathbb{E}[\hat{f}_X(x)]}_{\circledast} - f_X(x)$. Let's look closely for $\circledast$:

$$\begin{aligned}
\mathbb{E}[\hat{f}_X(x)] &= \mathbb{E}\left[\frac{1}{\cancel{n}h}\cancel{\sum_{i=1}^{n}}\mathcal{K}\left(\frac{X_i - x}{h}\right)\right] && (3.7)\\
&= \frac{1}{h}\mathbb{E}\left[\mathcal{K}\left(\frac{X_i - x}{h}\right)\right] \longleftarrow \text{by identical distribution \& linearity} && (3.8)\\
&= \frac{1}{h}\int \mathcal{K}\left(\frac{\xi - x}{h}\right) f_X(\xi)d\xi \longleftarrow \text{let } u = \frac{\xi - x}{h};\ du = \frac{1}{h}d\xi && (3.9)\\
&= \int \mathcal{K}(u)\, f_X(x + hu)du && (3.10)\\
&= \int \mathcal{K}(u) \underbrace{\left[f_X(x) + \frac{(hu)^1}{1!}f'_X(x) + \frac{(hu)^2}{2!}f''_X(x) + \mathrm{O}\left((hu)^3\right)\right]}_{\text{Taylor Expansion}} du && (3.11)\\
&= f_X(x)\underbrace{\int \mathcal{K}(u)du}_{=1} + hf'_X(x)\underbrace{\int u\mathcal{K}(u)du}_{=0} + \frac{h^2}{2}f''_X(x)\int u^2\mathcal{K}(u)du + \mathrm{o}(h^2) && (3.12)\\
&= f_X(x) + \frac{h^2}{2}f''_X(x)\int u^2\mathcal{K}(u)du + \mathrm{o}(h^2) && (3.13)
\end{aligned}$$

where Equation (3.11) holds by Taylor expansion. So, the bias is then:

$$\begin{aligned}
\text{Bias}\left(\hat{f}_X(x)\right) &= \mathbb{E}[\hat{f}_X(x)] - f_X(x) && (3.14)\\
&= \cancel{f_X(x)} + \frac{h^2}{2}f''_X(x)\int u^2\mathcal{K}(u)du + \mathrm{o}(h^2) - \cancel{f_X(x)} && (3.15)\\
&= \frac{\mathbf{h^2}}{2}f''_X(x)\int u^2\mathcal{K}(u)du + \mathrm{o}(h^2) && (3.16)
\end{aligned}$$

Note that if the curvature of the density: $f''_X(x) \neq 0$, then $\text{Bias}\left(\hat{f}_X(x)\right) \sim \mathbf{h^2}$ as $h \to 0$. □

Remark. Later we'll see a small h gives us smaller bias, but yields larger variance.

Question. How many times of Taylor Expansion we need to perform?
Answer. Until the first non-zero moment of density. In this case, we TE twice. See Spring24 TA Handout 11 Q2(a) for Higher–order Kernels (TE 4 times) & Q1(a) (TE 1 time).

Lemma 3.2 (Variance KDE)**.** Suppose $(X_i)_{i=1}^n \overset{iid}{\sim} f_X$. If:

① $\left\| f''' \right\|_\infty < \infty$, and

② $\int u^3 \mathcal{K}(u) du < \infty$

Then, as $h \to 0$ (i.e., choosing small h), the variance of density estimator is:

$$\text{Var}\left(\hat{f}_X(x)\right) = \frac{1}{\mathbf{nh}} f_X(x) \int \mathcal{K}(u)^2 du + \text{o}(\frac{1}{nh}) \tag{3.17}$$

Remark. *The proof details were left as exercises and ended up in Spring 2024 Final. I am not sure I completed it correctly but here is what I put on the exam.*

Proof. Similarly, by definition of $\text{Var}\left(\hat{f}_X(x)\right)$, we have:

$$\begin{aligned}
\text{Var}\left(\hat{f}_X(x)\right) &= \text{Var}\left(\frac{1}{nh}\sum_{i=1}^n \mathcal{K}\left(\frac{X_i - x}{h}\right)\right) && (3.18)\\
&= \frac{1}{n^2h^2}\text{Var}\left(\sum_{i=1}^n \mathcal{K}\left(\frac{X_i - x}{h}\right)\right) \longleftarrow \text{by independent} && (3.19)\\
&= \frac{1}{nh^2}\text{Var}\left(\mathcal{K}\left(\frac{X_i - x}{h}\right)\right) \longleftarrow \text{by identical} && (3.20)\\
&= \frac{1}{nh^2}\left[\underbrace{\mathbb{E}\left[\mathcal{K}\left(\frac{X_i - x}{h}\right)^2\right]}_{\equiv \mathcal{A}} - \underbrace{\mathbb{E}\left[\mathcal{K}\left(\frac{X_i - x}{h}\right)\right]^2}_{\equiv \mathcal{B}}\right] \quad (\bigstar) && (3.21)
\end{aligned}$$

Let's derive $\mathcal{A}$ and $\mathcal{B}$ separately:

$$\begin{aligned}
\mathcal{A} &\equiv \mathbb{E}\left[\mathcal{K}\left(\frac{X_i - x}{h}\right)^2\right] && (3.22)\\
&= \int \mathcal{K}\left(\frac{\xi - x}{h}\right)^2 f_X(\xi) d\xi \longleftarrow \text{let } u = \frac{\xi - x}{h};\ du = \frac{1}{h} d\xi && (3.23)\\
&= h \int \mathcal{K}\left(u\right)^2 f_X(x + hu) du && (3.24)\\
&= h \int \mathcal{K}\left(u\right)^2 \left[f_X(x) + \frac{(hu)^1}{1!} f'_X(x) + \text{O}\left((hu)^2\right)\right] du && (3.25)\\
&= h f_X(x) \int \mathcal{K}(u)^2 du + h f'_X(x) \underbrace{\int u \mathcal{K}(u) du}_{=0} + \text{o}(h) && (3.26)\\
&= h f_X(x) \int \mathcal{K}(u)^2 du + \text{o}(h) && (3.27)
\end{aligned}$$

And,

$$\mathcal{B} \equiv \mathbb{E}\left[\mathcal{K}\left(\frac{X_i - x}{h}\right)\right]^2 \tag{3.28}$$

$$= \left[\int \mathcal{K}\left(\frac{\xi - x}{h}\right) f_X(\xi) d\xi\right]^2 \longleftarrow \text{let } u = \frac{\xi - x}{h};\ du = \frac{1}{h} d\xi \tag{3.29}$$

$$= \left[h \int \mathcal{K}(u)\, f_X(x + hu) du\right]^2 \tag{3.30}$$

$$= \left[h \int \mathcal{K}(u) \left[f_X(x) + \frac{(hu)^1}{1!} f_X'(x) + \mathrm{O}\left((hu)^2\right)\right] du\right]^2 \tag{3.31}$$

$$= \left[h f_X(x) + \mathrm{o}(h)\right]^2 \tag{3.32}$$

$$= \mathrm{O}(h^2) \tag{3.33}$$

At Eqn (3.25) and (3.31) we perform Taylor expansions just as in **Bias KDE**.
So now (★) becomes:

$$\frac{1}{nh^2}\left[\mathbb{E}\left[\mathcal{K}\left(\frac{X_i - x}{h}\right)^2\right] - \mathbb{E}\left[\mathcal{K}\left(\frac{X_i - x}{h}\right)\right]^2\right] = \frac{1}{nh^2}\left[h f_X(x) \int \mathcal{K}(u)^2 du + \mathrm{o}(h) + \mathrm{O}(h^2)\right]$$

$$= \frac{1}{\mathbf{nh}} f_X(x) \int \mathcal{K}(u)^2 du + \mathrm{o}(\frac{1}{nh}) \tag{3.34}$$

Note that as $h \to 0$, $\mathrm{Var}(\hat{f}_X(x)) \nearrow \infty$. □

Fact 3.3 (Bias–Variance trade–off)**.** Now the trade–off should be obvious:

- Bias $\left(\hat{f}_X(x)\right) = \frac{\mathbf{h^2}}{2} f_X''(x) \int u^2 \mathcal{K}(u) du + \mathrm{o}(h^2) \nearrow 0$ as $h \to 0$
- Var $\left(\hat{f}_X(x)\right) = \frac{1}{\mathbf{nh}} f_X(x) \int \mathcal{K}(u)^2 du + \mathrm{o}(\frac{1}{nh}) \nearrow \infty$ as $h \to 0$ (fixed n)

So, it's either (small bias, large variance) ↔ (large bias, small variance).

Remark. See Harold's notes for MSE and optimal bandwidth selection ($h^{\text{opt}} \sim n^{\frac{-1}{5}}$), discussion of parametrics vs nonparametrics, and results of Consistency & AN for nonparametrics.

References

Hansen, B. E. (2022). Econometrics. Princeton University Press. https://users.ssc.wisc.edu/~bhansen/econometrics/