非参数统计学和机器学习：基本思想、方法与相互关系

洪永淼
中国科学院数学与系统科学研究院
中国科学院大学经济与管理学院
2021年4月20日

目录

问题

- 问题1：什么是**泰勒级数展开**（Taylor Series Expansion）？
- 问题2：什么是**傅里叶级数展开**（Fourier Series Expansion）？
- 问题3：什么是**样本均值 $\bar{X}_n$**？

假设 $\{X_i\}_{i=1}^n$ 是一个独立同分布（IID）随机样本，其样本均值定义为

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

(1) $E(\bar{X}_n) = \mu$

(2) $\mathrm{var}(\bar{X}_n) = \frac{\sigma^2}{n} \to 0 \text{ as } n \to \infty$

导论

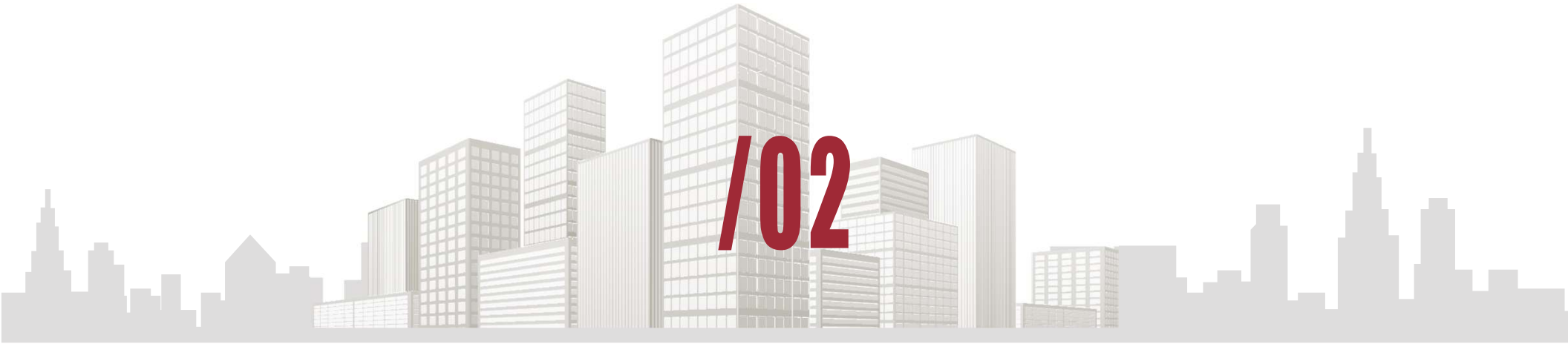
/01

1. 导论

- 非参数分析（**Nonparametric Analysis**）是统计学和计量经济学的一个重要方法，应用广泛
  - 什么是“非参”？
  - “非参”有什么用处?
  - 如何使用“非参”？
  - 如何解释“非参”？
  - “非参”与机器学习的关系是什么?

- We will provide a **unified approach** to viewing various nonparametric methods and their relationships with machine learning.

非参数分析的重要性

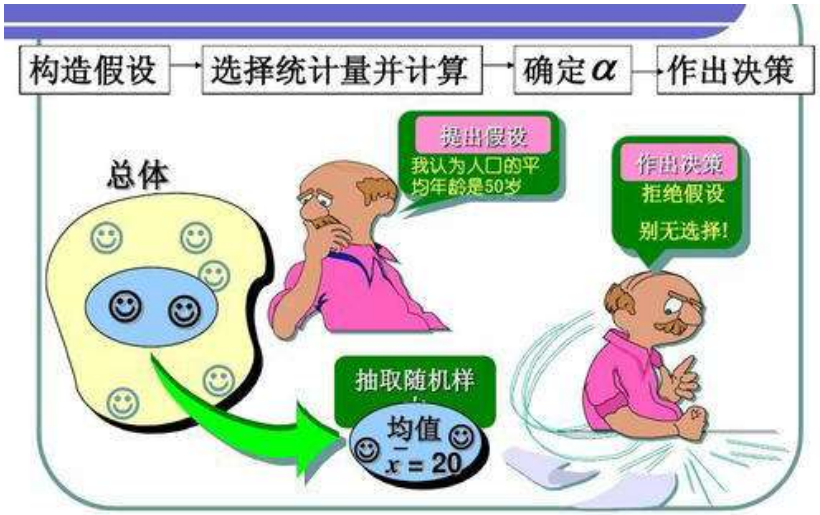


/02

2. 非参数分析的重要性

什么是非参数分析?

- 非参数分析是相对于参数分析（Parametric Analysis）而存在的
- 为了说明什么是非参数分析及其作用，我们首先考察经济学中一个参数分析例子



2. 非参数分析的重要性

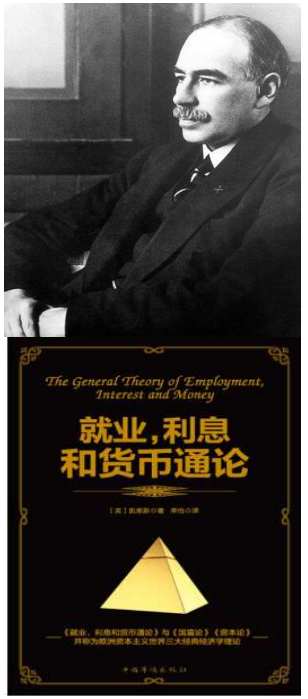
- **例1 [消费函数]：**凯恩斯乘数效应（Multiplier Effect）理论的一个核心概念

  ➢凯恩斯理论可简化为以下两个方程式：

  （1）National Income Identity: $Y_t = C_t + I_t + G_t + E_t$

  （2）**Consumption Function: $C_t = \alpha + \beta Y_t + \varepsilon_t$**

  其中 ε_t 代表除收入外所有其他因素对消费 C_t 的影响



2. 非参数分析的重要性

➢假设 ε_t 满足约束条件$E(\varepsilon_t|Y_t)=0$，因此消费 C_t 的条件均值是收入 Y_t 的线性函数：

$$E(C_t|Y_t)=\alpha+\beta Y_t$$

其一阶导数

$$\frac{dE(C_t|Y_t)}{dY_t}=\beta=\text{Marginal Propensity to Consume (MPC)}$$

这里参数 β 可解释为边际消费倾向

2. 非参数分析的重要性

➢从方程（1）和（2），可得政府公共支出的**乘数效应**

$$\frac{\partial Y_t}{\partial G_t} = \frac{1}{1-\beta} = 4 \quad \text{if} \quad \beta = 0.75$$

乘数效应的大小取决于 MPC，即 β 值

2. 非参数分析的重要性

- 假设消费函数是收入的二项式：

$$C_t = \alpha + \beta Y_t + \gamma Y_t^2 + \varepsilon_t$$

其中 $E(\varepsilon_t | Y_t) = 0$，则

$$\boldsymbol{E(C_t|Y_t) = \alpha + \beta Y_t + \gamma Y_t^2,} \qquad \mathbf{MPC} = \frac{\boldsymbol{dE(C_t|Y_t)}}{\boldsymbol{dY_t}} = \boldsymbol{\beta + 2\gamma Y_t}$$

可以看到，MPC 不再是一个常数，而与收入水平 Y_t 有关

2. 非参数分析的重要性

- 如果真实的消费函数是收入的二次项，但我们假设一个线性消费模型

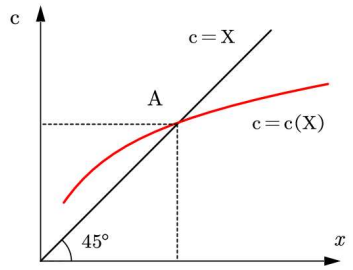
$$C_t = \alpha + \beta Y_t + \varepsilon_t$$

此时 $E(\varepsilon_t|Y_t) \neq 0$，线性模型是一个误设模型，参数 $\beta \neq$ MPC

- $E(\varepsilon_t|Y_t) = 0$ 是消费函数模型正确设定的条件
- MPC 取决于消费函数的函数形式

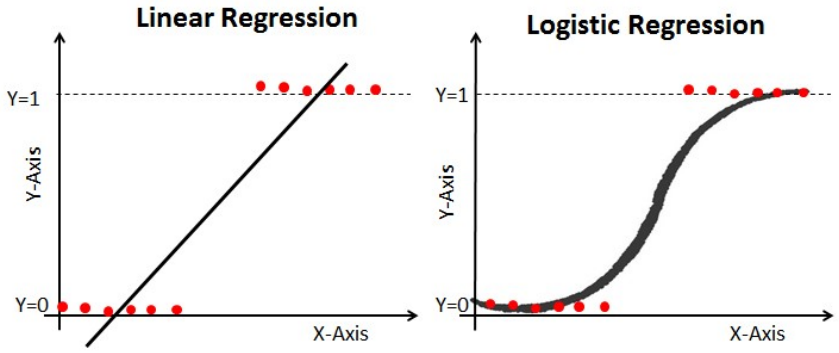
2. 非参数分析的重要性

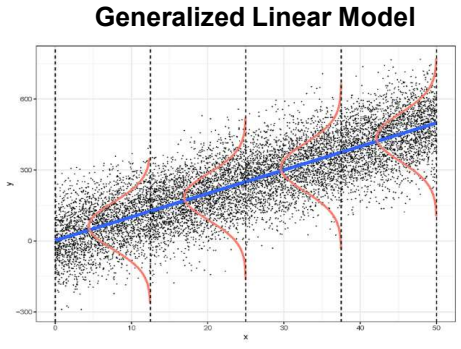
- **参数方法**：假设消费函数是收入的线性函数或某个具体的已知函数形式，其中包含少数（低维）未知参数，这叫做参数方法（Parametric Approach），相应的模型叫做参数模型（Parametric Model）。参数分析是统计学与计量经济学最常用的分析方法
- 计量经济学分析就是估计未知参数值，推断其统计显著性，进而判断其经济重要性，赋予经济解释，并提出政策建议等应用
- 判断准则：5% 显著性水平，P-值

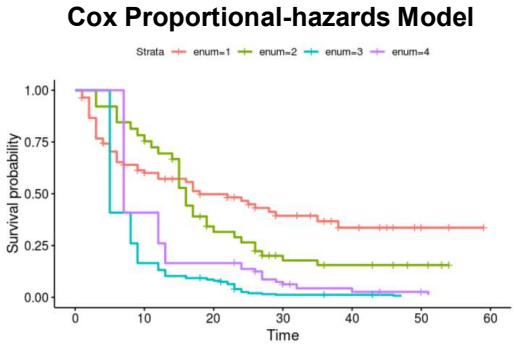


2. 非参数分析的重要性

- 常见的参数模型
  - ➢ 线性回归模型
  - ➢ 逻辑回归模型
  - ➢ 广义线性回归模型
  - ➢ Cox's (1972) 比例危险模型

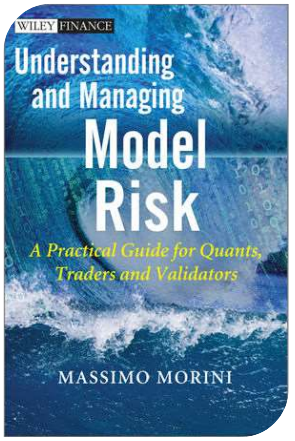






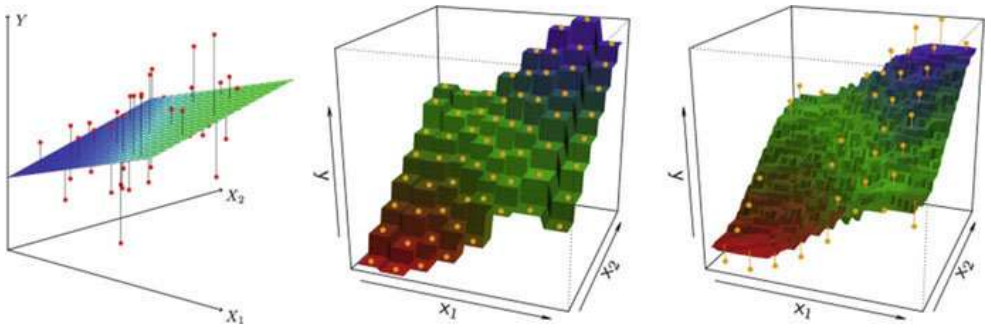
2. 非参数分析的重要性

- 任何参数模型都有误设的可能性，如函数形式误设、存在遗漏变量或结构变化等
- **模型误设（Model Misspecification）会导致什么后果?**
  - 可能无法一致估计真实参数值
  - 参数的经济解释（如 MPC，弹性系数等）不再有效
  - 政策建议误导（如 Multiplier Effect Prediction, COVID-19 Cases Prediction）
  - 模型风险（如定价误差而导致严重经济损失）

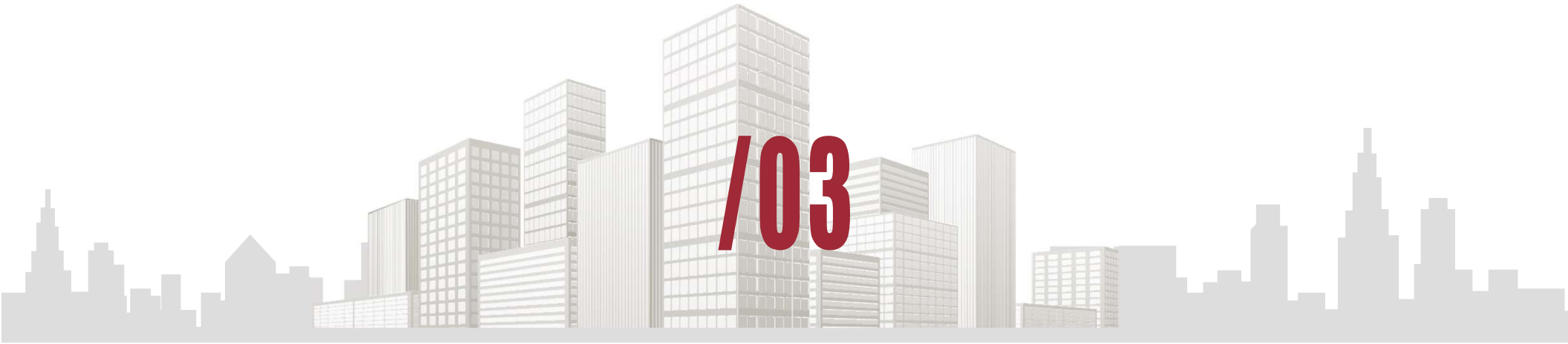


2. 非参数分析的重要性

- 应对之道
  - 统计学：Diagnostic Testing and Model Specification Testing
  - 非参数分析：不假设具体函数形式



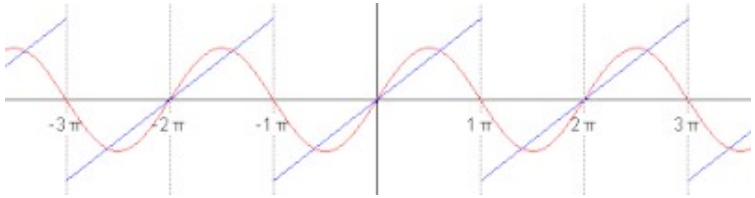
什么是非参数分析
(Nonparametric Analysis)

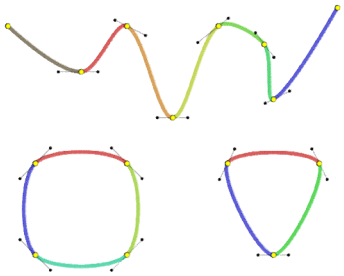


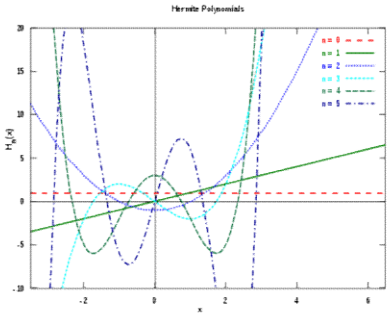
/03

3. 什么是非参数分析（Nonparametric Analysis）

- **非参数分析：不假设具体函数形式，让数据告诉真实的函数形式**
- 重要非参数方法：
  - 级数近似（Series Approximation）
    - ✓ Fourier Series
    - ✓ Polynomials
    - ✓ Hermite Polynomials
  - 样条平滑法（Spline Smoothing）
  - 核平滑法（Kernel Smoothing）
  - 局部多项式平滑法（Local Polynomial Smoothing）







3. 什么是非参数分析（Nonparametric Analysis）

非参数分析方法的发展简史

1946-1948

Spectral Density Estimation

1956, 1962

Rosenblatt (1956) -Parzen (1962): Kernel Probability Density Estimation

1964

Nadaraya (1964)-Watson (1964): Regression Estimation

1970-1990

- Series
- Splines
- Local Polynomial
- Artificial Neural Network

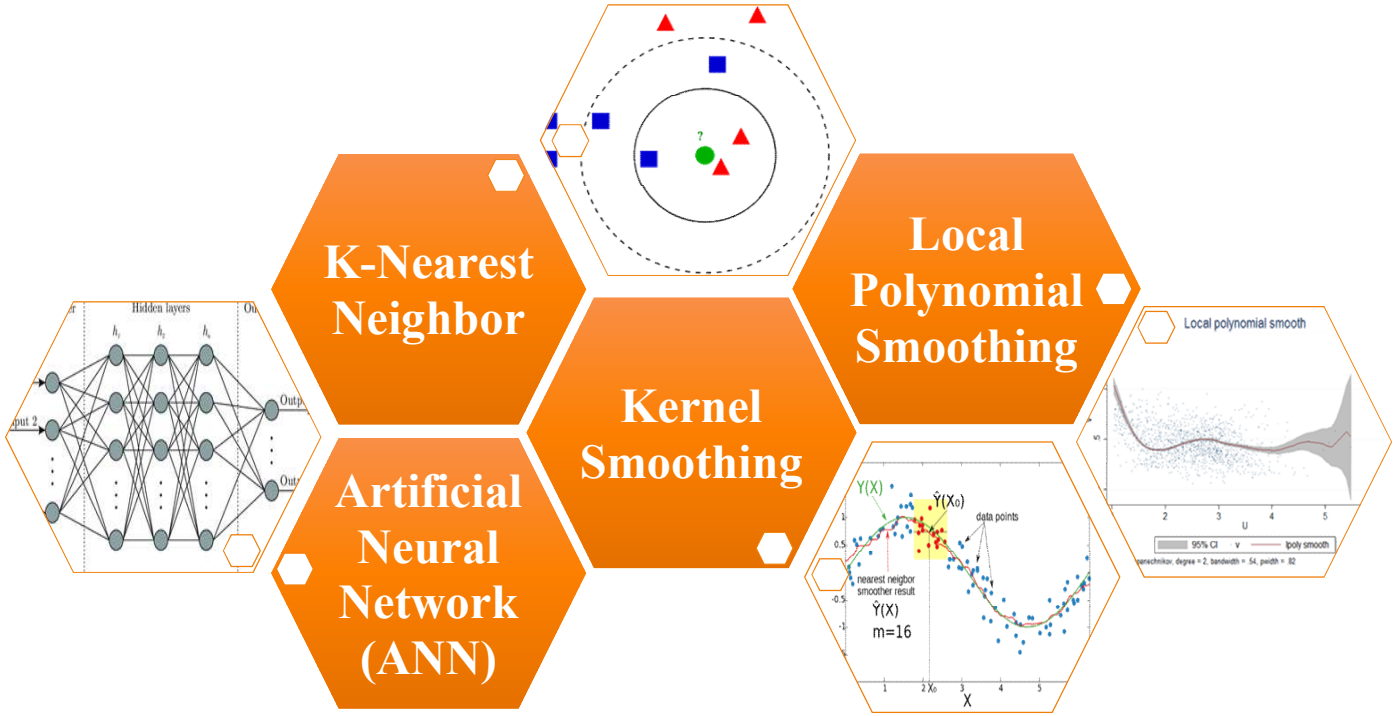
$$h(w) = \int_{-\pi}^{\pi} \hat{I}(w) W_T(\lambda - w) d\lambda, w \in [-\pi, \pi]$$

平滑（Smoothing）：非参数分析的基本思想

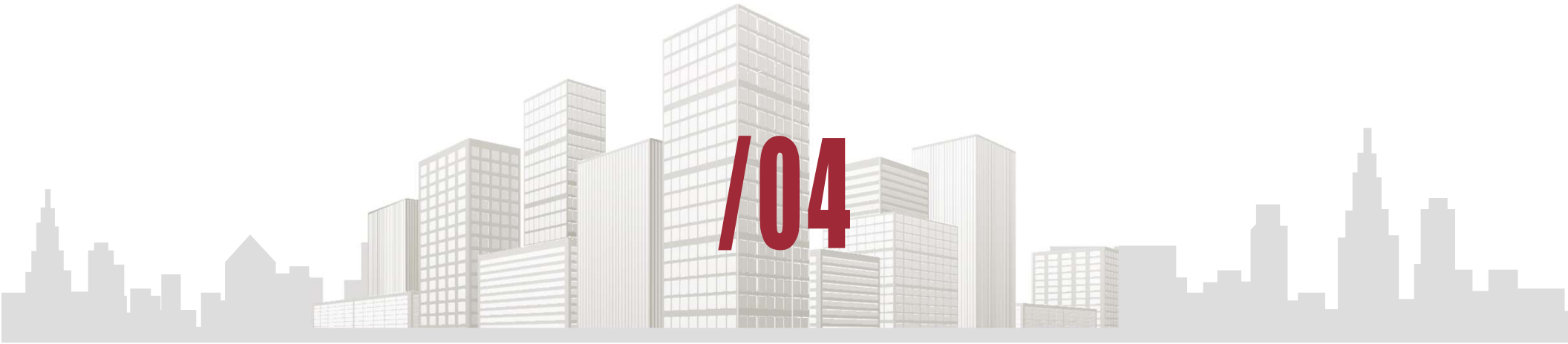
3. 什么是非参数分析（Nonparametric Analysis）

两大类非参数分析方法

- 全局平滑（Global Smoothing）和局部平滑（Local Smoothing）



全局平滑法
(Global Smoothing)



/04

4. 全局平滑法（Global Smoothing）

- 什么是全局平滑法?

数学基础 任何平方可积函数（$\int_{-\infty}^{\infty} f^2(x)dx < \infty$）可表示为傅里叶级数的加权线性组合

$$\boldsymbol{f(x) = \sum_{j=0}^{\infty} \beta_j \varphi_j(x)},$$

where $\{\varphi_j(x)\}$ is a sequence of complete basis functions.

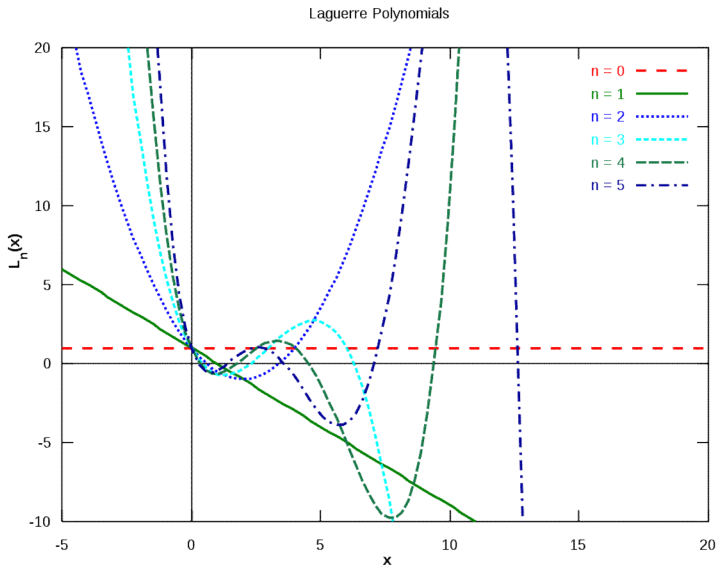
- 很多情况下，基函数 $\{\varphi_j(x)\}$ 是正交归化（Orthonormal）函数:

$$\int_{-\infty}^{\infty} \varphi_j(x)\varphi_l(x)dx = \delta_{jl} = \begin{cases} 1 & \text{if } j = l \\ 0 & \text{if } j \neq l \end{cases},$$

where δ_{jl} is called the Kronecker delta.

4. 全局平滑法（Global Smoothing）

- 基函数重要例子：
  - **Fourier Basis**
  - **Polynomials**
  - **Laguerre Polynomials**
  - **Legendre Polynomials**
  - **Hermite Polynomials**



- For series density estimation, see Cenzov (1962), Wahba (1975), Walter (1977)
- For series regression estimation, see Gallant, Newey

4. 全局平滑法（Global Smoothing）

- **例1**：　　$f(x) = x^2, \quad -\pi \le x \le \pi$

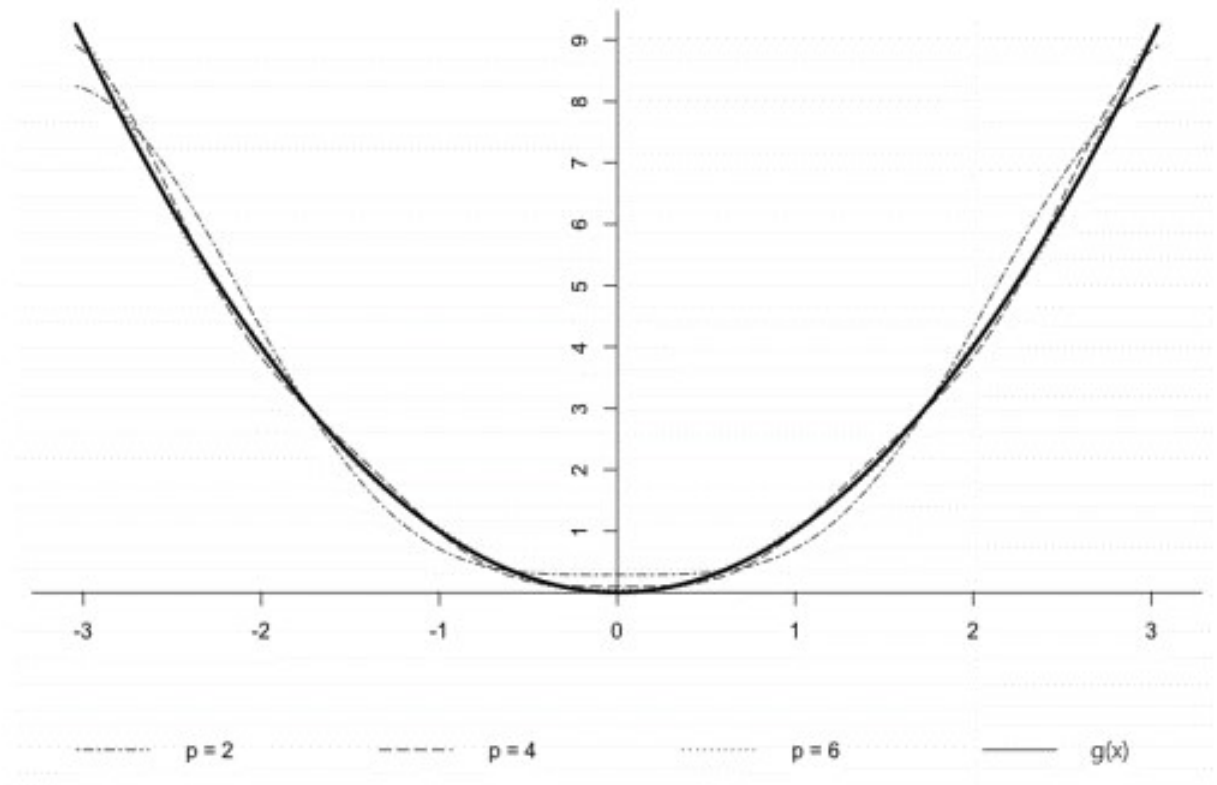
$$f(x) = \frac{\pi^2}{3} - 4\left[\cos(x) - \frac{\cos(2x)}{2^2} + \frac{\cos(3x)}{3^2} - \cdots\right]$$

$$= \frac{\pi^2}{3} - 4\sum_{j=1}^{\infty}(-1)^{j-1}\frac{\cos(jx)}{j^2}$$

$$= \sum_{j=0}^{\infty}\beta_j\varphi_j(x)$$

其中 $\beta_j = -\frac{(-1)^{j-1}}{j^2} \to 0$ as $j \to \infty$，即高阶系数趋于零

4. 全局平滑法（Global Smoothing）



Fourier Series Approximation to the Quadratic Function

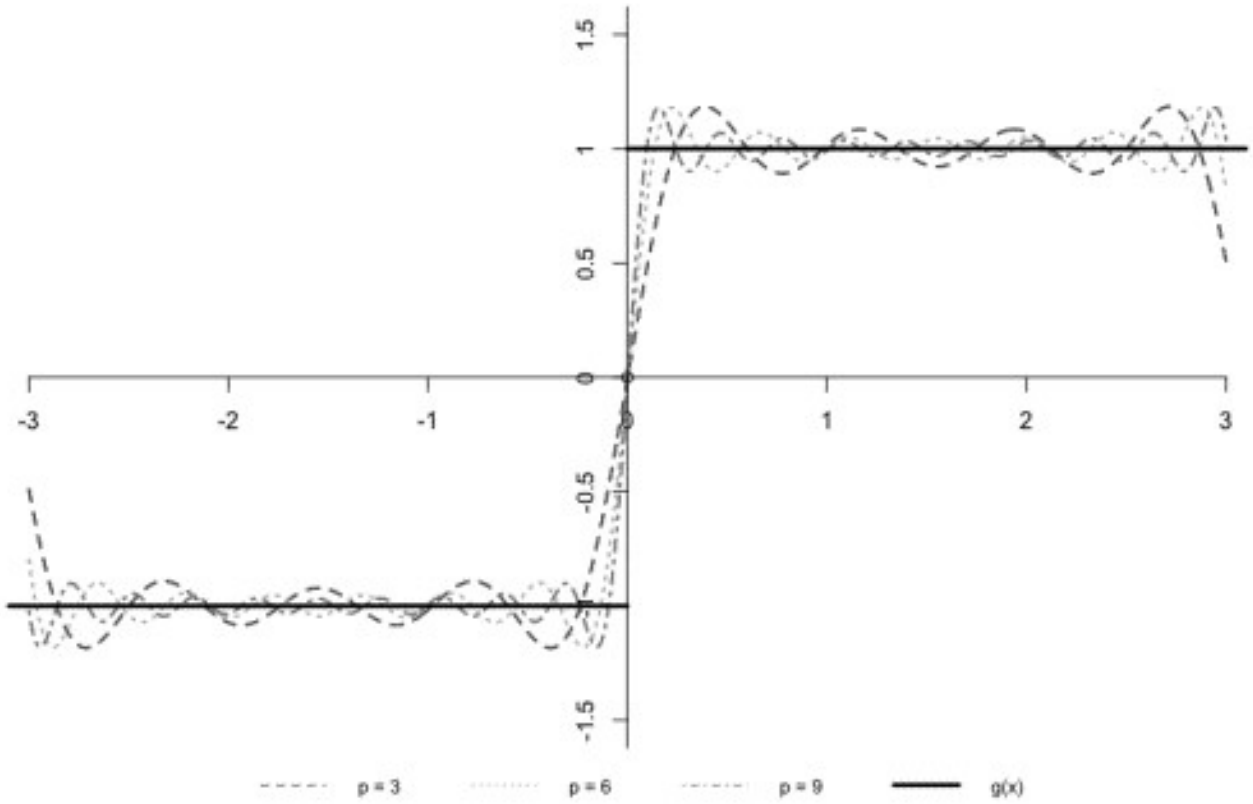
4. 全局平滑法（Global Smoothing）

- **例2**：$f(x)=\begin{cases}1 & if\ 0<x\leq 3\\ 0 & if\ x=0\\ -1 & if\ -3\leq x<0\end{cases}$

$$f(x)=\frac{4}{\pi}\left[\sin(x)+\frac{\sin(3x)}{3}+\frac{\sin(5x)}{5}+\cdots\right]=\frac{4}{\pi}\sum_{j=0}^{\infty}\frac{\sin[(2j+1)x]}{(2j+1)}$$

其中 $\beta_j=\frac{4}{\pi}\times\frac{1}{2j+1}\to 0$ as $j\to\infty$，即高阶系数趋于零

4. 全局平滑法（Global Smoothing）



Fourier Series Approximation to the Step Function

4. 全局平滑法（Global Smoothing）

- **级数回归（Series/Sieve Regression）**

$$Y_i = \sum_{j=0}^{p} \beta_j \varphi_j(X_i) + \varepsilon_{pi}, \qquad i = 1, 2, \ldots, n$$

其中 $p = p(n) \to \infty$, $\frac{p}{n} \to 0$ as $n \to \infty$。相应估计量

$$\hat{\gamma}(x) = \sum_{j=0}^{p} \hat{\beta}_j \varphi_j(x),$$

其中 $\hat{\beta} = (\varphi'\varphi)^{-1}\varphi' Y$, $\hat{\beta}$ 是 $(p+1) \times 1$ 向量, $\varphi'\varphi$ 是 $(p+1) \times (p+1)$ 矩阵

4. 全局平滑法（Global Smoothing）

- **均方误差（Mean Square Error）**

$$\begin{aligned}\mathrm{IMSE}\,[\hat{\gamma}(x)] &= \int E\,[\hat{\gamma}(x)-\gamma(x)]^2\,dx \\ &= \mathrm{var}[\hat{\gamma}(x)] + \mathrm{Bias}^2[\hat{\gamma}(x)] \\ &= \mathrm{O}\left(\frac{p}{n}+p^{-s}\right), \quad \mathrm{s}>0\end{aligned}$$

4. 全局平滑法（Global Smoothing）

- $p = p(n)$ is the maximum truncation order of series. It is called a **smoothing parameter** because it affects the degree of smoothness of the series estimator $\hat{r}(x)$. The estimator $\hat{r}(x)$ is called a smoother.
  - $p \to \infty$, $\text{Bias} = \sum_{j=p+1}^{p} \beta_j \varphi_j(x) \to 0$
  - $\frac{p}{n} \to 0$, $\text{Variance} \to 0$

? 问题：使用什么基函数（basis functions）$\{\varphi_j(x)\}$？

4. 全局平滑法（Global Smoothing）

- 为什么金融计量经济学常用 Hermite Polynomials？
  - Ait-Sahalia (2002), Gallant & Tauchen (1996)

? 问题：什么是 Hermite Polynomials？

- Hermite polynomials are a sequence of orthogonal polynomials of $\{H_j(x)\}$ such that

$$\int_{-\infty}^{\infty} H_j(x)H_l(x)e^{-\frac{x^2}{2}}dx = \delta_{jl}$$

where $H_j(x) = (-1)^j e^{\frac{x^2}{2}} \frac{d^j}{dx^j}\left(e^{-\frac{x^2}{2}}\right), -\infty < x < \infty.$

4. 全局平滑法（Global Smoothing）

- The first few Hermite polynomials are:

$$H_0(x) = 1,$$

$$H_1(x) = x,$$

$$H_2(x) = x^2 - 1,$$

$$H_3(x) = x^3 - 3x,$$

$$H_4(x) = x^4 - 6x^2 + 3$$

4. 全局平滑法（Global Smoothing）

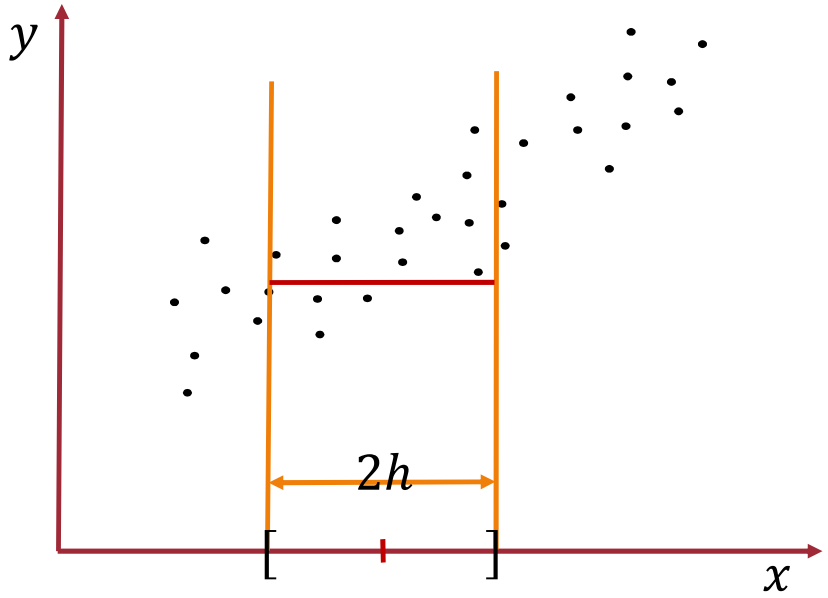
- 金融市场的经验典型特征事实：厚尾分布
  - Benchmark: Normal Distribution
- 非参数模型的参数 $\{\beta_j\}$ 是未知函数 $\mathrm{r}(x)$ 在基函数 $\varphi_j(x)$ 的投影系数，一般没有经济解释

局部平滑法
(Local Smoothing)

5. 局部平滑法（Local Smoothing）

基本思想（Basic Idea of Local Smoothing）

- **Fitting data in a small interval** which vanishes to a degenerate interval as the sample size $n \to \infty$.



5. 局部平滑法（Local Smoothing）

Kernel Probability Density Estimation (Rosenblatt, 1956; Parzen, 1960)

- Suppose $\{X_i\}_{i=1}^n$ is an IID random sample from an unknown **probability density function $f(x)$** with support $[a, b]$, where $a < b$.
- Then a kernel estimator for $f(x)$ at a given point $x \in [\mathrm{a}, \mathrm{b}]$ is

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x),$$

where

$$K_h(X_i - x) = \frac{1}{h}K\left(\frac{X_i - x}{h}\right),$$

$K: [-1, 1] \to \mathbb{R}^+$ is a kernel function, and $h = h(n) \to 0$ as $n \to \infty$ is a bandwidth.

5. 局部平滑法（Local Smoothing）

- **Second Order Kernel $K(\cdot)$**: Any prespecified symmetric probability density function. For example,
  - Gaussian kernel:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}, \qquad -\infty < u < \infty$$

  - Uniform kernel:

$$K(u) = \frac{1}{2}\mathbb{1}(|u| \leq 1)$$

where $\mathbb{1}(\cdot)$ is an indicator function, taking value $\mathbb{1}$ if $|u| \leq 1$, and zero otherwise.

5. 局部平滑法（Local Smoothing）

Special Case [直方图 (Histogram)]

- When the kernel $K(\cdot)$ is a **uniform kernel**, i.e.,

$$K(u) = \frac{1}{2}\mathbb{1}(|u| \leq 1),$$

the kernel density estimator $\hat{f}(x)$ becomes the well-known **histogram** with bin-size equal to $2h$:

$$\hat{f}(x) = \frac{1}{2nh}\sum_{i=1}^{n}\mathbb{1}\left(|X_i - x| \leq h\right).$$

5. 局部平滑法（Local Smoothing）

- Intuitively, the histogram is a standardized relative frequency over each small interval that is centered at point x and has a window size $2h$. As $n \to \infty$, it converges in probability to the unknown probability density $f(x)$ at point at x.

5. 局部平滑法（Local Smoothing）

局部平滑法的数学基础：泰勒级数展开（Taylor Series Expansion）

$$\text{Bias} = E\hat{f}(x) - f(x)$$

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

$$= \frac{1}{n}E\left[K\left(\frac{X_i - x}{h}\right)\right] - f(x) \quad \text{（根据同分布性质）}$$

$$= \frac{1}{n}\int_a^b K\left(\frac{y-x}{h}\right)f(y)dy - f(x) \quad \text{（对 } X_i \text{ 的分布积分）}$$

$$= \int_{\frac{a-x}{h}}^{\frac{b-x}{h}} K(u)f(x+hu)du - f(x) \quad \text{（变量变换 u = } (y-x)/h\text{）}$$

$$\cong \int_{-1}^{1} K(u)\left[f(x) + huf'(x) + \frac{(hu)^2}{2}f''(x)\right]du - f(x) \quad \text{（二阶泰勒展开）}$$

$$= f(x)\int_{-1}^{1}\boldsymbol{K(u)du} + hf'(x)\int_{-1}^{1}\boldsymbol{uK(u)du} + \frac{h^2}{2}f''(x)\int_{-1}^{1}\boldsymbol{u^2K(u)du} - f(x)$$

$$= \frac{h^2}{2}f''(x)\int_{-1}^{1}u^2K(u)du$$

问题：What is the bias of $\hat{f}(x)$?

5. 局部平滑法（Local Smoothing）

Kernel Regression Estimation (Nadaraya, 1964; Watson, 1964)

- The Nadaraya-Watson estimator is a local weighted sample mean

$$\hat{r}(x) = \arg\min_{r} \sum_{i=1}^{n} (Y_i - r)^2 K_h(x - X_i)$$

- FOC

$$\hat{r}(x) = \frac{\hat{m}(x)}{\hat{f}(x)} = \sum_{i=1}^{n} \hat{w}_i(x) Y_i$$

其中权重

$$\hat{w}_i(x) = \frac{K_h(x - X_i)}{\sum_{i=1}^{n} K_h(x - X_i)},$$

$K_h(x - X_i) = \frac{1}{h} K(\frac{x - X_i}{h})$, $K(\cdot)$ 是一个核函数（Kernel Function）

5. 局部平滑法（Local Smoothing）

- **The basic idea of local smoothing** is equivalent to the procedure of finding a **local weighted least squares estimate** (e.g., Härdle, 1990, pp.20).

5. 局部平滑法（Local Smoothing）

Special Case [回归图 (Regressogram)]

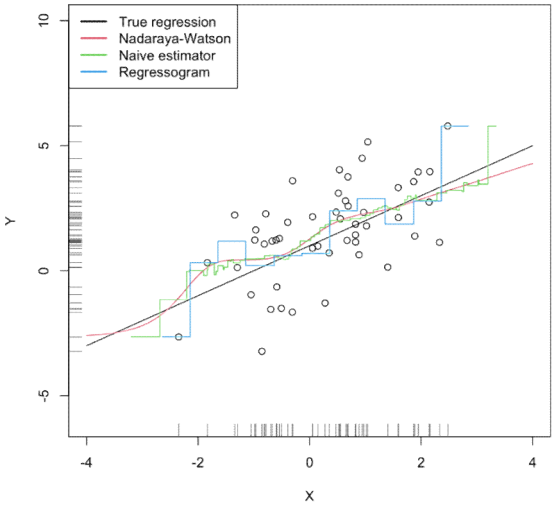
- When the kernel is a **uniform** function, i.e.,

$$K(u) = \frac{1}{2}\mathbb{1}(|u| \leq 1),$$

the Nadaraya-Watson estimator is simply a **local sample mean**, that is,

$$\hat{r}(x) = \frac{\sum_{i=1}^{n} Y_i \mathbb{1}(|X_i - x| \leq h)}{\sum_{i=1}^{n} \mathbb{1}(|X_i - x| \leq h)},$$

where the denominator is the number of observations that fall into the small interval $[x - h, x + h]$.



5. 局部平滑法（Local Smoothing）

- The local sample mean estimator is called the **regressogram** by Tukey (1961) to accentuate its relationship to the histogram.
- The regressogram is the local sample mean of $\{Y_i\}$ for which the values of the corresponding explanatory variable $\{X_i\}$ fall into a disjoint bin that is centered at point x and has a window size $2h$ (Tukey, 1947).
- Convergence in MSE of the regressogram to the true regression function has been shown in Collomb (1977) and Lecoutre (1983, 1984).

5. 局部平滑法（Local Smoothing）

- **Tradeoff between variance and squared bias**:

$$\text{MSE}\,\hat{r}(x) = O\left(\frac{1}{nh} + h^2\right)$$

  - $h \to 0$ as $n \to \infty$, Bias of $\hat{r}(x) \to 0$
  - $nh \to \infty$ as $n \to \infty$, Variance of $\hat{r}(x) \to 0$

- **Boundary Bias Problem**
  - The bias in the boundary region is “larger” in order of magnitude than the bias in the interior region, due to the asymmetric coverage of data in the boundary regions.

5. 局部平滑法（Local Smoothing）

Local Polynomial Smoothing

- The Nadaraya-Watson estimator is a **local constant** estimator, i.e., using a constant to predict observations in a small interval.

? 问题：Why not use a polynomial to fit the data in a small interval?

5. 局部平滑法（Local Smoothing）

- The idea of using a **local polynomial**: Katkovnik (1979, 1983, 1985) and Lejeune (1985):

$$\min_{\alpha} \sum_{i=1}^{n} (Y_i - Z_i'\alpha)^2 K_h(x - X_i)$$

where $Z_i = (1, X_i - x, \ldots, (X_i - x)^p)' = (p+1) \times 1$

$$\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_p)' = (p+1) \times 1$$

- when $p = 0$, it becomes a local constant smoothing.
- when $p = 1$, it becomes a local linear smoothing.

5. 局部平滑法（Local Smoothing）

问题：什么叫 Local Polynomial Smoothing？

- A polynomial is fitted to the data in a small neighborhood which is centered at point x and has a window size $2h$, where $h \to 0$ as $n \to \infty$.

5. 局部平滑法（Local Smoothing）

- Locally WLS: $\boldsymbol{\hat{\alpha} = (Z'WZ)^{-1}Z'WY}$
  - $p \geq 1,\ Bias = O(h^2)\ \forall\ x \in [a,b]$
  - $\hat{\alpha}$ has an equivalent kernel representation:

$$\boldsymbol{\hat{\alpha}_v = \sum_{i=1}^{n} \hat{w}_v\left(\frac{X_i - x}{n}\right) Y_i,\ \hat{w}_v(u) = \frac{\tilde{K}_v(u)}{Thf(x)}[1 + o_p(1)]}$$

其中 $\tilde{K}_v(u)$ 是一个核函数，而 $f(x)$ 是 X_i 的概率密度函数。

  - Locally WLS:

$$\hat{\alpha} = (\hat{\alpha}_0, \hat{\alpha}_1, \ldots, \hat{\alpha}_p)' \xrightarrow{p} \left(\gamma(x), \gamma'(x), \ldots, \frac{\gamma^{(p)}(x)}{p!}\right)' \text{ as } n \to \infty$$

5. 局部平滑法（Local Smoothing）

- The Nadaraya-Watson estimator is a special case of local polynomial smoothing, with $p = 0$
  - $Bias \neq O(h^2)$ in the boundary regions
  - $\int_{-\tau}^{1} uK(u)du \neq 0$, if $0 \leq \tau < 1$, so the term of $O(h)$ in the bias does not vanish to zero.

5. 局部平滑法（Local Smoothing）

- For Local Polynomial Smoothing with $p \geq 1$
  - $\int_{-\tau}^{1} \widetilde{K}_v(u)du = 1$ and $\int_{-\tau}^{1} u^j \widetilde{K}_v(u)du = 0$ for $1 \leq j \leq p, \forall\ 0 \leq \tau \leq 1$
  - The equivalent kernel $\widetilde{K}_v(\cdot)$ is automatically adapted to the boundary region, due to the nature of least squares estimation!
  - $Bias = O(h^2)$ in the boundary regions, the same order of magnitude as in the interior region. This is in sharp contrast with the Nadaraya-Watson estimator (or local constant estimator).

5. 局部平滑法（Local Smoothing）

比较：Global Polynomial Approximation

- White (1980, *International Economic Review*)

$$\min_{\alpha} \sum_{i=1}^{T} (Y_i - \alpha' Z_i)^2$$

where $Z_i = (1, X_i - x, \ldots, (X_i - x)^p)'$

- White (1980) 证明：当 $n \to \infty$ 时，

$$\text{OLS } \hat{\alpha} \nrightarrow \left(\gamma(x), \gamma'(x), \ldots, \frac{\gamma^{(p)}(x)}{p!} \right)$$

逆向思维非常重要！

5. 局部平滑法（Local Smoothing）

K-最近邻法（K-Nearest Neighbor Smoothing）

问题：What is **K-Nearest Neighbor** (KNN) Smoothing?

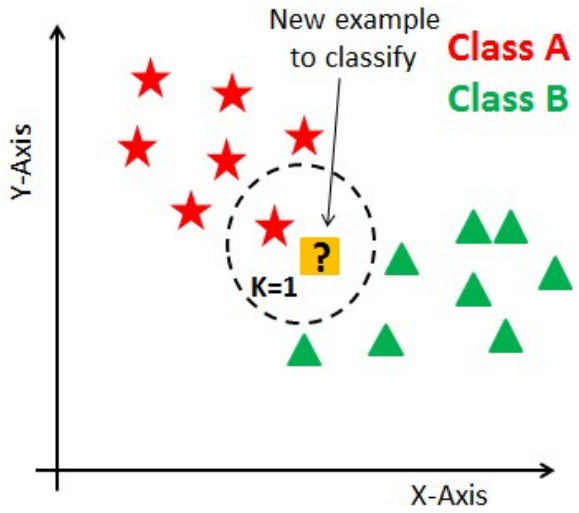
- The KNN estimator is an average of the observations of $\{Y_i\}$ in a neighbor for which the values of $\{X_i\}$ which are among the K-nearest neighbors of the point x in the Euclidean distance.
- The original KNN estimator is an equally weighted average of K observations in a varying neighborhood, so it can be viewed as a regressogram in the neighborhood.

5. 局部平滑法（Local Smoothing）

- The key difference between kernel smoothing and KNN smoothing is that kernel smoothing focuses on a fixed interval that is centered at x and has window size $2h$, while KNN smoothing considers a varying neighborhood:
  - For the variance of the KNN estimator to vanish to zero as $n \to \infty$, one must let K (the sample size in the K-nearest neighbor) increase.
  - For the bias of the KNN estimator to vanish to zero as $n \to \infty$, K must increase at a slower rate than the sample size n, so that the interval containing the point x shrinks as $n \to \infty$.

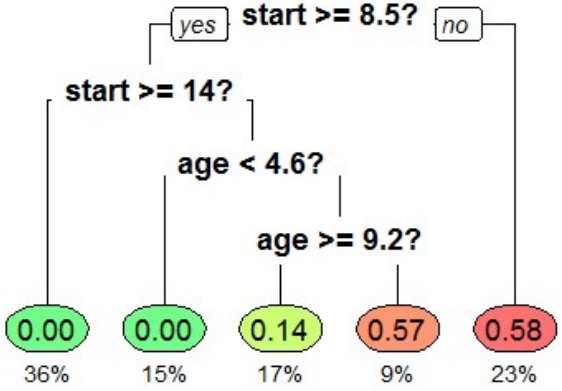
5. 局部平滑法（Local Smoothing）

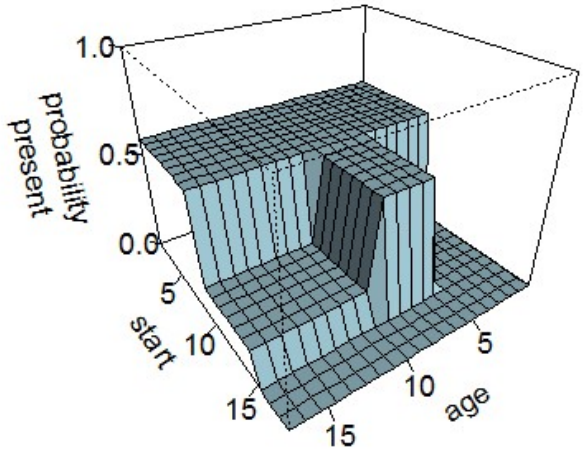
- The KNN estimator can be generalized to be a weighted average in a varying neighborhood. See Loftsgaarden & Quesenberry (1965), Hart (1967), Stone (1977), Mack (1981), Devroye (1978), and Györfi (1981).

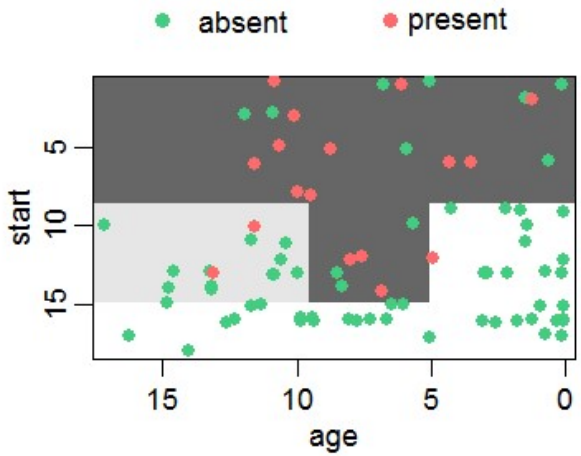


5. 局部平滑法（Local Smoothing）

回归树（Regression Tree）







5. 局部平滑法（Local Smoothing）

- A regression tree takes the form (Gordon & Olshen, 1980) of

$$\hat{r}(x) = \sum_{j=1}^{p} \widehat{C_j} \mathbb{1}(x \in N_j)$$

where $\{\widehat{C_j}\}$ are constants and $\{N_j\}$ are disjoint hyper-rectangles with sides parallel to the coordinate axes such that $U_{j=1}^{p} N_j = \mathbb{R}^d$, where d is the dimension of X_i.

5. 局部平滑法（Local Smoothing）

- The best least squares estimator of a tree is

$$\widehat{C}_j = \frac{1}{n_j}\sum_{\{i:\, X_i \in N_j\}} Y_i,$$

where n_j is the number of observations whose values of $\{X_i\}$ fall into the region N_j.

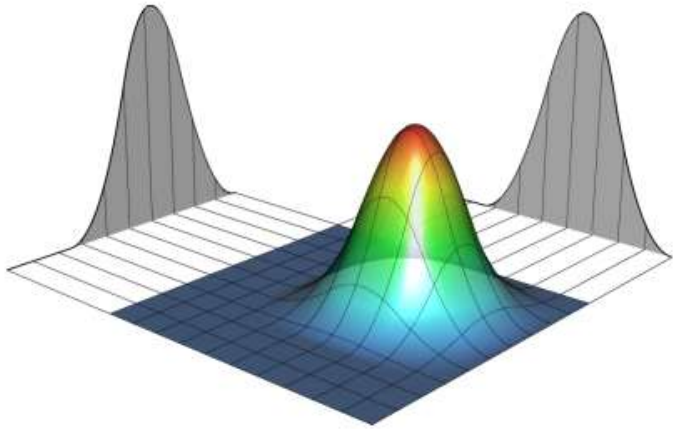
? 问题：How to construct the disjoint hyper-rectangles $\{N_j\}_{j=1}^{p}$?

- **A regression tree estimator is the average of $\{Y_i\}$ whose predictor X_i's values fall into $\{N_j\}_{j=1}^{p}$**. Therefore, it is a regressogram in each region N_j. See Breiman *et al.* (1984, 2004).

5. 局部平滑法（Local Smoothing）

总结：几乎所有非参数平滑估计量均可以表示为一个 Local Weighted Sample Mean of $\{Y_i\}_{i=1}^n$

- 不同的非参数方法体现在确定不同邻域（neighbors）和不同权重（weights）



5. 局部平滑法（Local Smoothing）

全局平滑法和局部平滑法的共同点

- An **orthogonal series estimator** (global smoothing) can be represented as a locally weighted average of $\{Y_i\}$ with the weighting function $W_{pi}(x)$ depending on the basis functions and the smoothing parameter p which is the maximum truncation order of the basis functions (e.g., Härdle, 1990).
- As Härdle (1990, p.59) points out, a **spline smoothing estimator** can be written as

$$\hat{r}(x) = n^{-1} \sum_{i=1}^{n} W_{\lambda i}(x) Y_i,$$

for some weight function $W_{\lambda i}(x)$.

5. 局部平滑法（Local Smoothing）

全局平滑法和局部平滑法的区别在哪里?

- 全局平滑法只需要估计一次，便可获得未知函数在整个支撑上的所有估计值
- 局部平滑法需要重新估计，才能获得函数在不同点 x 的估计值

? 问题：什么时候用 Global Smoothing？什么时候用 Local Smoothing？

5. 局部平滑法（Local Smoothing）

- Global Smoothing 应用例子：
  - Ait-Sahalia (2002, *Econometrica*)
  - Gallant and Tauchen (1996, *Econometric Theory*)
  - Hong & White (1995, *Econometrica*): Series Estimator
  - Cui, Hong & Li (2020, *Journal of Econometrics*): Spline Estimator

5. 局部平滑法（Local Smoothing）

总结：非参数分析的基本假设、思想与方法

（1）Key Assumption: Data generating process (DGP) is an **unknown stochastic process**.

（2）**Model-free**：用于未知函数形式、非线性情景

（3）Criterion: Mean Squared Error (MSE)

（4）关键是如何控制平滑程度（即如何选取平滑参数值）？

Trade-off between variance and squared bias

5. 局部平滑法（Local Smoothing）

（5）收敛速度（Convergence Rate）比较慢，需要比较大的样本（n）

- ✓ For kernel multivariate density estimation, the optimal convergence rate

$$MSE\left[\hat{f}(x), f(x)\right] \propto n^{-\frac{4}{4d+1}},$$

where d is the dimension of X_i.

（6）解释变量 X_i 事先给定；存在**维度灾难**（Curse of Dimensionality）

（7）非参数模型的可解释性：系数一般没有经济意义的解释

5. 局部平滑法（Local Smoothing）

? 问题：什么时候用参数方法？什么时候用非参数方法？

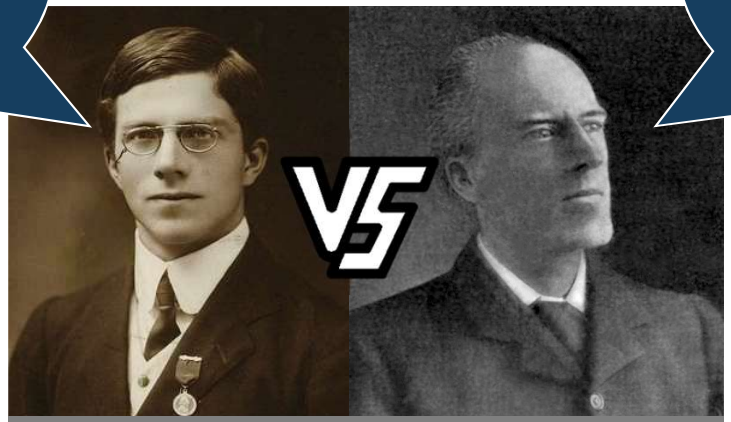
- 从 Bias-Variance Tradeoff 的视角看
  - 参数估计量的方差小（估计精确），但是偏差可能很大（当模型误设时）
  - 另一方面，非参数方法由于其灵活性，偏差较小，但估计量的方差较大，虽然其 MSE 总的来说较小

5. 局部平滑法（Local Smoothing）

20世纪 Pearson VS Fisher 的争论

Nonparametric approach gives **generally poor efficiency**

More concerned about **specification bias**

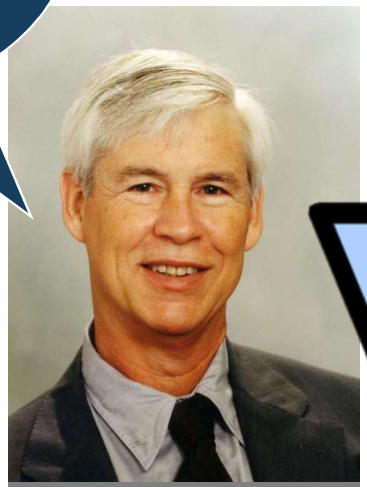


R. A. Fisher　　Karl Pearson

5. 局部平滑法（Local Smoothing）

计量经济学家的争论

From **specific** to **general** approach



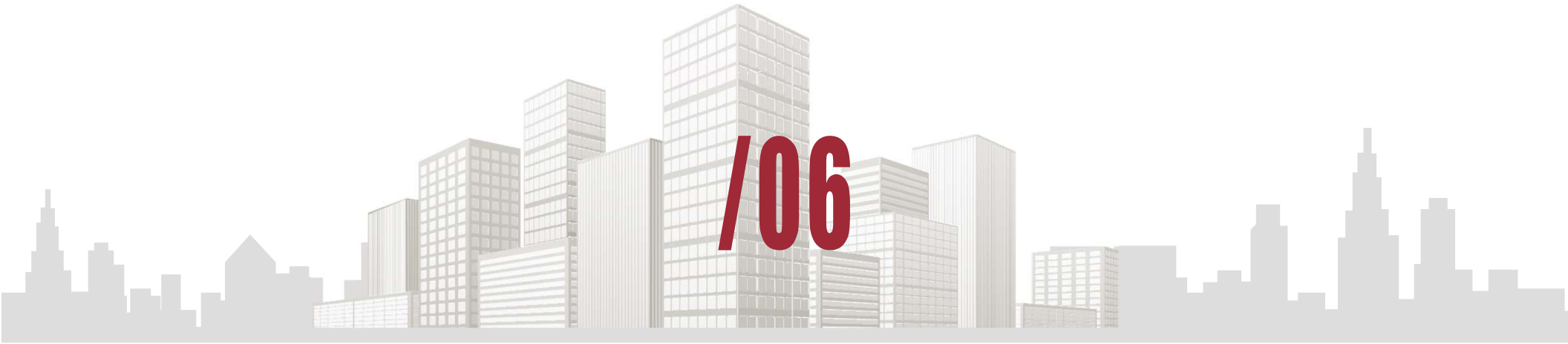
Robert Engle

VS

David Hendry

From **general** to **specific** approach

非参数方法与机器学习



/06

6. 非参数方法与机器学习

什么是机器学习（Machine Learning）？

- Arthur Samuel 在1959年提出，但大规模应用是大数据出现之后
- 机器学习是**基于大数据的自动学习计算机算法**
- 机器学习也是 **Model-free**
- 机器学习可分两大类：监督学习（Supervised Learning）和无监督学习（Unsupervised Learning）
- 机器学习的主要用途是基于大数据的**样本外预测（包括分类）**

6. 非参数方法与机器学习

“横看成岭侧成峰，远近高低各不同”

6. 非参数方法与机器学习

- **例 [回归预测与正则化 (Regularization)]**
  - 给定数据 $\{Y_i, X_i'\}_{i=1}^n$，考虑使用一个线性回归模型进行预测：

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i'\beta)^2$$

其中 $X_i'\beta = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}$

  - **OLS估计方法**：

$$\hat{\beta} = (X'X)^{-1}X'Y$$

  - 如果线性回归模型是正确设定，则 OLS $\hat{\beta}$ 是无偏估计量
  - 当解释变量 X_i 的个数很大时，解释变量之间可能存在共线性（multicollinearity）或近似共线性（near multicollinearity），此时 $X'X$ 的逆不存在或几乎不存在，导致OLS估计量不存在或不稳定

6. 非参数方法与机器学习

- **例 [回归预测与正则化 (Regularization)]**
  - **岭回归（Ridge Regression）**方法：在这种情形下，可以考虑

$$\min_{\{\beta_j\}} \sum_{i=1}^{n} (Y_i - X_i'\beta)^2 + \lambda \sum_{j=0}^{p} \beta_j^2$$

其估计量为

$$\hat{\beta} = (X'X + \lambda I)^{-1} X'Y$$

这个估计量 $\hat{\beta}$ 不再是无偏估计量，但其解存在且比较稳定。从本质上说，岭回归通过约束未知参数值的大小，以牺牲无偏性换取方差的显著减少，从而改进预测效果

6. 非参数方法与机器学习

- **例 [回归预测与正则化 (Regularization)]**
  - **LASSO 方法**：但是，岭回归并没有降低维数。假设高维解释变量存在稀疏性（sparsity），即只有少数解释变量的系数不为零。在这种情形下，可考虑

$$\min_{\{\beta_j\}} \sum_{i=1}^{n} (Y_i - X_i'\beta)^2 + \lambda \sum_{j=0}^{p} |\beta_j|$$

这个方法称为 LASSO，可识别重要解释变量，并将众多不重要的解释变量的系数直接设为零，从而达到约简模型的目的

6. 非参数方法与机器学习

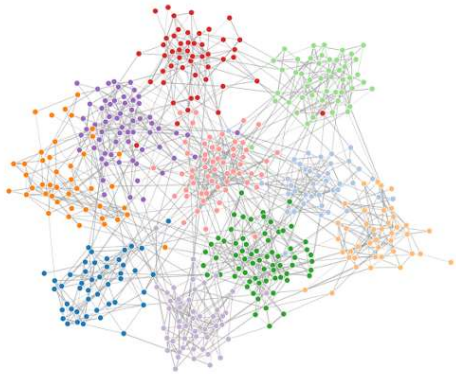
- 机器学习的基本思想类似于 LASSO 方法。但是有两个不同：
  - 机器学习一般情况下不用线性回归模型： Model-free
  - Risk of overfitting: 为了改进样本外预测精确度，机器学习将数据分为训练数据（training data）和测试数据（test data），其中训练数据用于确定算法结构，而预测数据用于检验样本外预测效果

6. 非参数方法与机器学习

- 什么是机器学习（Machine Learning）？
  - ➢ 机器学习是通过挖掘大数据（训练数据）中的系统性特征与变量之间的统计关系（如相关性），然后进行样本外预测。
  - ➢ 机器学习的本质：包括正则化的**数学优化 + 计算机算法**

✓ $\text{Data}\begin{cases}\text{Trainning Data} + \text{Penalty} \rightarrow \text{Algorithm} \\ \text{Test Data}\end{cases}$

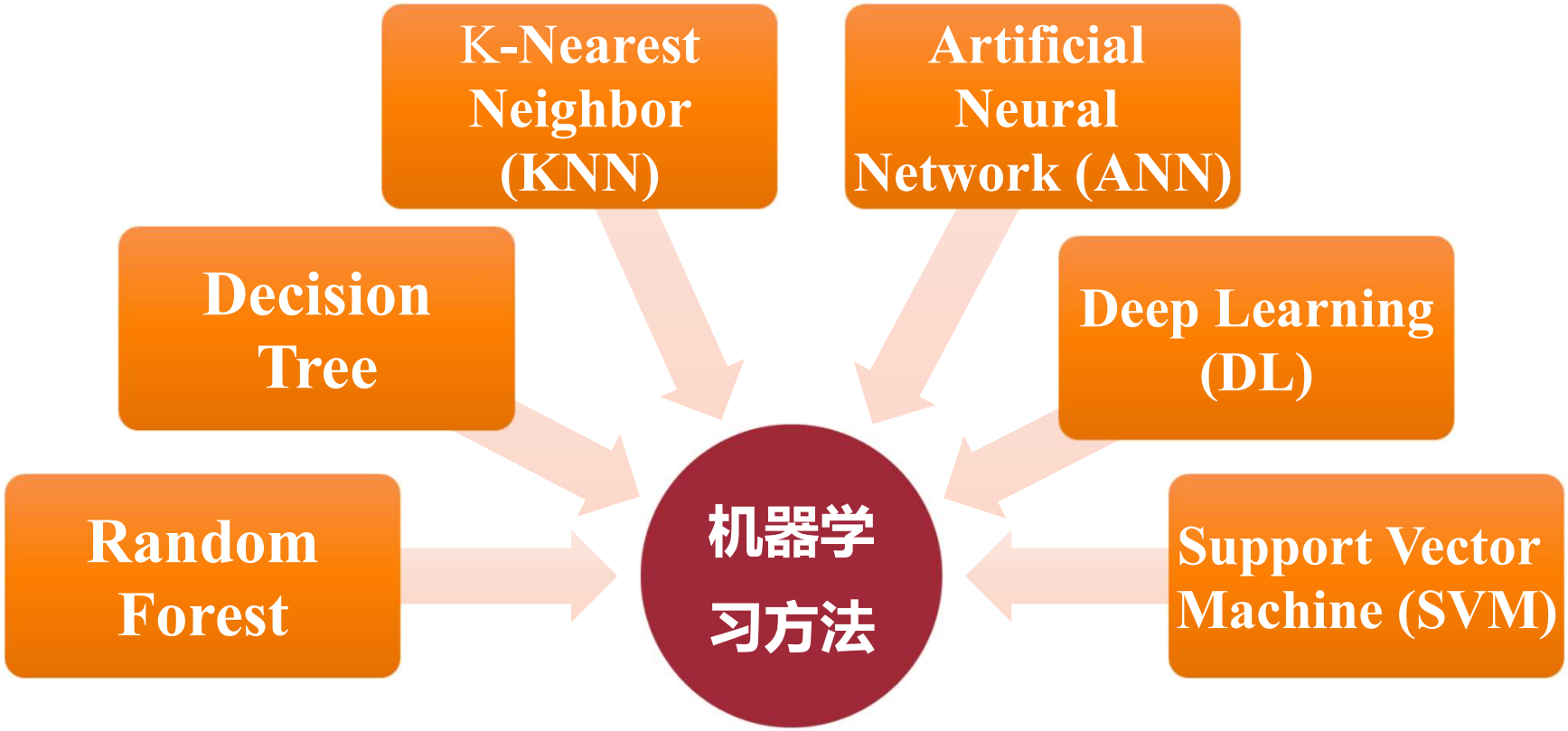
✓ $\min \text{SSR} + \text{Penalty}$



6. 非参数方法与机器学习

6. 非参数方法与机器学习

- 几种重要的机器学习方法



6. 非参数方法与机器学习

人工神经网络（ANN）

- ANN is motivated from cognitive science, particularly from the physiology of nerve cells.
- Suppose $\boldsymbol{X_i}$ is a ***d*-dimensional** predicting vector. The d components of X_i are called **inputs**, which are connected to multiple hidden units. At each hidden unit, the inputs are weighted to form a **linear combination** $(\beta_j' X_i)$, where the weight for each input is called a connecting strength.

  Different hidden units have different linear combinations $\left\{\beta_j' X_i\right\}_{j=1}^{p}$.

6. 非参数方法与机器学习

- These different linear combinations $\left\{\beta_j' X_i\right\}_{j=1}^{p}$ are transformed by the same **activation function** $g(\cdot)$ which is **a nonlinear mapping**.
- A popular example of the activation function is the logistic function

$$g(u) = \frac{1}{1+e^{-u}}.$$

- 问题：什么是 Activation Function $g(\cdot)$？其作用是什么？ See **Stinchcombe & White (1989)**

6. 非参数方法与机器学习

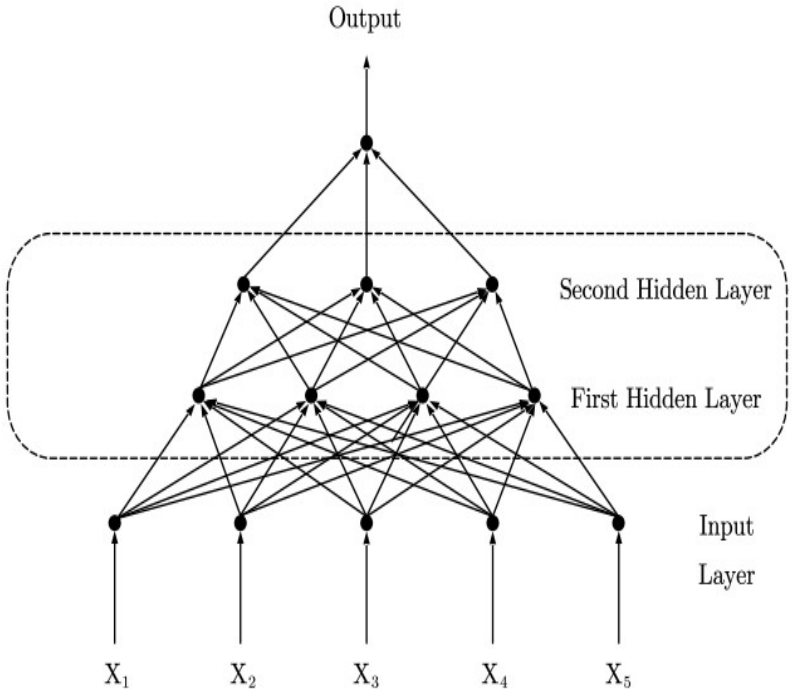
- The nonlinear output $g(\beta_j' X_i)$ of each hidden unit is called a multilayer perceptron. They are weighted again to form a linear combination which is then transformed by a second activation function $h(\cdot)$:

$$Y_i = h\left[\sum_{j=1}^{p} \alpha_j g(\beta_j' X_i)\right] + \varepsilon_i.$$

This is called a **single hidden layer** neural network. The second activation function $h(\cdot)$ can be a linear mapping.

6. 非参数方法与机器学习

- If there are more than one hidden unit at the second hidden layer, and different hidden units at the second hidden layer have different combinations, which are further transformed by the second activation function and are formed as a linear combination. This is called a multiple hidden layer neural network (**deep learning**).



6. 非参数方法与机器学习

诺贝尔经济学奖得主 Thomas Sargent (2018)：

“**Artificial intelligence** is, first of all, some **gorgeous rhetoric**. Artificial intelligence is actually **statistics**, but with a very gorgeous phrase, in fact, is statistics.”

2018世界科技创新论坛

- 更具体地说，机器学习是基于大数据的正则化非参数统计预测方法
- 计算机算法对实现数学优化问题非常重要

6. 非参数方法与机器学习

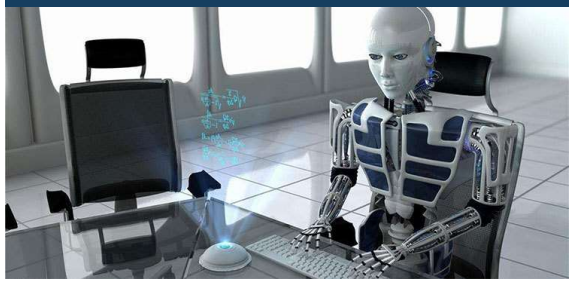
- 机器学习在经济中的应用实例

算法交易（85%，外汇市场）

商务智能（Business Intelligence）

自动驾驶（Autonomous Vehicle）

机器人代替人工



- 速记
- 文字翻译
- 同声传译
- 会计
- 新闻写作
- 论文写作

- 信用卡审批
- 小额贷款审批

机器学习改变商业模式

6. 非参数方法与机器学习

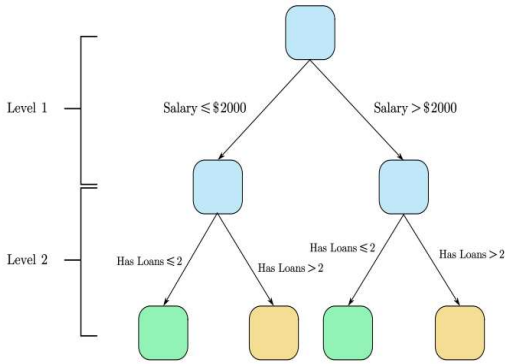
- 在很多实际应用中，机器学习的预测大都比较准确，但是机器学习本身像个黑箱（Black Box），其解释性（特别是因果解释）一直是一个难题
- 为什么需要**可解释性**?
  - **[案例1]** 如果机器学习拒绝了信用卡申请，需要向申请者说明拒绝原因

6. 非参数方法与机器学习

- **机器学习的基本假设**：Data Generating Process (DGP) 是一个未知的随机过程，且可能存在大量潜在的解释变量
- 非参数分析与机器学习在方法论上是一样的，都是 **Model-free**。同时，由于使用正则化，机器学习可以有效处理高维解释变量的问题
- **大多数机器学习方法是正则化的非参数统计方法的算法版本**
  - KNN, Regression Tree, Random Forest, ANN

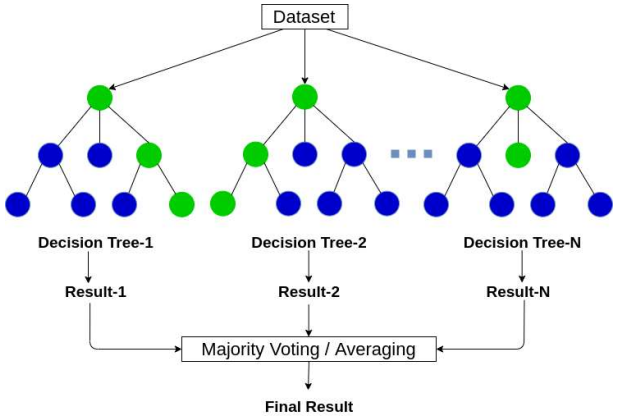
6. 非参数方法与机器学习

- 非参数统计学可以从理论上帮助理解机器学习的本质，奠定其数理统计学基础
- 特别是，非参数方法可以从统计学视角解释为什么机器学习方法预测精准
  - ➢ 例：KNN, Decision Tree, Random Forests, ANN
    - ✓ Lai (1977): Consistency of **KNN**
    - ✓ Breiman (2004): Consistency of **Decision Tree**



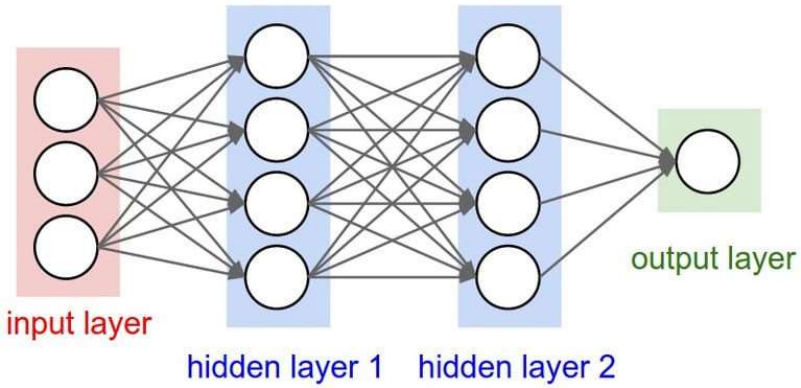
6. 非参数方法与机器学习

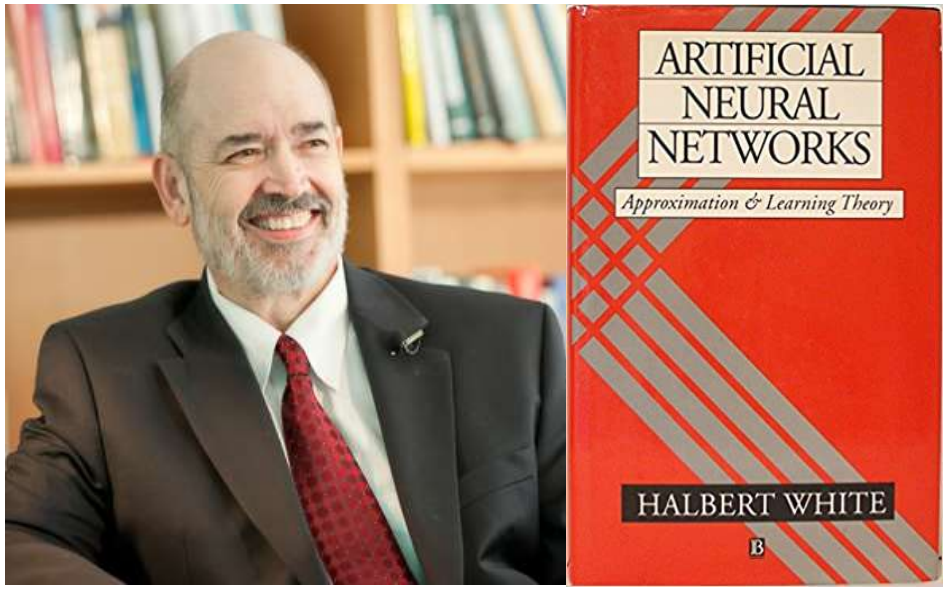
- 非参数方法可以解释为什么一些机器学习方法预测精准
  - ➢ 例：KNN, Decision Tree, Random Forests, ANN
    - ✓ Biau, Devroye & Lugosi (2008), Scornet, Biau & Vert (2015): Consistency of **Random Forests**
    - ✓ White (1989, *JASA*)： Consistency of **ANN**



6. 非参数方法与机器学习

- 计量经济学家 Halbert White 在 ANN 的基础理论方面，做出了重大原创性贡献





6. 非参数方法与机器学习

ANN Universal Approximation

- **White** (**1989**, 1992) show that even a single hidden layer neural network has the **universal approximation property**: It can approximate any nonlinear function to an arbitrary degree of accuracy with a suitable number of hidden units.
- The process of parameter estimation is called training the network, which can be obtained by minimizing the sum of squared residuals. The algorithms to implement this minimization task has been a challenge in the literature.

结论

/07

7. 结论

问题：什么是**非参数方法?**

- 非参数平滑法包括全局平滑法和局部平滑法
- 非参数分析的数学基础是傅里叶级数展开和泰勒级数展开
- 非参数分析的关键是如何选择平滑参数

7. 结论

问题：**机器学习等于非参数方法吗**?

- 非参数方法和机器学习在 DGP 假设基础和函数形式方面类似，它们均假设 DGP 是一个**未知的 Stochastic Process**，均是 **Model-free**
- 机器学习在**变量选择与降维**方面比非参数方法灵活
  - 非参数方法假定 X_i 的维数 d 是固定的且一般情形下是低维的，而由于使用了正则化，机器学习方法允许 X_i 的维度很大，甚至超过样本容量 n，这是机器学习和非参数方法最大的一个区别

7. 结论

- 由于大数据拥有高维潜在解释变量，这些解释变量存在不同程度的共线性（Multicollinearity），因此机器学习的最优预测模型可能具有**模型不确定性**，即一个数据“微扰”可能会导致最优预测模型的显著变化
- 从统计学视角看，机器学习是基于大数据的正则化非参数统计预测方法
- 机器学习是分析大数据的一种重要方法，特别是高维大数据

7. 结论

- 机器学习侧重于计算机算法的研究与应用
- 机器学习和统计学（不局限于非参数分析）的交叉可以产生统计学和计量经济学的新研究领域和研究方向，如**统计学习（Statistical Learning）**

THANKS