Modern Statistics **Lecture 13 - 04/03/2025**

Nonparametric Inference

Lecturer:Xiangyu Chang *Scribe: Yuxiao Liu, Shuyi Mai*

Edited by: Zhihong Liu

1 Recall

Parametric Inference

- Suppose we observe independent data $\{x_i\}_{i=1}^n$, the distribution of these data can be obtained through a large number of observations (prior information) (e.g. $N(\mu, \sigma^2)$, Uin(0,θ),Ber(p)),parameter information is inferred from the data(e.g. $\theta = (\mu, \sigma^2)$...).

$$MLE \Rightarrow \hat{\theta} \Rightarrow F_\theta.$$

- Suppose we observe pairs of data $\{(x_i, y_i)\}_{i=1}^n$.

$$r : x \in X \to y \in Y, \quad MSE : min_r \mathbb{E}[y - r(x)]^2.$$

  In order to find the regression function r(x)=$\mathbb{E}(Y|X = x)$, suppose:

$$r_\beta(x) = \beta_0 + \beta_1 x \quad \text{or} \quad r_\beta(x) = X^T \beta.$$

  then apply MLE,LS to infer parameters ($r_{\hat{\beta}(x)}$).

All of the above are **generative models**.

2 Nonparametric Inference

Definition 1 *Nonparametric inference refers to statistical techniques that use data to infer unknown quantities of interest while making as few assumptions as possible.*

$$\{X_i\}_{i=1}^n \Rightarrow F_X(x),$$

the specific distribution function is unknown.

Empirical Distribution Function (EDF)

Definition 2 *For i.i.d. samples $\{X_i\}_{i=1}^n$,*
the EDF is:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(X_i \le x),$$

where $\mathbb{I}(X_i \leq x) = \begin{cases} 1 & \textit{if } X_i \leq x, \\ 0 & \textit{otherwise.} \end{cases}$

Derivation 1

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(\mathbb{I}(X \leq x) = 1) = \mathbb{E}[\mathbb{I}(X \leq x)],$$

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq x).$$

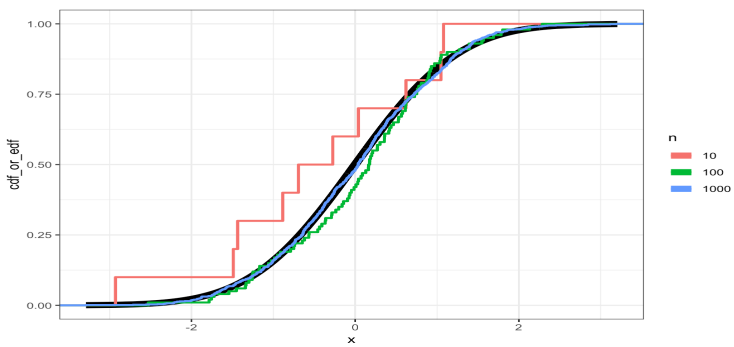

Figure 1. EDF

Theorem & Proof

1. **Unbiasedness**: $\mathbb{E}[F_n(x)] = F_X(x)$.
   proof:
   $$\mathbb{E}[F_n(x)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\mathbb{I}(X_i \leq x)] = \mathbb{P}(X \leq x) = F_X(x).$$
2. **Variance**: $\mathbb{V}(F_n(x)) = \frac{F_X(x)[1-F_X(x)]}{n}$.
   proof:
   $$\mathbb{V}(F_n(x)) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}(\mathbb{I}(X_i \leq x)) = \frac{1}{n}\mathbb{V}(\mathbb{I}(X \leq x)),$$
   $$\begin{aligned} \mathbb{V}(\mathbb{I}(X \leq x)) &= \mathbb{E}[\mathbb{I}^2(X \leq x)] - \mathbb{E}[\mathbb{I}(X \leq x)]^2 \\ &= \mathbb{P}(X \leq x) - (\mathbb{P}(X \leq x))^2 \\ &= F_X(x) - F_X^2(x) \\ &= F_X(x)[1 - F_X(x)]. \end{aligned}$$
3. **Consistency**: By Glivenko-Cantelli theorem, $F_n(x) \xrightarrow{P} F_X(x)$ as $n \to \infty$.
   proof: For any $\epsilon > 0$:
   $$\mathbb{P}\left(|F_n(x) - F_X(x)| \geq \epsilon\right) \leq \frac{\mathrm{Var}(F_n(x))}{\epsilon^2} = \frac{F_X(x)(1 - F_X(x))}{n\epsilon^2} \xrightarrow{n\to\infty} 0.$$

3 Density Estimation

Histogram Density Estimation

Steps For i.i.d. samples $\{X_i\}_{i=1}^n$,the PDF is f,domain is [0,1].

1. **Bin Construction**: Partition the domain into m bins of width $h = \frac{1}{m}$.
   $B_1 = [0, \frac{1}{m}), B_2 = [\frac{1}{m}, \frac{2}{m}), ...B_m = [\frac{m-1}{m}, 1]$.

2. **Count Observations**: Let n_j be the number of samples in the j-th bin.

3. **Probabilistic estimation** :
$$\hat{p_j} = \frac{n_j}{n}.$$

4. **Density Estimate**:
$$\hat{f}_n(x) = \frac{\hat{p_j}}{n} \quad \text{if } x \in B_j.$$
$$\hat{f}_n(x) = \frac{1}{h} \sum_{j=1}^{m} \hat{p_j} \mathbb{I}(x \in B_j).$$

Motivation

1. $p_j = \int_{B_j} f(x)dx = f(x^*)h, x^* \in B_j$(mean value theorem).

2. $\mathbb{E}[\hat{p}_j] = \frac{\mathbb{E}(n_j)}{n} = \frac{[\sum_{i=1}^n \mathbb{I}(x_i \in B_j)]}{n} = \frac{\sum_{i=1}^n \mathbb{E}[\mathbb{I}(x_i \in B_j)]}{n} = \frac{np_j}{n} = p_j$.

3. $\mathbb{E}[\hat{f}_n(x)] = \frac{1}{h}\mathbb{E}(\hat{p_j}) = \frac{p_j}{h} = f(x^*)$(as $m \to \infty, x \sim x^*$).

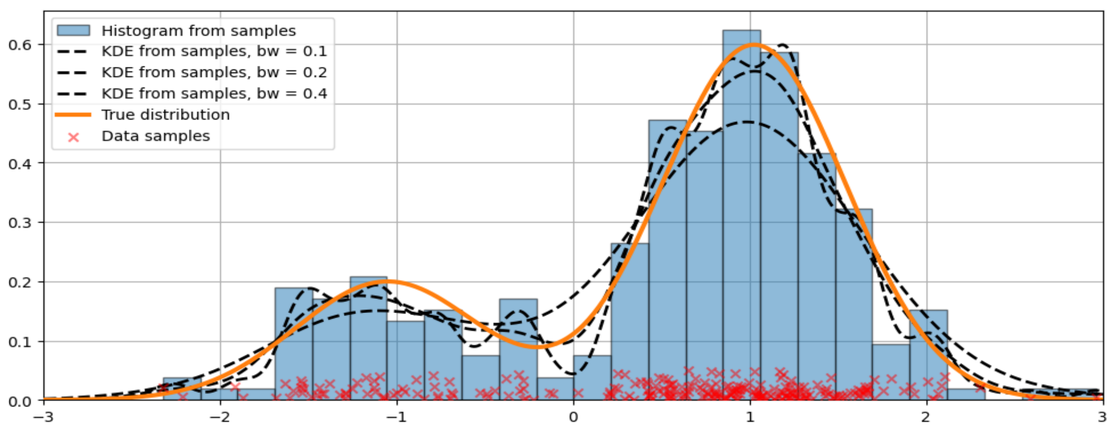

Figure 2. Density Estimation

Kernel Density Estimation (KDE)

Derivation 2 According to histogram:

$$\begin{aligned}
\hat{f}_n(x) &= \frac{1}{h}\sum_{j=1}^{m}\hat{p}_j\mathbb{I}(x \in B_j) \\
&= \frac{1}{h}\sum_{j=1}^{m}\frac{n_j}{n}\mathbb{I}(x \in B_j) \\
&= \frac{1}{nh}\sum_{j=1}^{m}\sum_{i=1}^{n}[\mathbb{I}(x_i \in B_j)\mathbb{I}(x \in B_j)] \\
&= \frac{1}{nh}\sum_{i=1}^{n}[\sum_{j=1}^{m}\mathbb{I}(x_i \in B_j)\mathbb{I}(x \in B_j) \\
\hat{f}_n(x) &= \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x - X_i}{h}\right),
\end{aligned}$$

where K is a kernel function.

Different Kernel Functions

$$\text{Kernel Functions} = \begin{cases}
(1) Gaussian & k(x) = \frac{1}{\sqrt{2\pi}}\exp{(-\frac{x^2}{2})}; \\
(2) Uniform & k(x) = \frac{1}{2}\mathbb{I}[-1,1]; \\
(3) Epanechnikov & k(x) = \frac{3}{4}\max{\{1 - x^2, 0\}}; \\
...
\end{cases}$$

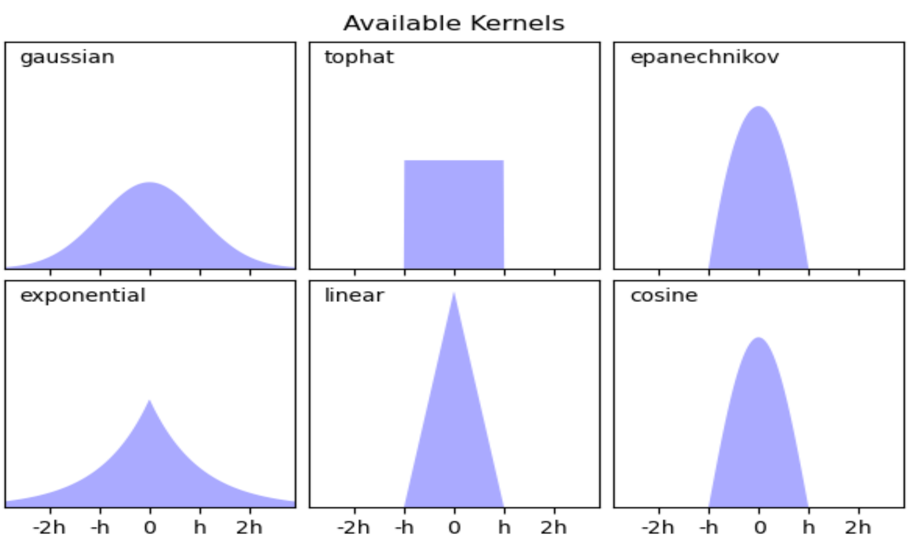

Figure 3. Different Kernel Function

Property of Kernel Functions

1. $\int_{\mathbb{R}} k(x)dx = 1$.
2. $\int_{\mathbb{R}} xk(x)dx = 0 \Rightarrow k(x) = k(-x)$.
3. $\lim_{x\to+\infty} k(x) = \lim_{x\to-\infty} k(x) = 0$.

4 Mean Integrated Squared Error (MISE)

Definition 3

$$MISE = \mathbb{E}\left[\int (\hat{f}_n(x) - f(x))^2 dx\right] = \int Bias^2(\hat{f}_n(x))dx + \int \mathbb{V}(\hat{f}_n(x))dx,$$

$$Bias(x) = \mathbb{E}[\hat{f}_n(x)] - f(x), \quad \mathbb{V}(x) = \mathbb{E}[(\hat{f}_n(x) - \mathbb{E}(\hat{f}_n(x)))^2].$$

Derivation 3

$$\begin{aligned} D(\hat{f}_n(x), f(x)) &= \int\limits_{\mathbb{R}} (\hat{f}_n(x) - f(x))^2 dx, \\ MISE = R(\hat{f}_n(x), f(x)) = \mathbb{E}[D(\hat{f}_n(x), f(x))] &= \int\limits_{\mathbb{R}^n}\int\limits_{\mathbb{R}} (\hat{f}_n(x) - f(x))^2 dx dF(x), \\ R(\hat{f}_n(x), f(x)) &= \int_{\mathbb{R}} \mathrm{Bias}^2(\hat{f}_n(x))dx + \int_{\mathbb{R}} \mathbb{V}(\hat{f}_n(x))dx. \end{aligned}$$

MISE for Density Estimation

MISE for Histogram Density Estimation

Theorem 1 *If f is an L-Lipschitz function, $max_{x\in[0,1]} f(x) \le M$, then*

$$bias(\hat{f}_n(x)) \le Lh, \quad \mathbb{V}(\hat{f}_n(x)) \le \frac{M}{nh} + \frac{M^2}{n},$$

where $\hat{f}_n(x)$ is the histogram density estimation of f.

$$\begin{aligned} MISE &= \int_{\mathbb{R}} Bias^2(\hat{f}_n(x))dx + \int_{\mathbb{R}} \mathbb{V}(\hat{f}_n(x)) \\ &= L^2h^2 + \frac{M}{nh} + \frac{M^2}{n}, \end{aligned}$$

Find minimizing MISE $MISE \Rightarrow h_{opt} = O(n^{-1/3})$, leading to $MISE = O(n^{-2/3})$.

MISE for KDE

Theorem 2 *If f is an L-Lipschitz function, then*

$$\mu_k^2 = \int x^2 k(x)dx, \quad \sigma_k^2 = \int h^2(x)dx,$$

$$bias(\hat{f}_n(x)) = \frac{1}{2}h^2 f''(x)\mu_k^2 + O(h^2), \quad \mathbb{V}(\hat{f}_n(x)) = \frac{1}{nh}f(x_0)\sigma_k^2 + O(\frac{1}{nh}),$$

where $\hat{f}_n(x)$ is the kernel density estimation of f.

$$MISE \approx \frac{1}{4}h^4 \int (f''(x))^2 dx + \frac{1}{nh}\int K^2(u)du.$$

Find minimizing MISE $MISE = 0 \Rightarrow h_{opt} = O(n^{-1/5})$, leading to $MISE = O(n^{-4/5})$.

Density Estimation	Convergence speed
Histogram	$O(n^{-2/3})$
Kernel	$O(n^{-4/5})$