Modern Statistics

Xiangyu Chang

May 5, 2026

Abstract

To be undated.

1 Lecture 16: Nonparametric Inference II

The previous lecture introduced the core concepts of nonparametric inference: the empirical distribution function (EDF), the histogram and kernel density estimators, and the MISE framework for evaluating their performance. In this lecture, we first provide the detailed step-by-step derivations of the bias, variance, and MISE for both the histogram and the KDE—filling in the computations behind the results stated earlier. We then move beyond density estimation to **nonparametric regression**, where we estimate $r(x) = \mathbb{E}[Y \mid X = x]$ without assuming a parametric form, using partition estimators, k-nearest neighbors, and kernel regression. Finally, we introduce the **bootstrap**, a powerful resampling technique for constructing confidence intervals when the sampling distribution of an estimator is analytically intractable.

1.1 Recap: EDF and Density Estimation

We briefly recall the key definitions from the previous lecture.

Definition 1.1 (Empirical Distribution Function)**.** Given i.i.d. samples $\{X_i\}_{i=1}^n \sim F$, the EDF is $F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}(X_i \leqslant x)$.

Proposition 1.2 (Unbiasedness of the EDF)**.** *The EDF is unbiased:* $\mathbb{E}[F_n(x)] = F(x)$.

Proof. $\mathbb{E}[F_n(x)] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathbb{I}(X_i \leqslant x)] = \frac{1}{n}\sum_{i=1}^n \Pr(X_i \leqslant x) = \frac{1}{n} \cdot n\, F(x) = F(x)$. ■

Definition 1.3 (Histogram Density Estimator)**.** For i.i.d. samples $\{X_i\}_{i=1}^n$ with density f, partition the support into m bins $B_j = [b_{j-1}, b_j)$ of equal width $h = b_j - b_{j-1}$, and define

$$\widehat{f}_n(x) = \frac{1}{h}\sum_{j=1}^m \widehat{p}_j\, \mathbb{I}(x \in B_j), \qquad \widehat{p}_j = \frac{n_j}{n}, \qquad n_j = \sum_{i=1}^n \mathbb{I}(X_i \in B_j).$$

1.2 Detailed Analysis: Histogram Density Estimator

We now derive the bias, variance, and MISE of the histogram estimator in detail.

1.2.1 Bias

For $x \in B_j$,

$$\mathbb{E}[\widehat{f}_n(x)] = \frac{1}{h}\,\mathbb{E}[\widehat{p}_j] = \frac{1}{h}\,\mathbb{P}\mathrm{r}(X \in B_j) = \frac{1}{h}\int_{B_j} f(u)\,\mathrm{d}u.$$

Let x^* be the midpoint of B_j. By Taylor expansion:

$$f(u) = f(x^*) + f'(x^*)(u - x^*) + \cdots \quad \Longrightarrow \quad \int_{B_j} f(u)\,\mathrm{d}u \approx h\,f(x^*),$$

since the linear term integrates to zero by symmetry around x^*. Therefore $\mathbb{E}[\widehat{f}_n(x)] \approx f(x^*)$, and the bias satisfies

$$|\mathrm{Bias}(\widehat{f}_n(x))| = |\,\mathbb{E}[\widehat{f}_n(x)] - f(x)| = |f(x^*) - f(x)| \leqslant L\,|x^* - x| \leqslant Lh,$$

where the last inequality uses the L-Lipschitz condition on f and $|x - x^*| \leqslant h$.

1.2.2 Variance

For $x \in B_j$, $\widehat{f}_n(x) = n_j/(nh)$, so:

$$\widehat{f}_n(x) = \frac{n_j}{nh}, \qquad \mathrm{Var}(\widehat{f}_n(x)) = \frac{1}{n^2h^2}\,\mathrm{Var}(n_j).$$

Since $n_j \sim \mathrm{Bin}(n, p_j)$ with $p_j = \mathbb{P}\mathrm{r}(X \in B_j) \approx h\,f(x^*)$:

$$\mathrm{Var}(n_j) = n\,p_j(1 - p_j) \approx n\,h\,f(x^*)(1 - h\,f(x^*)).$$

Therefore,

$$\mathrm{Var}(\widehat{f}_n(x)) = \frac{p_j(1-p_j)}{nh^2} = \frac{f(x^*)(1 - h\,f(x^*))}{nh} \leqslant \frac{f(x^*)}{nh} \leqslant \frac{M}{nh},$$

where $M = \max_x f(x)$ and we used $1 - hf(x^*) \leqslant 1$.

Expanding the full computation step by step:

$$\begin{aligned}
\mathrm{Var}(\widehat{f}_n(x)) &= \mathrm{Var}\left(\frac{\widehat{p}_j}{h}\right) = \frac{1}{h^2}\,\mathrm{Var}(\widehat{p}_j) = \frac{1}{h^2}\,\mathrm{Var}\left(\frac{n_j}{n}\right) \\
&= \frac{1}{n^2h^2}\,\mathrm{Var}\left(\sum_{i=1}^{n} \mathbb{I}(X_i \in B_j)\right) = \frac{1}{n^2h^2}\sum_{i=1}^{n}\mathrm{Var}(\mathbb{I}(X_i \in B_j)) \\
&= \frac{p_j(1-p_j)}{nh^2} \leqslant \frac{p_j}{nh^2} = \frac{f(x^*)\,h}{nh^2} = \frac{f(x^*)}{nh} \leqslant \frac{M}{nh}.
\end{aligned}$$

1.2.3 MISE

Proposition 1.4 (MISE of the Histogram)**.** *If f is L-Lipschitz with $\max f \leqslant M$, then* MISE $= O(n^{-2/3})$ *with optimal bandwidth $h_{\mathrm{opt}} = O(n^{-1/3})$.*

Proof. From the bias and variance bounds:

$$\text{MISE} = \int \text{Bias}^2(\widehat{f}_n(x))\,\mathrm{d}x + \int \text{Var}(\widehat{f}_n(x))\,\mathrm{d}x \leqslant L^2h^2 + \frac{M}{nh}.$$

To find the optimal h, split $\frac{M}{nh}$ into two equal parts and apply the AM-GM inequality to three terms:

$$L^2h^2 + \frac{M}{2nh} + \frac{M}{2nh} \geqslant 3\sqrt[3]{L^2h^2 \cdot \frac{M}{2nh} \cdot \frac{M}{2nh}} = 3\sqrt[3]{\frac{L^2M^2}{4n^2}}.$$

Equality holds when $L^2h^2 = \frac{M}{2nh}$, i.e., $h^3 = \frac{M}{2nL^2}$, giving

$$h_{\text{opt}} = \left(\frac{M}{2nL^2}\right)^{1/3} = O(n^{-1/3}), \qquad \text{MISE} = O(n^{-2/3}).$$

■

1.3 Detailed Analysis: Kernel Density Estimation

The KDE produces smoother density estimates than the histogram. We now derive its bias, variance, and MISE in detail.

1.3.1 Definition and Kernel Properties

Definition 1.5 (Kernel Density Estimator)**.** The kernel density estimator with bandwidth $h > 0$ and kernel function K is

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

where K satisfies:

- **Normalization**: $\int_{\mathbb{R}} K(u)\,\mathrm{d}u = 1$.
- **Symmetry**: $K(u) = K(-u)$.
- **Decay**: $\lim_{|u|\to\infty} K(u) = 0$.

We define the second moment $\mu_K^2 \overset{\text{def}}{=} \int u^2K(u)\,\mathrm{d}u < \infty$ and the roughness $\sigma_K^2 \overset{\text{def}}{=} \int K^2(u)\,\mathrm{d}u < \infty$.

1.3.2 Bias Derivation

Assume f is twice continuously differentiable. The bias of $\widehat{f}_h$ at a point x_0 is:

$$\begin{aligned}
\text{Bias}(x_0) &= \mathbb{E}[\widehat{f}_h(x_0)] - f(x_0) = \frac{1}{nh}\sum_{i=1}^{n}\mathbb{E}\left[K\left(\frac{x_0 - X_i}{h}\right)\right] - f(x_0) \\
&= \frac{1}{h}\,\mathbb{E}\left[K\left(\frac{x_0 - X}{h}\right)\right] - f(x_0) \\
&= \frac{1}{h}\int_{\mathbb{R}} K\left(\frac{x_0 - x}{h}\right) f(x)\,\mathrm{d}x - f(x_0).
\end{aligned}$$

Substituting $y = \frac{x-x_0}{h}$ (so $x = x_0 + hy$, $\mathrm{d}x = h\,\mathrm{d}y$) and using $K(-y) = K(y)$:

$$= \frac{1}{h}\int_{\mathbb{R}} K(-y)\, f(x_0 + hy)\, h\,\mathrm{d}y - f(x_0) = \int_{\mathbb{R}} K(y)\, f(x_0 + hy)\,\mathrm{d}y - f(x_0).$$

Writing $f(x_0) = f(x_0)\int K(y)\,\mathrm{d}y = \int K(y)\, f(x_0)\,\mathrm{d}y$ (since $\int K = 1$):

$$= \int_{\mathbb{R}} K(y)\,[f(x_0 + hy) - f(x_0)]\;\mathrm{d}y.$$

Applying the Taylor expansion $f(x_0 + hy) = f(x_0) + f'(x_0)\,hy + \frac{1}{2}f''(x_0)\,h^2y^2 + O(h^3)$:

$$\begin{aligned}
&= \int_{\mathbb{R}} K(y)\left[f'(x_0)\,hy + \frac{f''(x_0)\,h^2y^2}{2} + O(h^3)\right]\mathrm{d}y \\
&= f'(x_0)\,h\underbrace{\int y\,K(y)\,\mathrm{d}y}_{=0\text{ (symmetry)}} + \frac{f''(x_0)\,h^2}{2}\underbrace{\int y^2K(y)\,\mathrm{d}y}_{=\mu_K^2} + O(h^3) \\
&= \frac{f''(x_0)\,h^2\,\mu_K^2}{2} + O(h^3).
\end{aligned}$$

1.3.3 Variance Derivation

Since the X_i are i.i.d.:

$$\begin{aligned}
\mathrm{Var}(\widehat{f}_h(x_0)) &= \frac{1}{n^2h^2}\sum_{i=1}^{n}\mathrm{Var}\left(K\left(\frac{x_0 - X_i}{h}\right)\right) = \frac{1}{nh^2}\;\mathrm{Var}\left(K\left(\frac{x_0 - X}{h}\right)\right) \\
&\leqslant \frac{1}{nh^2}\;\mathbb{E}\left[K^2\left(\frac{x_0 - X}{h}\right)\right],
\end{aligned}$$

where we used $\mathrm{Var}(Y) \leqslant \mathbb{E}[Y^2]$. Substituting $y = \frac{x-x_0}{h}$ and using $K(-y) = K(y)$:

$$= \frac{1}{nh^2}\int_{\mathbb{R}} K^2\left(\frac{x_0 - x}{h}\right) f(x)\,\mathrm{d}x = \frac{1}{nh}\int_{\mathbb{R}} K^2(y)\, f(x_0 + hy)\,\mathrm{d}y.$$

Expanding $f(x_0 + hy) = f(x_0) + f'(x_0)\,hy + O(h^2)$:

$$\begin{aligned}
&= \frac{1}{nh}\left[f(x_0)\int K^2(y)\,\mathrm{d}y + f'(x_0)\,h\underbrace{\int y\,K^2(y)\,\mathrm{d}y}_{=0\text{ (symmetry)}} + O(h^2)\right] \\
&= \frac{f(x_0)\,\sigma_K^2}{nh} + o\left(\frac{1}{nh}\right).
\end{aligned}$$

1.3.4 MISE

Combining the bias and variance results, the pointwise MSE at x_0 is:

$$\mathrm{MSE}(x_0) = \mathrm{Bias}^2(x_0) + \mathrm{Var}(\widehat{f}_h(x_0)) = \frac{(f''(x_0))^2\,h^4\,\mu_K^4}{4} + \frac{f(x_0)\,\sigma_K^2}{nh}.$$

We apply the AM-GM inequality. Splitting the second term into four equal parts gives five terms:

$$\begin{aligned}\mathrm{MSE}(x_0) &= \frac{(f''(x_0))^2\, h^4\, \mu_K^4}{4} + 4 \cdot \frac{f(x_0)\, \sigma_K^2}{4nh} \\ &\geqslant 5\sqrt[5]{\frac{(f''(x_0))^2\, h^4\, \mu_K^4}{4} \cdot \left(\frac{f(x_0)\, \sigma_K^2}{4nh}\right)^4}.\end{aligned}$$

Equality holds when all five terms are equal:

$$\frac{(f''(x_0))^2\, h^4\, \mu_K^4}{4} = \frac{f(x_0)\, \sigma_K^2}{4nh} \quad \Longrightarrow \quad h^5 = \frac{f(x_0)\, \sigma_K^2}{n\, \mu_K^4\, (f''(x_0))^2},$$

which gives $h_{\mathrm{opt}} = O(n^{-1/5})$ and $\mathrm{MSE} = O(n^{-4/5})$.

Integrating over x yields the MISE:

$$\mathrm{MISE} \approx \frac{h^4\, \mu_K^4}{4} \int (f''(x))^2\, \mathrm{d}x + \frac{\sigma_K^2}{nh},$$

which is also minimized at $h = O(n^{-1/5})$ with $\mathrm{MISE} = O(n^{-4/5})$.

1.4 Nonparametric Regression

In parametric regression we model $r(x) = \mathbb{E}[Y \mid X = x]$ as a linear function $r_{\boldsymbol{\beta}}(x) = \boldsymbol{x}^\top \boldsymbol{\beta}$. **Nonparametric regression** instead estimates $r(x)$ directly from the data without specifying a functional form.

Given data $\{(x_i, y_i)\}_{i=1}^n$ with $y_i = r(x_i) + \varepsilon_i$, the nonparametric regression estimator takes the general weighted-average form:

$$r_n(x) = \sum_{i=1}^n v_{n,i}(x)\, y_i,$$

where the normalized weights are defined as

$$w_{n,i}(x) \stackrel{\mathrm{def}}{=} w(x, x_1, \ldots, x_n), \qquad v_{n,i}(x) \stackrel{\mathrm{def}}{=} \frac{w_{n,i}(x)}{\sum_{j=1}^n w_{n,j}(x)},$$

so that $\sum_{i=1}^n v_{n,i}(x) = 1$. Different choices of the weight function w yield different estimators.

Example 1.6 (Partition Estimator)**.** Partition the input space $\mathcal{X} = \bigcup_{j=1}^M B_j$ into disjoint regions with $B_i \cap B_j = \varnothing$ for $i \neq j$. For $x \in B_j$, the partition estimator averages the responses in that region:

$$r_n(x) = \frac{\sum_{i=1}^n \mathbb{I}(x_i \in B_j)\, y_i}{\sum_{i=1}^n \mathbb{I}(x_i \in B_j)}, \qquad x \in B_j.$$

Example 1.7 (k-Nearest Neighbors (KNN))**.** For a fixed query point x, order the training points by distance:

$$\|x_{(1)} - x\| \leqslant \|x_{(2)} - x\| \leqslant \cdots \leqslant \|x_{(n)} - x\|,$$

where $x_{(i)}$ denotes the i-th nearest neighbor. The KNN estimator averages the k nearest responses:

$$r_n(x) = \frac{1}{k} \sum_{i=1}^{k} y_{(i)}.$$

The parameter k plays a role analogous to the bandwidth h: small k gives low bias but high variance, and large k gives high bias but low variance.

Example 1.8 (Nadaraya–Watson Kernel Estimator)**.** Using a kernel function K and bandwidth h, the Nadaraya–Watson estimator is:

$$r_n(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}.$$

For instance, with the uniform kernel $K(u) = \frac{1}{2}\mathbb{I}(|u| \leqslant 1)$, only observations with $x_i \in [x - h,\ x + h]$ receive non-zero weight, making this a local averaging estimator.

1.5 Bootstrap Method

All the inference methods studied so far—confidence intervals, hypothesis tests—rely on knowing (or analytically deriving) the sampling distribution of the estimator. The **bootstrap** offers a computational alternative: approximate the sampling distribution by resampling from the observed data.

1.5.1 Classical Confidence Interval via CLT

Let $X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F$ with $\mu = \mathbb{E}[X_i]$ and $\sigma^2 = \mathrm{Var}(X_i) < \infty$. Define

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i, \qquad \widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2.$$

By the Central Limit Theorem,

$$\frac{\bar{X}_n - \mu}{\widehat{\sigma}/\sqrt{n}} \xrightarrow{d} N(0,1),$$

so an approximate $(1-\alpha)$ confidence interval for μ is

$$\left(\bar{X}_n - z_{\alpha/2} \frac{\widehat{\sigma}}{\sqrt{n}},\ \ \bar{X}_n + z_{\alpha/2} \frac{\widehat{\sigma}}{\sqrt{n}} \right).$$

1.6 Summary and Outlook

This lecture provided detailed analyses and introduced new methods for nonparametric inference:

1. **Histogram density estimation**: Detailed derivation of bias ($\leqslant Lh$), variance ($\leqslant M/(nh)$), and MISE ($= O(n^{-2/3})$ with $h_{\text{opt}} = O(n^{-1/3})$) via the AM-GM inequality.
2. **Kernel density estimation**: Step-by-step derivation of bias ($= \frac{1}{2}h^2 f''(x_0)\mu_K^2 + O(h^3)$), variance ($= f(x_0)\sigma_K^2/(nh) + o(1/(nh))$), and MISE ($= O(n^{-4/5})$ with $h_{\text{opt}} = O(n^{-1/5})$).

3. **Nonparametric regression**: Weighted-average estimators for $r(x) = \mathbb{E}[Y \mid X = x]$ without parametric assumptions—partition estimators, k-nearest neighbors, and the Nadaraya–Watson kernel estimator.

Together with the previous lecture, we now have a complete nonparametric toolkit that complements the parametric methods developed earlier in the course. These methods trade some statistical efficiency for robustness when parametric assumptions are questionable.

References