MATH-517: Assignment 3

Alexia Paratte

2025-10-05

Theoretical Exercise

Local Linear Regression as a Linear Smoother

Setup: We observe i.i.d. samples (X_i, Y_i) from the model

$$Y_i = m(X_i) + \varepsilon_i, \qquad i \in \{1, \dots, n\}$$

and we estimate m at a target point $x \in \mathbb{R}$ via the local linear estimator we defined in class:

$$(\hat{\beta}_0(x), \hat{\beta}_1(x)) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} \Big(Y_i - \beta_0 - \beta_1(X_i - x)\Big)^2 K\Big(\tfrac{X_i - x}{h}\Big)$$

where K is a kernel and $h > 0$ a bandwidth. The fitted value is denoted $\hat{m}(x) = \hat{\beta}_0(x)$.
We also define

$$S_{n,k}(x) = \frac{1}{nh} \sum_{i=1}^{n} (X_i - x)^k K\Big(\tfrac{X_i - x}{h}\Big), \quad k \in \{0, 1, 2\}$$

$$T_{n,k}(x) = \frac{1}{nh} \sum_{i=1}^{n} (X_i - x)^k K\Big(\tfrac{X_i - x}{h}\Big) Y_i, \quad k \in \{0, 1\}$$

$$K_i := K\Big(\tfrac{X_i - x}{h}\Big), \quad i = 1, \dots, n$$

and

$$\Delta(x) = S_{n,0}(x) S_{n,2}(x) - S_{n,1}(x)^2$$

1. Show that $\hat{m}(x)$ can be expressed as a weighted average of the observations

Let

$$X = \begin{bmatrix} 1 & X_1 - x \\ \vdots & \vdots \\ 1 & X_n - x \end{bmatrix}, \quad W = \mathrm{diag}(K_1, \dots, K_n), \quad y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$

By weighted least squares, we get

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (X^\top W X)^{-1} X^\top W y$$

Computation yields

$$X^\top W X = \begin{bmatrix} \sum_i K_i & \sum_i (X_i - x) K_i \\ \sum_i (X_i - x) K_i & \sum_i (X_i - x)^2 K_i \end{bmatrix} = nh \begin{bmatrix} S_{n,0} & S_{n,1} \\ S_{n,1} & S_{n,2} \end{bmatrix}$$

$$X^\top W y = \begin{bmatrix} \sum_i K_i Y_i \\ \sum_i (X_i - x) K_i Y_i \end{bmatrix} = nh \begin{bmatrix} T_{n,0} \\ T_{n,1} \end{bmatrix}$$

Hence, with $\Delta = \Delta(x)$,

$$(X^\top W X)^{-1} = \frac{1}{nh} \cdot \frac{1}{\Delta} \begin{bmatrix} S_{n,2} & -S_{n,1} \\ -S_{n,1} & S_{n,0} \end{bmatrix}$$

Therefore,

$$\hat{\beta} = \frac{1}{\Delta} \begin{bmatrix} S_{n,2} & -S_{n,1} \\ -S_{n,1} & S_{n,0} \end{bmatrix} \begin{bmatrix} T_{n,0} \\ T_{n,1} \end{bmatrix}, \qquad \hat{m}(x) = \hat{\beta}_0(x) = \frac{S_{n,2} T_{n,0} - S_{n,1} T_{n,1}}{\Delta}$$

Expanding $T_{n,0}, T_{n,1}$ shows

$$\hat{m}(x) = \sum_{i=1}^{n} w_{ni}(x) Y_i$$

so $\hat{m}(x)$ is a linear smoother. $\square$

2. Explicit weights $w_{ni}(x)$

From the previous definitions and calculations:

$$\widehat{m}(x) = \frac{S_{n,2}\,\frac{1}{nh}\sum_i K_i Y_i - S_{n,1}\,\frac{1}{nh}\sum_i (X_i - x) K_i Y_i}{\Delta} = \sum_{i=1}^{n} \left[\frac{1}{nh} \cdot \frac{K_i (S_{n,2} - (X_i - x) S_{n,1})}{\Delta} \right] Y_i$$

Thus,

$$\boxed{w_{ni}(x) = \frac{1}{nh} \frac{K\left(\frac{X_i - x}{h}\right)\left(S_{n,2}(x) - (X_i - x) S_{n,1}(x)\right)}{S_{n,0}(x) S_{n,2}(x) - S_{n,1}(x)^2}}$$

These weights depend only on x, X_i, K, h (not on the Y_i's).

3. Prove that $\sum_{i=1}^{n} w_{ni}(x) = 1$

We sum the weights:

$$\sum_{i=1}^{n} w_{ni}(x) = \frac{1}{nh} \cdot \frac{1}{\Delta} \left(S_{n,2} \sum_{i=1}^{n} K_i - S_{n,1} \sum_{i=1}^{n} (X_i - x) K_i \right).$$

But

$$\sum_i K_i = nh\, S_{n,0}, \qquad \sum_i (X_i - x) K_i = nh\, S_{n,1}.$$

So,

$$\sum_{i=1}^{n} w_{ni}(x) = \frac{1}{nh} \cdot \frac{1}{\Delta} \left(S_{n,2} \left(nh S_{n,0} \right) - S_{n,1} \left(nh S_{n,1} \right) \right) = \frac{S_{n,2} S_{n,0} - S_{n,1}^2}{\Delta} = \frac{\Delta}{\Delta} = 1 \quad \square$$

Conclusion

The local linear estimator can be written as

$$\widehat{m}(x) = \sum_{i=1}^{n} w_{ni}(x) Y_i,$$

with weights $w_{ni}(x)$ given in Section 2. By Section 3, they sum to one and do not depend on the responses Y_i.

Practical exercise

Aim and setup

We estimate the optimal **global** bandwidth for a local linear smoother under homoscedastic noise and a quartic (biweight) kernel via the AMISE rule

$$\hat{h}_{\text{AMISE}} = n^{-1/5}\left(\frac{35\,\hat{\sigma}^2\,|\text{supp}(X)|}{\hat{\theta}_{22}}\right)^{1/5}, \qquad \hat{\theta}_{22} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N}\left[\hat{m}_j''(X_i)\right]^2 \mathbf{1}_{\{X_i \in \mathcal{X}_j\}},$$

with

$$\hat{\sigma}^2 = \frac{1}{n-5N}\sum_{i=1}^{n}\sum_{j=1}^{N}(Y_i - \hat{m}_j(X_i))^2 \mathbf{1}_{\{X_i \in \mathcal{X}_j\}}.$$

Here the sample is split into N contiguous blocks in x: in each block j we fit the quartic OLS $y_i = \beta_{0j} + \beta_{1j}x_i + \beta_{2j}x_i^2 + \beta_{3j}x_i^3 + \beta_{4j}x_i^4 + \varepsilon_i$ to obtain $\hat{m}_j$ and $\hat{m}_j''(x) = 2\hat{\beta}_{2j} + 6\hat{\beta}_{3j}x + 12\hat{\beta}_{4j}x^2$. Data are generated as $X \sim \text{Beta}(\alpha, \beta)$, $Y = m(X) + \varepsilon$ with $m(x) = \sin\big((x/3 + 0.1)^{-1}\big)$ and $\varepsilon \sim \mathcal{N}(0,1)$. The support length is $|\text{supp}(X)| = 1$.

Results

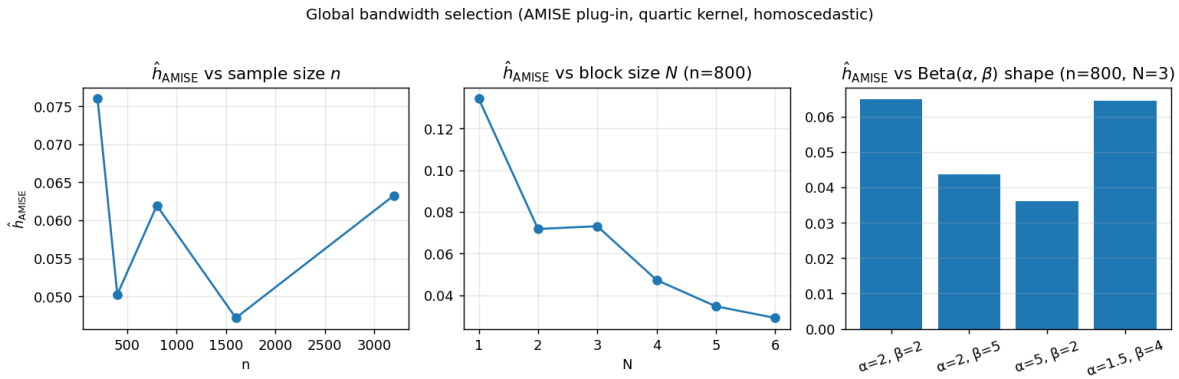

Figure 1: Global bandwidth selection (AMISE plug-in, quartic kernel, homoscedastic).
Left: $\hat{h}_{\text{AMISE}}$ versus sample size n.
Middle: $\hat{h}_{\text{AMISE}}$ versus block size N (with $n = 800$).
Right: $\hat{h}_{\text{AMISE}}$ for several Beta(α, β) shapes (with $n = 800, N = 3$).

Discussion

- **Effect of sample size** n (left panel).
  The AMISE formula predicts $h_{\text{AMISE}} \propto n^{-1/5}$, therefore a slow decrease with n. In finite samples the plug-in estimates $\hat{\theta}_{22}$ and $\hat{\sigma}^2$ add variability. It explains the non-monotone pattern (e.g., the bump at large n). Overall the scale is consistent with an $n^{-1/5}$ trend: larger n tends to allow a **smaller** bandwidth.
- **Effect of block size** N (middle panel).
  Increasing N decreases $\hat{h}_{\text{AMISE}}$ in this run. *Intuition*: with more blocks we fit more local polynomials, which typically increases the estimated curvature term $\hat{\theta}_{22}$; since $h \propto \hat{\theta}_{22}^{-1/5}$, a larger curvature estimate yields a **smaller** bandwidth.
  Practically there is a trade-off: too small $N \to$ biased curvature (oversmoothing); too large $N \to$ noisy fits and the denominator $n - 5N$ in $\hat{\sigma}^2$ gets small. A moderate N (e.g., 3–5 here) is a sensible choice.
- **Effect of the covariate distribution** (right panel).
  The value of $\hat{h}_{\text{AMISE}}$ changes with Beta(α, β). When the Beta shape concentrates mass near the boundaries (e.g., $\alpha = 2, \beta = 5$ or $\alpha = 1.5, \beta = 4$), the effective curvature weighting $\int (m''(x))^2 f_X(x)\, dx$ changes and $\hat{\theta}_{22}$ differs, which shifts the optimal bandwidth. In this run, symmetric Beta$(2, 2)$ and the right-skewed Beta$(1.5, 4)$ give the **largest** $\hat{h}_{\text{AMISE}}$, while Beta$(5, 2)$ (mass near 1) gives the **smallest**.

Summary: The dominant driver is the **sample size** via the $n^{-1/5}$ rate. The **block size** N tunes the curvature/variance plug-ins and should remain moderate (with $5N < n$). The **covariate distribution** affects the optimal bandwidth through θ_{22}; skewed designs can lead to noticeably different $\hat{h}_{\text{AMISE}}$.

Script

The full script we ran for this report can be found here.