MATH-517: Assignment 3

Julien Nicolay

October 2025

=======

Theoretical exercise

Definition of the assumptions

We are given i.i.d. samples

$$(x_i, y_i), \quad i = 1, \dots, n,$$

from the model

$$y_i = m(x_i) + \varepsilon_i.$$

Our goal is to estimate the regression function

$$m(x) = \mathbb{E}[Y \mid X = x].$$

For a fixed point $x \in \mathbb{R}$, the *local linear estimator* is obtained by solving the weighted least squares problem

$$(\hat{\beta}_0(x), \hat{\beta}_1(x)) = \arg \min_{\beta_0, \beta_1 \in \mathbb{R}} \sum_{i=1}^{n} \Big(y_i - \beta_0 - \beta_1 (x_i - x) \Big)^2 K\Big(\frac{x_i - x}{h} \Big),$$

where K is a kernel function and $h > 0$ is a bandwidth.

From the Taylor expansion in **Local Polynomial Regression**,

$$\hat{\beta}_j(x) = \frac{m^{(j)}(x)}{j!}, \qquad j \in \{0, ..., p\}$$

Thus, the local linear estimator identifies

$$\hat{m}(x) = \hat{\beta}_0(x), \qquad \hat{m}'(x) = \hat{\beta}_1(x).$$

corresponding to the case of a local polynomial of degree $p = 1$.

Matrix formulation

Define

$$X = \begin{bmatrix} 1 & (x_1 - x) \\ 1 & (x_2 - x) \\ \vdots & \vdots \\ 1 & (x_n - x) \end{bmatrix}, \qquad W = \operatorname{diag}(K(\tfrac{x_i - x}{h}))$$

and let $y = (y_1, \dots, y_n)^\top$.

The weighted least squares solution is

$$\hat{\beta}(x) = (X^\top W X)^{-1} X^\top W y.$$

Hence,

$$\hat{m}(x) = \hat{\beta}_0(x) = e_1^\top \hat{\beta}(x) = e_1^\top (X^\top W X)^{-1} X^\top W y,$$

where $e_1 = (1,0)^\top$.

Explicit expression of the weights

$$X^\top W X = \begin{pmatrix} \sum_{i=1}^n K(\frac{x_i - x}{h}) & \sum_{i=1}^n (x_i - x)\, K(\frac{x_i - x}{h}) \\ \sum_{i=1}^n (x_i - x)\, K(\frac{x_i - x}{h}) & \sum_{i=1}^n (x_i - x)^2\, K(\frac{x_i - x}{h}) \end{pmatrix},$$

and

$$X^\top W y = \begin{pmatrix} \sum_{i=1}^n K(\frac{x_i - x}{h})\, y_i \\ \sum_{i=1}^n (x_i - x)\, K(\frac{x_i - x}{h})\, y_i \end{pmatrix}.$$

By using the matrix inversion formula

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix},$$

and introducing

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

we obtain

$$(X^\top W X)^{-1} = \frac{1}{(nh)^2 \Big(S_{n,0}(x) S_{n,2}(x) - \big(S_{n,1}(x)\big)^2\Big)} \begin{pmatrix} nh\, S_{n,2}(x) & -nh\, S_{n,1}(x) \\ -nh\, S_{n,1}(x) & nh\, S_{n,0}(x) \end{pmatrix}.$$

Then the estimator can be written as

$$\hat{m}(x) = \frac{S_{n,2}(x)\left(\frac{1}{nh}\sum_{i=1}^n K\left(\frac{x_i-x}{h}\right) y_i\right) - S_{n,1}(x)\left(\frac{1}{nh}\sum_{i=1}^n (x_i-x)\, K\left(\frac{x_i-x}{h}\right) y_i\right)}{S_{n,0}(x)\, S_{n,2}(x) - \left(S_{n,1}(x)\right)^2}.$$

Equivalently, this shows that

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

with weights

$$w_{n,i}(x) = \frac{\frac{1}{nh} K\left(\frac{x_i-x}{h}\right) \left[S_{n,2}(x) - (x_i-x)\, S_{n,1}(x)\right]}{S_{n,0}(x)\, S_{n,2}(x) - \left(S_{n,1}(x)\right)^2}.$$

Sum of the weights

Since

$$\sum_{i=1}^n K\left(\frac{x_i-x}{h}\right) = nh\, S_{n,0}(x), \qquad \sum_{i=1}^n (x_i-x)\, K\left(\frac{x_i-x}{h}\right) = nh\, S_{n,1}(x),$$

we have

$$\begin{aligned}
\sum_{i=1}^n w_{n,i}(x) &= \sum_{i=1}^n \frac{\frac{1}{nh} K\left(\frac{x_i-x}{h}\right) \left[S_{n,2}(x) - (x_i-x)S_{n,1}(x)\right]}{S_{n,0}(x)S_{n,2}(x) - (S_{n,1}(x))^2} \\
&= \frac{1}{nh}\, \frac{S_{n,2}(x)\, \sum_{i=1}^n K\left(\frac{x_i-x}{h}\right) - S_{n,1}(x)\, \sum_{i=1}^n (x_i-x)\, K\left(\frac{x_i-x}{h}\right)}{S_{n,0}(x)S_{n,2}(x) - (S_{n,1}(x))^2} \\
&= \frac{1}{nh}\, \frac{S_{n,2}(x)\,(nh\, S_{n,0}(x)) - S_{n,1}(x)\,(nh\, S_{n,1}(x))}{S_{n,0}(x)S_{n,2}(x) - (S_{n,1}(x))^2} \\
&= \frac{S_{n,2}(x)S_{n,0}(x) - (S_{n,1}(x))^2}{S_{n,0}(x)S_{n,2}(x) - (S_{n,1}(x))^2} \\
&= 1.
\end{aligned}$$

Practical exercise

1. Aim of the simulation

The objective of this study is to investigate the behavior of the bandwidth parameter h_{AMISE} and the optimal number of blocks N_{opt}, selected via Mallows' C_p criterion, in a nonparametric regression framework where the regression function is nonlinear and the covariate follows a Beta distribution.

More precisely, the simulation study aims to assess how both model parameters and hyperparameters influence the estimation process. The model parameters include the regression function $m(x)$ and the error variance σ^2. The hyperparameters correspond to the methodological and design choices that affect the estimator's behavior, namely the bandwidth h, the number of blocks N, and the shape parameters α, β of the Beta distribution $X \sim \text{Beta}(\alpha, \beta)$, which determines the design density of the covariate X. Through this analysis, we seek to understand how these quantities jointly affect the bias–variance trade-off and the resulting optimal smoothing level in nonparametric estimation.

---

2. Model and setting

We consider the model

$$Y = m(X) + \varepsilon, \qquad m(x) = \sin((x/3 + 0.1)^{-1}), \qquad X \sim \text{Beta}(\alpha, \beta), \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Different values of (α, β) are explored to assess the influence of the shape of the covariate distribution on the estimates of h_{AMISE} and N_{opt}. The error variance is fixed to $\sigma^2 = 1$, and several sample sizes n are considered.

---

3. Methodology

The range of X is divided into N blocks of approximately equal size, based on the empirical quantiles of the sample. Initially, I partitioned the range of X into blocks of equal width, but this approach caused numerical issues when using extreme values of α and β in the simulation as there was not enough points in several intervals. After consulting Ruppert et al. (1995), I noted that "the partition can be formed by either dividing the range into equally sized blocks or by dividing the data into equal-sized subsamples. The second option, which we use in our simulations, has the advantage of adapting better to nonuniform designs and decreasing the chance of overfitting." Following this recommendation, I chose to divide the data into equal-sized subsamples rather than equal-width intervals, ensuring greater stability for highly skewed Beta distributions.

Within each block, a **polynomial regression of degree 4** is fitted, which allows estimating both the regression function $m(x)$ and its second derivative $m''(x)$.

From these blockwise fits, we estimate the quantities

$$\hat{\theta}_{22} = \frac{1}{n} \sum_i \big(m''(x_i)\big)^2, \qquad \hat{\sigma}^2 = \frac{\text{RSS}}{n - 5N},$$

where RSS denotes the residual sum of squares over all fitted blocks.

---

4. Bandwidth and Mallows' criterion

With $\text{supp}(X) = [0, 1]$, the AMISE-optimal bandwidth is estimated as

$$h_{\text{AMISE}} = n^{-1/5} \left(\frac{35\,\hat{\sigma}^2}{\hat{\theta}_{22}} \right)^{1/5},$$

and the optimal number of blocks N_{opt} is obtained by minimizing Mallows' criterion:

$$C_p(N) = \frac{\text{RSS}_N}{\text{RSS}_{N_{\max}}/(n - 5N_{\max})} - (n - 10N).$$

This provides a data-driven trade-off between model complexity (through N) and the residual fit quality.

---

5. Simulation design

The simulation procedure is repeated across multiple configurations:

- **Sample sizes:** $n \in \{200, 500, 1000\}$
- **Beta distributions:** $(\alpha, \beta) \in \{(5,2), (2,2), (1,2), (0.5,2), (0.1,2)\}$
- **Repetitions:** $R = 100$

For each replication, we compute: the sequence $h_{\mathrm{AMISE}}(N)$ for $N = 1, \dots, N_{\max}$, the corresponding $C_p(N)$ values, and identify N_{opt} as the minimizer of C_p.

---

6. Results

(a) Behavior of h_{AMISE} as a function of N

Figure 1 shows the estimated bandwidth h_{AMISE} as a function of the number of blocks N, for a representative configuration $(n = 500, \alpha = 2, \beta = 2)$.

The h_{AMISE} decreases when the number of blocks N increases. According to the theoretical expression $h_{\mathrm{AMISE}}(N) \propto \left(\frac{\hat{\sigma}^2(N)}{\hat{\theta}_{22}(N)}\right)^{1/5}$, the optimal bandwidth depends on how the variance estimate $\hat{\sigma}^2(N)$ and the curvature term $\hat{\theta}_{22}(N)$ evolve with the number of blocks N. When N increases, the blocks become smaller, which allows a finer estimation of the curvature and so, $\hat{\theta}_{22}(N)$ increases. $\hat{\theta}_{22}(N)$ grows faster than $\hat{\sigma}^2(N)$ as we fixed σ^2 to 1, so the ratio $\frac{\hat{\sigma}^2}{\hat{\theta}_{22}}$ decreases, and therefore h_{AMISE} decreases. Therefore, increasing N reduces the bias but comes at the cost of higher variance.

(b) Evolution of N_{opt} with the sample size

Figure 2 reports the evolution of the optimal block number N_{opt} as the sample size n increases.

As expected, N_{opt} grows with n, since larger samples allow finer partitioning of the data points without introducing excessive variance as mentionned above. In other words, when there is more data in the sample, each block still contains enough observations to provide stable local estimates, reducing the bias while keeping the variance under control.

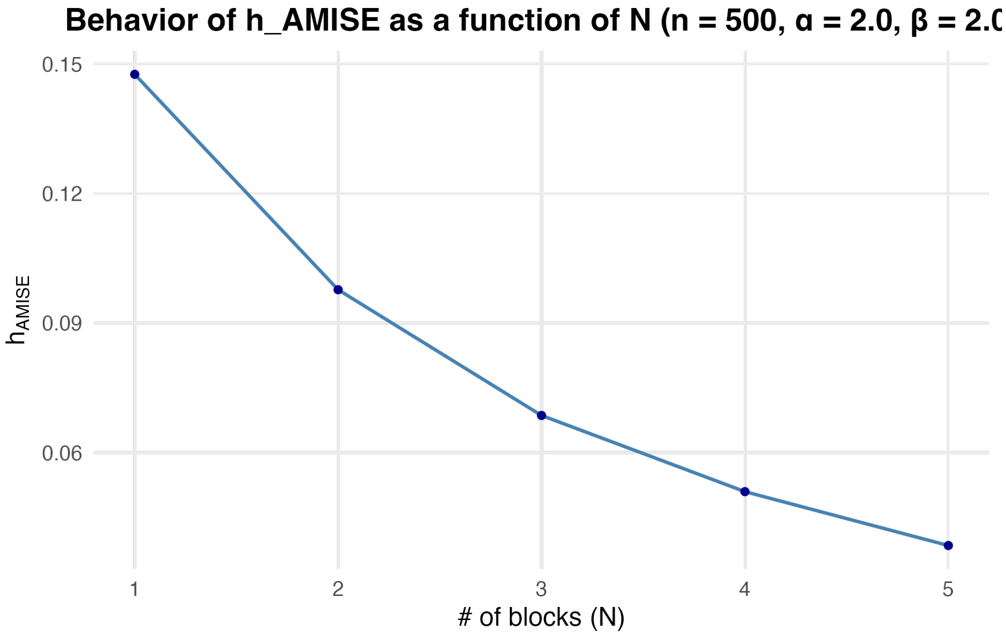

Figure 1

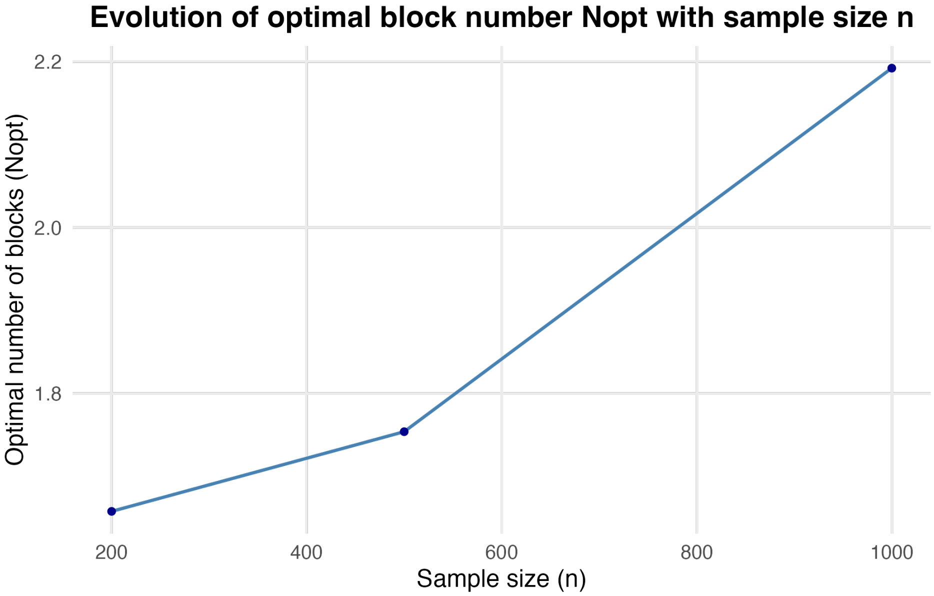

Figure 2

(c) Influence of the Beta shape

Figure 3 presents the mean estimated values of h_{AMISE} across the different Beta distributions.

The results show that the estimated bandwidth is sensitive to the distribution of X. When α increases while β remains fixed, the distribution of X becomes more concentrated toward one side of the support, typically toward 1 when $\alpha > \beta$ and toward 0 when $\alpha < \beta$.

When the distribution shifts toward 0, most of the data accumulate near the boundary, where the function $m(x)$ has higher curvature. This leads to a large increase in the estimated $\widehat{\theta}_{22}$ term, which captures the squared curvature of $m(x)$. Since $h_{\text{AMISE}} \propto \left(\frac{\hat{\sigma}^2}{\widehat{\theta}22}\right)^{1/5}$, this strong rise in $\widehat{\theta}_{22}$ causes the optimal bandwidth to decrease, resulting in finer local smoothing near the boundary.

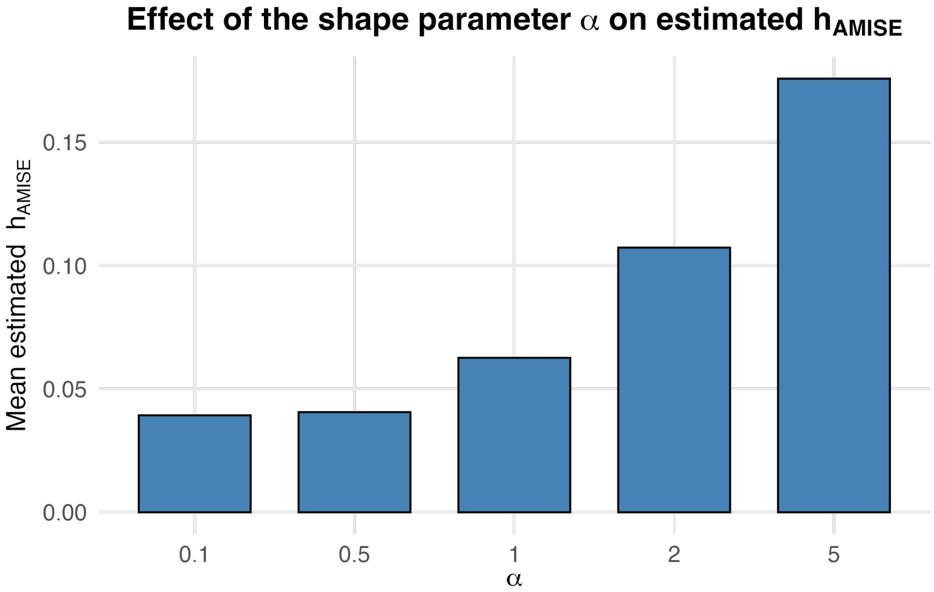

Figure 3

---

7. Discussion and conclusion

The simulations confirm several classical insights in nonparametric estimation:

1. The **number of blocks** N plays a crucial role in the estimation of the curvature term $\widehat{\theta}_{22}$. Mallows' criterion provides a stable, data-driven approach for selecting it.
2. The decrease of h_{AMISE} with larger N reflects the classical bias–variance trade-off.

3. As the sample size n increases, N_{opt} also increases, since larger datasets enable finer partitioning that reduces bias while maintaining stable variance.

4. The distribution of X impacts directly the variation of the number of observations along the different regions in the support of X. When the distribution of X shifts towards 0 with $\alpha < \beta$, data concentrate near regions of high curvature, increasing $\widehat{\theta}_{22}$ and thus reducing h_{AMISE}.