Assignment 3

Student: Leonardo Tonelli

MATH 517 - Statistical Computation and Visualisation

October 4, 2025

École Polytechnique Fédérale de Lausanne

1 Theoretical Exercise

We are given i.i.d. samples $(x_i, y_i), i = 1, \dots n$ from the model:

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

where $x_i \in \mathbb{R}$. The local linear regression estimator at a point x is defined by:

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

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

1.1 Showing $\hat{m}(x)$ is a Weighted Average

Let's define the design matrix and weight matrix:

- $W_x = \text{diag}\left(K\left(\frac{X_1 - x}{h}\right), \dots, K\left(\frac{X_n - x}{h}\right)\right)$
- $X_x = \begin{bmatrix} 1 & X_1 - x \\ \vdots & \vdots \\ 1 & X_n - x \end{bmatrix}$

The weighted least squares solution is:

$$\hat{\beta}(x) = (X_x^\top W_x X_x)^{-1} X_x^\top W_x Y$$

Since $\hat{m}(x) = \hat{\beta}_0(x)$, we extract only the first component by multiplying with $e_1 = (1, 0)^\top$:

$$\hat{m}(x) = e_1^\top (X_x^\top W_x X_x)^{-1} X_x^\top W_x Y$$

.

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

$$w_{ni}(x) = e_1^\top (X_x^\top W_x X_x)^{-1} X_x^\top W_x e_i$$

and e_i is the i-th standard basis vector to isolate the ith element of the weight vector. These weights depend only on $\{X_i\}$, K, h, and x, not on $\{Y_i\}$.

1.2 Explicit Expression for Weights

To show the equivalence of that expression with the weights just found we need to make explicit the vectorized solution for the weight vector and rewrite it in the S notation.
Thus, let's compute $X_x^\top W_x X_x$:

$$X_x^\top W_x X_x = \begin{bmatrix} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) & \sum_{i=1}^{n} (X_i - x) K\left(\frac{X_i - x}{h}\right) \\ \sum_{i=1}^{n} (X_i - x) K\left(\frac{X_i - x}{h}\right) & \sum_{i=1}^{n} (X_i - x)^2 K\left(\frac{X_i - x}{h}\right) \end{bmatrix}$$

Using the notation:

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

We recognize:

$$X_x^\top W_x X_x = nh \begin{bmatrix} S_{n,0}(x) & S_{n,1}(x) \\ S_{n,1}(x) & S_{n,2}(x) \end{bmatrix}$$

The inverse is:

$$(X_x^\top W_x X_x)^{-1} = \frac{1}{nh[S_{n,0}(x) S_{n,2}(x) - S_{n,1}^2(x)]} \begin{bmatrix} S_{n,2}(x) & -S_{n,1}(x) \\ -S_{n,1}(x) & S_{n,0}(x) \end{bmatrix}$$

Since $\hat{m}(x) = \hat{\beta}_0(x) = e_1^\top \hat{\beta}(x)$ where $e_1 = (1, 0)^\top$, we get:

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

The weights become:

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

1.3 Proof That Weights Sum to 1

We need to show $\sum_{i=1}^{n} w_{ni}(x) = 1$.

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

Using the definitions of $S_{n,k}(x)$:

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

Substituting these expressions:

$$\begin{aligned}\sum_{i=1}^{n} w_{ni}(x) &= \frac{S_{n,2}(x)(nhS_{n,0}(x)) - S_{n,1}(x)(nhS_{n,1}(x))}{nh[S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)]} \\ &= \frac{nh[S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)]}{nh[S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)]} \\ &= 1\end{aligned}$$

This completes the proof that the weights sum to 1.

2 Practical Exercise

Aim of the Study

This simulation study investigates how the asymptotically optimal bandwidth h_{AMISE} for local linear regression depends on various factors: sample size n, block size N used in parameter estimation, and the distribution of covariates X through Beta distribution parameters (α, β). The function to estimate will be referenced multiple times, then we put a figure here displaying it in a plot, for reference:

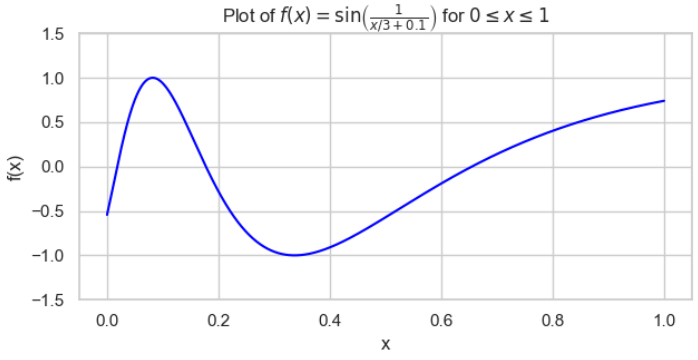


Figure 1: Function to estimate

Simulation Design and Implementation

The study is implemented in python, and its key components are:

- **Data Generation**:

  - Covariates X generated from Beta(α, β) distribution
  - Response $Y = m(X) + \epsilon$ where $m(x) = \sin\left((x/3 + 0.1)^{-1}\right)$ and $\epsilon \sim N(0, 1)$
  - Multiple iterations (5 per configuration) ensure statistical stability
  - Seeds are managed for reproducibility purposes

- **Quantile-based Blocking Implementation**: The blocks are not taken with the same length, but based off quantiles of the distribution of our covariates X. In this way each block has the same amount of points, facilitating fitting $\hat{m}(x)$ for each block in undersampled examples of our covariates and sparse Beta distribution.
  - Partitioned into N blocks using `np.quantile(X, np.linspace(0, 1, N+1))`
  - Each block contains approximately equal number of observations: block$_j = [q_{j-1}, q_j]$
  - blocks with < 5 observations skipped to ensure reliable polynomial fitting

- **Parameter Estimation Procedure**:
  - Polynomial $m_j(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$ fitted in each block using sklearn built in functions
  - Second derivative computed analytically: $m_j''(x) = 2\beta_2 + 6\beta_3 x + 12\beta_4 x^2$
  - $\hat{\theta}_{22} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N}[m_j''(X_i)]^2 \mathbf{1}_{\{X_i \in \text{block}_j\}}$
  - $\hat{\sigma}^2 = \frac{1}{n-5N}\sum_{i=1}^{n}\sum_{j=1}^{N}[Y_i - m_j(X_i)]^2 \mathbf{1}_{\{X_i \in \text{block}_j\}}$

- **Bandwidth Calculation**:

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

- **Simulations**: The simulations are run in **three** different data collection loops. In the first one, the bandwidth h_{AMISE} is collected for different pairs of Beta parameters and for different sample sizes leaving N as fixed at the optimal value computed with the C_p criterion. The second loop instead (used to study the effect of blocks number N), compute the optimal bandwidth for varying sample size n and varying Beta distributions. Lastly, a simulation is run to evaluate the relationship between optimal N and n, this compute the N optimal for varying sizes of the samples and different beta distributions, without recomputing the bandwidth h_{AMISE}. In all three cases, the computations is repeated 5 times with different data samples, to ensure robustness of results and complete visualization.

Optimization Procedure When they are studied the effect of the Beta distribution parameters and the sample size n, it is chosen a fixed optimal block size N using Mallow's C_p criterion, as suggested in the text of the exercise. Then it is chosen a block number N such that the following criterion is minimized:

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

where $N_{\max} = \min(n/20, 10)$ ensures sufficient data per block. The implementation also includes robust error handling for edge cases.

Results and Discussion

Effect of Sample Size (n) The plot in Figure 2 clearly demonstrates the theoretical relationship $h_{\text{AMISE}} \propto n^{-1/5}$. As the sample size increases from 250 to 2000, the **optimal bandwidth decreases** gradually for all Beta distributions. However, the decrease occurs at a slower rate for the Beta(5,2) distribution. This can be explained by the behavior of the true function $m(x)$, shown in Figure 1, which is relatively simple and linear on the right side of its support. Because the Beta(5,2) distribution draws most covariates X from this region, the sampled data presents a function that appears simpler. Consequently, the optimal bandwidth does not require the same degree of refinement to capture complex local features, which explains its slower rate of decrease as the sample size grows.

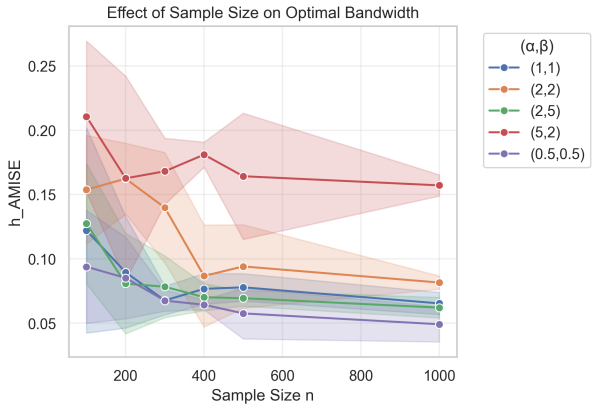


Figure 2: Impact of sample size n on optimal bandwidth h_{AMISE}

Behavior of h_{AMISE} as N Grows The graph in Figure 3 reveals a clear **inverse relationship** between the optimal bandwidth h_{AMISE} and the block size N. For all covariate distributions studied, the bandwidth decreases gradually as the number of blocks **increases**. This relationship is mathematically grounded in the formula for h_{AMISE}, which depends linearly on the estimate $\hat{\sigma}^2$, a quantity that itself has an inverse relationship with N. Additionally, the model's performance within each block improves as N increases, this decreasese the MSE at each block and hence the $\hat{\sigma}^2$.

To illustrate this **intuitively**, let's consider the case of high fixed amount of data $n = 1000$. For small N, each local estimate $\hat{m}_j$ is already a good fit for the data within its large block, as the model has ample data. However, as N increases and the number of data points per block decreases, the degree 4 polynomial fit becomes increasingly flexible relative to the reduced data, allowing it to capture finer, local variations. This gradually improves the fit, reducing residuals and thus $\hat{\sigma}^2$, until it potentially leads to overfitting in the limit. This mechanism explains the observed decrease in $\hat{\sigma}^2$ and, consequently, in h_{AMISE}.

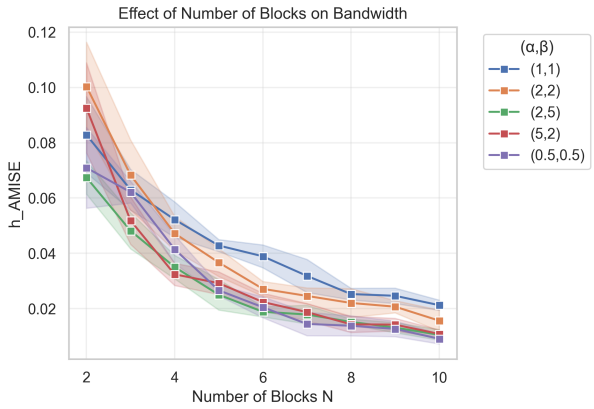
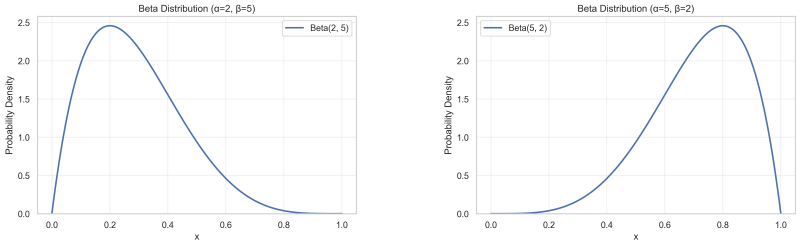


Figure 3: Impact of block numbers N on optimal bandwidth h_{AMISE}

Bandwidth h_{AMISE} with different sampling distributions Figure 5 displays the average optimal bandwidth h_{AMISE} (over 5 iterations) for different sample sizes n, where the block size N is selected by minimizing Mallow's C_p. An interesting pattern emerges for the Beta(2,5) and Beta(5,2) distributions, where the optimal bandwidth for one is consistently much smaller than the other. As the accompanying density plots illustrate (Figures 4a, 4b), these two distributions are **skewed** in opposite directions. Consequently, the covariate sampling they produce features regions with a **substantial difference in the concentration** of data points.



(a) $Beta(2, 5)$ (b) $Beta(5, 2)$

The optimal bandwidth h_{AMISE} is much lower for the right-skewed distribution (Beta(5,2)) than for the left-skewed one (Beta(2,5)). This phenomenon could be explained by the **geometry** of the true function $m(x)$ in Figure 1. The function exhibits higher curvature on the left side of its domain compared to the more linear behavior on the right. When the covariate sample is concentrated on the right (Beta(5,2)), the local fits of $\hat{m}$ primarily capture this simple, linear relationship, which is reflected in a smaller optimal bandwidth. This results in an estimator that is simpler but may be less accurate globally. Conversely, when samples are concentrated on the left (Beta(2,5)), the optimal bandwidth must be larger to accommodate the complex local curvature of the underlying function, leading to a final estimate that is more likely to be wiggly and overfit to the specific sample.

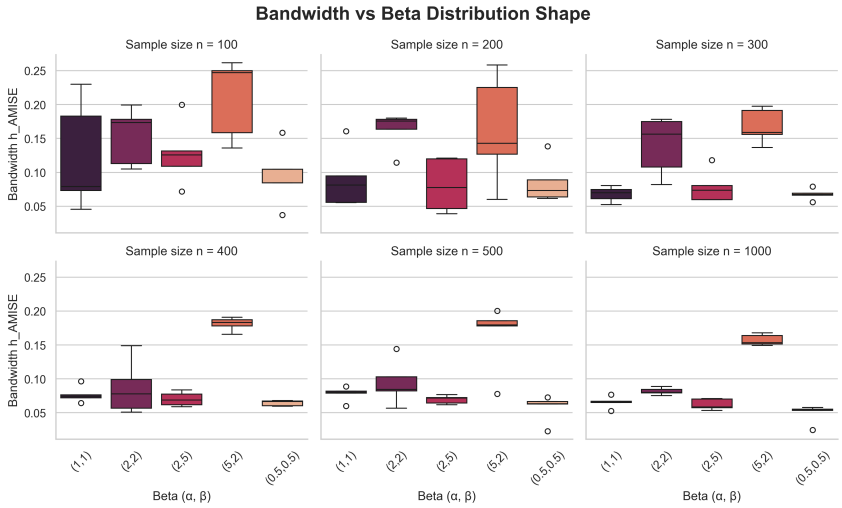


Figure 5: Impact of Beta distribution on optimal bandwidth h_{AMISE}

Should N depend on n? Why? Figure 6 illustrates the relationship between the optimal block count N and the sample size n. The results show a **slow linear relationship** between the two in most of the covariates sampling settings. This could signal that N should depend on n linearly. We can try to provide an interpretion to this result. With smaller samples, a lower N is optimal to ensure each block contains enough data points for a stable polynomial fit. As n increases, we can afford to use a larger N, which allows the local regressions to adapt more flexibly to regional variations in the data, thereby reducing bias in the block models. However, this relationship is **not** linear **indefinitely**. The complexity of the true function $m(x)$ imposes an upper bound; beyond a certain sample size, increasing N further provides risk of overfitting, leading to the observed stabilization. According to this reasoning, the block size N should be chosen in consideration of the sample size n to balance the bias-variance trade-off effectively.

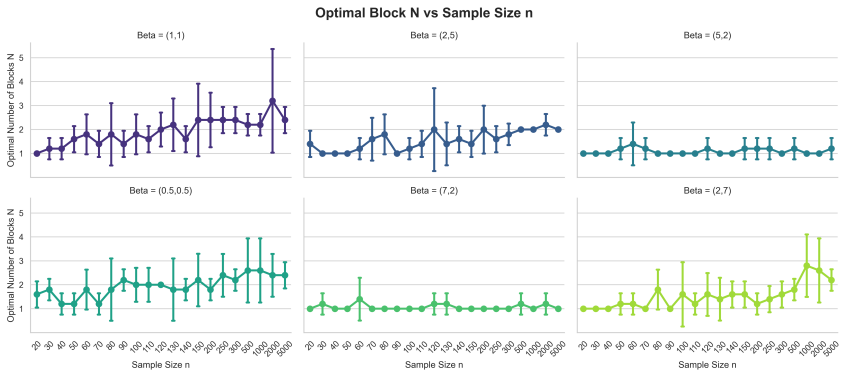


Figure 6: Relationship between optimal N and the sample size N

In the plots, the left-skewed Beta distributions demonstrates a **constant** optimal block number around $N^* = 1$ across different sample sizes. This result can be explained by the fact that this distribution primarily samples covariates from the right region of the function's support, where $m(x)$ exhibits simple, linear behavior. In this region, a single global degree 4 polynomial provides a good fit, resulting in a very low Residual Sum of Squares (RSS) without needing localized models. This case illustrates that the linear relationship between n and N^* observed for other distributions is not universal. The optimal block count is ultimately dictated by the interplay between sample size and the local complexity of the function within the data-dense regions of the covariate space.

Conclusion

This simulation study confirms that the optimal bandwidth h_{AMISE} is **sensitive** to the sample size n, the block parameter N, and the covariate distribution. The results validate the theoretical $n^{-1/5}$ scaling and demonstrate that the data distribution's shape directly influences bandwidth selection by highlighting regions of differing functional complexity. Furthermore, the block size N should scale with n to optimize the bias-variance trade-off, but this relationship is bounded by the inherent complexity of the underlying function.