MATH-517: Assignment 3

Noemi Ortona

2025-09-05

1. Theoretical exercise

1.1

In order to prove that the local linear regression estimator, $\hat{m}(x)$, belongs to the class of **linear smoothers**, we need to prove that the estimator can be written as a weighted average of the observations Y_i, where the weights $w_{ni}(x)$ depend on the predictor variables X_i, the target point x, the kernel function K, and the bandwidth h, but **not** on the response variables Y_i.

The Minimization Problem

Let's start analyzing the function that we have to minimize in order to find the coefficients of the local linear regression estimator, $\hat{\beta}_0(x)$ and $\hat{\beta}_1(x)$.

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1(X_i - x))^2 \, K\left(\frac{X_i - x}{h}\right)$$

- The term $(Y_i - \beta_0 - \beta_1(X_i - x))$ is the residual for the i-th observation with respect to the local line at point x.
- The term $K\left(\frac{X_i - x}{h}\right)$ is the weight assigned by the kernel function. This weight is large when X_i is "close" to x and small when it is "far".

Deriving the Normal Equations

To find the values of β_0 and β_1 that minimize L, we take the partial derivatives with respect to each parameter and set them to zero. For notational simplicity, let's define $k_i(x) = K\left(\frac{X_i - x}{h}\right)$.

$$\frac{\partial L}{\partial \beta_0} = \sum_{i=1}^{n} -2\left(Y_i - \beta_0 - \beta_1(X_i - x)\right) k_i(x) = 0$$

$$\frac{\partial L}{\partial \beta_1} = \sum_{i=1}^{n} -2(X_i - x)\left(Y_i - \beta_0 - \beta_1(X_i - x)\right) k_i(x) = 0$$

Solving the System by Substitution

To make the system more manageable, we use the following summary notations for the weighted sums:

- $S_0 = \sum_{i=1}^{n} k_i(x)$
- $S_1 = \sum_{i=1}^{n} (X_i - x) k_i(x)$
- $S_2 = \sum_{i=1}^{n} (X_i - x)^2 k_i(x)$
- $T_0 = \sum_{i=1}^{n} Y_i k_i(x)$
- $T_1 = \sum_{i=1}^{n} Y_i (X_i - x) k_i(x)$

The system becomes:

$$\begin{cases} \hat{\beta}_0 S_0 + \hat{\beta}_1 S_1 = T_0 \\ \hat{\beta}_0 S_1 + \hat{\beta}_1 S_2 = T_1 \end{cases}$$

$$\hat{\beta}_1 S_1 = T_0 - \hat{\beta}_0 S_0 \implies \hat{\beta}_1 = \frac{T_0 - \hat{\beta}_0 S_0}{S_1}$$

$$\hat{\beta}_0 S_1 + \left(\frac{T_0 - \hat{\beta}_0 S_0}{S_1}\right) S_2 = T_1$$

$$\hat{\beta}_0 S_1^2 + T_0 S_2 - \hat{\beta}_0 S_0 S_2 = T_1 S_1$$

$$\hat{\beta}_0 (S_1^2 - S_0 S_2) = T_1 S_1 - T_0 S_2$$

$$\hat{\beta}_0 = \frac{T_1 S_1 - T_0 S_2}{S_1^2 - S_0 S_2} = \frac{T_0 S_2 - T_1 S_1}{S_0 S_2 - S_1^2}$$

Expressing the Estimator as a Weighted Average

Now that we have solved for $\hat{\beta}_0$, we substitute the definitions of T_0 and T_1 back into the solution:

$$\hat{m}(x) = \hat{\beta}_0 = \frac{\left(\sum_{i=1}^n Y_i k_i(x)\right) S_2 - \left(\sum_{i=1}^n Y_i (X_i - x) k_i(x)\right) S_1}{S_0 S_2 - S_1^2}$$

We can rewrite the numerator by factoring out Y_i and $k_i(x)$:

$$\hat{m}(x) = \sum_{i=1}^n Y_i \underbrace{\left[\frac{k_i(x)\left(S_2 - (X_i - x)S_1\right)}{S_0 S_2 - S_1^2}\right]}_{w_{ni}(x)}$$

As required, these weights **depend only on the target point** x, the data points X_i, the kernel function K, and the bandwidth h. They do **not** depend on the response values Y_i. This completes the proof that local linear regression is a **linear smoother**.

1.2

From the first part of the exercise, we derived the weight expression using the sum notation S_k, here we defined $S_k = \sum_{j=1}^n (X_j - x)^k K\left(\frac{X_j - x}{h}\right)$.

$$w_{ni}(x) = \frac{K\left(\frac{X_i - x}{h}\right)\left(S_2 - (X_i - x)S_1\right)}{S_0 S_2 - S_1^2} \quad (*),$$

The new notation provided in the exercise is:

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

$$\implies \mathbf{S_k = nh \cdot S_{n,k}(x)}$$

Substitution into the Weight Expression

We will now substitute this relationship into our original weight equation $(*)$.

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

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

We can cancel the common factor nh from the numerator and denominator. This leaves us with the final, explicit expression for the weights:

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

1.3

We need to prove that $\sum_{i=1}^{n} w_{ni}(x) = 1$, that is a property of any weighted average.

The denominator, $S_0S_2 - S_1^2$, is a constant with respect to the summation index i, so we can factor it out:

$$\sum_{i=1}^{n} w_{ni}(x) = \frac{1}{S_0S_2 - S_1^2}\sum_{i=1}^{n}\left[K\left(\frac{X_i - x}{h}\right)(S_2 - (X_i - x)S_1)\right]$$

Let's expand the sum in the numerator by distributing the kernel term:

$$\sum_{i=1}^{n}\left[S_2 \cdot K\left(\frac{X_i - x}{h}\right) - S_1 \cdot (X_i - x)K\left(\frac{X_i - x}{h}\right)\right]$$

We can split this into two separate sums:

$$= \sum_{i=1}^{n} S_2 \cdot K\left(\frac{X_i - x}{h}\right) - \sum_{i=1}^{n} S_1 \cdot (X_i - x)K\left(\frac{X_i - x}{h}\right)$$

Since S_2 and S_1 are also constants with respect to the index i, we can pull them out of their respective sums:

$$= S_2\left(\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\right) - S_1\left(\sum_{i=1}^{n}(X_i - x)K\left(\frac{X_i - x}{h}\right)\right)$$

Now, we can recognize the sums in the parentheses. By their very definition:

- $\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) = S_0$
- $\sum_{i=1}^{n} (X_i - x) K\left(\frac{X_i - x}{h}\right) = S_1$

Substituting these back into our expression, the numerator becomes:

$$S_2 \cdot S_0 - S_1 \cdot S_1 = S_0 S_2 - S_1^2$$

We now place the simplified numerator back over the original denominator:

$$\sum_{i=1}^{n} w_{ni}(x) = \frac{S_0 S_2 - S_1^2}{S_0 S_2 - S_1^2} = 1$$

2. Practical exercise

The goal is to perform a simulation study to assess the impact of some parameters/hyperparameters on the optimal bandwidth h_AMISE.

2.1 How does h_AMISE behave when N grows? Can you explain why?

To answer this question, we must understand how the block size N influences the estimation of the unknown quantities σ^2 and θ_{22}. The parameter N controls the complexity of the model used to estimate these quantities.

A larger value of N means the data is split into more, smaller blocks. Within each small block, the fitted quartic polynomial will adapt more closely to the local data points, resulting in a more flexible overall model. This increased flexibility leads to larger values for the estimated second derivatives, $\hat{m}_j''(x)$.

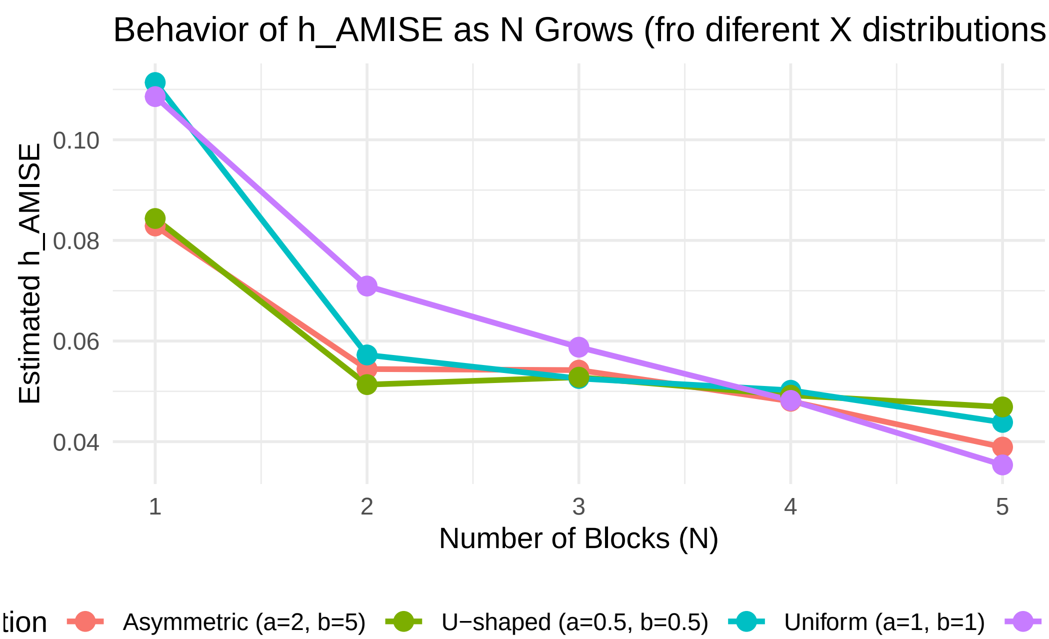
Since $\hat{\theta}_{22}$ is calculated as the average of the squared second derivatives, a more wiggly model (larger N) will produce a larger $\hat{\theta}_{22}$.

Looking at the formula for h_{AMISE}:

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

we can see that θ_{22} is in the denominator. Therefore, as N increases, $\hat{\theta}_{22}$ increases, and consequently the estimated h_{AMISE} decreases.

To verify this, we run a single simulation with a large sample size (n=2000) and calculate the estimated h_{AMISE} for each possible value of N from 1 to N_{max}.



Findings: The plot above confirms our theoretical expectation. The estimated value of h_{AMISE} is a decreasing function of the number of blocks, N, used for the pilot estimation.

Depending on how the the number of observations varies between different regions h_{AMISE} decrease differently, but decreases anyway.

Should N depend on n? Why?

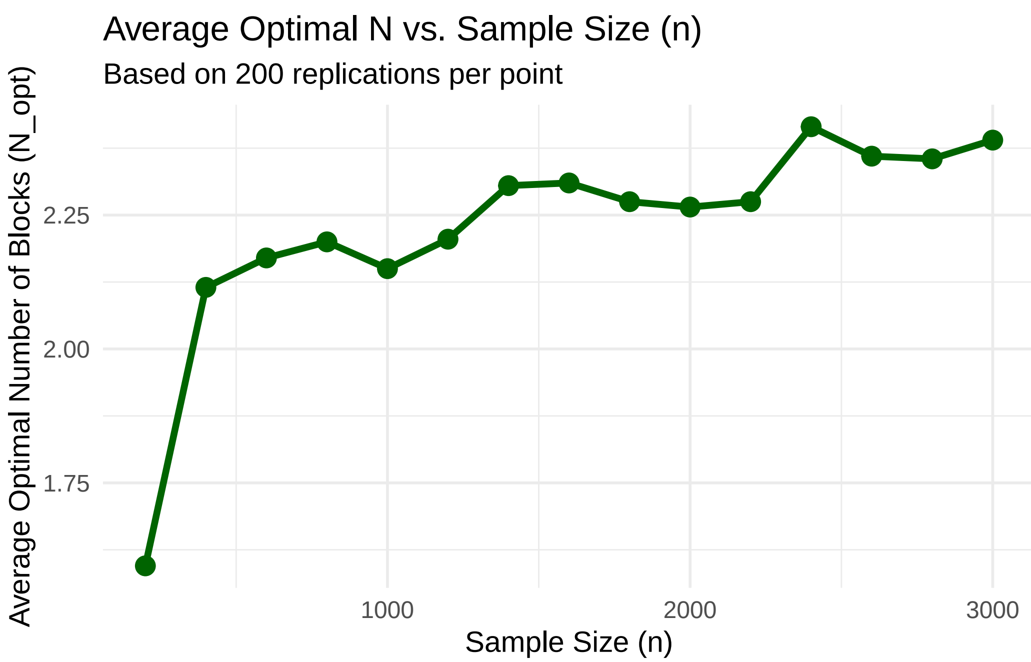
Yes, the optimal choice of N **must depend on** n. This is a **bias-variance trade-off**.

- **Small** n: With a small sample size, using a large N would result in very few data points per block. This would make the polynomial fit in each block highly unstable and variable. To control this high variance, a smaller N is preferred.
- **Large** n: With a large sample size, we can afford to use a larger N. Each block will still contain enough data for a stable fit. A larger N allows the pilot model to be more flexible and capture the local features of the true regression function $m(x)$ more accurately, leading to a less biased estimate of θ_{22}.

To demonstrate this relationship empirically, we conduct a comprehensive simulation study:

- **Define a Range of Sample Sizes:** We test a series of datasets with increasing sample sizes (n). Specifically, we start with $n = 200$ and increase the sample size in steps of 200 until we reach $n = 3000$.

- **Perform Multiple Replications:** For each single sample size n, we generate **200 different, independent datasets**. This step is crucial because the optimal N found for any single dataset can be influenced by the specific random sample of data points. By running 200 trials for the same n, we can average out this randomness.
- **Find the Optimal N:** In each of these trials, we apply the `find_optimal_N` function to determine the N_{opt} that minimizes Mallows's C_p for that specific dataset.
- **Average the Results:** After completing the 200 trials for a given n, we calculate the **average of the 200 N_{opt} values** found (the average is a relaiable estimate of the best number of blocks for that sample size)
- **Visualize the Trend:** We plot these averaged values against their corresponding sample sizes



Findings: The plot clearly shows an increasing relationship. As the sample size n grows, the optimal number of blocks N chosen by Mallows's C_p also tends to increase.

From the graph, we notice that the optimal number of blocks grows quickly at first; this is because for smaller sample sizes, increasing N from one to two provides a large reduction in bias that strongly outweighs the Mallows's $C_p$penalty for the added model complexity

What happens when the number of observations varies a lot between different regions in the support of X? How is this linked to the parameters of the Beta distribution?

The number of observations in different regions is determined by the probability density function $f_X(x)$ of the covariate.

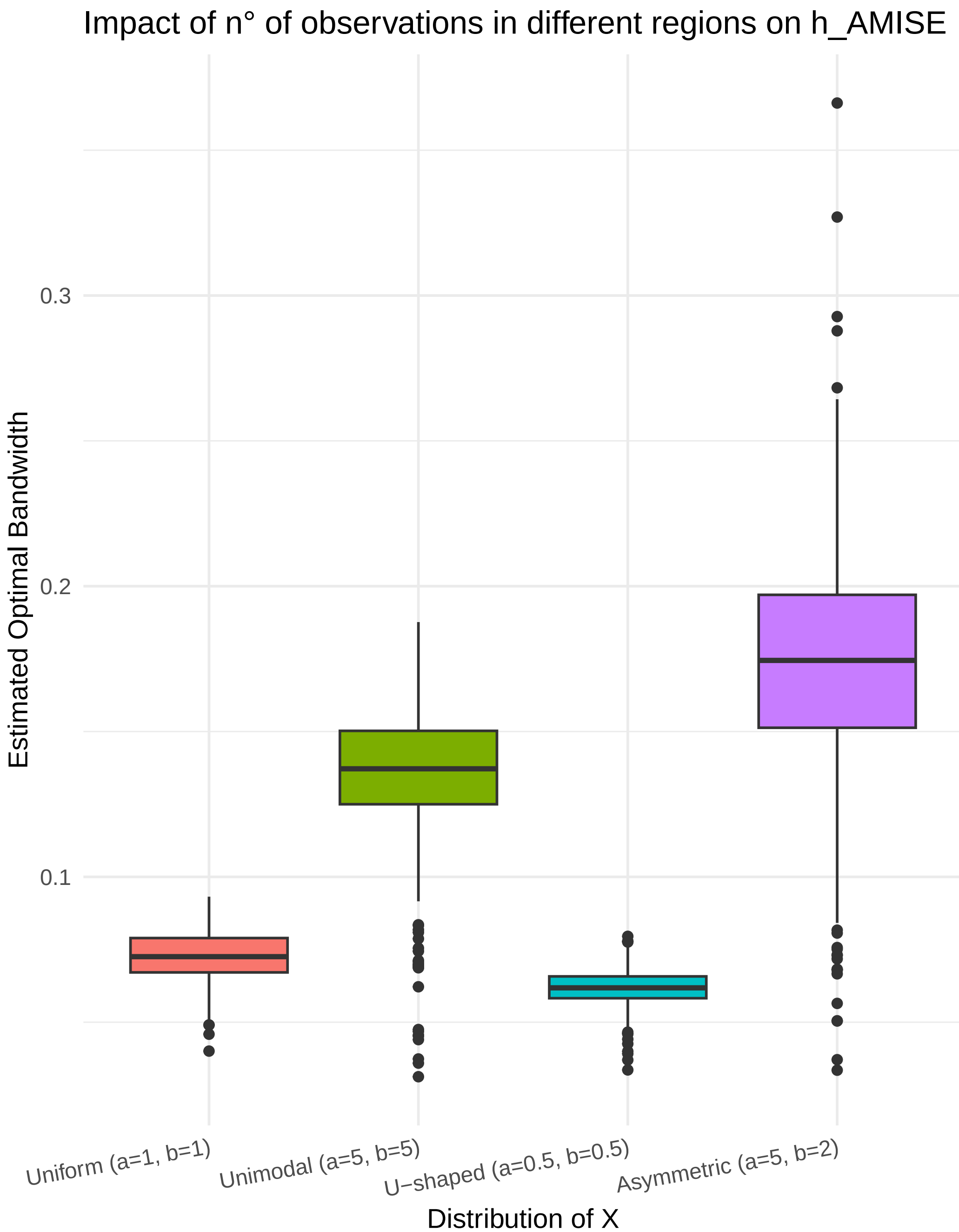
This density has a significant and direct impact on θ_{22} , that is inversely proportional to global optimal bandwidth h_{AMISE}

$$\theta_{22} = \int m''(x)^2 f_X(x)dx$$

This is not a simple average of the squared curvature, but a weighted average, where the weights are the density values $f_X(x)$ themselves. This means that regions where data is dense (high $f_X(x)$) contribute far more to the final value θ_{22} than regions where data is sparse.

The parameters α and β of the Beta distribution are known as shape parameters; their values and their relationship to each other entirely determine the shape of the distribution's density curve.

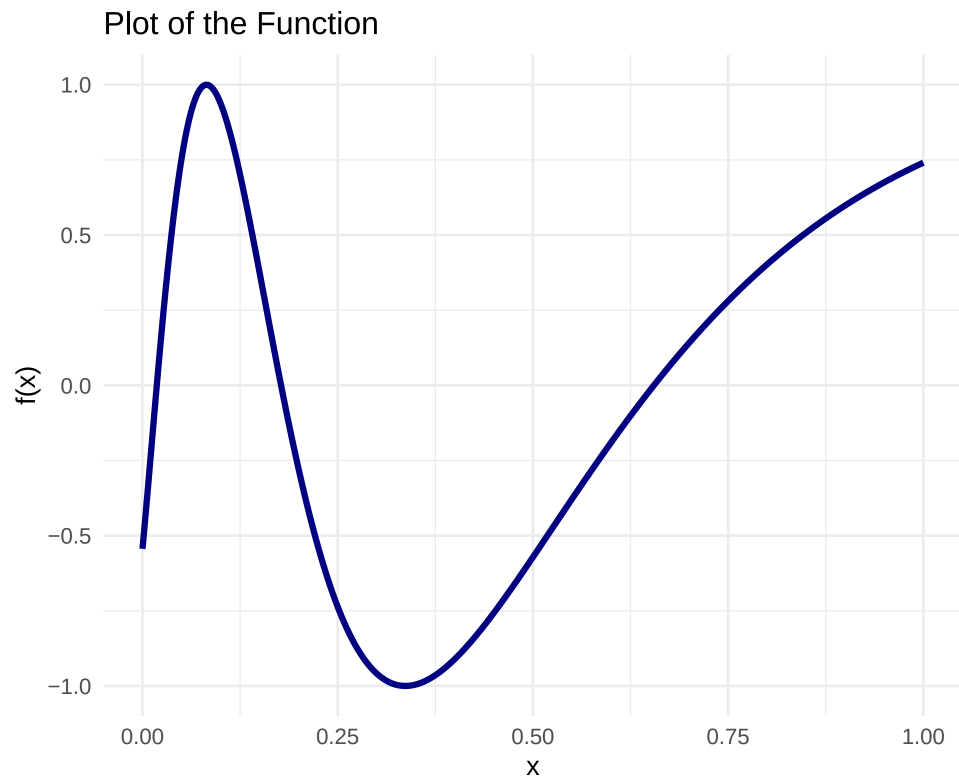
To analise this, for each shape (e.g., Uniform, Unimodal), we perform 200 replications; in each replication, a new random sample of 500 points is generated, and its corresponding optimal h_{AMISE} is calculated. This process yields a collection of 200 h_{AMISE} values for each distribution type. We then use a boxplot to visualize these collections, to compare the median bandwidth and the overall variability of the estimates for each scenario.

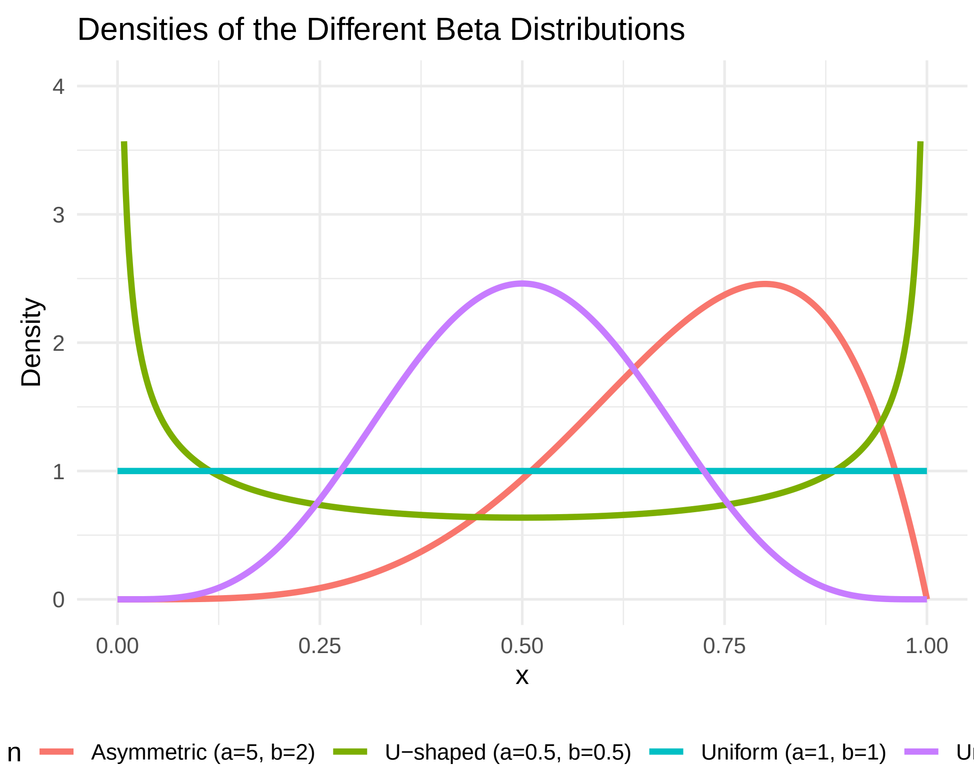


From the box-plots we can observe that:

- The **Uniform** (`a=1, b=1`) and **U-shaped** (`a=0.5, b=0.5`) distributions have very small median values for h_{AMISE} . This suggests that in the regions where these distributions have high data density (across the whole support for Uniform, and at the edges for U-shaped), the function's curvature $m''(x)$ is relatively high. The algorithm detects this high "wiggliness" in the data-rich regions, leading to a large θ_{22} and thus a small h_{AMISE} to capture these details.

- The **Unimodal** (`a=5, b=5`) distribution, which concentrates data in the center of the support, have a larger median bandwidth. This implies that in the central region (around x=0.5), the true function $m(x)$ is relatively smooth (has low curvature). Because most of the data is in this smooth region, the weighted average θ_{22} is smaller, leading the algorithm to select a larger bandwidth.
- The **Asymmetric** (`a=5, b=2`) distribution produces the highest median bandwidth and the **highest variance**. This distribution concentrates data on the right side of the support. The large bandwidth suggests that the function is smoothest in this data-rich region. The high variance indicates that the estimation of h_{AMISE} is unstable under this condition. This happens because the algorithm gets conflicting information: it sees a smooth function where data is plentiful but must also account for the sparse, potentially more complex regions, leading to inconsistent estimates across different random samples.


Plot of the Function
f(x)
1.0
0.5
0.0
−0.5
−1.0
0.00
0.25
0.50
0.75
1.00
x


Densities of the Different Beta Distributions
Density
4
3
2
1
0
0.00
0.25
0.50
0.75
1.00
x
n
Asymmetric (a=5, b=2)
U−shaped (a=0.5, b=0.5)
Uniform (a=1, b=1)