MATH-517: Assignment 3

Luca Civila

2024-05-09

Theoretical exercise

Overview

This report reconstructs the derivation of the **local linear kernel regression** estimator. We show that minimizing a kernel–weighted quadratic loss yields $\hat{\beta}(x) = (\hat{\beta}_0(x), \hat{\beta}_1(x))$, and that the fitted value at x can be written as a kernel smoother $\widehat{m}(x) = \hat{\beta}_0(x) = \sum_{i=1}^{n} w_{in}(x) y_i$ with explicit weights $w_{in}(x)$. In the last part, once we found our weights, we demonstrate that they sum up to 1

Problem setup

Let $(X_i, Y_i)_{i=1}^n$ be observations.
Fix a target point $x \in \mathbb{R}$, a bandwidth $h > 0$, and a kernel function $K(\cdot)$. Define

$$K_i \equiv K\left(\frac{X_i - x}{h}\right), \qquad u_i \equiv X_i - x$$

The local linear regression estimator at a point x is defined by:

$$(\hat{\beta}_0(x), \hat{\beta}_1(x)) = \arg\min_{\beta_0, \beta_1 \in \mathbb{R}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 u_i)^2 K_i$$

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

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

Let's define the the weight matrix W, the design matrix X and the vector Y:

$$X = \begin{bmatrix} 1 & u_1 \\ \vdots & \vdots \\ 1 & u_n \end{bmatrix} \in \mathbb{R}^{n\times 2}, \qquad W = \text{diag}(K_1, ..., K_n) \in \mathbb{R}^{n\times n}, \qquad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

We know that the solution of a weighted least squares (WLS) solution is:

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

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

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

Being a scalar, our $\hat{m}(x)$ can be seen as the result of an inner product (according to the dimension of the our terms) between two vectors: $e_1^\top (X^\top W X)^{-1} X^\top W$ and Y. This leads to the fact that our weights $\hat{m}(x)$ can be rewrite as:

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

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

and $e_i = (0, ..., 0, 1, 0, ..., 0)^\top$ standard basis vector to isolate the i-th element of the weight vector. As we can see, the weights w_{in} do not depend on Y but only on $\{X_i\}$, K_i, and x.

Explicit Expression for Weights

To rewrite $\hat{m}(x)$ in a simpler expression, we need to compute explicit solution of the WLS and rewrite it in the S notation.

Thus, let's compute $X^\top W X$:

$$X^\top W X = \begin{bmatrix} \sum_{i=1}^n K_i & \sum_{i=1}^n u_i K_i \\ \sum_{i=1}^n u_i K_i & \sum_{i=1}^n u_i^2 K_i \end{bmatrix}$$

Using the notation:

$$S_{n,k}(x) = \frac{1}{nh} \sum_{i=1}^n u_i^k K_i$$

We recognize:

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

In our case the matrix is 2×2 so the inverse is quite simple to compute:

$$(X^\top W X)^{-1} = \frac{1}{nh[S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)]} \begin{bmatrix} S_{n,2}(x) & -S_{n,1}(x) \\ -S_{n,1}(x) & S_{n,0}(x) \end{bmatrix} \in \mathbb{R}^{2\times 2};$$

and

$$X^\top W = \begin{bmatrix} K_1 & \dots & K_n \\ K_1 u_1 & \dots & K_n u_n \end{bmatrix} \in \mathbb{R}^{2\times n}.$$

Now we have all the ingredients to rewrite Equation 1:

$$w_{in}(x) = \frac{1}{\Delta}\{S_{n,2}(x) - S_{n,1}(x)u_i\}K_i, \qquad \text{where} \quad \Delta = \frac{1}{nh[S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)]}$$

Now by expliciting all the quantities, we obtain:

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

Proof That Weights Sum to 1

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

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

Substituting the expressions for $S_{n,k}(x)$:

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

Substituting these into the equation for the sum of weights:

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

$$= \frac{nh[S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)]}{nh[S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)]}$$

$$= 1$$

This completes the proof that the weights sum to 1.

Practical Exercise

Aim of the Study

This simulation study aims to investigate the behavior of the optimal bandwidth h_{AMISE} in local polynomial regression under varying conditions, with particular focus on understanding how different factors influence bandwidth selection. The primary objective is to analyze the relationship between three key elements: sample size n, block partitioning strategy N, and covariate distribution characteristics (governed by $Beta$ distribution parameters α and β).

The study employs a challenging regression function $m(x) = sin\left(\left(\frac{x}{3} + 0.1\right)^{-1}\right)$ that exhibits markedly different behavior across its domain, displaying rapid oscillations in the left region and smooth characteristics in the right region. This design allows us to examine how bandwidth selection adapts to local function complexity.

Specifically, the research addresses three fundamental questions:

How does the optimal bandwidth h_{AMISE} respond to increasing block partitioning N?

What is the relationship between sample size and optimal block partitioning?

How does the distribution of covariates (especially through asymmetric Beta distributions) affect bandwidth selection?

By systematically varying these parameters and employing Mallow's Cp for optimal block size selection, we discuss how each factor influences the optimal bandwidth.

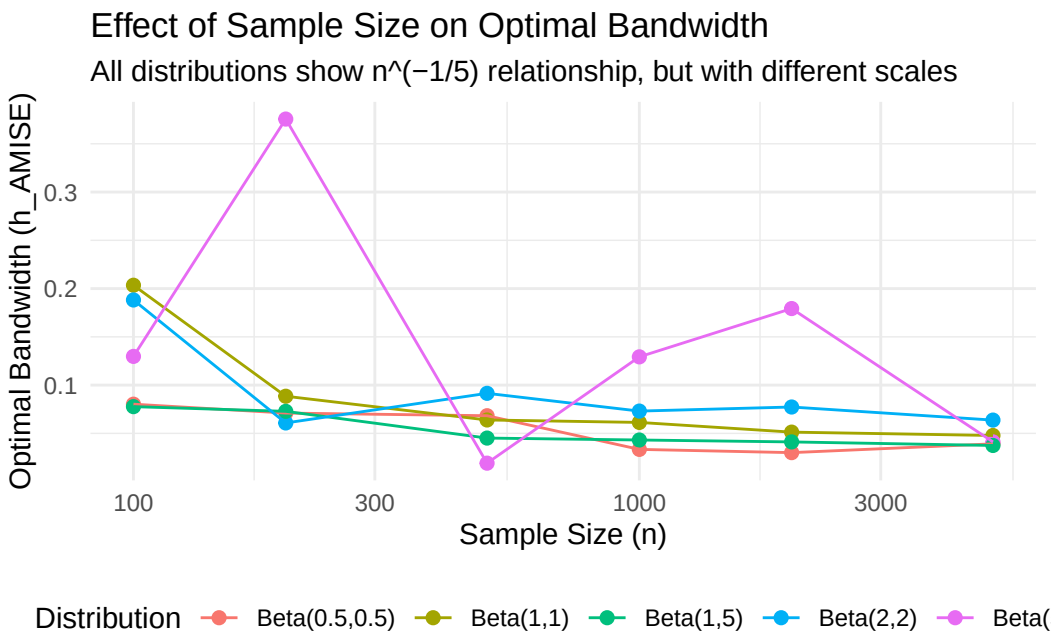
Simulation Design and Quantities

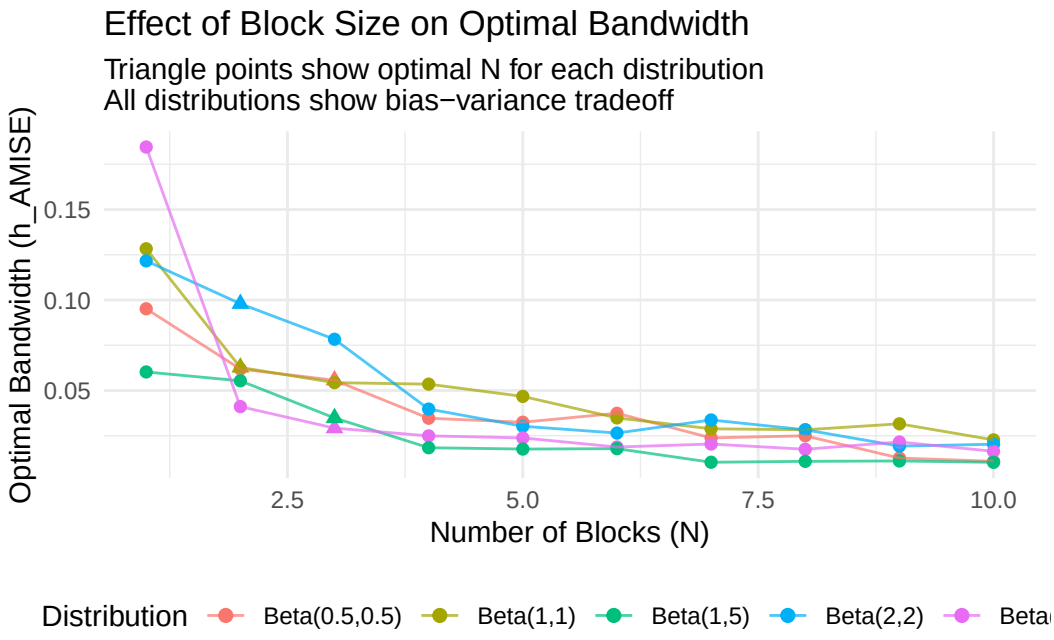
The study employs the following key quantities:

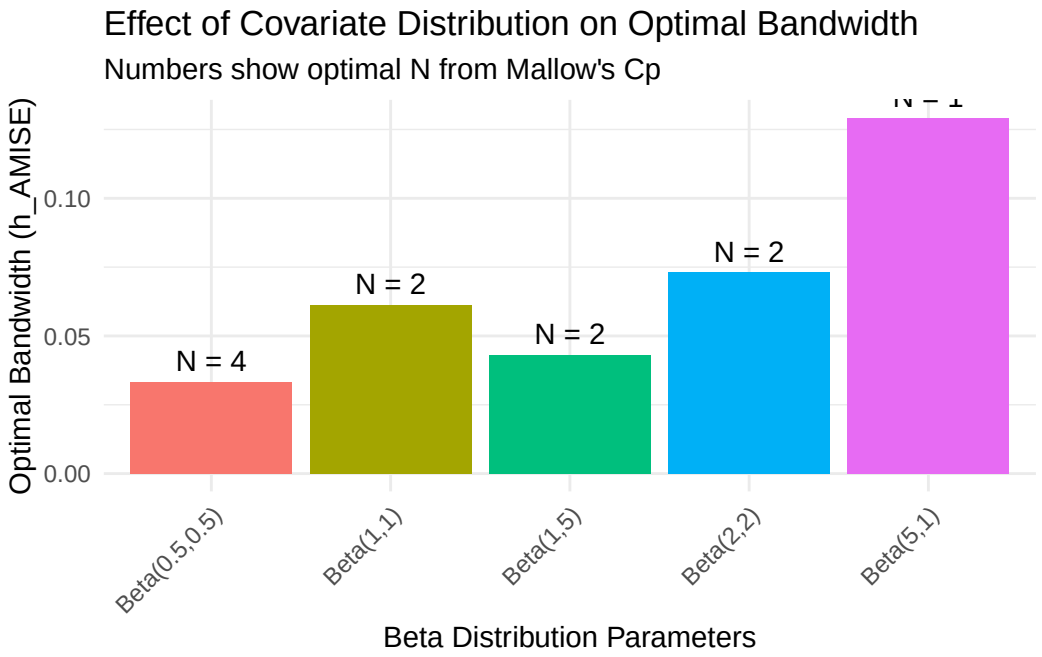
- **Fixed parameters**:
  - True regression function: $m(x) = sin\left(\left(\frac{x}{3} + 0.1\right)^{-1}\right)$
  - Error variance: $\sigma^2 = 1$
  - Kernel: Quartic (biweight) kernel
  - Covariate support: $[0, 1]$ (Beta distribution)
- **Varying parameters**:
  - Sample size $n \in \{100, 200, 500, 1000, 2000, 5000\}$
  - Block size $N \in \{1, \ldots, 10\}$ for $n = 1000$
  - Beta parameters $(\alpha, \beta) \in \{(0.5, 0.5), (1, 1), (2, 2), (5, 1), (1, 5)\}$
- **Estimated quantities**:
  - $\hat{\theta}_{22}$: Estimated squared second derivative term
  - $\hat{\sigma}^2$: Estimated error variance
  - h_{AMISE}: Optimal bandwidth from asymptotic theory

The bandwidth formula is:

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


Effect of Sample Size on Optimal Bandwidth
All distributions show n^(−1/5) relationship, but with different scales
Optimal Bandwidth (h_AMISE)
0.3
0.2
0.1
100
300
1000
3000
Sample Size (n)
Distribution
Beta(0.5,0.5)
Beta(1,1)
Beta(1,5)
Beta(2,2)
Beta(


Effect of Block Size on Optimal Bandwidth
Triangle points show optimal N for each distribution
All distributions show bias−variance tradeoff
Optimal Bandwidth (h_AMISE)
0.15
0.10
0.05
2.5
5.0
7.5
10.0
Number of Blocks (N)
Distribution
Beta(0.5,0.5)
Beta(1,1)
Beta(1,5)
Beta(2,2)
Beta(



Results and Discussion

Relationship Between Sample Size and Block Size: The fundamental relationship between sample size n and optimal block size N stems from the bias-variance tradeoff in local polynomial estimation. With larger sample sizes, we can partition the data into more blocks without risking overfitting because each block maintains sufficient data points for stable parameter estimation. The Mallow's Cp criterion effectively captures this relationship by penalizing both overly simplistic models (too few blocks, leading to high bias) and overly complex models (too many blocks, leading to high variance). As n increases, the optimal N typically grows because we can afford more localized fitting while maintaining estimation stability within each block.

Behavior of h_{AMISE} **with Respect to Sample Size**: The plot of h_{AMISE} versus sample size n reveals the expected theoretical relationship across all distributions: optimal bandwidth decreases as sample size increases, following the characteristic $n^{-\frac{1}{5}}$ rate. However, the absolute values and rate of decrease vary significantly by distribution type.

Impact of Distribution Characteristic on Bandwidth Selection: For right-skewed distributions like Beta(5,1), where data concentrates in smooth regions, we observe relatively larger bandwidths that decrease gradually with n. This pattern occurs because the smooth nature of the real $m(x)$ in high-density regions permits substantial smoothing even with limited data.

Conversely, for left-skewed distributions like Beta(1,5), where data clusters in the oscillatory left region, we see smaller initial bandwidths with steeper decreases as n grows. The complex local structure in high-density areas demands fine resolution, and additional samples significantly improve our ability to estimate this curvature.

The uniform Beta(1,1) and bell-shaped Beta(2,2) distributions exhibit intermediate behavior, while the U-shaped Beta(0.5,0.5) presents the most challenging case, requiring careful bandwidth selection to balance boundary effects at both extremes.

Conclusion

The simulation confirms theoretical predictions: optimal bandwidth decreases with sample size, depends critically on proper block size selection, and must adapt to covariate distribution characteristics.