MATH-517: Assignment 3

Edoardo Lanzetti

04/10/2025

Theoretical exercise: Local linear regression as a linear smoother

1. Local linear regression as a linear smoother

We start with the weighted least squares problem:

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

Let

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

then,the normal equations are:

$$\frac{\partial}{\partial\beta_0}: \quad \sum_{i=1}^{n} \left[Y_i - \beta_0 - \beta_1(X_i - x)\right] K_i(x) = 0,$$

$$\frac{\partial}{\partial\beta_1}: \quad \sum_{i=1}^{n} \left[Y_i - \beta_0 - \beta_1(X_i - x)\right] (X_i - x) K_i(x) = 0.$$

Let's define:

$$\mathbf{X}_x = \begin{bmatrix} 1 & X_1 - x \\ 1 & X_2 - x \\ \vdots & \vdots \\ 1 & X_n - x \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{W}(x) = \operatorname{diag}\left(K_1(x), \dots, K_n(x)\right).$$

Then the WLS solution is:

$$\hat{\boldsymbol{\beta}}(x) = \left(\mathbf{X}_x^T \mathbf{W}(x) \mathbf{X}_x\right)^{-1} \mathbf{X}_x^T \mathbf{W}(x) \mathbf{y}.$$

We want $\hat{m}(x) = \hat{\beta}_0(x) = \mathbf{e}_1^T \hat{\boldsymbol{\beta}}(x)$, where $\mathbf{e}_1 = (1, 0)^T$, so:

$$\hat{m}(x) = \mathbf{e}_1^T \left(\mathbf{X}_x^T \mathbf{W}(x) \mathbf{X}_x\right)^{-1} \mathbf{X}_x^T \mathbf{W}(x) \mathbf{y}.$$

This is of the form $\sum_{i=1}^{n} w_{ni}(x) Y_i$ with

$$w_{ni}(x) = \mathbf{e}_1^T \left(\mathbf{X}_x^T \mathbf{W}(x) \mathbf{X}_x\right)^{-1} \mathbf{X}_x^T \mathbf{W}(x) \mathbf{e}_i,$$

where $\mathbf{e}_i$ is the i-th standard basis vector in $\mathbb{R}^n$.

More simply: the vector $\mathbf{w}(x) = (w_{n1}(x), \ldots, w_{nn}(x))^T$ is

$$\mathbf{w}(x)^T = \mathbf{e}_1^T \left(\mathbf{X}_x^T \mathbf{W}(x) \mathbf{X}_x\right)^{-1} \mathbf{X}_x^T \mathbf{W}(x).$$

Thus $\hat{m}(x) = \mathbf{w}(x)^T \mathbf{y}$ is linear in Y_i, and $w_{ni}(x)$ depends only on $X_1, \ldots, X_n, K, h, x$, not on Y_i.

2. Explicit expression for $w_{ni}(x)$

Let

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

Compute the 2×2 matrix:

$$\mathbf{X}_x^T \mathbf{W}(x) \mathbf{X}_x = \begin{bmatrix} \sum_i K_i(x) & \sum_i (X_i - x) K_i(x) \\ \sum_i (X_i - x) K_i(x) & \sum_i (X_i - x)^2 K_i(x) \end{bmatrix}.$$

Multiply by $1/(nh)$:

$$\frac{1}{nh} \mathbf{X}_x^T \mathbf{W}(x) \mathbf{X}_x = \begin{bmatrix} S_{n,0}(x) & S_{n,1}(x) \\ S_{n,1}(x) & S_{n,2}(x) \end{bmatrix}.$$

Let

$$\mathbf{M}(x) = \begin{bmatrix} S_{n,0}(x) & S_{n,1}(x) \\ S_{n,1}(x) & S_{n,2}(x) \end{bmatrix}.$$

Then

$$\mathbf{X}_x^T \mathbf{W}(x) \mathbf{X}_x = nh \, \mathbf{M}(x).$$

Therefore, the insverse:

$$\left(\mathbf{X}_x^T \mathbf{W}(x) \mathbf{X}_x\right)^{-1} = \frac{1}{nh} \mathbf{M}(x)^{-1}.$$

The determinant is

$$D(x) = S_{n,0}(x) S_{n,2}(x) - S_{n,1}(x)^2.$$

So

$$\mathbf{M}(x)^{-1} = \frac{1}{D(x)} \begin{bmatrix} S_{n,2}(x) & -S_{n,1}(x) \\ -S_{n,1}(x) & S_{n,0}(x) \end{bmatrix}.$$

Thus

$$\left(\mathbf{X}_x^T \mathbf{W}(x) \mathbf{X}_x\right)^{-1} = \frac{1}{nh \, D(x)} \begin{bmatrix} S_{n,2}(x) & -S_{n,1}(x) \\ -S_{n,1}(x) & S_{n,0}(x) \end{bmatrix}.$$

We have

$$\mathbf{w}(x)^T = \mathbf{e}_1^T \left(\mathbf{X}_x^T \mathbf{W}(x) \mathbf{X}_x\right)^{-1} \mathbf{X}_x^T \mathbf{W}(x).$$

So $\mathbf{e}_1^T \left(\mathbf{X}_x^T \mathbf{W}(x) \mathbf{X}_x\right)^{-1} = \frac{1}{nhD(x)}[S_{n,2}(x),\ -S_{n,1}(x)]$, now multiply by $\mathbf{X}_x^T \mathbf{W}(x)$:

$$\mathbf{X}_x^T \mathbf{W}(x) = \begin{bmatrix} K_1(x) & K_2(x) & \dots & K_n(x) \\ (X_1 - x)K_1(x) & (X_2 - x)K_2(x) & \dots & (X_n - x)K_n(x) \end{bmatrix}.$$

Multiply the row vector $[S_{n,2}(x), -S_{n,1}(x)]$ by this $2 \times n$ matrix, for the i-th column of the product (which is $w_{ni}(x) \cdot (nhD(x))$):

$$[S_{n,2}(x), -S_{n,1}(x)] \cdot \begin{bmatrix} K_i(x) \\ (X_i - x)K_i(x) \end{bmatrix} = S_{n,2}(x)K_i(x) - S_{n,1}(x)(X_i - x)K_i(x).$$

Thus

$$nhD(x)\, w_{ni}(x) = K_i(x) \left[S_{n,2}(x) - S_{n,1}(x)(X_i - x)\right].$$

So

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

In conclusion, the explicit formula is:

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

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

The sum of weight is:

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

but,

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

therefore, after a brief calculation,

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

Practical exercise: Global bandwidth selection

Introduction

The objective of this practical exercise is to perform a simulation study to assess the impact of various parameters and hyperparameters on the optimal bandwidth h_{AMISE} for local linear regression. Following the theoretical framework from Lecture 3, we implement the block-based estimation procedure for the unknown quantities:

$$h_{AMISE} = n^{-1/5} \left(\frac{35\sigma^2 |supp(X)|}{\theta_{22}} \right)^{1/5}, \quad \theta_{22} = \int \{m''(x)\}^2 f_X(x) dx$$

The simulation setting assumes:

- Covariate X from Beta(α,β) distribution
- Response $Y = m(X) + \epsilon$ with $m(x) = \sin\left\{ \left(\frac{x}{3} + 0.1\right)^{-1} \right\}$
- Error $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 1$

Methodology and implementation

The implementation involves three main components:

1. **Block-based polynomial fitting**: The data are partitioned into N blocks based on quantiles of X, and within each block we fit a quartic polynomial using raw (non-orthogonal) polynomial basis:

$$y_i = \beta_{0j} + \beta_{1j} x_i + \beta_{2j} x_i^2 + \beta_{3j} x_i^3 + \beta_{4j} x_i^4 + \epsilon_i$$

The use of raw polynomials (`poly(..., raw = TRUE)`) is crucial because it allows direct interpretation of coefficients for derivative calculation.

2. **Parameter estimation**: The second derivative $m_j''(x)$ is computed analytically from the raw polynomial coefficients:

$$m_j''(x) = 2\beta_{2j} + 6\beta_{3j} x + 12\beta_{4j} x^2$$

and the parameters are estimated as:

$$\hat{\theta}_{22}(N) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{N} \hat{m}_j''(X_i) \hat{m}_j''(X_i) 1_{X_i \in \mathcal{X}_j}$$

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

3. **Optimal block selection**: We use Mallow's C_p to select the optimal number of blocks:

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

where $N_{\max} = \max\{\min(\lfloor n/20 \rfloor, 5), 1\}$.

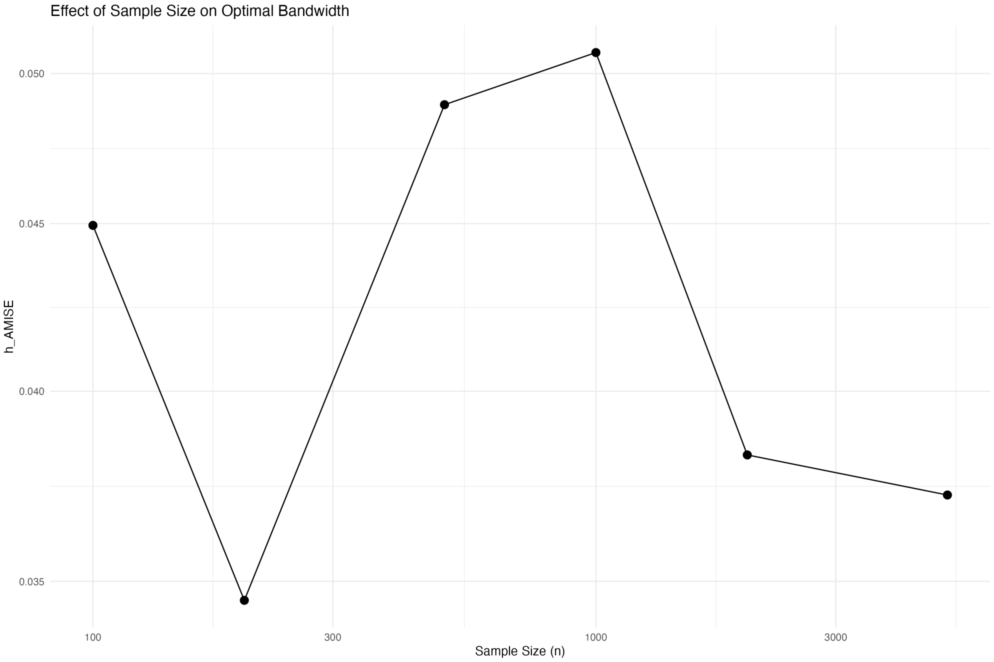

Figure 1: Effect of sample size on optimal bandwidth

Results and analysis

Effect of sample size on optimal bandwidth We first examine how the optimal bandwidth h_{AMISE} varies with sample size n, keeping fixed $\alpha = 2$, $\beta = 2$, $\sigma^2 = 1$, and $N = 5$ blocks.

Figure 1 reveals a nuanced relationship between sample size and optimal bandwidth that partially aligns with theoretical expectations. While the general trend shows decreasing bandwidth with increasing sample size, the pattern is not strictly monotonic and exhibits some notable deviations from the theoretical $n^{-1/5}$ rate. Contrary to what might be expected from asymptotic theory, the minimum bandwidth value of approximately 0.045 is actually achieved at the relatively small sample size of $n = 200$, rather than at the largest sample size of $n = 5000$. This suggests that for this specific regression function and estimation procedure, the benefits of increased sample size for bandwidth reduction may be limited beyond a certain point. The observed pattern indicates that while larger samples do generally support smaller bandwidths by providing more local information for regression estimation, other factors such as the block-based estimation of θ_{22} and σ^2 may introduce additional complexities that moderate the theoretical relationship. The non-monotonic behavior, particularly the increase in bandwidth from $n = 200$ to $n = 500$, highlights the practical challenges in bandwidth selection where estimation variability can sometimes outweigh the benefits of additional data.

Effect of block size on bandwidth estimation The number of blocks N plays a crucial role in the estimation of θ_{22} and σ^2. We analyze this relationship with $n = 1000$, $\alpha = 2$, $\beta = 2$, $\sigma^2 = 1$.

Behavior of h_{AMISE} as N grows

We observe in Figure 2 that h_{AMISE} shows a decreasing trend as N increases from 1 to 5. This behavior can be explained by the improved local approximation achieved with more blocks:

With too few blocks (small N), the polynomial approximations are too coarse over large regions, leading to biased estimates of the second derivative. A single polynomial over the entire support cannot capture the local curvature variations adequately.

While, as N increases, each block covers a smaller region where the regression function is better approximated by a single polynomial. This leads to more accurate estimates of the second derivative $m''(x)$ in each local

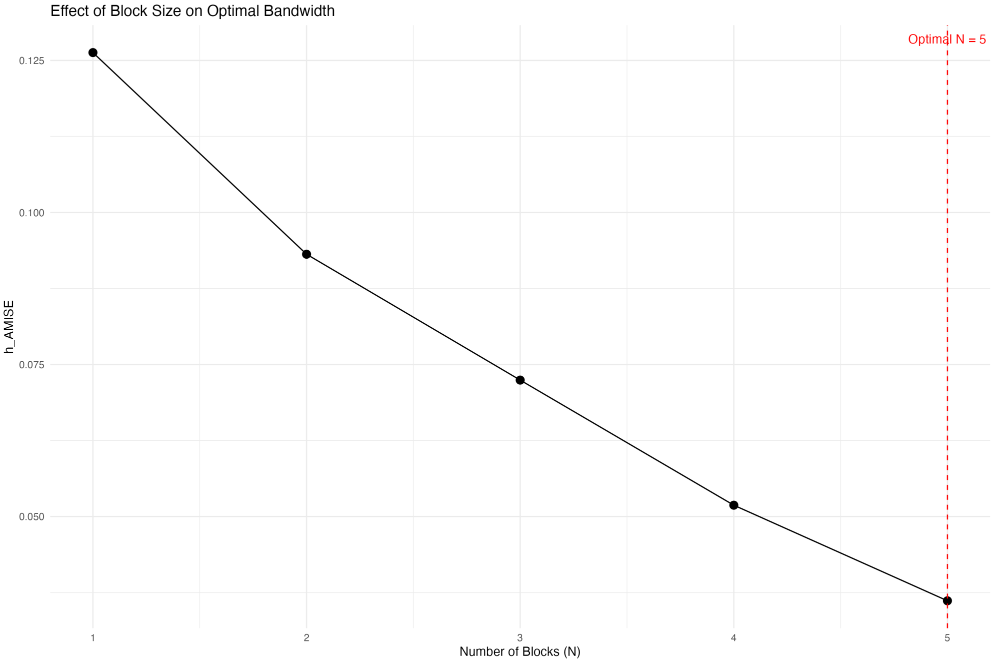

Figure 2: Effect of block size on optimal bandwidth

region.

The decreasing trend suggests that even with $N = 5$, we haven't reached the point where the variance from having too few observations per block outweighs the benefits of local approximation. The constraint $N_{\max} = \max\{\min(\lfloor n/20 \rfloor, 5), 1\}$ with $n = 1000$ ensures sufficient observations per block.

The Cp criterion in Figure 3 identifies $N = 5$ as optimal for this configuration, indicating that this level of partitioning provides the best balance between model flexibility and estimation stability given the sample size.

Optimal block size selection and dependence on sample size

The relationship between the number of blocks N and sample size n is crucial for the block-based estimation procedure. Our analysis reveals that N should indeed depend on n for several important reasons. Larger sample sizes enable finer partitioning while maintaining adequate observations per block, with $n = 1000$ and $N = 5$ providing approximately 200 observations per block for stable quartic polynomial estimation. The regression function $m(x) = \sin\left\{\left(\frac{x}{3} + 0.1\right)^{-1}\right\}$ exhibits varying curvature across its domain, and more blocks allow better adaptation to local features, though this benefit is limited by the available data in each block. The implemented constraint $N_{\max} = \max\{\min(\lfloor n/20 \rfloor, 5), 1\}$ ensures a minimum of 20 observations per block, preventing numerical instability in polynomial fitting while allowing adaptive complexity based on sample size.

Impact of covariate distribution on bandwidth selection

The distribution of the covariate X plays a fundamental role in bandwidth selection through both the density weighting in θ_{22} and the practical implementation of block-based estimation. Figure 4 demonstrates how the shape of the covariate distribution significantly influences the estimated optimal bandwidth. When $\alpha \approx \beta$, the Beta distribution is symmetric around 0.5, providing relatively uniform coverage of the support interval $[0, 1]$ and leading to consistent bandwidth estimates in the range 0.03-0.04, as all regions contribute similarly to the estimation of θ_{22}. As the distribution becomes asymmetric, with $\alpha > \beta$ concentrating observations near $x = 1$ or $\alpha < \beta$ concentrating near $x = 0$, the estimation quality becomes region-dependent, favoring areas with higher data density while potentially compromising estimation in sparse regions.

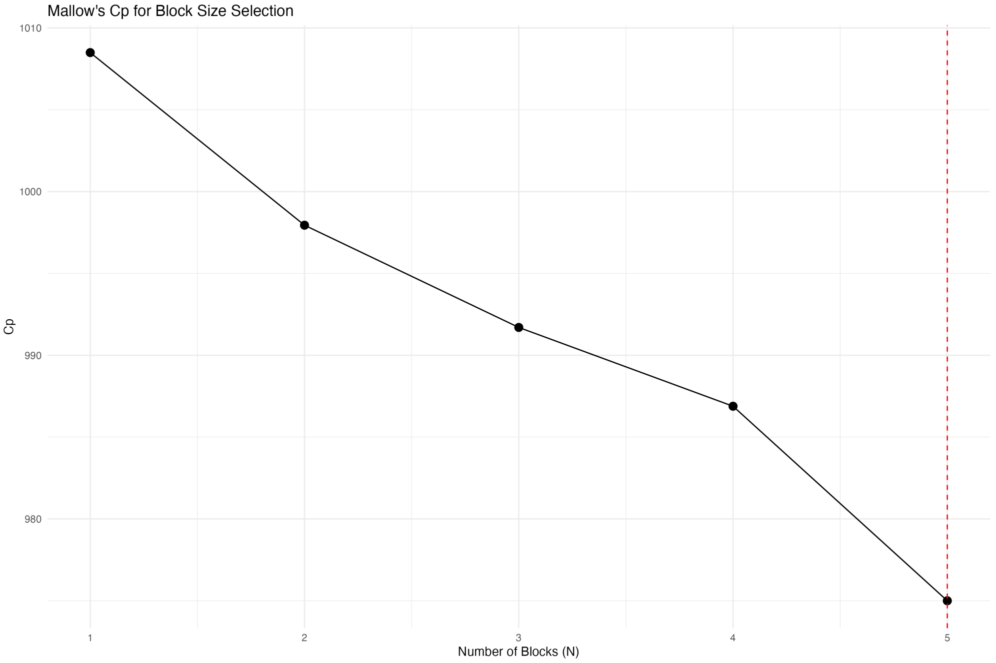

Figure 3: Mallow's Cp for block size selection

The lower-right region of the heat map, characterized by large α and small β values, reveals a fundamental limitation of the block-based estimation approach. When the Beta distribution becomes extremely concentrated near $x = 1$ due to these parameter choices, the block partitioning strategy encounters significant difficulties. The quantile-based blocking creates regions in the left portion of the support that contain very few or no observations, leading to failure in polynomial fitting. The use of raw polynomials exacerbates this issue, as the polynomial terms x^2, x^3, and x^4 become highly correlated when the data are concentrated in a small region. This multicollinearity produces numerically unstable coefficient estimates, particularly for the higher-order terms that are essential for computing the second derivative. Consequently, the estimates of $m''(x)$ become unreliable, preventing meaningful computation of θ_{22} and thus h_{AMISE}. This phenomenon illustrates the dual impact of covariate distribution on the estimation procedure, where regions with sparse data not only compromise the local polynomial estimation but also receive reduced weighting in the final bandwidth calculation, creating a compound effect that manifests as the observed "grey zone" in the parameter space.

Robustness analysis with replications

To verify the stability of our findings, we repeated the analyses with 200 replications for each configuration. The error bars in these plots represent ± 1 standard deviation around the mean, indicating the variability of estimates across different random samples rather than confidence intervals for the mean.

Figure 5 provides a more comprehensive view of the relationship between sample size and optimal bandwidth through 200 replications. The mean h_{AMISE} values demonstrate a clear stable trend, this improved consistency emerges because the replication process averages out the random fluctuations that caused irregularities in the single-sample analysis. The ± 1 standard deviation bars reveal substantial variability in bandwidth estimates, particularly for smaller sample sizes. As sample size increases to $n = 5000$, the variability decreases, with the bars becoming considerably narrower. This reduction in dispersion confirms that larger samples provide more stable and reliable bandwidth selection, as predicted by statistical theory. The replication analysis thus resolves the apparent contradictions observed in the single-sample results, demonstrating that while individual samples may show irregular patterns, the underlying relationship between sample size and optimal bandwidth does indeed follow a clear pattern.

Figure 6 demonstrates that $N = 5$ achieves optimal performance on average across the 200 replications. The

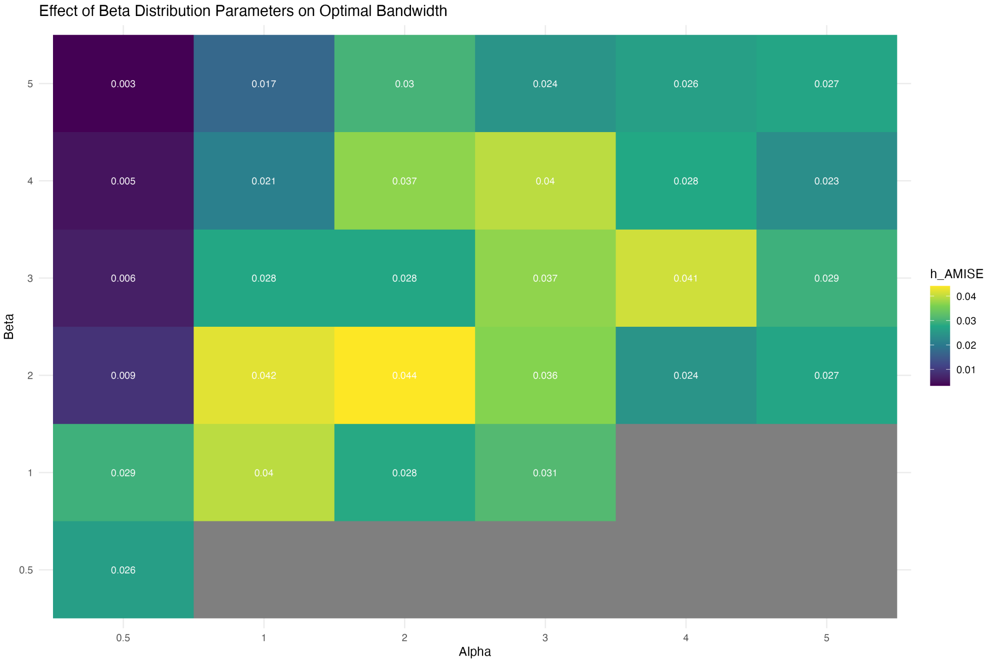

Figure 4: Effect of Beta distribution parameters on optimal bandwidth

error bars, representing ±1 standard deviation, reveal that the variability in h_{AMISE} estimates is minimized at this block size, indicating that $N = 5$ provides the most stable and reliable bandwidth selection. This reduced variability suggests that the estimation procedure reaches an optimal balance between bias and variance when using five blocks. The increased variability observed at both extremes of the block size spectrum reflects fundamental limitations of the estimation approach. With very small N (particularly $N = 1$), the single polynomial must approximate the entire regression function, leading to substantial bias in the second derivative estimates as the polynomial cannot adequately capture local curvature variations. This bias manifests as inconsistent h_{AMISE} values across different samples. Conversely, with larger N values approaching the maximum, each block contains fewer observations, increasing the variance of the polynomial coefficient estimates. The limited data in each block makes the second derivative estimates highly sensitive to random sampling variations, resulting in unstable bandwidth selection. The stability at $N = 5$ emerges because this configuration provides sufficient data within each block (approximately 200 observations with $n = 1000$) for reliable polynomial fitting, while allowing adequate localization to capture the varying curvature of the true regression function. This balance ensures that the second derivative estimates are both accurate and consistent across different random samples, leading to the observed minimization of variability in the optimal bandwidth estimates.

Figure 7 shows consistent selection of $N = 5$ as optimal. The error bars indicate moderate variability in Cp values across replications, but the minimum at $N = 5$ remains clear and stable.

Figure 8 reveals even more clearly the problematic region where estimation fails. While in the stable regions ($\alpha \approx \beta$), the bandwidth estimates are quite consistent across replications, while in the extreme parameter combinations, the estimation becomes highly unstable or impossible.

Important distinction: It is crucial to note that the error bars in these replication plots represent ±1 standard deviation of the estimates themselves, not confidence intervals for the mean. This means they show the actual dispersion of h_{AMISE} values we would expect to see in practice when applying this bandwidth selection method to different datasets from the same data-generating process. The relatively wide error bars for certain configurations highlight the inherent variability in data-driven bandwidth selection.

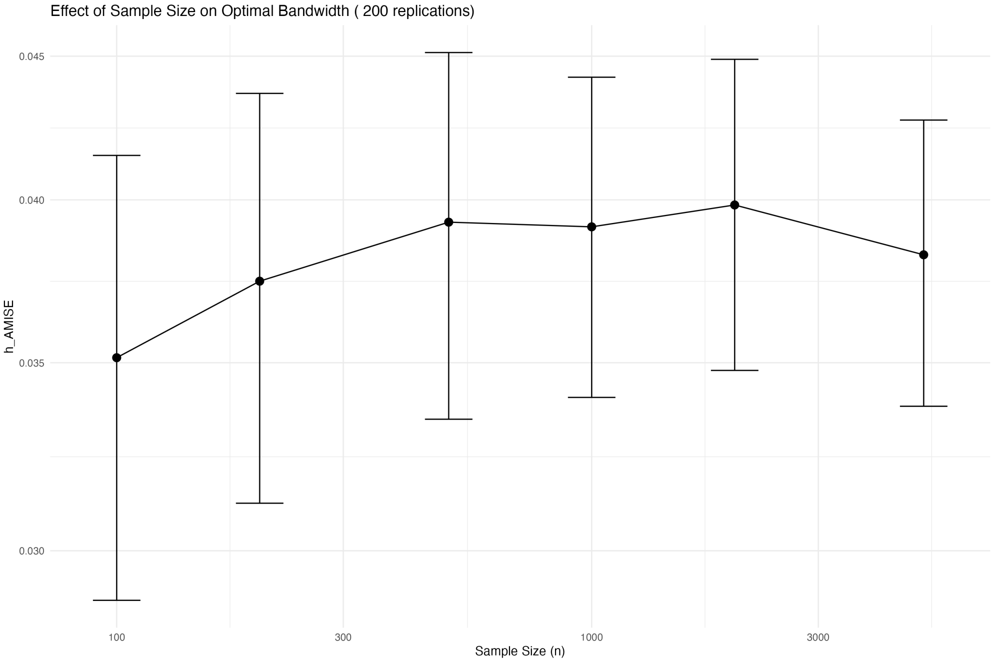

Figure 5: Effect of sample size on optimal bandwidth (200 replications)

Conclusion

In conclusion, our simulation study highlights that the optimal bandwidth h_{AMISE} is not solely determined by asymptotic theory, but rather by the interplay of several practical factors. Larger sample sizes generally support more stable bandwidths, yet the relationship is not strictly monotonic due to estimation variability. The block-based procedure proves effective in improving local approximation, with an intermediate number of blocks (e.g., N=5) providing the best trade-off between bias and variance. Finally, the distribution of the covariate X plays a central role: symmetric and balanced designs yield consistent estimates, whereas strongly skewed distributions challenge the stability of the method. These findings underline both the strengths and the limitations of data-driven bandwidth selection in applied regression settings.

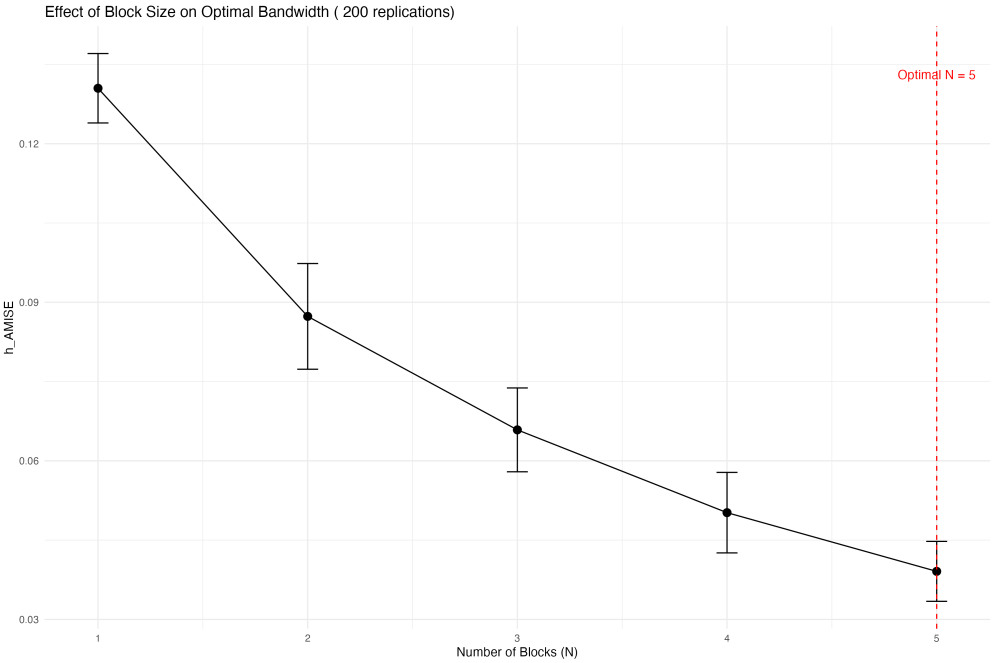

Figure 6: Effect of block size on optimal bandwidth (200 replications)

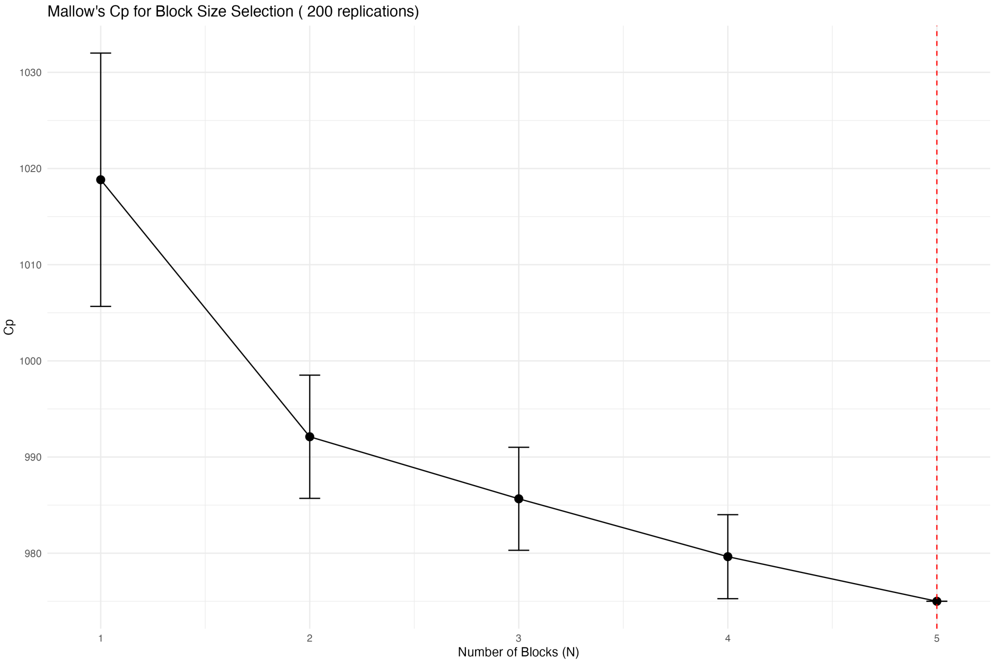

Figure 7: Mallow's Cp for block size selection (200 replications)

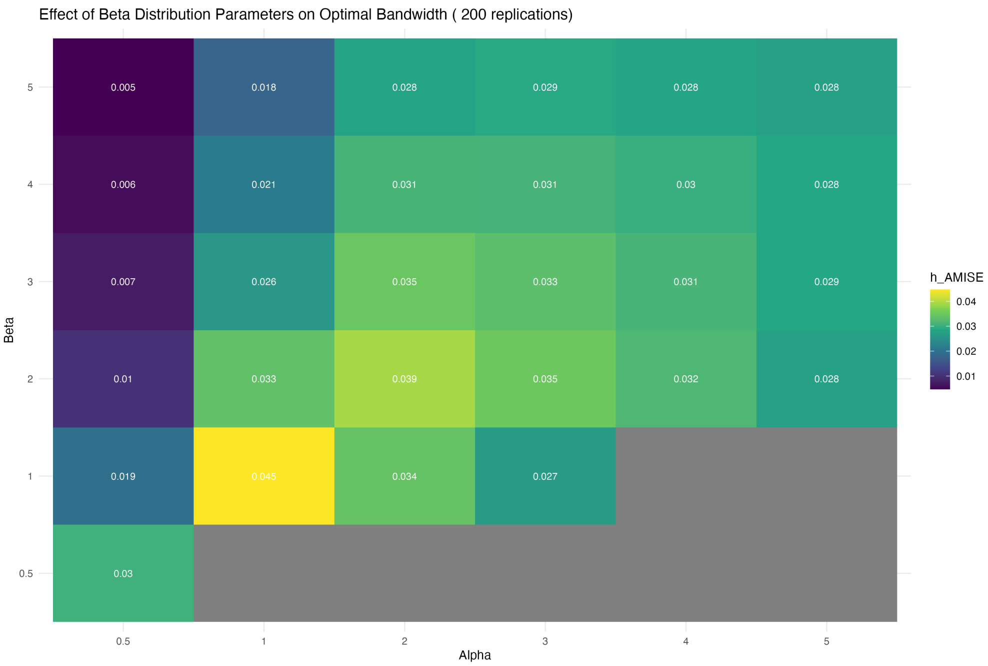

Figure 8: Effect of Beta parameters on optimal bandwidth (200 replications)