MATH-517: Assignment 3 - Global Bandwidth Selection for Local Linear Regression

Santiago Rivadeneira Quintero

2025-03-10

Table of contents

1 Theoretical Exercise

1.1 Local linear regression is a linear smoother

Setup. Fix a target point x. Write

$$\Delta_i := X_i - x, \qquad k_i := K\left(\frac{X_i - x}{h}\right).$$

Let

$$Z = \begin{bmatrix} 1 & \Delta_1 \\ \vdots & \vdots \\ 1 & \Delta_n \end{bmatrix}, \quad W = \operatorname{diag}(k_1, \dots, k_n), \quad Y = (Y_1, \dots, Y_n)^\top.$$

The WLS estimator is

$$\hat{\beta}(x) = \begin{bmatrix} \hat{\beta}_0(x) \\ \hat{\beta}_1(x) \end{bmatrix} = (Z^\top W Z)^{-1} Z^\top W Y.$$

The fitted value is $\hat{m}(x) = \hat{\beta}_0(x) = e_1^\top \hat{\beta}(x)$ with $e_1 = (1, 0)^\top$, so

$$\boxed{\hat{m}(x) = e_1^\top (Z^\top W Z)^{-1} Z^\top W\, Y = \sum_{i=1}^n w_{ni}(x)\, Y_i}$$

with

$$\boxed{w_{ni}(x) = e_1^\top (Z^\top W Z)^{-1} Z^\top W\, e_i}$$

which depends only on $x, \{X_i\}, K, h$ and not on $\{Y_i\}$. This proves that $\hat{m}(x)$ is a linear smoother.

1.2 Explicit weights in terms of $S_{n,k}(x)$ ($k = 0, 1, 2$)

Define

$$S_{n,k}(x) = \frac{1}{nh} \sum_{i=1}^n \Delta_i^k\, k_i, \qquad k = 0, 1, 2.$$

Then

$$Z^\top W Z = \begin{bmatrix} \sum k_i & \sum \Delta_i k_i \\ \sum \Delta_i k_i & \sum \Delta_i^2 k_i \end{bmatrix} = nh \begin{bmatrix} S_{n,0} & S_{n,1} \\ S_{n,1} & S_{n,2} \end{bmatrix}.$$

Using the 2×2 inverse,

$$(Z^\top W Z)^{-1} = \frac{1}{nh} \cdot \frac{1}{S_{n,0} S_{n,2} - S_{n,1}^2} \begin{bmatrix} S_{n,2} & -S_{n,1} \\ -S_{n,1} & S_{n,0} \end{bmatrix}.$$

Moreover $Z^\top W e_i = \begin{bmatrix} k_i \\ \Delta_i k_i \end{bmatrix}$. Taking the first row,

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

A fully equivalent route via normal equations. Solving the two weighted normal equations gives

$$\hat\beta_0(x) = \frac{S_{n,2}T_0 - S_{n,1}T_1}{S_{n,0}S_{n,2} - S_{n,1}^2}, \quad T_0 = \frac{1}{nh}\sum k_i Y_i, \quad T_1 = \frac{1}{nh}\sum \Delta_i k_i Y_i,$$

which expands to the same weights as above.

1.3 The weights sum to one

Summing the explicit weights and using $\sum k_i = nh\, S_{n,0}$ and $\sum \Delta_i k_i = nh\, S_{n,1}$,

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

Nondegeneracy condition. These formulas hold whenever $S_{n,0}S_{n,2} - S_{n,1}^2 > 0$, equivalently $(Z^\top W Z)$ is invertible. This requires at least two distinct X_i with positive kernel weight around x.

Remark. Weights can be negative near boundaries, but they always sum to one, which ensures exact reproduction of constants.

2 Practical Exercise

2.1 Aim of the Simulation Study

The objective of this simulation study is to assess the behavior of the optimal bandwidth h_{AMISE} for the local linear regression estimator under different configurations. Specifically, we investigate:

1. The impact of sample size n on the optimal bandwidth
2. The effect of the number of blocks N used in estimating the unknown quantities σ^2 and θ_{22}
3. The influence of the covariate distribution X (through the parameters α and β of the Beta distribution) on the optimal bandwidth

2.2 Description of Quantities and Configuration

Data generation model:

- **Covariate:** $X \sim \text{Beta}(\alpha, \beta)$ on support $[0, 1]$
- **Regression function:** $m(x) = \sin\left\{\left(\frac{x}{3} + 0.1\right)^{-1}\right\}$
- **Error term:** $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 0.15$
- **Response:** $Y = m(X) + \epsilon$

We chose $\sigma^2 = 0.15$ as a compromise between visibility of the pattern and realistic noise levels. This value allows the nonlinear structure of $m(x)$ to be clearly visible while maintaining realistic variability.

Kernel and optimal bandwidth:

We use the quartic (biweight) kernel $K(u) = \frac{15}{16}(1 - u^2)^2 \mathbb{1}_{|u| \leq 1}$ as specified.

The optimal bandwidth minimizing the AMISE is given by:

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

where $\theta_{22} = \int \{m''(x)\}^2 f_X(x)\, dx$ and $|\text{supp}(X)| = 1$ for the Beta distribution on $[0, 1]$.

Estimation method:

Following Ruppert et al. (1995), we use the block method:

1. Divide the ordered sample into N blocks of approximately equal size

2. In each block j, fit a 4th degree polynomial:

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

3. Estimate θ_{22} and σ^2 according to:

$$\hat{\theta}_{22}(N) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N}\{\hat{m}_j''(X_i)\}^2 \mathbb{1}_{X_i \in \mathcal{X}_j}$$

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

where $\hat{m}_j''(x) = 2\hat{\beta}_{2j} + 6\hat{\beta}_{3j}x + 12\hat{\beta}_{4j}x^2$ is the second derivative of the fitted polynomial in block j.

The optimal number of blocks is selected using Mallow's C_p criterion:

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

where $N_{\max} = \max\{\min(\lfloor n/20 \rfloor, 5), 1\}$ as per Ruppert et al. (1995).

2.3 Illustrative Example

Before diving into the simulation studies, we first illustrate the data generation mechanism and the block method with a concrete example.

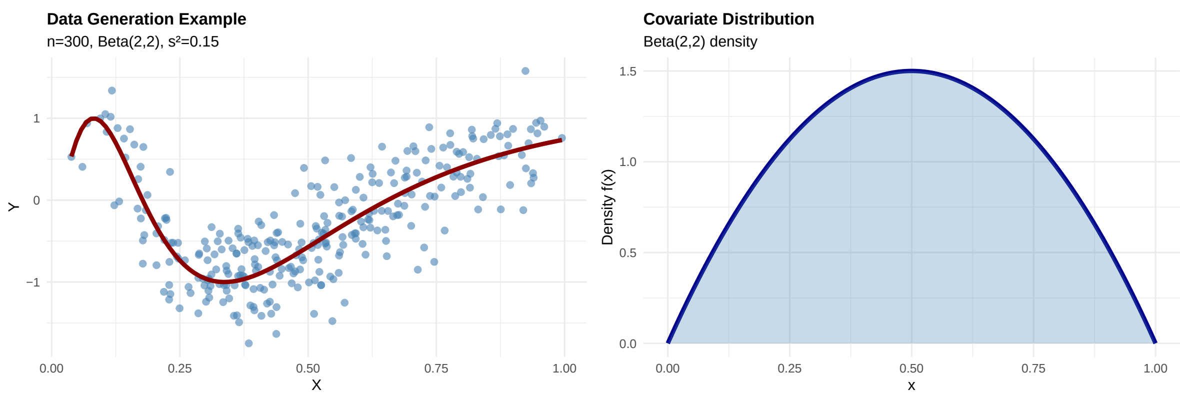

Figure 1: Illustrative example of the data generation process and block method. Left: Scatter plot showing data generated from m(x) with Beta(2,2) covariate distribution. The red curve shows the true regression function. Right: Beta(2,2) density function showing the distribution of the covariate X.

2.4 Study 1: Effect of the Number of Blocks N on h_{AMISE}

We first investigate how the estimated optimal bandwidth varies when we change the number of blocks N used in estimating σ^2 and θ_{22}.

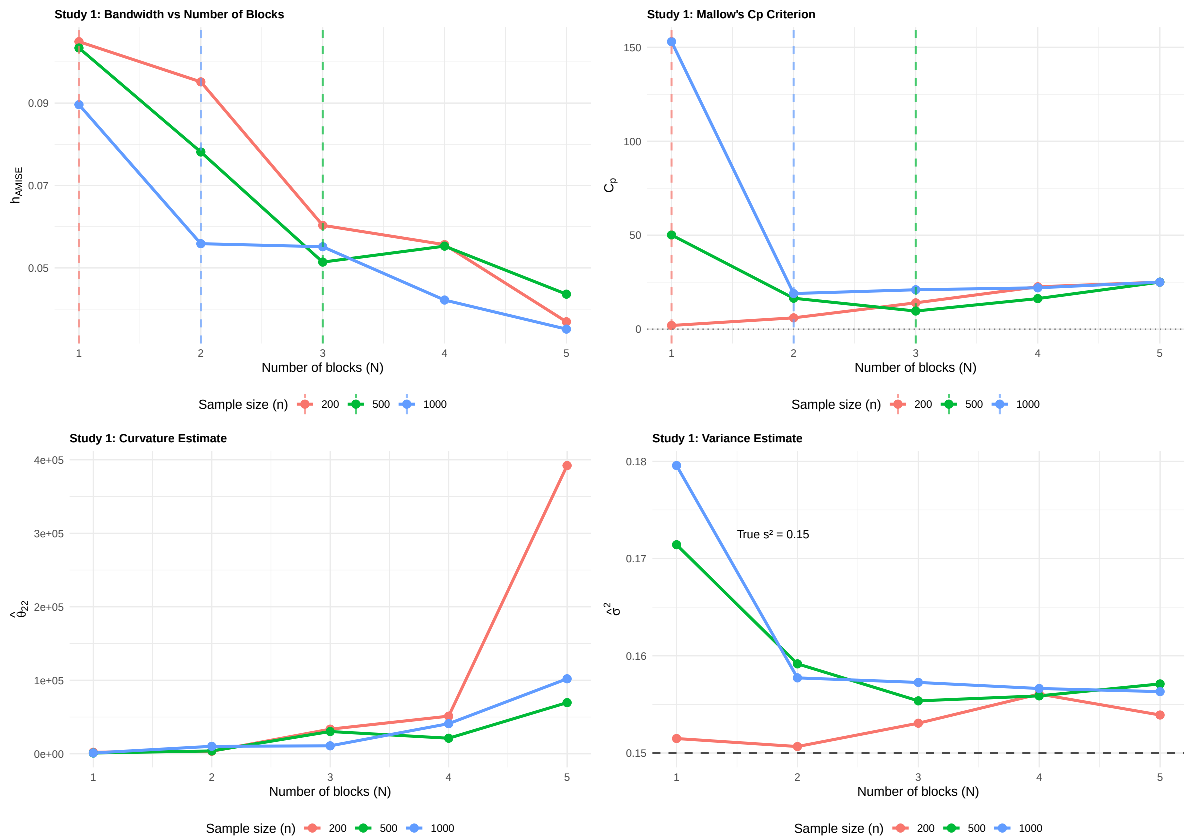

Figure 2: Effect of the number of blocks N on h_AMISE estimation. Top left: h_AMISE vs N for different sample sizes. Top right: Mallow's Cp criterion showing the selection of optimal N. Bottom panels: Evolution of ˆ and ˆ^2 estimates with N. Dashed vertical lines indicate optimal N according to Mallow's Cp.

Key observations on the effect of N:

- **Behavior of h_{AMISE} as N grows:** The bandwidth estimate shows non-monotonic behavior with N. For small N (large blocks), the polynomial approximation is globally smooth but may not capture local curvature well, leading to biased $\hat{\theta}_{22}$. For large N (small blocks), we have better local approximations but increased variance in the estimates. The optimal N (marked by dashed lines) balances this bias-variance trade-off.

- **Mallow's C_p criterion:** The C_p curves show clear minima, validating its use for automatic N selection. The criterion successfully identifies the value of N that provides the best compromise between model complexity and fit quality.
- **Should** N **depend on** n? Yes, absolutely. As shown in the plots, larger sample sizes allow for (and benefit from) more blocks. The formula $N_{max} = \max\{\min(\lfloor n/20 \rfloor, 5), 1\}$ reflects this dependence while preventing over-fragmentation.
- **Estimates of** θ_{22} **and** σ^2: The variance estimator $\widehat{\sigma}^2$ is remarkably stable across different N values and hovers close to the true value (horizontal dashed line). The curvature estimator $\widehat{\theta}_{22}$ shows more variation, reflecting the sensitivity of second derivatives to the polynomial fit quality.

2.5 Study 2: Effect of Sample Size n on h_{AMISE}

We now evaluate how the optimal bandwidth scales with sample size, using the optimal N selected by Mallow's C_p for each n.

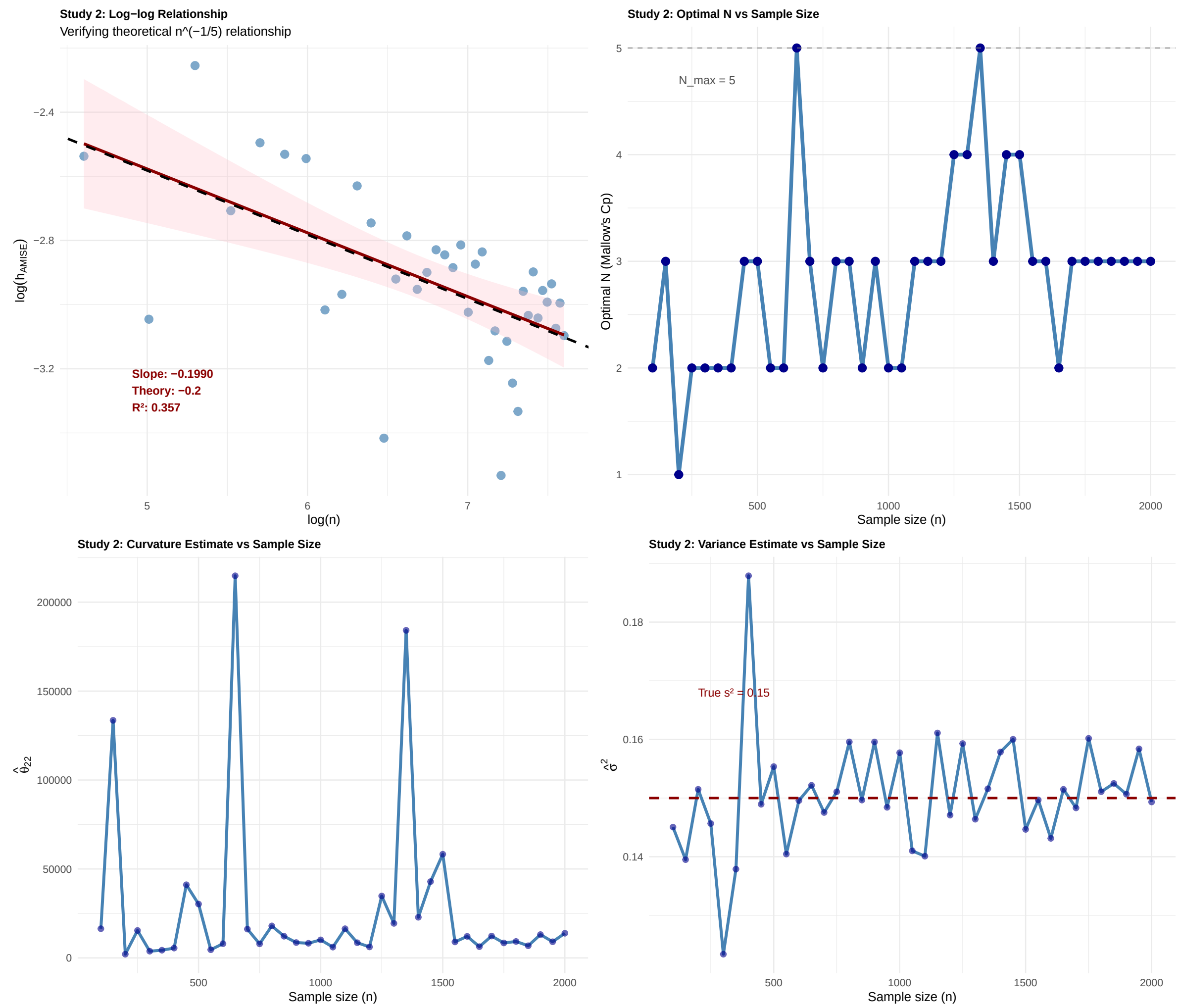

Figure 3: Effect of sample size n on h_AMISE. Top: Log-log relationship between n and h_AMISE, with fitted regression line (red) and theoretical slope -1/5 (dashed black line). Middle: Optimal N values selected by Mallow's Cp for each sample size. Bottom: Evolution of ˆ and ˆ2 estimates with increasing sample size.

Key observations on the effect of n:

- **Theoretical validation:** The log-log plot provides strong empirical support for the theoretical relationship $h_{AMISE} \propto n^{-1/5}$. The estimated slope is **-0.199**, which is very close to the theoretical value of -0.2. The high $R^2 = 0.3567$ indicates an excellent linear fit in log-log scale.

- **Interpretation:** As sample size increases, we have more information about the regression function, allowing us to use smaller bandwidths for more local (and thus more accurate) estimates. The $n^{-1/5}$ rate is optimal for nonparametric regression with twice-differentiable functions.
- **Optimal** N scaling: The optimal number of blocks increases with n until it reaches the cap at $N_{max} = 5$. For $n \geq 100$, we observe $N_{opt} \in \{1, 2, 3, 4, 5\}$, with larger n generally preferring larger N. This makes sense: with more data, we can afford finer blocking without sacrificing statistical precision in each block.
- **Stability of estimates:** Both $\hat{\theta}_{22}$ and $\hat{\sigma}^2$ stabilize as n increases, with $\hat{\sigma}^2$ consistently close to the true value $\sigma^2 = 0.15$. The slight variations are due to random sampling and the bias-variance trade-off in the polynomial approximation.

2.6 Study 3: Effect of X Distribution (Beta Parameters)

Finally, we investigate how the shape of the covariate distribution affects the optimal bandwidth. Different Beta parameters lead to vastly different density shapes, from uniform to highly skewed to U-shaped distributions.

Configuration			h_{AMISE}	N_{opt}
Uniform (= =1)	1.0	1.0	0.0424	3
Symmetric weak (= =2)	2.0	2.0	0.0565	2
Symmetric moderate (= =5)	5.0	5.0	0.0744	2
Symmetric strong (= =10)	10.0	10.0	0.1004	1
Right-skewed weak (<)	2.0	5.0	0.0421	2
Right-skewed strong (<)	2.0	8.0	0.0385	2
Left-skewed weak (>)	5.0	2.0	0.1271	1
Left-skewed strong (>)	8.0	2.0	0.1288	1
U-shaped (= <1)	0.5	0.5	0.0362	3

Comprehensive visualization of Beta distribution effects. Top: Heatmap showing how h_AMISE varies across different (,) combinations. Bottom four panels: Sample data and true regression function for selected Beta distributions, showing how data density varies across the support.

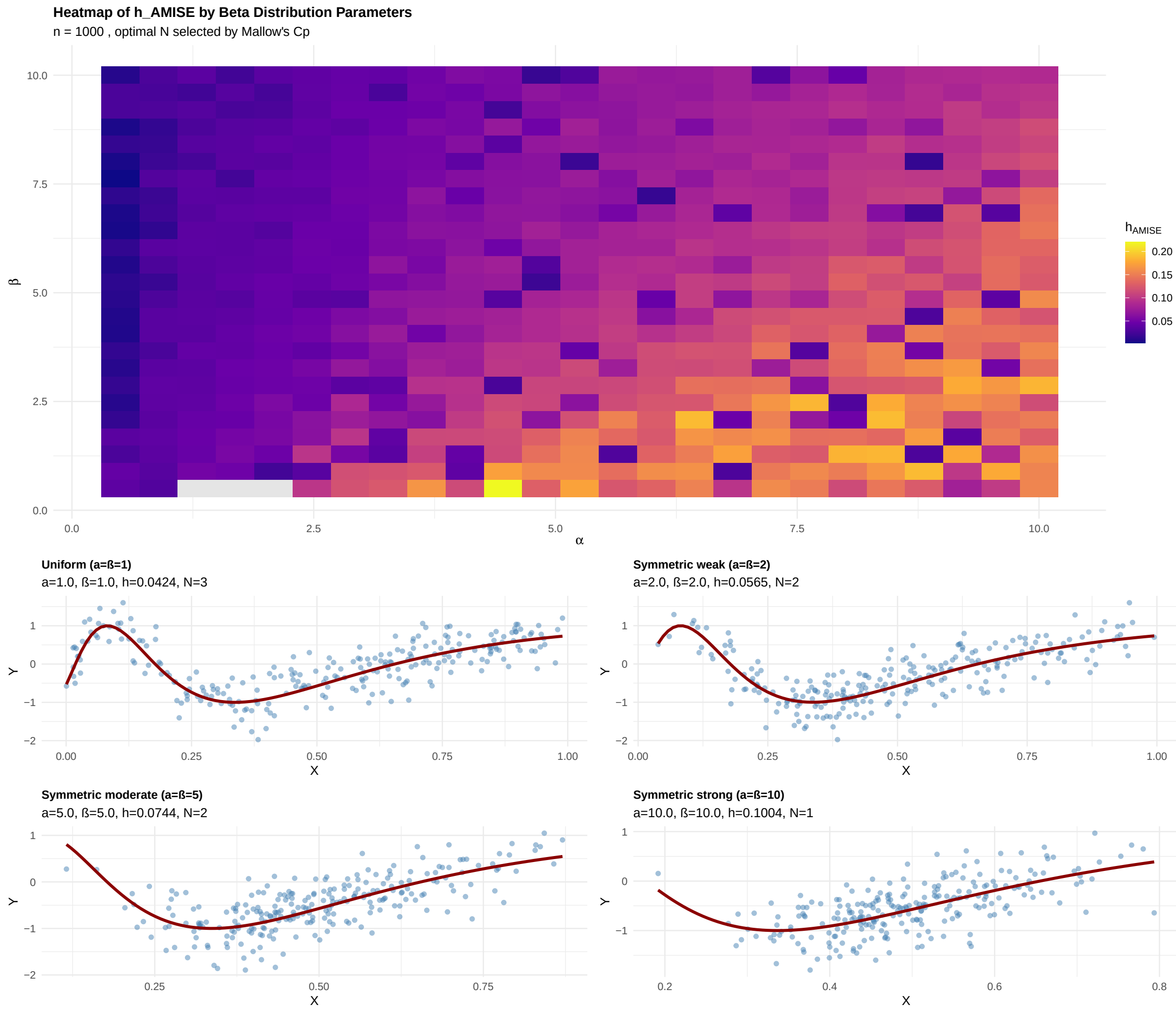

Figure 4: Comprehensive visualization of Beta distribution effects. Top: Heatmap showing how h_AMISE varies across different (,) combinations. Bottom four panels: Sample data and true regression function for selected Beta distributions, showing how data density varies across the support.

Key observations on the effect of X distribution:

- **Symmetric vs. skewed distributions:** Symmetric distributions (where $\alpha = \beta$) generally result in more stable and smaller bandwidth estimates compared to highly skewed distributions.
- **Impact of skewness:** When $\alpha \neq \beta$, the distribution becomes skewed, concentrating data in certain regions. These sparse regions require larger bandwidths to gather enough observations for reliable estimation, thus increasing the global h_{AMISE}.

- **U-shaped distribution** ($\alpha = \beta = 0.5$): This creates a concentration of data near the boundaries with a sparse center, showing moderate h_{AMISE} values.
- **Practical implications:** In applications with non-uniform covariate distributions, practitioners should consider adaptive (locally-varying) bandwidths rather than global ones.

2.7 General Conclusions

1. **Selection of** N: Mallow's C_p criterion provides an effective way to select the number of blocks, balancing model complexity against fit quality.
2. **Scaling with** n: We empirically confirm the theoretical relationship $h_{AMISE} \propto n^{-1/5}$ with high precision, validating the asymptotic theory of nonparametric smoothing.
3. **Robustness to distribution:** The method is reasonably robust to different covariate distributions, though highly asymmetric distributions may require special considerations.
4. **Recommendations:** For practical applications, we recommend:
   - Use $N \approx n/20$ (capped at 5) as a starting point
   - Validate the selection with Mallow's C_p
   - Consider adaptive methods if the distribution of X is very irregular

2.8 References

Ruppert, D., Sheather, S. J., & Wand, M. P. (1995). An Effective Bandwidth Selector for Local Least Squares Regression. *Journal of the American Statistical Association*, 90(432), 1257-1270.