Assignment 3

1 Theoretical Exercise: Local Linear Regression as a Linear Smoother

1.1 Exercise

The local linear regression estimator at point x has the following form:

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

The design matrix and weight matrix are:

$$\mathbf{X}_x = \begin{bmatrix} 1 & X_1 - x \\ 1 & X_2 - x \\ \vdots & \vdots \\ 1 & X_n - x \end{bmatrix} \in \mathbb{R}^{n \times 2}$$

$$\mathbf{W}(x) = \text{diag}\left(K\left(\frac{X_1 - x}{h}\right), K\left(\frac{X_2 - x}{h}\right), \dots, K\left(\frac{X_n - x}{h}\right)\right) \in \mathbb{R}^{n \times n}$$

We aim to find the $\hat{\boldsymbol{\beta}}$ such sthat:

$$\hat{\boldsymbol{\beta}}(x) = \arg\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}_x\boldsymbol{\beta})^\top \mathbf{W}(x)(\mathbf{Y} - \mathbf{X}_x\boldsymbol{\beta}) \tag{2}$$

The solution is given by:

$$\hat{\boldsymbol{\beta}}(x) = (\mathbf{X}_x^\top \mathbf{W}(x)\mathbf{X}_x)^{-1}\mathbf{X}_x^\top \mathbf{W}(x)\mathbf{Y} \tag{3}$$

Since $\hat{m}(x) = \hat{\beta}_0(x) = \mathbf{e}_1^\top \hat{\boldsymbol{\beta}}(x)$ where $\mathbf{e}_1 = (1, 0)^\top$, we have:

$$\hat{m}(x) = \mathbf{e}_1^\top (\mathbf{X}_x^\top \mathbf{W}(x)\mathbf{X}_x)^{-1}\mathbf{X}_x^\top \mathbf{W}(x)\mathbf{Y} \tag{4}$$

The vector of weights is:

$$\mathbf{w}(x) = \mathbf{W}(x)\mathbf{X}_x(\mathbf{X}_x^\top \mathbf{W}(x)\mathbf{X}_x)^{-1}\mathbf{e}_1 \tag{5}$$

Then:

$$\hat{m}(x) = \mathbf{w}(x)^\top \mathbf{Y} = \sum_{i=1}^{n} w_i(x) Y_i \tag{6}$$

The weights $w_i(x)$ depend only on X_i, K, h, and x, but not on Y_i.

1.2 Exercise

The objective is now to write $w_i(x)$ as an expression of the following forms:

$$S_{n,0}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \tag{7}$$

$$S_{n,1}(x) = \frac{1}{nh}\sum_{i=1}^{n} (X_i - x) K\left(\frac{X_i - x}{h}\right) \tag{8}$$

$$S_{n,2}(x) = \frac{1}{nh}\sum_{i=1}^{n} (X_i - x)^2 K\left(\frac{X_i - x}{h}\right) \tag{9}$$

Compute the matrix products:

$$\mathbf{X}_x^\top \mathbf{W}(x)\mathbf{X}_x = \begin{bmatrix} \sum_{i=1}^n K_i & \sum_{i=1}^n (X_i - x)K_i \\ \sum_{i=1}^n (X_i - x)K_i & \sum_{i=1}^n (X_i - x)^2 K_i \end{bmatrix} \tag{10}$$

$$= nh \begin{bmatrix} S_{n,0}(x) & S_{n,1}(x) \\ S_{n,1}(x) & S_{n,2}(x) \end{bmatrix} \tag{11}$$

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

The inverse is:

$$(\mathbf{X}_x^\top \mathbf{W}(x)\mathbf{X}_x)^{-1} = \frac{1}{nh[S_{n,0}(x)S_{n,2}(x) - S_{n,1}(x)^2]} \begin{bmatrix} S_{n,2}(x) & -S_{n,1}(x) \\ -S_{n,1}(x) & S_{n,0}(x) \end{bmatrix} \tag{12}$$

The i-th weight is:

$$w_{ni}(x) = \frac{1}{nh[S_{n,0}(x)S_{n,2}(x) - S_{n,1}(x)^2]} \left[S_{n,2}(x) - S_{n,1}(x)(X_i - x)\right] K\left(\frac{X_i - x}{h}\right) \tag{13}$$

1.3 Exercise

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

$$\sum_{i=1}^{n} w_{ni}(x) = \frac{1}{nh[S_{n,0}S_{n,2} - S_{n,1}^2]} \sum_{i=1}^{n} \left[S_{n,2} - S_{n,1}(X_i - x)\right] K_i \tag{14}$$

$$= \frac{1}{nh[S_{n,0}S_{n,2} - S_{n,1}^2]} \left[S_{n,2}\sum_{i=1}^{n} K_i - S_{n,1}\sum_{i=1}^{n}(X_i - x)K_i\right] \tag{15}$$

Using the definitions:

$$\sum_{i=1}^{n} K_i = nhS_{n,0} \tag{16}$$

$$\sum_{i=1}^{n} (X_i - x)K_i = nhS_{n,1} \tag{17}$$

Therefore:

$$\sum_{i=1}^{n} w_{ni}(x) = \frac{1}{nh[S_{n,0}S_{n,2} - S_{n,1}^2]} [S_{n,2}S_{n,0} - S_{n,1}S_{n,1}] \tag{18}$$

$$= \frac{nh}{nh[S_{n,0}S_{n,2} - S_{n,1}^2]} (S_{n,0}S_{n,2} - S_{n,1}^2) = 1 \tag{19}$$

2 Practical Exercise

2.1 Relationship Between Sample Size and Block Size

The key link between the sample size n and the optimal block size N arises from the bias-variance tradeoff inherent in local polynomial estimation. As the sample size increases, the data can be divided into more blocks without overfitting, since each block still contains enough observations for reliable parameter estimation. Mallow's C_p criterion captures this balance by penalizing models that are either too simple (few blocks, leading to high bias) or too complex (many blocks, resulting in high variance). Consequently, as n grows, the optimal choice of N tends to increase, allowing for more localized modeling while preserving estimation stability within blocks.

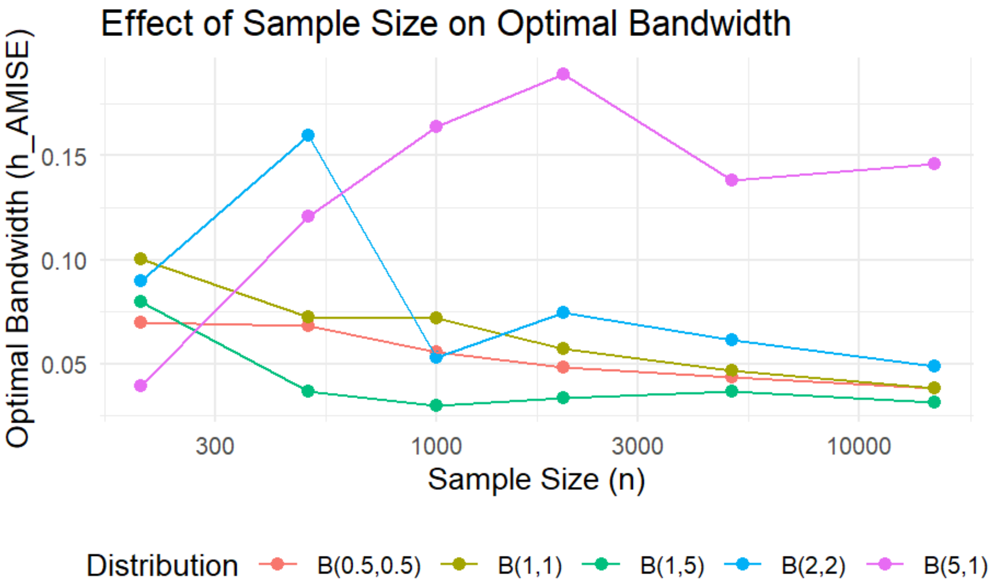

Figure 1: Relationship between sample size n and optimal bandwidth h_{AMISE}

The first graph shows how the sample size affects our choice of smoothing parameter. The most important thing to notice is that all the lines follow the same basic pattern: as we get more data points, we can use smaller bandwidths. This makes sense because with more data, we can afford to be more precise in our estimates without worrying too much about noise.

While most distributions show clear decreasing patterns, Beta(5,1) doesn't decay as much as the others. Its line is much flatter, suggesting that adding more data doesn't help reduce the bandwidth much for this distribution. This might be because the data is already so concentrated in one region that additional data points don't provide much new information about the function's behavior in other areas.

As we look at the far right of the graph (in case of very large n), all the lines start to come closer together. This tells us that when we have huge amounts of data, the advantages of some distributions over others become less important.

2.2 Graph 2: Block Size Selection

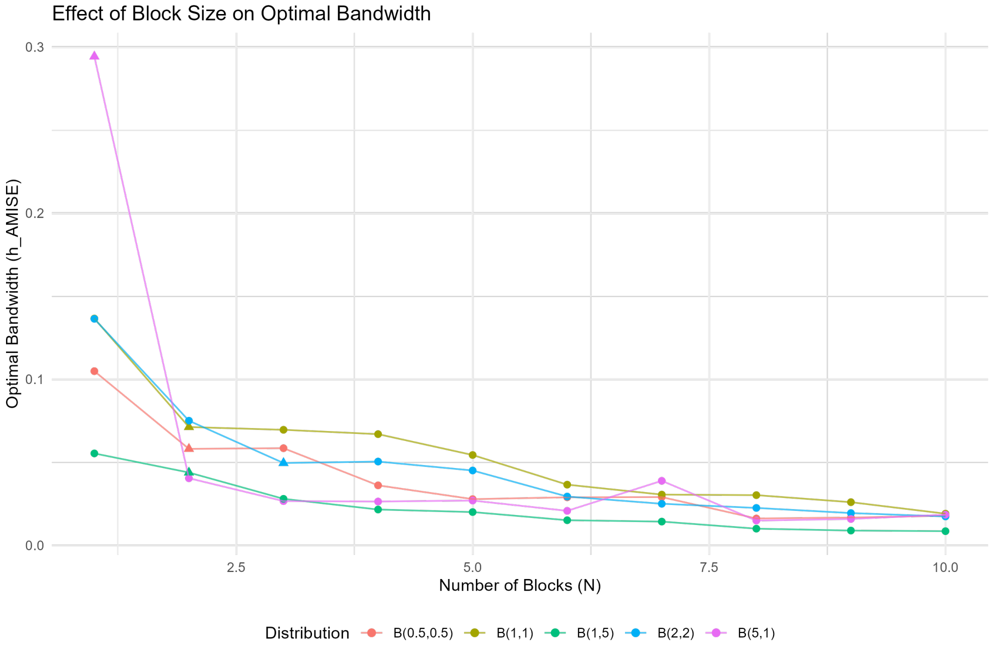

Figure 2: Effect of block count N on optimal bandwidth estimation

This second graph shows us how the number of blocks affects the bandwidth.

When we use very few blocks (just 1 or 2), all the distributions give us bandwidths that are too high. This happens because with so few blocks, we're oversimplifying the data.

Overall, increasing the number blocks (N) helps us get better local estimates of the noise level (σ^2) and the curvature ($m''(x)$). This reduces bias and can lead to a smaller optimal bandwidth. However, when we use more blocks, each block has fewer data points, which makes the estimates from each block less reliable (higher variance). In our simulation, we see that the optimal bandwidth tends to decrease as we use more blocks. This means that in our case, the benefit of reducing bias is more important than the problem of increased variance.

We conclude that the number of blocks should depend on the sample size (n). If we use too few blocks, the estimates are too smooth and may miss important details (bias). If we use too many blocks, the estimates become too noisy (variance). The best number of blocks balances these two issues. Since larger samples can handle more blocks without becoming too noisy, the optimal N should increase with n. This is why we use a data-driven method like Mallow's C_p to choose N based on the data.

2.3 Graph 3: Distributional Comparison of Optimal Bandwidths

This graph clearly shows that the pattern of our data makes a big difference in how much smoothing we need. The U-shaped data requires about twice as much smoothing as the most lopsided data.

What's interesting is that the shape of the data matters more than whether it's symmetric or not. The U-shaped data needs heavy smoothing because it has lots of points near the edges where our function behaves

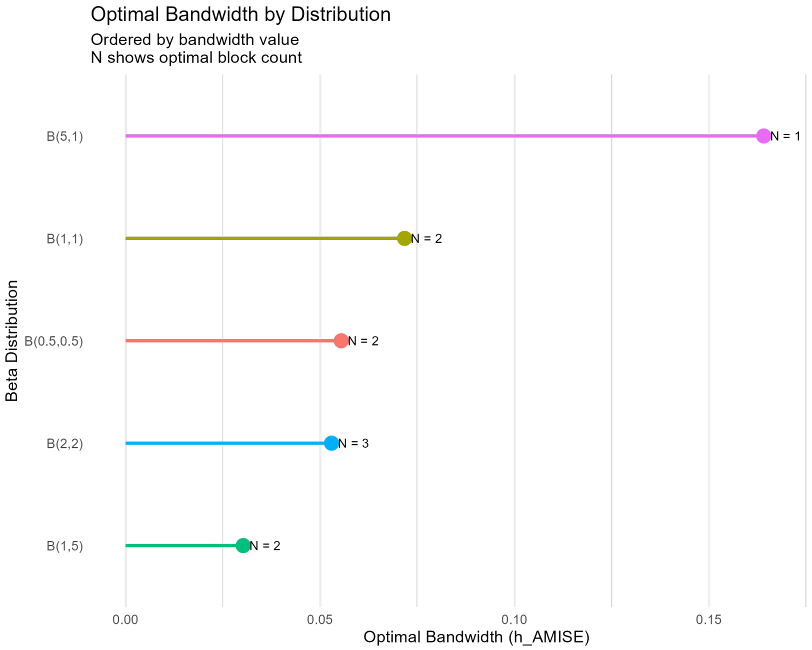

Figure 3: Optimal bandwidth by Beta distribution type ($n = 1000$)

wildly. On the other hand, the lopsided data needs very little smoothing because its points are concentrated in regions where the function is easier to estimate.

We can also see that data that needs heavy smoothing works better with fewer blocks, while data that needs light smoothing can handle more detailed blocking. This tells us that we can't use the same smoothing approach for all types of data - we have to adjust based on the specific pattern of our data.