MATH-517: Assignment 3

Giorgio Panizzutti

1. Theoretical Exercise

1.1.

The problem is a weighted least squares problem at x:

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

with

$$X = \begin{pmatrix} 1 & X_1 - x \\ \vdots & \vdots \\ 1 & X_n - x \end{pmatrix}, \quad W = \operatorname{diag}(K\left(\frac{X_1 - x}{h}\right), \dots, K\left(\frac{X_n - x}{h}\right)), \quad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}.$$

The problem can be rewritten as

$$(\hat{\beta}_0(x), \hat{\beta}_1(x)) = \arg\min_{\beta_0,\beta_1\in\mathbb{R}} (Y - X\beta)^\top W (Y - X\beta),$$

The weighted least squares solution is

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

Let $e_1 = (1, 0)^\top$. Then

$$\hat{m}(x) = \hat{\beta}_0(x) = e_1^\top (X^\top W X)^{-1} X^\top W Y = \sum_{i=1}^{n} w_{ni}(x)\, Y_i,$$

where $w_{ni}(x)$ is the i-th entry of the row vector $e_1^\top (X^\top W X)^{-1} X^\top W$.

1.2.

Using the notation

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

we can write

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

The inverse is

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

then $X^\top W$ is the $2 \times n$ matrix

$$X^\top W = \begin{pmatrix} K\left(\frac{X_1-x}{h}\right) & \cdots & K\left(\frac{X_n-x}{h}\right) \\ (X_1-x)K\left(\frac{X_1-x}{h}\right) & \cdots & (X_n-x)K\left(\frac{X_n-x}{h}\right) \end{pmatrix}.$$

the product $(X^\top W X)^{-1} X^\top W$ is then

$$\frac{1}{nh(S_{n,0}(x)S_{n,2}(x) - S_{n,1}(x)^2)} \begin{pmatrix} (S_{n,2}(x) - (X_1-x)S_{n,1}(x))K\left(\frac{X_1-x}{h}\right) & \cdots \\ (-S_{n,1}(x) + (X_1-x)S_{n,0}(x))K\left(\frac{X_1-x}{h}\right) & \cdots \end{pmatrix}$$

The i-th entry of the first row is then

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

1.3.

The sum of the weights is

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

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

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

2. Practical Exercise

The goal is to understand how the global bandwidth h_{AMISE}

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

behaves when we change specific quantities in the estimation: (i) the number of blocks N used to estimate σ^2 and θ_{22} with blockwise quartic OLS, (ii) the sample size n, and (iii) the shape of the covariate distribution $X \sim \mathrm{Beta}(\alpha, \beta)$. We keep $\sigma^2 = 1$ and $|\mathrm{supp}(X)| = 1$ since $X \in [0, 1]$. The regression curve is $m(x) = \sin((x/3 + 0.1)^{-1})$.

We generate i.i.d. samples $\{(X_i, Y_i)\}_{i=1}^n$ with $Y_i = m(X_i) + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. To estimate the unknown θ_{22} and σ^2 we split the sample into N blocks by quantiles of X so that each block has (almost) the same number of observations even when X is skewed. In each block j we fit a quartic polynomial by OLS

$$\hat{m}_j(x) = \beta_{0j} + \beta_{1j}x + \beta_{2j}x^2 + \beta_{3j}x^3 + \beta_{4j}x^4,$$

and then compute

$$\hat{\theta}_{22}(N) = \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^N \{\hat{m}_j''(X_i)\}^2 \mathbb{1}_{\{X_i \in \mathcal{X}_j\}}, \qquad \hat{\sigma}^2(N) = \frac{1}{n - 5N}\sum_{i=1}^n \sum_{j=1}^N \{Y_i - \hat{m}_j(X_i)\}^2 \mathbb{1}_{\{X_i \in \mathcal{X}_j\}}.$$

We plug $\hat{\sigma}^2(N)$ and $\hat{\theta}_{22}(N)$ in the formula for h_{AMISE}. When we need to choose N from data we use Mallow's C_p

$$C_p(N) = \frac{\mathrm{RSS}(N)}{\mathrm{RSS}(N_{\max})/(n - 5N_{\max})} - (n - 10N), \quad N_{\max} = \max\{\min(\lfloor n/20 \rfloor, 5), 1\},$$

and pick the minimizer over $N = 1, \ldots, N_{\max}$.

We average over $R = 200$ repetitions to get stable estimates of $\hat{h}_{AMISE}$.

Effect of the block count. Figure 1 shows, for $n = 2000$, the mean and standard deviation of $\hat{h}_{AMISE}$ as N increases for three shapes of X: Beta$(2, 2)$ (symmetric), Beta$(1, 4)$ (mass near 0), Beta$(4, 1)$ (mass near 1). In all cases $\hat{h}_{AMISE}$ decreases with N. This is expected: larger N makes each of the quartic polynomial fits concern a smaller region, which increases $\hat{\theta}_{22}$ (sum of squared second derivatives evaluated across the sample). Since $\hat{h}_{AMISE} \propto (\hat{\sigma}^2/\hat{\theta}_{22})^{1/5}$ and $\hat{\sigma}^2$ is fairly stable, h_{AMISE} goes down.

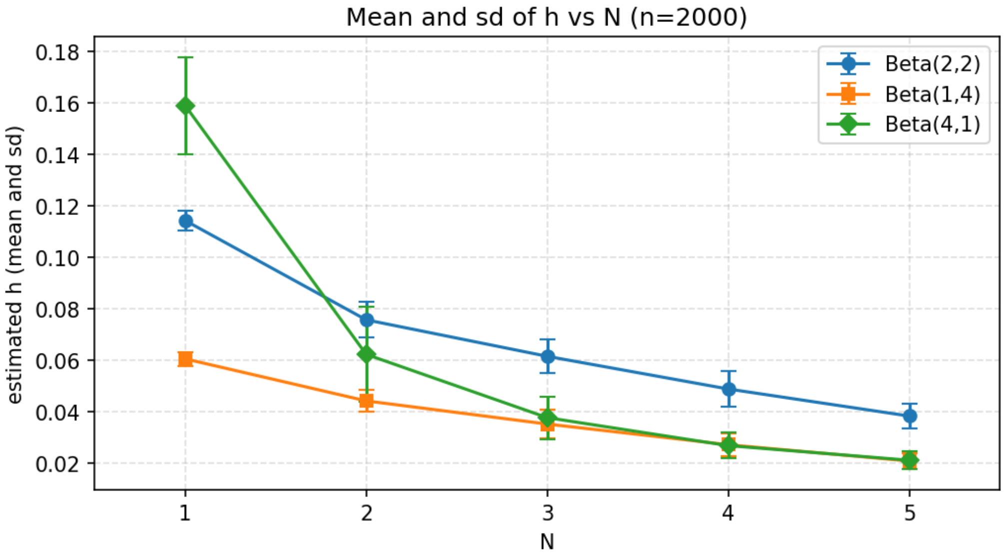

Figure 1: Mean with standard deviation of $\hat{h}_{AMISE}$ versus N for $n = 2000$ using three Beta distributions.

Does optimal N depend on n? Figure 2 reports the average N selected by C_p as a function of n for the same three Beta shapes, with bands showing the standard deviation across replications. There is a clear increasing trend: as n grows, the selected N increases slowly. In the symmetric case Beta$(2, 2)$, the average selected N grows roughly logarithmically with n, so the trend appears linear on a log scale. For the skewed cases (Beta$(1, 4)$ and Beta$(4, 1)$), the selected N grows less. There is a large difference between the two skewed cases: when the mass of X is near 0 (Beta$(1, 4)$) the selected N is much larger than when mass is near 1 (Beta$(4, 1)$). This is because m has higher curvature near 0, and more blocks may help to capture it better, with a larger n we get a larger number of samples across the whole support, covering better the less sampled regions. Also, having more values for each block helps to estimate the local function shape better, so greater n allows for larger optimal N.

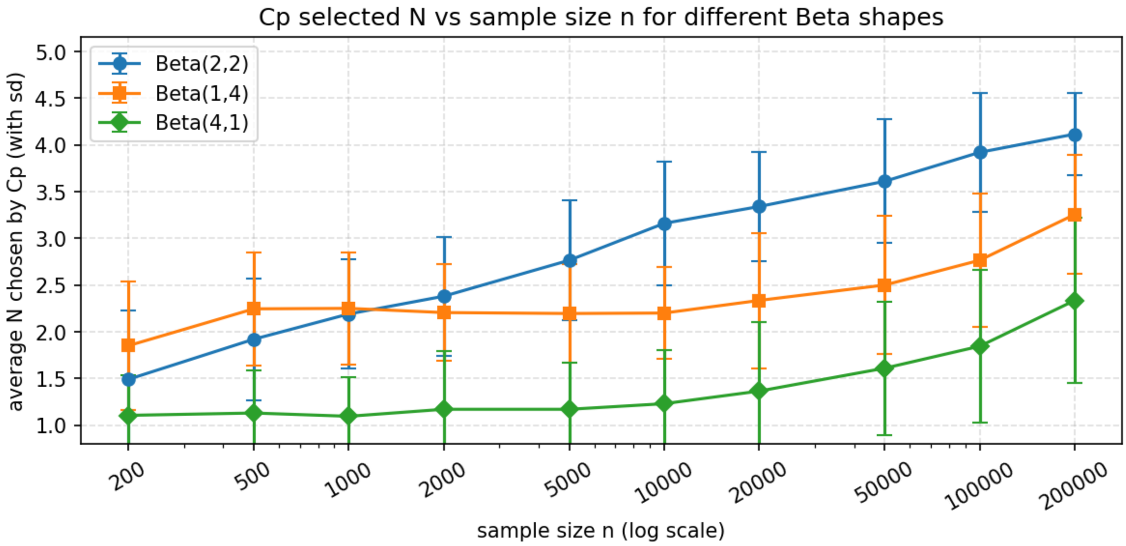

Figure 2: Average N selected by C_p versus n for Beta$(2, 2)$, Beta$(1, 4)$, and Beta$(4, 1)$; bands show the standard deviation across $R = 200$ repetitions.

Impact of the covariate Beta distribution. Figure 3 shows a heatmap of the mean $\hat{h}_{AMISE}$ over a grid of different values of α and β for $n = 1000$. The largest h_{AMISE} occur for large α and small β (mass near 1), while the smallest occur for small α and large β (mass near 0). This matches the geometry of m, curvature is highest near zero. When the density of X has more probability where $|m''(x)|$ is large, $\theta_{22} = \int m''(x)^2 f_X(x)\, dx$ increases and the optimal h decreases. When more mass is in flatter regions, θ_{22} is smaller and h_{AMISE} increases.

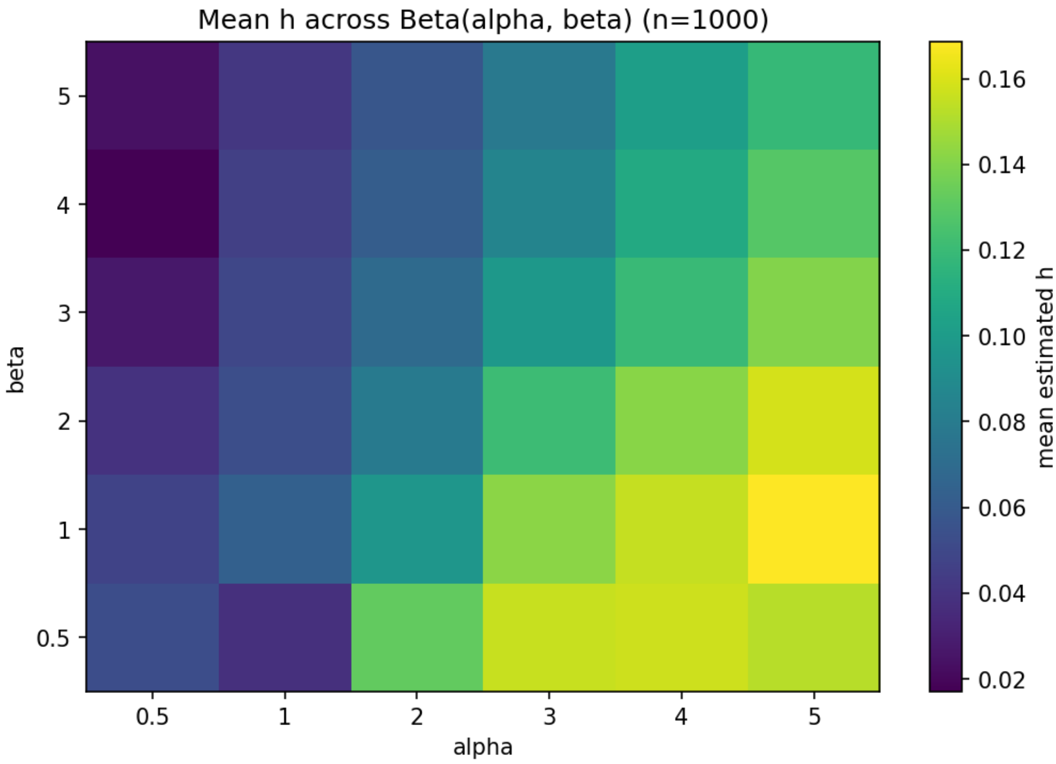

Figure 3: Mean $\hat{h}_{AMISE}$ on a grid of Beta(α, β) distributions for $n = 1000$.

Second derivative of m. To better see why h_{AMISE} varies with the distribution of X, we plot $m(x)$ and its second derivative $m''(x)$ (here computed numerically, but can be analytically derived) on $[0, 1]$. The curvature is highest near $x = 0$ and decreases toward $x = 1$, so $\theta_{22} = \int m''(x)^2 f_X(x)dx$ increases when more probability mass is placed near 0. This directly shows the impact of the shape of X on the estimate of the optimal bandwidth calculated using the $\hat{h}_{AMISE}$ formula where $\hat{\theta}_{22}$ is at the denominator and gets higher when the density of X has more probability where $|m''(x)|$ is large.

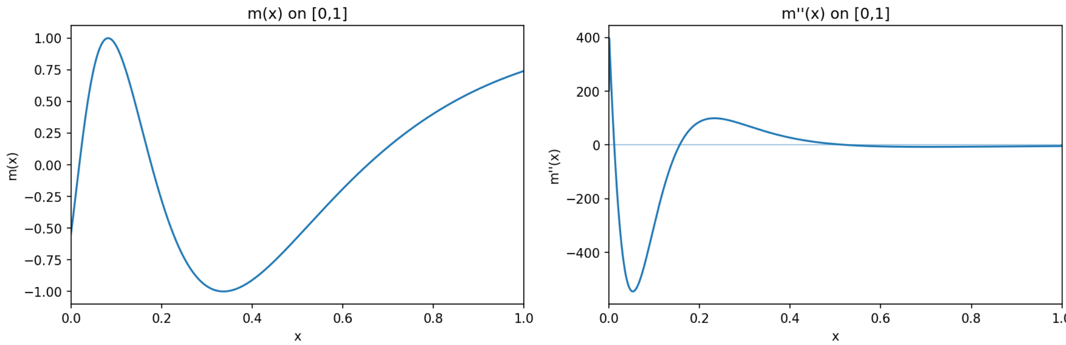

Figure 4: Regression function $m(x)$ (left) and its second derivative $m''(x)$ (right) on $[0, 1]$.