MATH-517: Assignment 2

Filippo Reina

2025-10-05

1. Theoretical exercise: Local linear regression as a linear smoother

Question

Recall our basic setup: we are given i.i.d. samples $(x_i, y_i), i = 1, \dots n$ from the model

$$y_i = m(x_i) + \epsilon_i, \quad i = 1, \dots n$$

and our goal is to estimate m with some function $\hat{m}$. Assume that each $x_i \in \mathbb{R}$ (i.e., the predictors are 1-dimensional).

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

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

where K is a kernel function and $h > 0$ is a bandwidth. The fitted value is then given by $\hat{m}(x) = \hat{\beta}_0(x)$ and we will show that such estimator belongs to the class of linear smoothers, so that

$$\hat{m}(x) = \sum_{i=1}^{n} w(x, x_i) \cdot y_i$$

for some choice of weights $w(x, x_i)$.

1. Show that $\hat{m}(x)$ can be expressed as a weighted average of the observations:

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

where the weights $w_{ni}(x)$ depend only on x, $\{X_i\}$, K, and h, but not on the Y_i's.

2. Using the notation

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

derive an explicit expression for $w_{ni}(x)$ in terms of $S_{n,0}(x), S_{n,1}(x), S_{n,2}(x)$, and the kernel.

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

Answer

1. Looking at Equation (1), we can write $\hat{\beta}$ as the solution to a weighted least-squares problem:

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

where we defined:

$$\mathbf{Y} = (Y_1, \cdots, Y_n)^\top, \qquad \mathbf{W} = \operatorname{diag}(K_1, \cdots, K_n), \qquad \mathbf{X} = \begin{bmatrix} 1 & X_1 - x \\ \vdots & \vdots \\ 1 & X_n - x \end{bmatrix}$$

and we denote $K_i = K\left((X_i - x)/h\right)$.

Since $\hat{m}(x) = \hat{\beta}_0$ is linear, there exist weights independent of Y_i such that:

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

2. We start by considering $\mathbf{X}^\top \mathbf{W} \mathbf{X}$.

$$\mathbf{X}^\top \mathbf{W} \mathbf{X} = \mathbf{X}^\top \begin{bmatrix} K_1 & K_1(X_1 - x) \\ \vdots & \vdots \\ K_n & K_n(X_n - x) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} K_i & \sum_{i=1}^{n} K_i (X_i - x) \\ \sum_{i=1}^{n} (X_i - x) K_i & \sum_{i=1}^{n} K_i (X_i - x)^2 \end{bmatrix}.$$

Looking at Equation (2), it follows that:

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

The determinant of that matrix is $S_{n,0}(x)S_{n,2}(x) - S_{n,1}(x)^2$, so the inverse is:

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

Proceding to $\mathbf{X}^\top \mathbf{W}$, what is simply a $2 \times n$ matrix:

$$\mathbf{X}^\top \mathbf{W} = \begin{bmatrix} K_1 & \cdots & K_n \\ (X_1 - x)K_1 & \cdots & (X_n - x)K_n \end{bmatrix}.$$

Notice that the first row of $(\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{W}$ is the row vector $(w_{ni}(x))_{i=1,\cdots,n}$.

Ignoring the normalizing constant, we can write the ith entry of $w_{ni}(x)$, $w_{ni}(x) = S_{n,2}(x)K_i - S_{n,1}(X_i - x)K_i$, and thus:

$$w_{ni}(x) = \frac{K\left(\frac{X_i - x}{h}\right)\left(S_{n,2}(x) - (X_i - x)S_{n,1}(x)\right)}{nh\left[S_{n,0}(x)S_{n,2}(x) - S_{n,1}(x)^2\right]}. \tag{3}$$

3. Looking at the Formula (3), it we need to show that $\sum_{i=1}^n w_{ni}(x) = 1$.

$$\sum_{i=1}^n w_{ni}(x) = \frac{1}{nh\left\{S_{n,0}(x)S_{n,2}(x) - S_{n,1}(x)^2\right\}} \left\{ S_{n,2}(x) \underbrace{\sum_{i=1}^n K_i}_{(a)} - S_{n,1}(x) \underbrace{\sum_{i=1}^n K_i(X_i - x)}_{(b)} \right\},$$

where:

$$(a) = nh \cdot S_{n,0}(x), \qquad (b) = nh \cdot S_{n,1}(x).$$

So, we can simplify the numerator to:

$$nh\left\{S_{n,2}(x)S_{n,0}(x) - S_{n,1}(x)S_{n,1}(x)\right\} = nh\left\{S_{n,0}(x)S_{n,2}(x) - S_{n,1}(x)^2\right\},$$

and it follows that:

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

2. Practical exercise: Global bandwidth selection

Assume that we have a sample $\{(X_i, Y_i)\}_{i=1}^n$ of i.i.d. random vectors and that we are interested in estimating the conditional expectation $m(x) = \mathbb{E}(Y \mid X = x)$. We consider here the local linear estimator $\hat{m}$, as defined in Slide 14 of Lecture 3.

Throughout the assignment, we will assume homoscedasticity, i.e., the local variance $\sigma^2(x) = \mathbb{V}(Y \mid X = x) \equiv \sigma^2$, as well as a quartic (biweight) kernel for $\hat{m}$. Under these assumptions, we know that the optimal bandwidth minimising the asymptotic mean integrated squared error is given by

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

where the two unknown quantities σ^2 and θ_{22} can be estimated by parametric OLS. For instance, one can

- Block the sample in N blocks and fit, in each block j, the model

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

  to obtain estimate

  $$\hat{m}_j = \hat{\beta}_{0j} + \hat{\beta}_{1j}x_i + \hat{\beta}_{2j}x_i^2 + \hat{\beta}_{3j}x_i^3 + \hat{\beta}_{4j}x_i^4$$

- Estimate the unknown quantities by

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

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

Task

The goal is to perform a simulation study to assess the impact of some parameters/hyperparameters on the optimal bandwidth h_{AMISE}. For instance, we will assume the following setting for the simulation study

- a covariate X from a beta distribution Beta (α, β)
- a response values $Y = m(X) + \epsilon$ where
  - the regression function m is given by $\sin\left\{\left(\frac{x}{3} + 0.1\right)^{-1}\right\}$

  - $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- fix σ^2 at some visually appealing value (e.g., $\sigma^2 = 1$ should be fine)

From there, estimate h_{AMISE} as described above and in Lecture 3 for different values of the following parameters/hyperparameters

- the sample size n (to assess the impact of the amount of available information),
- the block size N in the estimation of the unknown quantities σ^2 and θ_{22}, and
- the parameters α and β of the beta density of the covariate (to assess the impact of the shape of the distribution of the covariate).

Comment and report your findings using appropriate visualisation tools. Possible questions to address:

- How does h_{AMISE} behave when N grows? Can you explain why?
- Should N depend on n? Why?
- What happens when the number of observations varies a lot between different regions in the support of X? How is this linked to the parameters of the Beta distribution?

When assessing the effect of the sample size n or the density support of the covariate X on the optimal bandwidth h_{AMISE}, you can fix the value of N at an optimal value. This value could be considered as optimal in the sense that it minimizes the Mallow's C_p

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

where

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

and $N_{\max} = \max\{\min(\lfloor n/20 \rfloor, 5), 1\}$; see Ruppert et al. (1995) for choosing the optimal block size.

Results

Figure 1 gives an idea of the distribution of the data. $\sigma^2 = 0.5$ was chosen and kept fixed during the whole report.

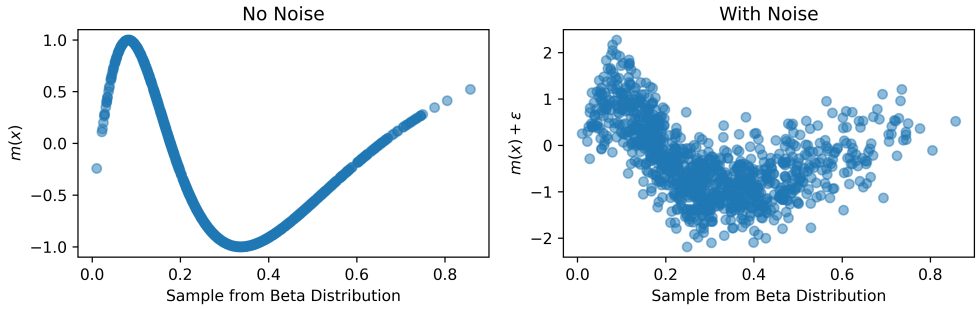

Figure 1: Scatter of the true $m(x)$ (left) versus noisy observations $m(x)+\varepsilon$ (right) for $n = 1,000$ draws from Beta$(2, 5)$, with $\sigma^2 = 0.5$.

Looking at Figure 2, we can see that as N increases, h_{AMISE} decreases.

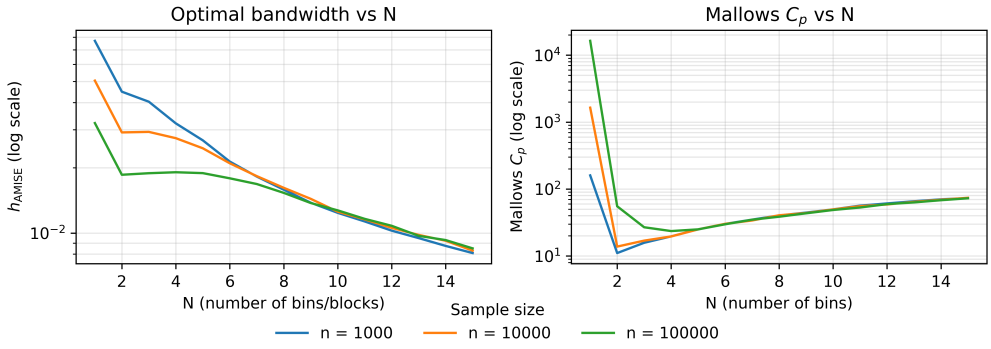

Figure 2: (Left) AMISE-bandwidth shrinks with more bins N; (right) Mallows C_p rises, both shown for $n \in \{1,000,\ 10,000\ 100,000\}$, $\alpha = 2$, $\beta = 5$ (mean across 100 simulations).

Intuitively, with very few blocks, quartic fits cover wide ranges and underestimate curvature, so $\hat{\theta}_{22}$ is too small and $\hat{h}_{AMISE}$ is large. Increasing N lets the fit track the curvature better, and $\hat{\theta}_{22}$ increases towards its target, leading to a drop in $\hat{h}_{AMISE}$.

with very few blocks the quartic fits are forced to cover large ranges of x, which leads to an underestimation of $\hat{\theta}_{22}$, pushing $\hat{h}$ upwards. As N increases, the fitted model tracks better the local curvature, leading to higher $\hat{\theta}_{22}$ and lower $\hat{\sigma}^2$, which translate to lower $\hat{h}$.

This trend is captured by Figure 3, that shows that our estimate $\hat{\sigma}$ is independent of N, while $\hat{\theta}_{22}$ increases when N increases.

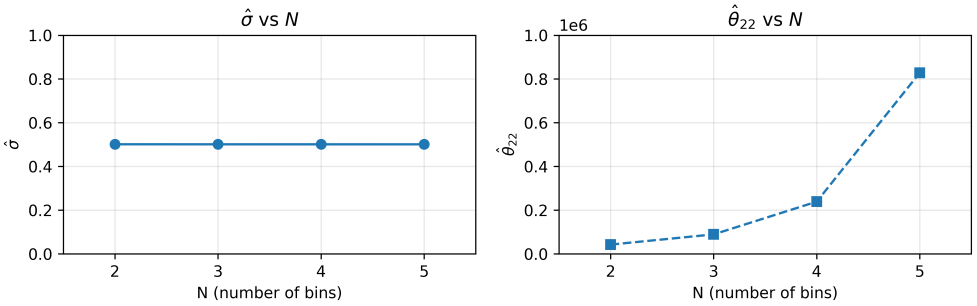

Figure 3: Mean of 100 runs, with $n = 1,000$, $\alpha = 2$, $\beta = 5$, $\sigma^2 = 0.5$.

Figure 2 also shows that, all else being equal, larger sample sizes n lead to a lower h_{AMISE}. Looking at the right panel, it seems that the optimal value number of bins N might depend on n. To analyze this further, we plot the optimal value of N, defined as the *argmin* of C_p, vs n (Figure 4). The figure shows that the optimal N increases logarithmically as n grows.

Intuitively, as we get more data we can choose narrower buckets that still contain plenty of points and allow to obtain a better fit for the model.

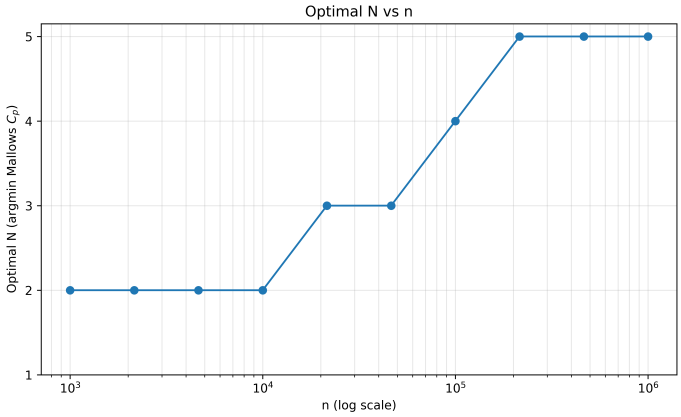

Figure 4: Optimal bin count chosen by Mallows C_p, for $\alpha = 2$, $\beta = 5$, averaged across 10 repetitions.

Figure 5 is a good way to visualize the result explained above. In this case, we fit the quartic models using different N and n. When $n = 1,000$, using $N = 5$ produces a wiggly result due to having few points in each bucket. But when $n = 1,000,000$, this effect is mitigated by the number of points in each bucket, and the fit with $N = 5$ is less wiggly.

Regarding the relation between N and the shape of the beta distribution, Figure 6 shows that when α is high and β is low h_{AMISE} is higher, while when α is low and β is high h_{AMISE} is lower.

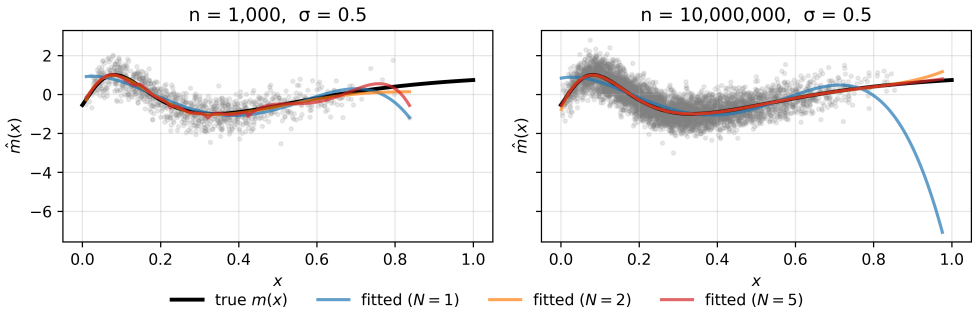

Figure 5: Piece-wise quartic fits compared to the true curve for $n = 1,000$ (left) and $n = 10,000,000$ (right); larger samples tighten the fit.

Since h_{AMISE} represents the global bandwidth, it tends to be higher than ideal in regions with lots of data, and lower than ideal in regions with few datapoints. Looking at the formula for θ_{22}, $\hat{m}''(x)$ is weighted by $f_X(x)$, meaning that regions with higher density are assigned an higher weight. Since $|m''(x)|$ is largest near 0 and smallest near 1:

- when the density of the distribution is higher around 0 (small α), $\hat{\theta}_{22}$ increases and thus $\hat{h}_{AMISE}$ decreases.
- when the density is higher around 1 (small β), $\hat{\theta}_{22}$ decreases and $\hat{h}_{AMISE}$ increases.

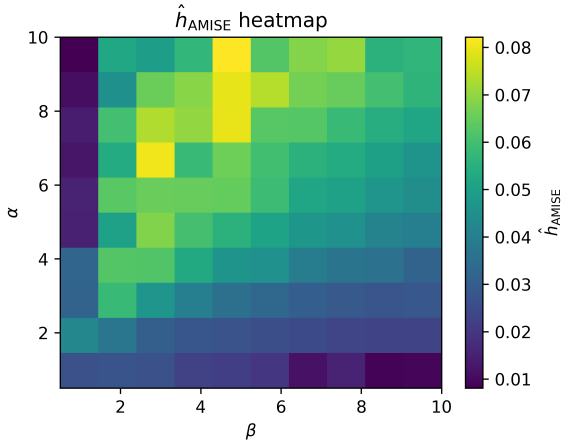

Figure 6: Heatmap of h_{AMISE} across (α, β) for $n = 10,000$ and $N = 2$.