MATH-517: Assignment 3

AUTHOR
Loan Bianchi

PUBLISHED
October 4, 2025

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

We begin by recalling the definition of our estimator :

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

Let us actually solve 1 and 2 at the same time. We know from slide 14 of the lecture notes that we want to reach that $\hat{m}(x) = \hat{\beta}_0 = \sum_{i=1}^{n} w_{ni}(x) Y_i$, for $w_{ni}(x) = \frac{1}{nh} \frac{K\left(\frac{x - X_i}{h}\right)(S_{n,2}(x) - (X_i - x)S_{n,1}(x))}{S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)}$, with $S_{n,k}(x) = \frac{1}{nh} \sum_{i=1}^{n} (X_i - x)^k K\left(\frac{X_i - x}{h}\right), \quad k = 0, 1, 2,$.

Now the real challenge is to understand how we get to that formula. Firstly, let us denote $f(\beta) = \sum_{i=1}^{n} \Big(Y_i - \beta_0 - \beta_1(X_i - x)\Big)^2 K\left(\frac{X_i - x}{h}\right)$, with $\beta = (\beta_0, \beta_1)$. Defining A as the matrix with the i th row being $A_i = (1, X_i - x)$ and D the diagonal matrix with entries $D_{ii} = K\left(\frac{X_i - x}{h}\right)$, we have $f(\beta) = (Y - A\beta)^\top D(Y - A\beta)$. As a composition of a linear function with a quadratic function, f is convex and we can differentiate and equate to 0 to find the minimizer.

We get the following system of equations :

$$-2 \sum_{i=1}^{n} \Big(Y_i - \hat{\beta}_0 - \hat{\beta}_1(X_i - x)\Big) K\left(\frac{X_i - x}{h}\right) = 0$$

$$-(X_i - x)2 \sum_{i=1}^{n} \Big(Y_i - \hat{\beta}_0 - \hat{\beta}_1(X_i - x)\Big) K\left(\frac{X_i - x}{h}\right) = 0$$

Or simplifying and rearranging :

$$\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i = \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)(\hat{\beta}_0 + \hat{\beta}_1(X_i - x)) \Leftrightarrow \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i = S_{n,0}(x)\hat{\beta}_0 + S_{n,1}(x)\hat{\beta}_1$$

$$\sum_{i=1}^{n} (X_i - x) K\left(\frac{X_i - x}{h}\right) Y_i = \sum_{i=1}^{n} (X_i - x) K\left(\frac{X_i - x}{h}\right)(\hat{\beta}_0 + \hat{\beta}_1(X_i - x)) \Leftrightarrow \frac{1}{nh} \sum_{i=1}^{n} (X_i - x) K\left(\frac{X_i - x}{h}\right) Y_i = S_{n,1}(x)\hat{\beta}_0 + S_{n,2}(x)\hat{\beta}_1$$

This can be rewritten as

$$\begin{pmatrix} S_{0,n}(x) & S_{1,n}(x) \\ S_{1,n}(x) & S_{2,n}(x) \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i \\ \frac{1}{nh} \sum_{i=1}^{n} (X_i - x) K\left(\frac{X_i - x}{h}\right) Y_i \end{pmatrix}.$$

As long as $S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)$ is non zero, we can use the inversion formula for a 2×2 matrix to get $\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \frac{1}{S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)} \begin{pmatrix} S_{2,n}(x) & -S_{1,n}(x) \\ -S_{1,n}(x) & S_{0,n}(x) \end{pmatrix} \begin{pmatrix} \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i \\ \frac{1}{nh} \sum_{i=1}^{n} (X_i - x) K\left(\frac{X_i - x}{h}\right) Y_i \end{pmatrix}$

Hence, we get $\hat{\beta}_0 = \sum_{i=1}^{n} w_{ni}(x) Y_i$, for $w_{ni}(x) = \frac{1}{nh} \frac{K\left(\frac{x - X_i}{h}\right)(S_{n,2}(x) - (X_i - x)S_{n,1}(x))}{S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)}$, which concludes 1 and 2.

Now, let us prove that these weights indeed sum up to 1.

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

$= nhS_{n,0}(x)S_{n,2}(x) - S_{n,1}(x) \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)(X_i - x) = nh(S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x))$, which concludes the third point.

2. Practical exercise: Global bandwidth selection

Important note : This analysis include an interactive shiny part. To make it work, it should be enough to open the .qmd file with Rstudio and simply click on "run document" (after having installed the shiny package if not already installed). In this provided pdf, we included static plots so that the user can see what it should look like, but it's supposed to be interactive.

(I tried it on two computers to make sure that everything was fine and it worked for both so I sincerely hope it will work).

The aim of this simulation study is to analyse the behavior of the optimal bandwidth h_{opt} under different conditions; with samples of different sizes n, with different number of blocks N used to estimates the quantities that we need, as well as with different distributions. In order to study this optimal bandwidth, we will make use of the beta distribution $beta(\alpha, \beta)$ which will reveal itself very useful for that simulation study, in the sense that varying parameters α and β leads to distributions that have density functions with various shapes.

This simulation was constructed in a way that allows the reader to capture a maximum of information out of the simplest possible plot. After considering many options, the one that seemed to be the most convincing was to make a shiny interactive plot, which simply shows a vertical line representing the value of h with h_{opt} evolving on it when varying the parameters. The reason for this quite surprising choice of a single line with a point is to make the reader focus on what really matters in that study and not get lost in many other values which can be relevant for other purposes but not for this study. Nonetheless, we decided to include the estimated $\hat{\sigma}^2$ next to the point, so that we can follow how the variance is estimated for the values of the parameters. It is also a way to be sure that our variance is well-estimated and that our value of h_{opt} is coherent.

This simulation study will have two parts: the first one will be meant to study how our bandwidth behaves when we vary α, β, and n, but we "fix" N to be optimal according to the formula given in the assignment task, i.e the N minimizing

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

where

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

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

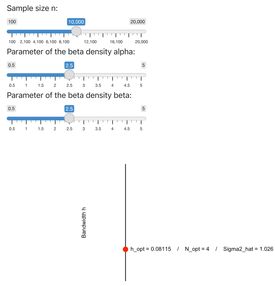
This first part is thought so that we can understand how these varying parameters impact the optimal bandwidth, knowing that we have a "good" N. Here, I say "good" because the paper mentioned (Ruppert et al. (1995)) tells us that $N > 5$ can also be taken into consideration when we have a regression function m with many oscillations, which should be the case in our set-ups when x is small... Now, for the simplicity of this study, we assume in the first part that $N \leq 5$.

In the second part of this study, we will still be able to make α, β, and n vary, but we can now also let N vary (and allows $N > 5$). It is meant to help us understand how the number of blocks in our estimation impacts the bandwidth for some fixed parameters α, β, and n. The idea is to fix these parameters (at values of the reader's choice) and make N vary to see how our point evolves along the line. We also thought that it could be interesting to show what the optimal N is next to the point so that we can know how far from the optimal value we get when we vary N. In this part, we also include $\hat{\theta}_{22}$ next to the point. The reason for that choice will be made clearer in the analysis below.

An important precision is that the simulation is made so that when we vary n, we don't have a new sample every time, but we add or remove observations from our previous sample. The idea behind this is to make it more relatable to real life, where we often have a fixed sample of observations, to which we can add some observations if we get new ones over time. The idea is then to study how the bandwidth evolves when we add those new observations.

Another precision is that if N doesn't divide the size of the sample, we distribute the remaining observations evenly among the blocks (so that each block gets at most one of these remaining observations)

First part of the simulation - optimal bandwidth h_{opt} for different values of α, β and n (N automatically set to the optimal value)



For the first part, we can see that if we fix α and β, our h_{opt} decreases when we make n grow. This makes sense as the more observations we have, the more points we have close to one another. We can thus make h smaller so that our kernel gives more importance to points close to x, as there should be plenty of them. On the contrary, if we have less observations, we might have more variance and giving a lot of importance only to very close points might lead to a dangerous game as there might not be many of them.

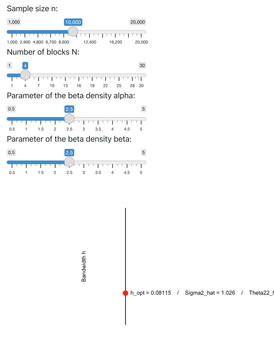
Now, let us fix n (say, at 10'000) and observe the behavior of h_{opt} when we make the density distribution of X vary.

If we fix $\alpha = 0.5$ and make β vary, we see that the bigger the β, the smaller the h_{opt}. This is because when β is big compared to α, the $beta$ distribution is skewed near 0. Hence, as all the observations are (for most of them) concentrated in a small interval, we need smaller h to captures the importance of all points close to each other, especially as our m function oscillates very quickly for x near 0. Interestingly, we see that when we set $\alpha = \beta = 0.5$, we get a high h_{opt}. A theory for this is that as our X_is are concentrated both near 0 and near 1, our m function oscillates very quickly for x near 0 but less quickly near 1, do we kind of need to have a higher h to estimate our function near 1, but that leads to a quite poor estimate, which we can see in the quite bad estimation for the variance.

If we fix $\beta = 0.5$ and make α vary, we see that the bigger the α, the bigger the h_{opt}. This makes sense as the more X_is there are near one, the less oscillations of our function m will intervene and we will need to give importance to X_is a bit further from our x to understand these "slow" oscillations.

If we fix $\alpha = \beta = 2.5$, we get an intermediate value, which again makes sense as most of our X_is with be near 0.5 and is thus in-between the two extremes mentioned earlier.

Second part of the simulation - optimal bandwidth h_{opt} for different values of α, β, n and N



For the second part, we see that no matter what value of the parameters n, α and β we choose, we get that the bigger the number of blocks, the smaller the h_{opt}. This could be explained by the fact that if we have a lot of blocks, our $\hat{\theta}_{22}$ will be bigger. Indeed, splitting into a lot of blocks with induce the fact that the linear regressions will be only on small parts of the sinus oscillations, making the curvature of the estimate bigger compared to if the regression were made over a lot of oscillations. A bigger $\hat{\theta}_{22}$ leads in turn to a smaller h_{opt} if we simply look at the formula for computing h_{opt}.