MATH-517: Assignment 3

Filipp Sekatski

5 October 2025

Theoretical exercise

We have i.i.d. samples (x_i, y_i) for $i = 1, \dots, n$ from the model

$$y_i = m(x_i) + \epsilon_i$$

and we want to estimate m with some function $\hat{m}$. For a Kernel K and a bandwidth $h > 0$, we define the local linear regression estimator at a point x by

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

We then have the fitted value $\hat{m}(x) = \hat{\beta}_0(x)$.

1) We want to find

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \arg\min_{\beta \in \mathbb{R}^2} (Y - X\beta)^T W (Y - X\beta)$$

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

Writing

$$f(\beta) = (Y - X\beta)^T W (Y - X\beta) = Y^T W Y - 2\beta^T X^T W Y + \beta^T X^T W X \beta$$

and setting the gradient to be zero (as usual for LS estimators), we find

$$2X^T W Y = 2X^T W X \hat{\beta} \implies \hat{\beta} = (X^T W X)^{-1} X^T W Y.$$

Note that W is positive definite so that $X^T W X$ is invertible if X has full rank. Hence we see that $\hat{m} = \hat{\beta}_0$ can be written as a weighted average of the observations

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

where $w_{ni}(x)$ depends only on x, X_i's, K and h.

2) Now we want to find the exact weights $w_{ni}(x)$. To simplify the notations, let $K_i = K\left(\frac{X_i - x}{h}\right)$ and $Z_i = X_i - x$ for $i = 1, \dots, n$. Also let $A_k = \sum_{i=1}^n Z_i^k K_i$ for $k = 0, 1, 2$. We have

$$\begin{aligned}
\hat{\beta} &= \left(\begin{bmatrix} 1 & \dots & 1 \\ Z_1 & \dots & Z_n \end{bmatrix} \operatorname{diag}(K_1, \dots, K_n) \begin{bmatrix} 1 & Z_1 \\ \vdots & \vdots \\ 1 & Z_n \end{bmatrix} \right)^{-1} \\
&\quad \begin{bmatrix} 1 & \dots & 1 \\ Z_1 & \dots & Z_n \end{bmatrix} \operatorname{diag}(K_1, \dots, K_n) Y \\
&= \left(\begin{bmatrix} K_1 & \dots & K_n \\ Z_1 K_1 & \dots & Z_n K_n \end{bmatrix} \begin{bmatrix} 1 & Z_1 \\ \vdots & \vdots \\ 1 & Z_n \end{bmatrix} \right)^{-1} \begin{bmatrix} K_1 & \dots & K_n \\ Z_1 K_1 & \dots & Z_n K_n \end{bmatrix} Y \\
&= \begin{bmatrix} \sum_{i=1}^n K_i & \sum_{i=1}^n K_i Z_i \\ \sum_{i=1}^n K_i Z_i & \sum_{i=1}^n K_i Z_i^2 \end{bmatrix}^{-1} \begin{bmatrix} K_1 & \dots & K_n \\ Z_1 K_1 & \dots & Z_n K_n \end{bmatrix} Y \\
&= \frac{1}{A_0 A_2 - A_1^2} \begin{bmatrix} A_2 & -A_1 \\ -A_1 & A_0 \end{bmatrix} \begin{bmatrix} K_1 & \dots & K_n \\ Z_1 K_1 & \dots & Z_n K_n \end{bmatrix} Y \\
&= \frac{1}{A_0 A_2 - A_1^2} \begin{bmatrix} A_2 K_1 - A_1 Z_1 K_1 & \dots & A_2 K_n - A_1 Z_n K_n \\ \dots & \dots & \dots \end{bmatrix} Y.
\end{aligned}$$

Hence by setting $S_{n,k}(x) = \frac{1}{nh} \sum_{i=1}^n (X_i - x)^k K\left(\frac{X_i - x}{h}\right) = \frac{1}{nh} A_i$ for $i = 0, 1, 2$, we get that

$$w_{nj}(x) = \frac{1}{nh} \frac{K_j \left(S_{n,2}(x) - S_{n,1}(x) Z_1 \right)}{S_{n,0}(x) S_{n,2}(x) - S_{n,1}(x)^2}$$

for each $j = 1, \dots, n$. That is

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

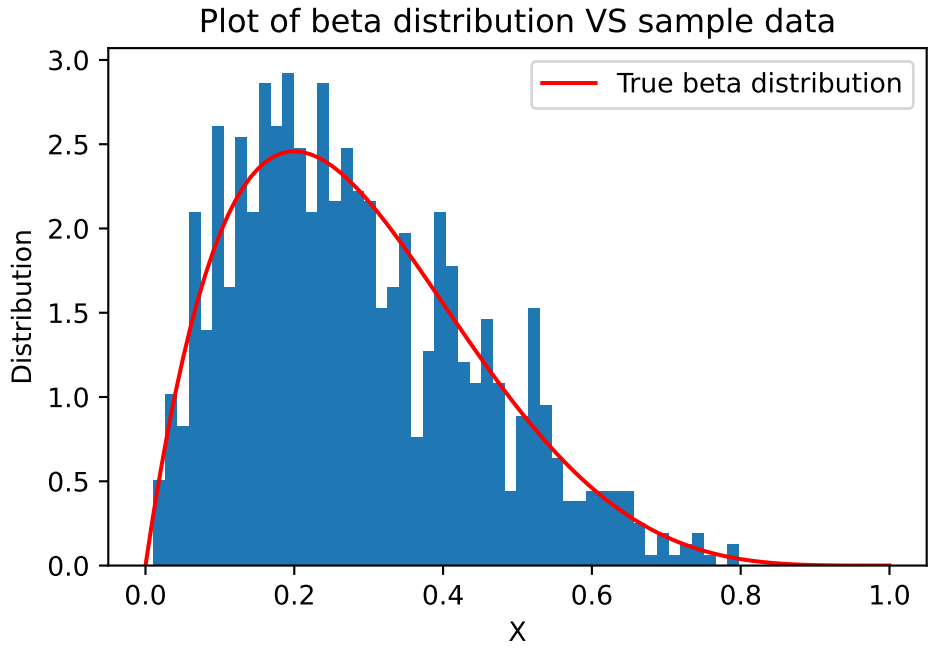
3) Finally we want to show that the weights satisfy $\sum_{i=1}^n w_{ni}(x) = 1$. For this we write

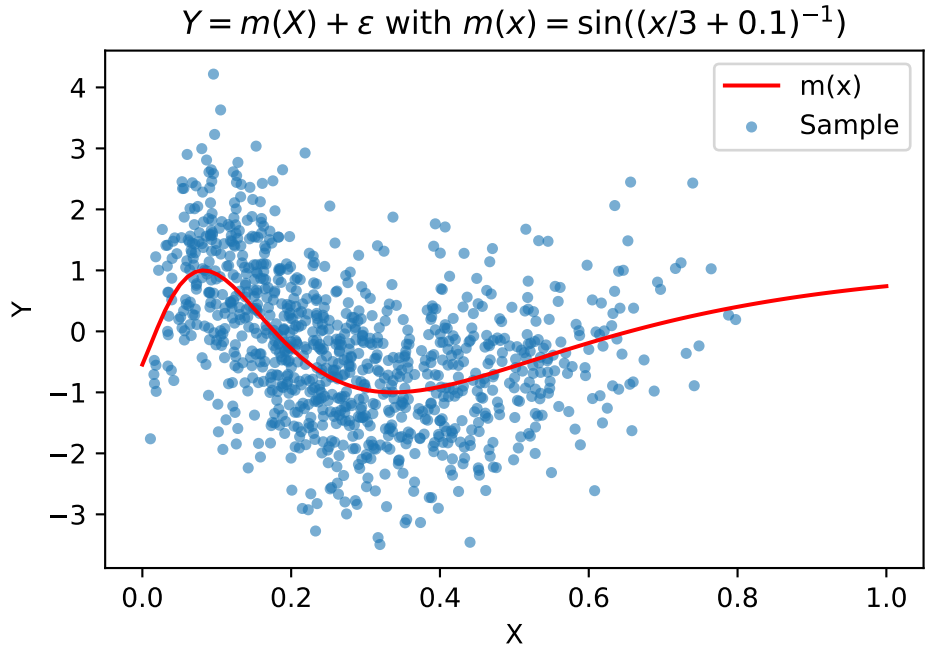
$$\begin{aligned}
\sum_{i=1}^n w_{ni}(x) &= \frac{1}{A_0 A_2 - A_1^2} \sum_{i=1}^n (K_i A_2 - K_i A_1 Z_i) \\
&= \frac{1}{A_0 A_2 - A_1^2} \left(A_2 \sum_{i=1}^n K_i - A_1 \sum_{i=1}^n K_i Z_i \right) \\
&= \frac{A_2 A_0 - A_1 A_1}{A_0 A_2 - A_1^2} = 1.
\end{aligned}$$

Practical exercise

We have $(X_i, Y_i)_{i=1}^n$ i.i.d. random vectors where $X_i \sim \text{Beta}(\alpha, \beta)$ and $Y_i = m(X_i) + \epsilon_i$ with $m(x) = \sin((x/3 + 0.1)^{-1})$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Here is an example of such data for

$n = 1000, \alpha = 2, \beta = 5$ and $\sigma^2 = 1$:





But now imagine that we don't know m and that we have just access to our sample. We could be interested in estimating $m(x) = \mathbb{E}[Y|X = x]$. Under assumptions of homoscedasticity and

quartic kernel, the optimal bandwidth mnimizing MISE asymptotically is given by

$$h_{\text{AMISE}} = n^{-0.2}\left(\frac{35\sigma^2|\operatorname{Supp}(X)|}{\theta_{22}}\right)^{0.2}$$

where $\theta_{22} = \int (m^{''}(x))^2 f_X(x)\,\mathrm{d}x$. Not having any information other than the sample, we estimate θ_{22} and σ^2 by performing a linear regression on the data. More precisely, we divide the data into N blocks $\chi_1, \dots, \chi_N$ and for each block $j = 1, \dots, n$ we fit 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 $\hat{m}_j(x_i) = \hat{\beta}_{0j} + \hat{\beta}_{1j}x_i + \hat{\beta}_{2j}x_i^2 + \hat{\beta}_{3j}x_i^3 + \hat{\beta}_{4j}x_i^4$. We then have the estimators

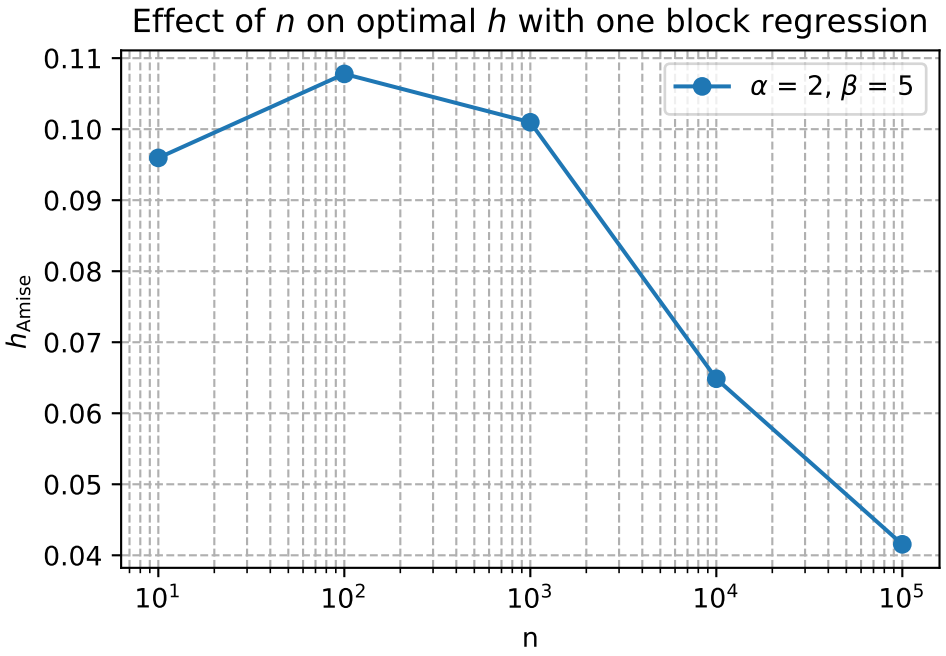
$$\hat{\theta}_{22}(N) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N}(\hat{m}_j^{''}(X_i))^2\{X_i \in \chi_j\}$$

and

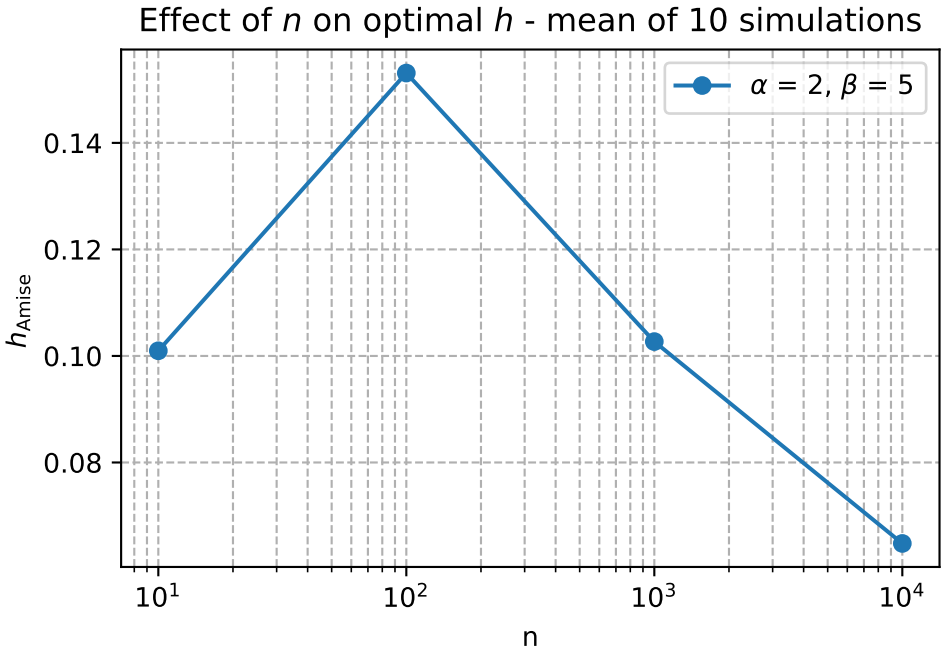
$$\hat{\sigma}^2(N) = \frac{1}{n-5N}\sum_{i=1}^{n}\sum_{j=1}^{N}(Y_i - \hat{m}_j(X_i)^2)\{X_i \in \chi_j\}$$

We are interested in the effect on h_{AMISE} and our estimators if we vary the size of the data n, the number of blocks N or the parameters of the true distribution α, β. From now on, we fix $\sigma^2 = 1$.

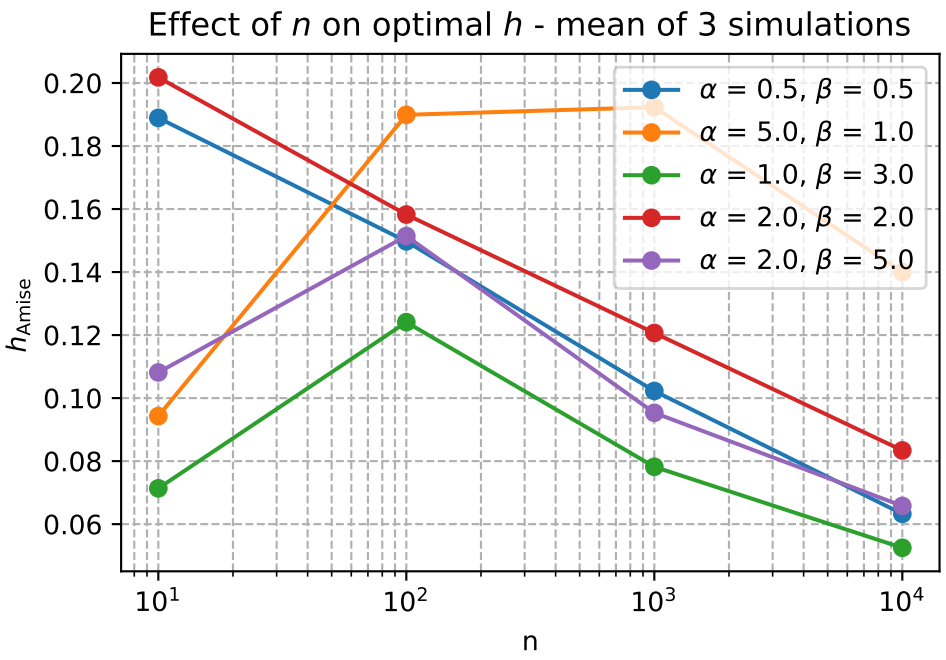
Let's first fix $\alpha = 2$ and $\beta = 5$ and fit the model with only $N = 1$ block. We will see what happens with one simulation as we vary the size of the sample n.



We see that in our simulated data, the optimal bandwidth for $n = 10$ is smaller than for $n = 100$. It may just be due to randomness and not really reflect a true trend, because 10 data points is very small. We do several simulations and compute the mean in order to correct for randomness.

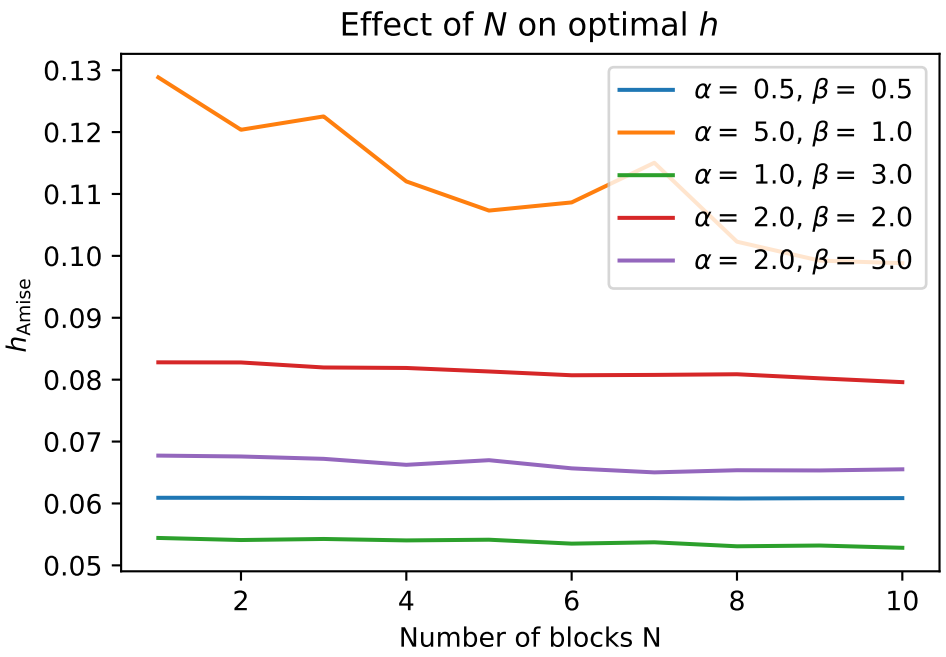


So in fact we do see that h increases at first before decreasing. Let's now look at what happens for different choices of parameters α and β. We make 3 simulations for each distribution.



The trend that we can see is that even if h seems to increase at first, it will eventually decrease as n grows for all distributions.

Now let's fix $n = 10000$ and vary the number of blocks N. We will look on the effect it has on h for the same simulated data.



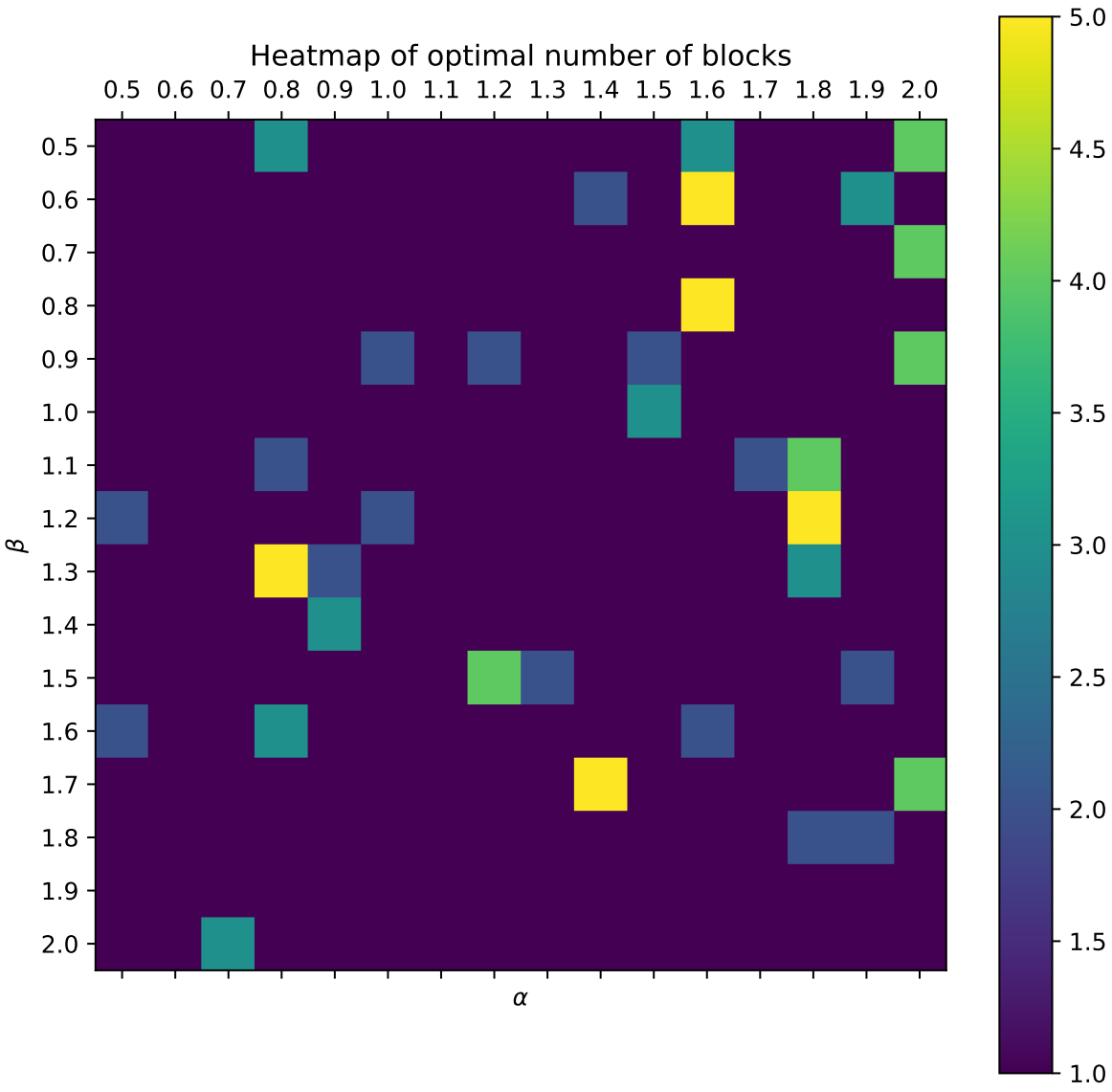
It seems that the number of blocks doesn't influence the optimal bandwidth very significatively. One could ask when is it actually more optimal to split the data into blocks. There is an optimal number of blocks N depending on n, that minimizes

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

where

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

and $N_{\max} = \max\{\min\{\lfloor n/20 \rfloor, 5\}, 1\} \leq 5$. So we see that we limit ourselves with at most 5 blocks. We can simulate some data for different distributions and compute the optimal N in each case. For $n = 1000$ and α, β ranging from 0.5 to 2.0 in steps of 0.1, we have the following optimal number of blocks:



We don't see any particular pattern. It seems that we cannot know a priori what will be the optimal number of blocks. But it does help in some cases to divide the data into blocks.

Finally we have a mean estimated σ^2 in our case.

```
On average, our estimated sigma squared is 1.0597476431275843.
```

We have seen how the different parameters affect the optimal bandwidth, and confirmed that dividing the data in blocks does help in some cases.