

MATH-517: Assignment 3

Duncan Bleich

2025-10-03

Theoretical exercise

First notice that we can rewrite

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

as

$$\left(\hat{\beta}_0(x), \hat{\beta}_1(x)\right) = \arg \min_{\beta_0, \beta_1 \in \mathbb{R}} \left\| \mathbf{W}^{1/2} (\vec{Y} - \mathbf{X}\vec{\beta}) \right\|^2 \quad (1)$$

where

$$\mathbf{W} = \text{diag} \left(K\left(\frac{X_1 - x}{h}\right), \dots, K\left(\frac{X_n - x}{h}\right) \right), \quad \vec{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix},$$

and

$$\mathbf{X} = \begin{pmatrix} 1 & X_1 - x \\ 1 & X_2 - x \\ \vdots & \vdots \\ 1 & X_n - x \end{pmatrix}.$$

This is a regular weighted least square problem, whose solution is given by

$$\hat{\beta} = \left[\mathbf{X}^T \mathbf{W} \mathbf{X} \right]^{-1} \mathbf{X}^T \mathbf{W} \vec{Y},$$

when $\left[\mathbf{X}^T \mathbf{W} \mathbf{X} \right]$ is invertible. Indeed, we can write (??) in a linear model form with

$$\vec{Y}' = \mathbf{W}^{1/2} \vec{Y}, \quad \mathbf{X}' = \mathbf{W}^{1/2} \mathbf{X},$$

whose least square estimator takes the form

$$\begin{aligned}\hat{\beta} &= [\mathbf{X}'^T \mathbf{X}']^{-1} \mathbf{X}'^T \vec{Y}' \\ &= [\mathbf{X}^T (\mathbf{W}^{1/2})^T \mathbf{W}^{1/2} \mathbf{X}]^{-1} \mathbf{X}^T (\mathbf{W}^{1/2})^T \mathbf{W}^{1/2} \vec{Y} \\ &= [\mathbf{X}^T \mathbf{W} \mathbf{X}]^{-1} \mathbf{X}^T \mathbf{W} \vec{Y}.\end{aligned}$$

This follows from the fact that $\mathbf{W} = \mathbf{W}^T$ and the matrix $\mathbf{W}^{1/2}$ with

$$(\mathbf{W}^{1/2})_{(i,j)} = \sqrt{\mathbf{W}_{(i,j)}}, \quad i, j \in \{1, \dots, n\}$$

is well-defined as \mathbf{W} has non-negative entries and as \mathbf{W} is symmetric, $\mathbf{W}^{1/2}$ is also symmetric. For this reasoning to work, we must assume \mathbf{X}' to be of full rank, i.e., of rank two in this particular case. Let $\mathbf{M} = [\mathbf{X}^T \mathbf{W} \mathbf{X}]^{-1} \mathbf{X}^T \mathbf{W} \in \mathbb{R}^{2 \times n}$, we get

$$\vec{\beta} = \mathbf{M} \vec{Y}$$

and since $\hat{m}(x) = \hat{\beta}_0(x)$, we are only interested in the first entry of $\vec{\beta}$. Define $e_1 = (1, 0)^T$, we have that

$$\begin{aligned}\hat{\beta}_0 &= e_1^T \mathbf{M} \vec{Y} \\ &= \sum_{i=1}^n \mathbf{M}_{(1,i)} Y_i,\end{aligned}$$

where $\mathbf{M}_{(1,i)}$, $i = 1, \dots, n$ depend on x, K, h and the X_j 's only, they do not depend on the Y_i 's. We can therefore express $\hat{m}(x)$ as a weighted average on the observations:

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

with $w_{ni}(x) = \mathbf{M}_{(1,i)}$ for $i = 1, \dots, n$. We now use the notation

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

in order to 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. To do so, we need to go back to our definition of \mathbf{M} :

$$\begin{aligned}
[\mathbf{X}^T \mathbf{W} \mathbf{X}]^{-1} &= \left[\begin{pmatrix} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) & \sum_{i=1}^n (X_i - x) K\left(\frac{X_i - x}{h}\right) \\ \sum_{i=1}^n (X_i - x) K\left(\frac{X_i - x}{h}\right) & \sum_{i=1}^n (X_i - x)^2 K\left(\frac{X_i - x}{h}\right) \end{pmatrix} \right]^{-1} \\
&= \begin{pmatrix} nh S_{n,0}(x) & nh S_{n,1}(x) \\ nh S_{n,1}(x) & nh S_{n,2}(x) \end{pmatrix}^{-1} \\
&= \frac{1}{nh[S_{n,0}S_{n,2} - S_{n,1}^2]} \begin{pmatrix} S_{n,2} & -S_{n,1} \\ -S_{n,1} & S_{n,0} \end{pmatrix}
\end{aligned}$$

and

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

From this, we get that

$$\mathbf{M}_{1,i} = w_{ni}(x) = \frac{S_{n,2}(x)K([X_i - x]/h) - S_{n,1}(X_i - x)K([X_i - x]/h)}{nh[S_{n,0}S_{n,2} - S_{n,1}^2]}.$$

Finally, summing over i gives

$$\begin{aligned}
\sum_{i=1}^n w_{ni} &= \frac{S_{n,2}(x) \sum_{i=1}^n K([X_i - x]/h) - S_{n,1}(x) \sum_{i=1}^n (X_i - x)K([X_i - x]/h)}{nh[S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)]} \\
&= \frac{nhS_{n,0}(x)S_{n,2}(x) - nhS_{n,1}(x)S_{n,1}(x)}{nh[S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)]} \\
&= 1.
\end{aligned}$$

Practical exercise

We consider a sample $\{(X_i, Y_i)\}_{i=1}^n$ of i.i.d. random vectors. Our goal is to estimate the conditional expectation

$$m(x) = \mathbb{E}[Y \mid X = x],$$

using the local linear estimator \hat{m} , as introduced in Lecture 3 (Slide 14). Throughout this study, we assume homoscedasticity, i.e.,

$$\sigma^2(x) = \text{Var}(Y \mid X = x) \equiv \sigma^2,$$

and we use a quartic (biweight) kernel for \hat{m} . Under these assumptions, the optimal bandwidth minimizing the asymptotic mean integrated squared error (AMISE) is

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

Here, the quantities σ^2 and θ_{22} are unknown and will be estimated using parametric OLS. To do so, we partition the data into N blocks. In each block j , we fit the polynomial regression model

$$Y_i = \beta_{0j} + \beta_{1j}X_i + \beta_{2j}X_i^2 + \beta_{3j}X_i^3 + \beta_{4j}X_i^4 + \epsilon_i,$$

which yields the fitted regression function

$$\hat{m}_j(X_i) = \hat{\beta}_{0j} + \hat{\beta}_{1j}X_i + \hat{\beta}_{2j}X_i^2 + \hat{\beta}_{3j}X_i^3 + \hat{\beta}_{4j}X_i^4.$$

From this, the unknown parameters can be estimated as

$$\hat{\theta}_{22}(N) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^N \left(\hat{m}_j''(X_i) \right)^2 \mathbf{1}_{X_i \in \mathcal{X}_j},$$

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

The purpose of the simulation study is to examine how different parameters and hyperparameters influence the estimation of the optimal bandwidth h_{AMISE} . We generate data according to the following process:

- Covariate: $X \sim \text{Beta}(\alpha, \beta)$.
- Response: $Y = m(X) + \epsilon$, where

$$m(x) = \sin\left(\left(\frac{x}{3} + 0.1\right)^{-1}\right), \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

For simplicity, we fix the noise variance at $\sigma^2 = 1$.

To assess the impact of the sample size n on our estimate of h_{AMISE} , we will look at the values $n \in \{200, 400, 800, 1600\}$ and fix $\alpha, \beta = 2$. For each fixed n , we will take N to be the value that minimizes the Mallows'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 $N_{\max} = \max\{\min(\lfloor n/20 \rfloor, 5), 1\}$.

From the expression of h_{AMISE} , we would expect that doubling the sample size n , would shrink \hat{h}_{AMISE} by $2^{-1/5}$. Let us now look at how our estimate \hat{h}_{AMISE} changes for different values of n :

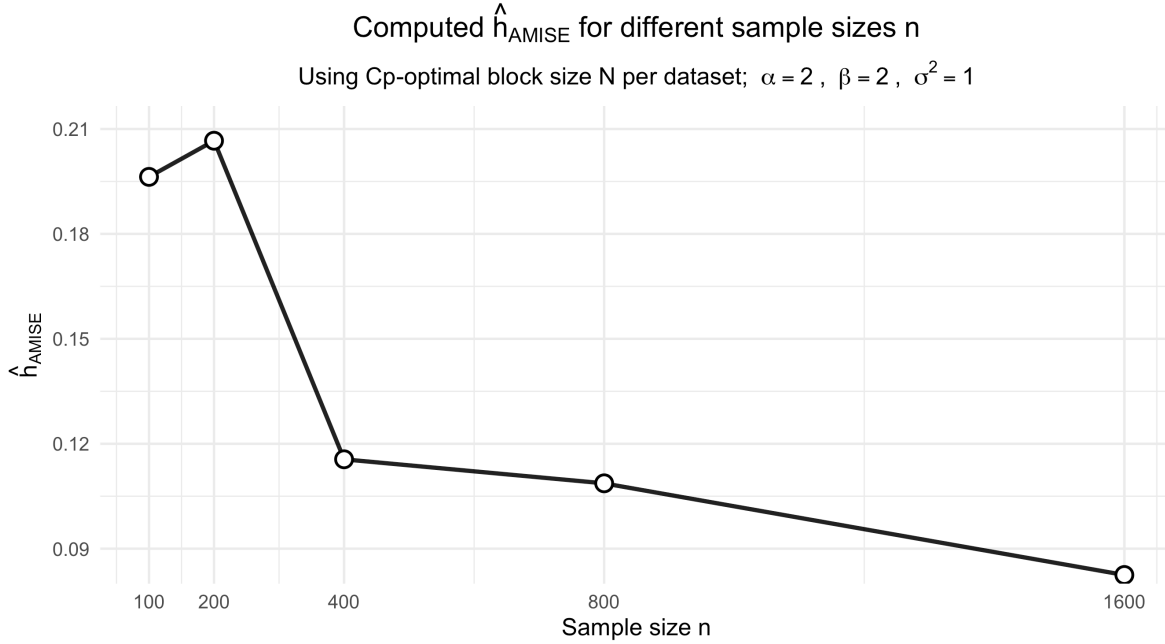


Figure 1: Estimated optimal bandwidth h versus different sample sizes $n=100, 200, 400, 800$ and 1600 . The estimated bandwidth seems to decrease as the sample size increases, which is what we expected would happen.

While it does look like the estimated bandwidth \hat{h}_{AMISE} shrinks as the sample size n grows, it is difficult to see where our theoretical $2^{-1/5}$ is. A more helpful way of looking at this would be by taking the log-log scale:

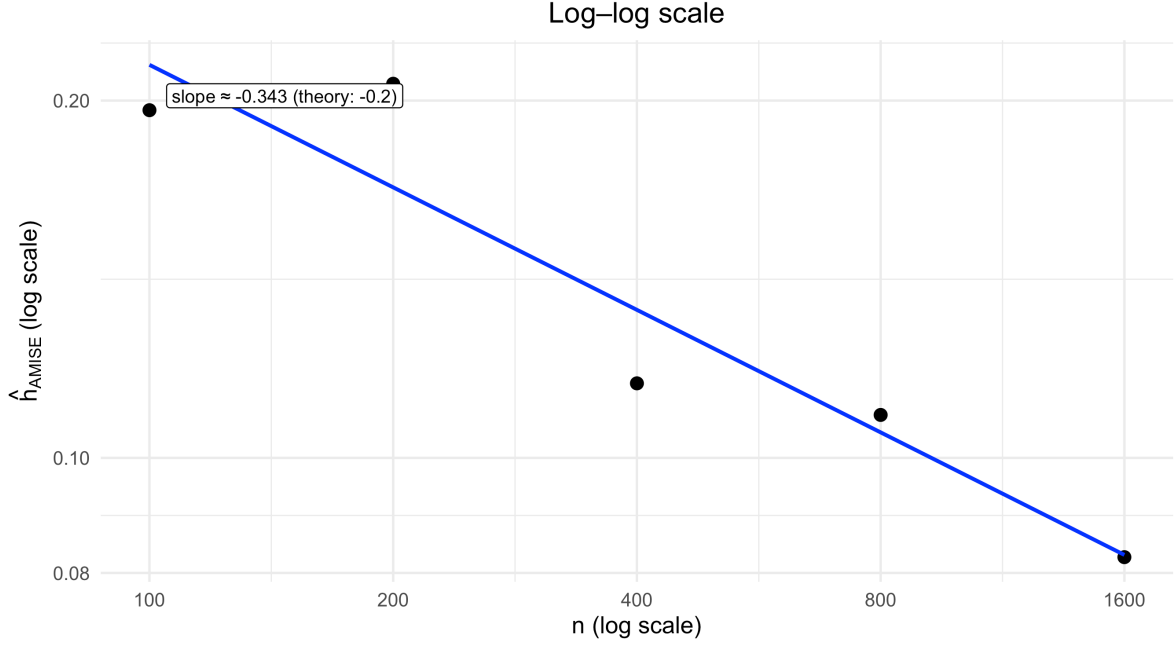


Figure 2: Log-log scaling of the plug-in bandwidth estimate \hat{h}_{AMISE} versus sample size n . Each point is an estimate computed for a given n (using the C_p -optimal block size N for that dataset). The straight line is an OLS fit in log-log space; its slope (shown on the plot) is close to the theoretical -0.2 , confirming the expected scaling $\hat{h}_{\text{AMISE}} \propto n^{-1/5}$. Axes are $\log(n)$ and $\log(\hat{h}_{\text{AMISE}})$.

We repeat the simulation $R = 400$ times for each sample size n . In each replicate, we resample (X, Y) , select the C_p -optimal block size N , and compute \hat{h}_{AMISE} .

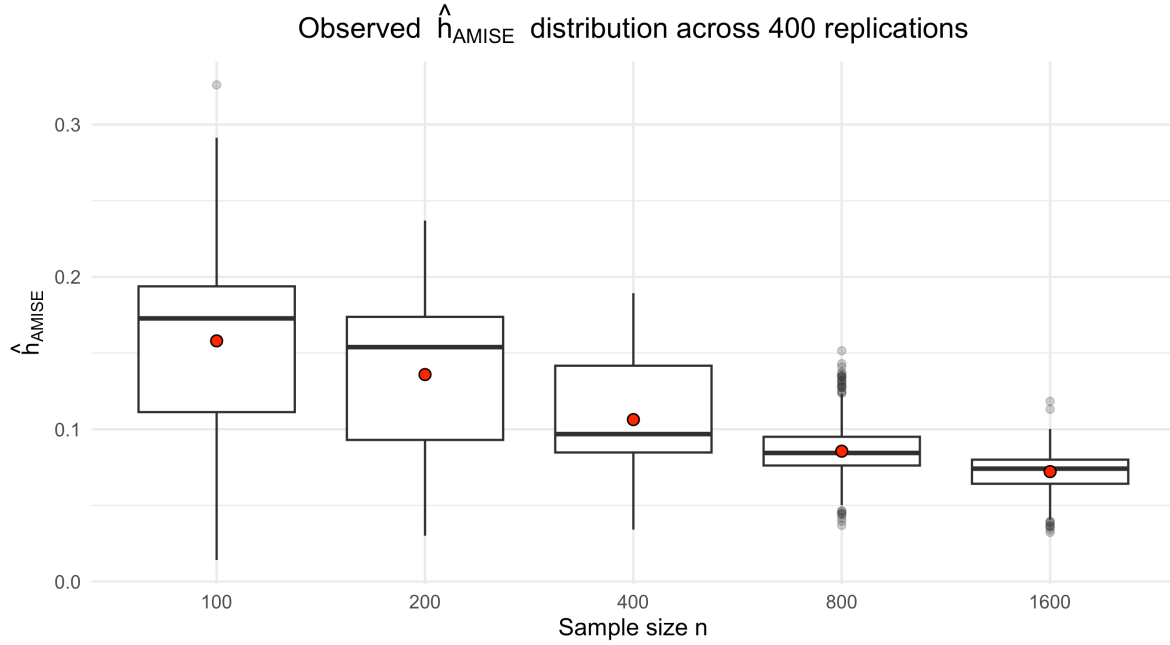


Figure 3: Boxplots of the plug-in bandwidth estimates \hat{h}_{AMISE} across $R = 400$ replications for each sample size $n \in \{200, 400, 800, 1600\}$. In each replicate we resample (X, Y) , choose the Cp-optimal block size N , and compute \hat{h}_{AMISE} . The overlaid dot indicates the mean across replications.

The boxplots above summarize the distribution across replications, which gives a clear view of variability. The red dots indicate the means across replications. Proceeding as before, we get the following plot on a log-log scale:

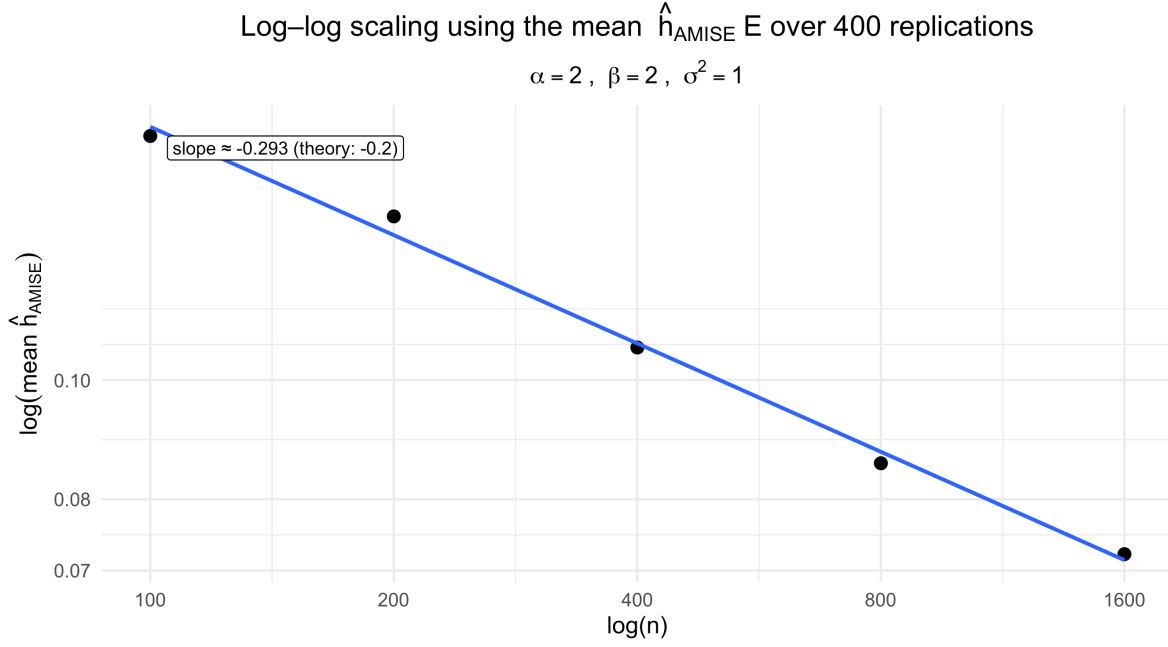


Figure 4: This plot shows $\log(\bar{\hat{h}}_{\text{AMISE}})$ versus $\log(n)$, where $\bar{\hat{h}}_{\text{AMISE}}$ is the mean over $R = 400$ replications at each sample size $n \in 200, 400, 800, 1600$. In each replicate we resample (X, Y) , choose the C_p -optimal block size N , and compute \hat{h}_{AMISE} . The straight line is an OLS fit in log-log space; its slope is close to the theoretical -0.2 , confirming $\hat{h}_{\text{AMISE}} \propto n^{-1/5}$.

This strengthens our previous assumption, namely, that $\hat{h}_{\text{AMISE}} \propto n^{-1/5}$. Let us now quickly look at how the optimal value of N evolves, when increasing n :

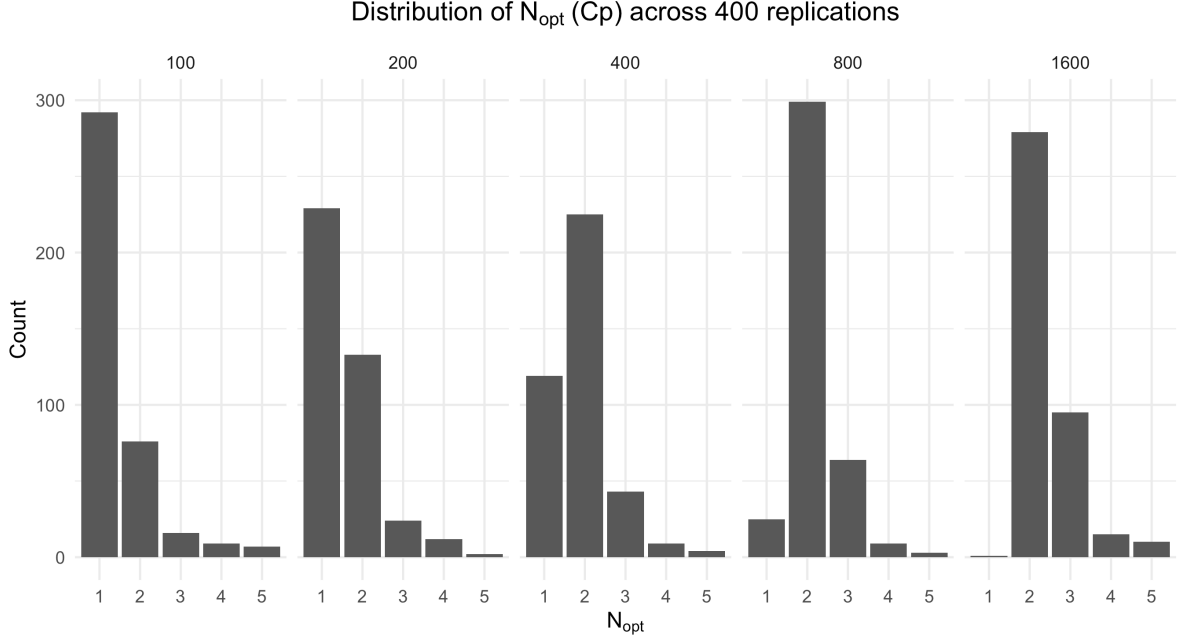


Figure 5: Estimated \hat{h}_{AMISE} vs. block size N by sample size n . Red points show the estimates computed from blockwise degree-4 fits for each block size N ; the dashed vertical line in each panel marks the value of N that minimizes Mallows' C_p .

From this, it seems like N should depend on the sample size n . Indeed, each block fit uses 5 parameters, so we require $n - 5N > 0$ and a sufficient number of observations *per block* (practically $\gtrsim 10$ – 15 in the sparsest block) for stable estimation. As n increases we can afford a larger N without starving blocks; empirically the C_p -optimal \hat{N}_{opt} increases slowly with n . We now want to look into the impact the chosen block size N , used in the estimation of θ_{22} and σ^2 , has on the estimate \hat{h}_{AMISE} . We will this time let the sample size take the values 200, 800, 3200 and 12800. Moreover, for each n , we let $N \in \{1, \dots, N_{\text{max}}\}$.