MATH-517: Assignment 3

Andriessen Balthazar

Invalid Date

Theoretical exercise

Question 1

Let

$$r_i := X_i - x, \qquad K_i := K\Big(\frac{X_i - x}{h}\Big), \qquad i = 1, \dots, n.$$

Define the local design matrix, weight matrix and response vector by

$$X = \begin{pmatrix} 1 & r_1 \\ \vdots & \vdots \\ 1 & r_n \end{pmatrix} \in \mathbb{R}^{n\times 2}, \qquad W = \operatorname{diag}(K_1, \dots, K_n) \in \mathbb{R}^{n\times n}, \qquad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \in \mathbb{R}^n.$$

The local linear estimator is the solution of the weighted least squares problem

$$(\hat\beta_0(x), \hat\beta_1(x)) = \arg\min_{\beta\in\mathbb{R}^2} (Y - X\beta)^\top W (Y - X\beta).$$

Assuming $X^\top W X$ is invertible, the minimizer is

$$\begin{pmatrix} \hat\beta_0(x) \\ \hat\beta_1(x) \end{pmatrix} = (X^\top W X)^{-1} X^\top W Y.$$

Hence the fitted value at x satisfies

$$\hat m(x) = \hat\beta_0(x) = e_1^\top (X^\top W X)^{-1} X^\top W Y,$$

where $e_1 = (1, 0)^\top$. Thus $\hat m(x)$ can be written as

$$\hat m(x) = \sum_{i=1}^n w_{n,i}(x)\, Y_i, \qquad w_{n,i}(x) = \big[e_1^\top (X^\top W X)^{-1} X^\top W\big]_i,$$

and the weights $w_{n,i}(x)$ depend only on x, X_i, K, h (not on the Y_i's).

Question 2

Using this notation

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

Let's introduce

$$T_{n,j}(x) = \frac{1}{nh}\sum_{i=1}^{n} r_i^j K_i Y_i, \qquad j = 0, 1.$$

The equation that $\hat{\beta}$ solves is

$$X^\top W X \beta = X^\top W Y,$$

which we can rewrite as

$$\begin{pmatrix} S_{n,0}(x) & S_{n,1}(x) \\ S_{n,1}(x) & S_{n,2}(x) \end{pmatrix} \begin{pmatrix} \hat{\beta}_0(x) \\ \hat{\beta}_1(x) \end{pmatrix} = \begin{pmatrix} T_{n,0}(x) \\ T_{n,1}(x) \end{pmatrix}.$$

With $D_n(x) := S_{n,0}(x)S_{n,2}(x) - S_{n,1}(x)^2$ the solution for $\hat{\beta}_0(x)$ is

$$\hat{\beta}_0(x) = \frac{S_{n,2}(x)\, T_{n,0}(x) - S_{n,1}(x)\, T_{n,1}(x)}{D_n(x)}.$$

Substituting the definitions of the $T_{n,j}$ and rearranging gives

$$\hat{\beta}_0(x) = \frac{1}{nh}\sum_{i=1}^{n} \frac{S_{n,2}(x) - S_{n,1}(x)\, r_i}{D_n(x)}\, K_i\, Y_i.$$

Thus, we finally get

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

Question 3

$$\begin{aligned}
\sum_{i=1}^{n} w_{n,i}(x) &= \frac{1}{D_n(x)}\sum_{i=1}^{n}\frac{1}{nh}K_i(S_{n,2} - r_i S_{n,1}) \\
&= \frac{1}{D_n(x)}\Big(S_{n,2}\cdot\frac{1}{nh}\sum_{i=1}^{n}K_i - S_{n,1}\cdot\frac{1}{nh}\sum_{i=1}^{n}r_i K_i\Big) \\
&= \frac{1}{D_n(x)}(S_{n,2}S_{n,0} - S_{n,1}^2) = \frac{D_n(x)}{D_n(x)} = 1.
\end{aligned}$$

Practical exercise

Aim

The practical aim of this simulation study is to investigate how the **data-driven bandwidth** for a local linear estimator of the regression function,

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

computed via the asymptotic AMISE-optimal formula (for a quartic/biweight kernel), behaves when we vary: (i) the **number of observations** n, (ii) the **number of blocks** N used to estimate the two plug-in quantities $\hat{\sigma}^2$ and $\hat{\theta}_{22}$, and (iii) the **shape of the covariate distribution** (Beta(α, β)).

Quantities and estimators considered

This section lists the main quantities entering the simulation and the estimators used in the study.

- **Data generating process (DGP).**
  - Covariate $X \sim \text{Beta}(\alpha, \beta)$ on $[0, 1]$.
  - Regression function used in the simulation: $m(x) = \sin\left((x^3 + 0.1)^{-1}\right)$.
  - Observation model: $Y = m(X) + \varepsilon$, with $\varepsilon \sim N(0, \sigma^2)$ and fixed noise variance σ^2 (we use $\sigma^2 = 1$ unless stated otherwise).
- **Blockwise quartic fits.**
  - For a chosen integer N we split the sample into N contiguous blocks by *sorted* X (approximately equal counts per block). In block j we fit a quartic polynomial

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

  - For each observation we evaluate the fitted quartic and its second derivative. Blocks with too few points fall back to simple defaults in the implementation.
- **Plug-in estimators.**

- The curvature integral is estimated by

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

  where $j(i)$ is the block containing observation i and $\hat{m}^{(2)}_j$ is the second derivative from the quartic fit in block j.
- The noise variance is estimated by the pooled residual sum of squares from the blockwise quartic fits, divided by $(n-5N)$ (the total residual degrees of freedom):

$$\hat{\sigma}^2(N) = \frac{\mathrm{RSS}_{\mathrm{blocks}}}{n-5N}.$$

- **AMISE-optimal bandwidth** (quartic / biweight kernel constants combined into the formula used in the app):

$$\hat{h}_{AMISE}(N) = n^{-1/5}\left(\frac{35\,\hat{\sigma}^2(N)\,|\mathrm{supp}(X)|}{\hat{\theta}_{22}(N)}\right)^{1/5},$$

  where $|\mathrm{supp}(X)|$ is the length of the empirical support (for Beta on $[0,1]$ this is numerically close to 1; we use $\max(X)-\min(X)$).

- **Mallows' C_p** for choosing N.

  For a grid of candidate N we compute

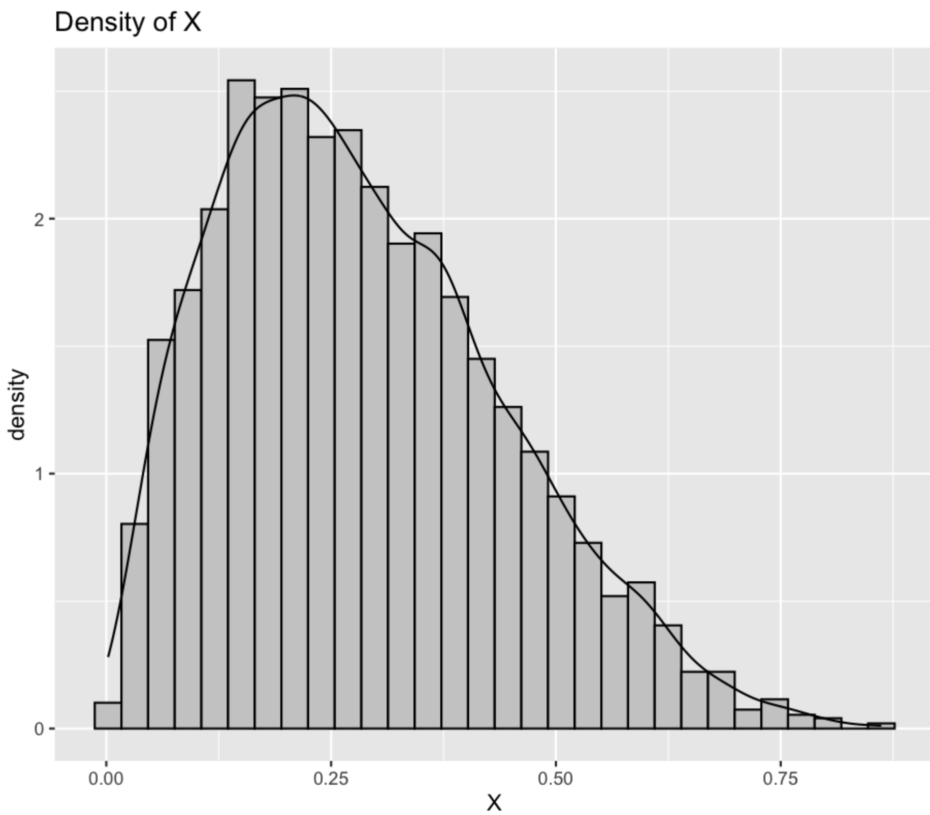
$$C_p(N) = \frac{\mathrm{RSS}(N)}{\mathrm{RSS}(N_{\max})/(n-5N_{\max})} - (n-10N),$$

  and select the N minimizing C_p; here $N_{\max} = \max\{\min(\lfloor n/20\rfloor, 5), 1\}$ by default (Ruppert et al., 1995).

- **Practical implementation details**
  - Blocks are contiguous in sorted-X order and approximately equal-size (remainder distributed on the first blocks).
  - Quartic fits use `lm()`; if a block has too few points the implementation falls back to a constant fit to avoid failures.
  - All computed diagnostics (per-N: $\hat{\theta}_{22}$, $\hat{\sigma}^2$, $\hat{h}$, block counts, RSS) are stored for later summarisation.

Results and visualisation

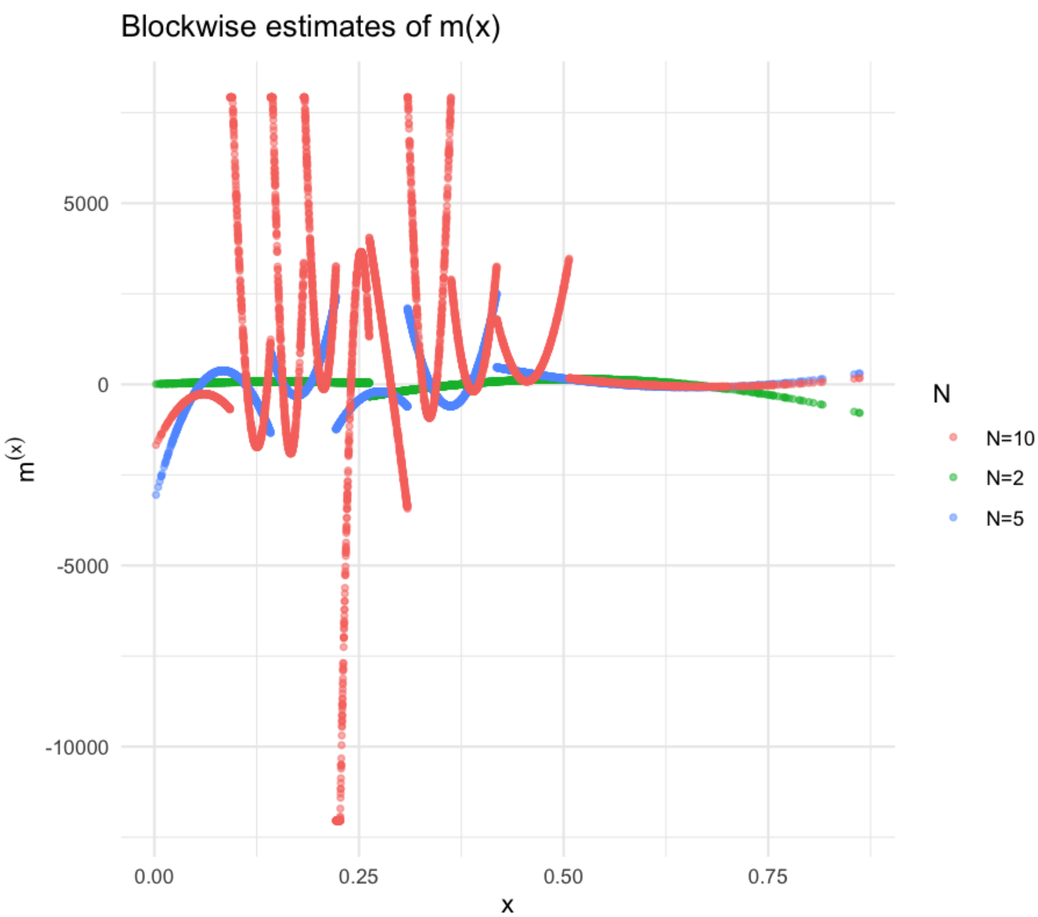
Figure 0 — Density of $X \sim \mathrm{Beta}(2,5) with n = 5000$



Interpretation.
We can observe that most values of (X) lie near 0, with a long right tail.

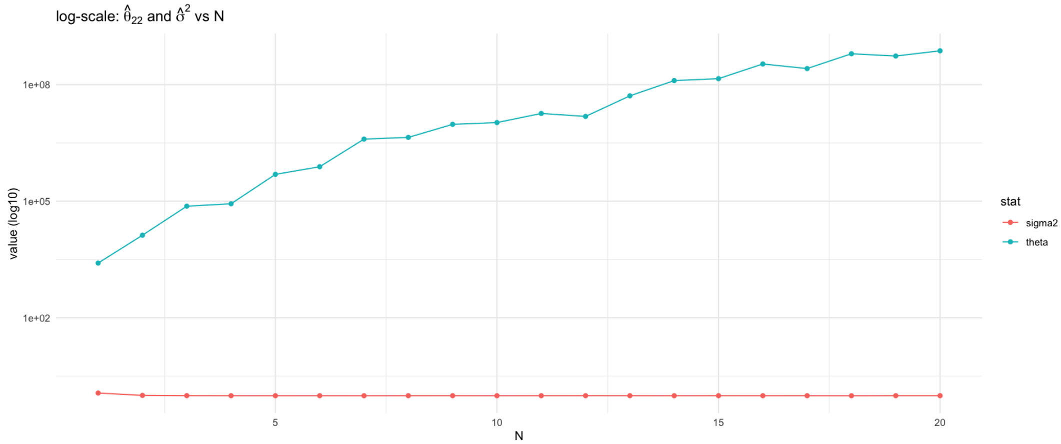
1) How does h_{AMISE} behave when N grows? Can you explain why?

Figure 1 — Blockwise overlays of $m''(x)$ for different N on the same grid



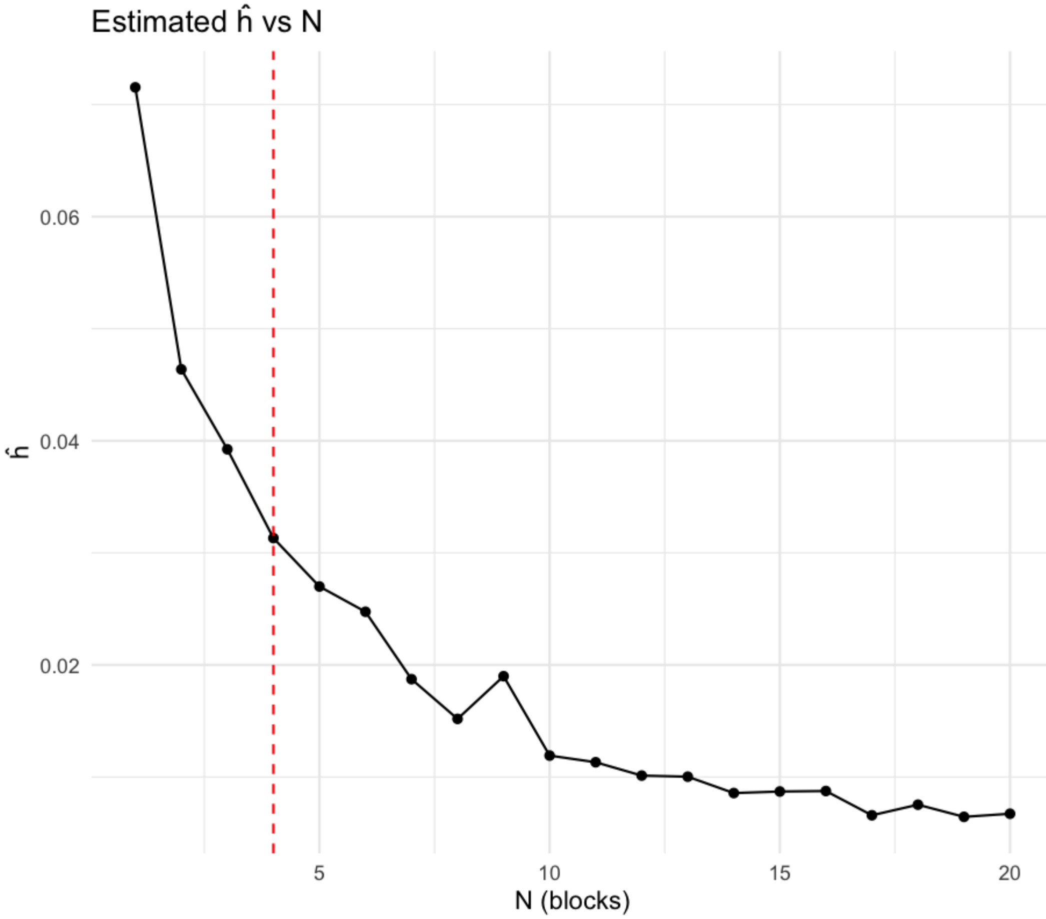
Interpretation.
When N is small, the blockwise $m''(x)$ estimates are smoother and tend to miss local curvature features.
As N grows, more structure appears, but variability increases.

Figure 2 — $\hat{\theta}_{22}(N)$ and $\hat{\sigma}^2(N)$ on log-scale



Interpretation.
As expected from Figure , $\hat{\theta}_{22}$ increases along with N, since it depends from the second derivative of m and by figure 1 we can see that there is a high variation it its curvature However $\hat{\sigma}^2$ stays roughly constant while $\hat{\theta}_{22}$ falls, this effect is reinforced.

Figure 3 — Estimated $\hat{h}_{AMISE}$ vs N with the optimal N by Mallows C_p



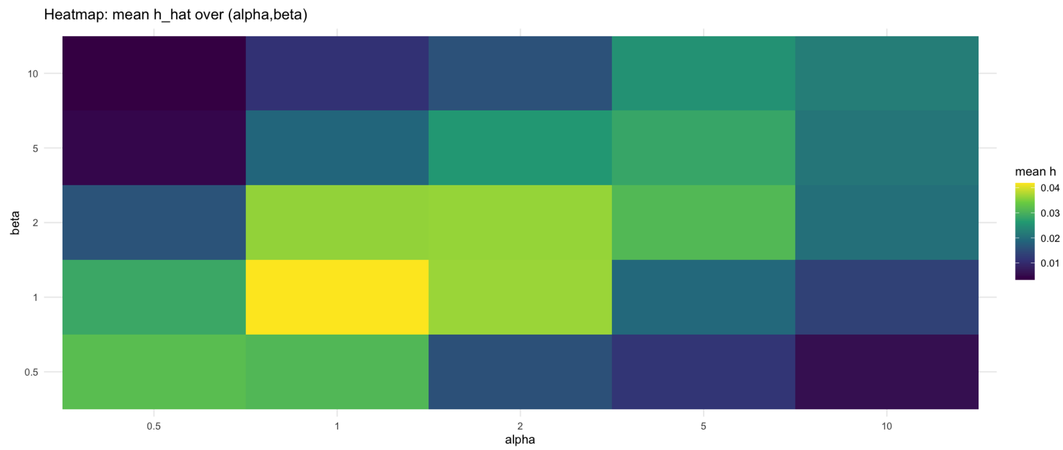
Interpretation. Thus as expected by the $\hat{h}$ formula and figure 2, $\hat{h}$ decreases as N becomes bigger.

2) Should N depend on n? Why?

Yes. For a quartic fit we need roughly at least 5 effective observations per block to estimate the 5 coefficients. More data allow more blocks while keeping the minimum block size adequate. In practice N should increase with n but not linearly, it's better to choose N by an information criterion such as C_p or via cross-validation.

What happens when the number of observations varies a lot between different regions in the support of X? How is this linked to the Beta parameters?

Figure 4 - Heatmap of the estimated mean h depending on different values of α and β



The Beta parameters (α, β) control that heterogeneity:

- **Small** α, β (e.g. $\alpha = \beta = 0.5$) $\Rightarrow$ U-shaped Beta (mass near both edges, sparse interior). Interior blocks tend to have few points and wide x-span $\to \hat{\theta}_{22}$ often small $\to$ larger mean $\hat{h}$. This explains bright cells for U-shaped parameter choices.

- **Large** $\alpha, \beta \Rightarrow$ mass concentrated near the center $\to$ block fits are more stable $\to$ larger $\hat{\theta}_{22}$ and smaller $\hat{h}$ (darker cells).

- **Skewed shapes** (one parameter small, the other large) produce a mix: some blocks very dense (good estimates), others very sparse (bad estimates). This increases the *variance* of blockwise estimates and often produces intermediate or variable mean $\hat{h}$ across replicates.